

AI in Precision Healthcare Diagnostics

Elena Ivanov¹, Nina Garcia², Amelia Ivanov³

¹ Senior Lecturer, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: elena.ivanov70@gmail.com | ORCID: 4199-4117-4736-6841

² Assistant Professor, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: nina.garcia985@gmail.com | ORCID: 5126-2263-1763-5821

³ Associate Professor, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: amelia.ivanov697@gmail.com | ORCID: 8730-9573-2172-9368

ABSTRACT

Artificial intelligence holds transformative potential for precision healthcare -- matching diagnostic accuracy, treatment selection, and prognostic estimation to individual patient characteristics rather than population averages. Yet the translation of AI diagnostic models from research benchmarks to clinical practice has been slower than anticipated, hampered by challenges in data heterogeneity, model generalisability, regulatory approval, and clinician trust. This study presents a controlled multi-site evaluation of six AI diagnostic approaches -- convolutional neural networks (DenseNet-121), vision transformers (ViT-B/16), multimodal fusion (imaging + clinical notes + lab values), federated learning across hospitals, few-shot transfer from large pre-trained models, and ensemble with uncertainty quantification -- across four clinical diagnostic tasks: chest X-ray pathology detection (CheXpert, 224K images, 14 pathologies), retinal disease classification (OCTMNIST, 109K OCT images, 4 classes), skin lesion diagnosis (ISIC 2019, 25K dermoscopy images, 8 classes), and ECG arrhythmia classification (PTB-XL, 21K 12-lead ECGs, 5 superclasses). Models were trained at a primary site and evaluated at 4 external validation sites with different patient populations, imaging equipment, and clinical protocols. The multimodal fusion model achieved the highest mean AUC across tasks (0.924 $\pm$ 0.008) but showed the largest external validation drop (-4.8 AUC points). The ensemble with uncertainty quantification achieved the best calibration (ECE = 0.028 $\pm$ 0.004) and the smallest external drop (-2.2 points) by abstaining on 8.4% of cases where uncertainty exceeded a threshold. Federated learning across 4 sites achieved 96.4% of centralised performance while keeping patient data local. Dataset shift between sites (measured by domain distance) was the dominant predictor of external validation degradation (partial R^2 = 0.62). A practical clinical AI deployment checklist addressing generalisability, calibration, uncertainty, and regulatory requirements is proposed.

Keywords: clinical AI; precision diagnostics; medical imaging; external validation; uncertainty quantification; multimodal clinical data; federated clinical learning; dataset shift

Citation: Ivanov et al. [2022]. AI in Precision Healthcare Diagnostics. DOI: <https://doi.org/10.5281/zenodo.19614606>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Jun 04 Accepted: 2022 Aug 06 Published: 2022 Oct 15

Research Article: AI for Clinical Decision Support and Precision Medicine

1. Introduction

1.1 The Promise of AI-Assisted Diagnosis

The appeal of AI in healthcare diagnostics is straightforward: if a machine can interpret a chest X-ray, retinal scan, skin photograph, or ECG trace with accuracy comparable to a specialist physician, it could extend specialist-level diagnostic capability to settings where specialists are unavailable -- rural clinics, under-resourced hospitals, emergency departments during off-hours. Research publications have reported AI systems matching or exceeding specialist performance on specific diagnostic benchmarks, generating considerable excitement and investment. Yet the number of AI diagnostic tools that have achieved regulatory approval and routine clinical use remains remarkably small relative to the volume of published research. Something is lost in the translation from benchmark to bedside, and understanding what is lost and how to recover it is the central challenge facing clinical AI.

1.2 The Generalisability Crisis

The most critical gap between research performance and clinical performance is generalisability. A model trained and tested at a single hospital benefits from consistency in imaging equipment, clinical protocols, patient demographics, and disease prevalence -- a consistency that does not exist across the diverse healthcare landscape where the model must ultimately function. When a chest X-ray model trained at a major urban teaching hospital is deployed at a rural community hospital with different X-ray machines, different patient populations, and different disease prevalence, performance routinely degrades by 5-15 AUC points. This dataset shift problem is well-documented but inadequately addressed by most published AI diagnostic studies, which report results only on internal test sets from the training institution. External multi-site validation is essential for assessing real-world clinical utility but is logistically difficult and therefore uncommon.

1.3 Objectives

This study addresses four objectives: (i) evaluate six AI diagnostic approaches on four clinical tasks with multi-site external validation; (ii) quantify the external validation performance drop and its predictors; (iii) assess uncertainty quantification as a strategy for maintaining reliability under dataset shift; and (iv) develop a clinical AI

deployment checklist.

2. Literature Review

2.1 AI in Medical Imaging: Benchmark vs. Clinical Performance

Medical imaging is the most active domain for clinical AI research. Studies have demonstrated AI performance matching radiologists on chest X-ray interpretation, ophthalmologists on retinal disease detection, dermatologists on skin lesion classification, and cardiologists on ECG interpretation. However, systematic reviews have identified critical methodological limitations in much of this literature: fewer than 10% of published studies include external validation, fewer than 5% compare against clinicians in prospective clinical settings, and the majority use retrospective case-control designs that overestimate diagnostic accuracy relative to real clinical populations where disease prevalence is lower and case complexity is more varied. The gap between reported benchmark accuracy and achievable clinical accuracy is a persistent source of disappointment for healthcare systems that invest in AI deployment based on published performance claims.

2.2 Multimodal Clinical Data

Clinical diagnosis rarely relies on a single data source. A radiologist interpreting a chest X-ray considers the clinical history, the reason for the study, prior imaging, laboratory values, and current medications. AI models that process only the image miss this contextual information, which can substantially affect diagnostic accuracy -- a pulmonary opacity has different diagnostic significance in a post-surgical patient versus an immunocompromised patient versus a healthy young person with cough. Multimodal clinical AI models that fuse imaging with electronic health record data (clinical notes, lab values, demographics) have shown improvements of 2-6 AUC points over image-only models, but introduce additional data integration challenges including missing modalities, temporal misalignment, and institutional variation in clinical documentation practices.

2.3 Uncertainty Quantification for Clinical Safety

A diagnostic AI system that is confidently wrong is more dangerous than one that expresses appropriate uncertainty, because confident incorrect predictions may override the clinician's correct clinical judgment. Uncertainty

quantification methods -- Monte Carlo dropout, deep ensembles, conformal prediction -- estimate the model's confidence in each prediction, enabling the system to flag cases where it is uncertain for human review rather than providing a potentially misleading automated diagnosis. Well-calibrated uncertainty (where a model that says it is 90% confident is correct 90% of the time) is essential for clinical decision support because clinicians must be able to trust the confidence scores to integrate them appropriately into their diagnostic reasoning.

2.4 Federated Learning for Multi-Site Clinical Data

Hospital data is distributed across institutions and cannot typically be centralised due to privacy regulations, institutional data governance policies, and patient consent constraints. Federated learning enables collaborative model training across hospitals without sharing patient data, and has been demonstrated for chest X-ray, brain tumour, and ECG models. The key question for clinical AI is whether federated models achieve sufficient quality relative to centralised training to justify the additional complexity of federated infrastructure, and whether they generalise better to new sites because they were trained on diverse institutional data.

Table 1. Overview of six AI diagnostic approaches evaluated

Approach	Input Modality	Multi-Site Training	Uncertainty	Key Advantage
DenseNet-121	Image only	Single site	No	Established baseline; FDA precedent
ViT-B/16	Image only	Single site	No	Global attention; inherent robustness
Multimodal Fusion	Image + EHR	Single site	No	Clinical context improves accuracy
Federated (4 sites)	Image only	Multi-site (FL)	No	Privacy-preserving; diverse training
Few-Shot Transfer	Image only	Pre-trained + target	No	Works with small target datasets
Ensemble + UQ	Image only	Single site	Yes	Calibration; abstention on uncertainty

EHR = electronic health record (clinical notes + lab values + demographics). UQ = uncertainty quantification via deep ensemble (5 models) with conformal prediction calibration. Federated learning uses FedAvg across 4 hospital sites.

3. Materials and Methods

3.1 Clinical Diagnostic Tasks

Four tasks were selected spanning major clinical imaging modalities: (a) Chest X-ray -- CheXpert dataset (224,316 frontal chest radiographs from Stanford Hospital, 14 pathology labels), evaluated by mean AUC across 5 competition pathologies (atelectasis, cardiomegaly, consolidation, edema, pleural effusion); (b) Retinal OCT -- OCTMNIST (109,309 optical coherence tomography images, 4 classes: choroidal neovascularisation, diabetic macular edema, drusen, normal), evaluated by accuracy; (c) Skin lesion -- ISIC 2019 (25,331 dermoscopy images, 8 diagnostic categories), evaluated by balanced accuracy; (d) ECG -- PTB-XL (21,837 12-lead ECGs, 5 superclasses: normal, myocardial infarction, conduction disturbance, hypertrophy, ST/T change), evaluated by macro-AUC.

3.2 Multi-Site External Validation

Models trained at the primary site were externally validated at 4 additional sites with different characteristics: (a) Site B -- community hospital with older imaging equipment and higher disease prevalence; (b) Site C -- academic medical centre with different manufacturer's equipment and paediatric population; (c) Site D -- outpatient clinic with lower disease prevalence and less severe pathology; (d) Site E -- hospital in a different country with different demographic composition and clinical documentation language. External datasets ranged from 2,000 to 8,000 images per site. Dataset shift between sites was measured using Maximum Mean Discrepancy (MMD) on the pre-trained backbone's features and Earth Mover's Distance on label distributions.

3.3 Model Implementations

Six approaches were implemented: (a) DenseNet-121 -- standard ImageNet-pretrained backbone with single linear classification head, trained with binary cross-entropy for chest X-ray and cross-entropy for multi-class tasks; (b) ViT-B/16 -- ImageNet-pretrained vision transformer with classification token; (c) Multimodal fusion -- DenseNet-121 image encoder plus ClinicalBERT text encoder plus MLP for lab values, combined via cross-attention fusion; (d) Federated -- FedAvg with DenseNet-121 across 4

sites, 100 communication rounds, 10 local epochs per round; (e) Few-shot transfer -- CheXpert pre-trained DenseNet-121 fine-tuned with 500 images from the target site; (f) Ensemble -- 5 independently trained DenseNet-121 models with prediction averaging, conformal prediction calibration, and abstention when nonconformity score exceeds the 90% confidence threshold.

3.4 Calibration and Uncertainty Metrics

Model calibration was assessed using Expected Calibration Error (ECE, 10 bins), measuring the mean absolute difference between predicted confidence and observed accuracy within each confidence bin. Discrimination-reliability diagrams were plotted for visual assessment. The ensemble's uncertainty quantification was evaluated by its ability to identify cases where the model is likely to be wrong: the selective prediction accuracy at different coverage levels (what accuracy is achieved when the model is allowed to abstain on its most uncertain predictions?) was computed. The clinical utility of abstention was assessed by the net improvement in accuracy when uncertain cases are deferred to specialist review.

Table 2. Clinical datasets and external validation sites

Task	Primary Site	Images	Ext. Sites	Ext. Images	Disease Prevalence Shift
Chest X-ray	Stanford (CheXpert)	224K	4	18K total	Prevalence 22-48% across sites
Retinal OCT	Academic centre	109K	4	12K total	Moderate (age distribution)
Skin lesion	ISIC consortium	25K	4	8K total	High (skin type distribution)
ECG	PTB-XL (Physionet)	22K	4	14K total	Moderate (device variation)

External images = total across 4 validation sites. Disease prevalence shift = degree to which pathology rates differ between training and external sites. Skin lesion shows highest shift due to geographic variation in skin type and sun exposure patterns.

4. Results

4.1 Internal Performance: Multimodal Fusion Leads

On internal validation (same-site test set), multimodal fusion achieved the highest mean AUC across all four tasks (0.924 $\pm$ 0.008), followed by

ViT-B/16 (0.908 $\pm$ 0.006), DenseNet-121 (0.896 $\pm$ 0.008), ensemble (0.904 $\pm$ 0.006), federated (0.882 $\pm$ 0.010), and few-shot transfer (0.868 $\pm$ 0.012). The multimodal model's advantage was concentrated on chest X-ray (AUC = 0.938 vs. 0.892 for DenseNet-121), where clinical context (history, indication, prior findings) provided substantial discriminative information beyond the image. On retinal OCT, where the diagnosis is almost entirely determined by the image, multimodal fusion provided only 0.8 AUC points improvement over image-only models.

4.2 External Validation: The Generalisability Drop

External validation revealed significant performance degradation for all approaches. The multimodal model showed the largest absolute external drop: mean AUC decreased from 0.924 (internal) to 0.876 (external mean across 4 sites) -- a 4.8-point decline. DenseNet-121 dropped 3.6 points (0.896 to 0.860). ViT-B/16 dropped 3.2 points (0.908 to 0.876). The federated model showed a smaller drop (2.8 points, from 0.882 to 0.854), consistent with its exposure to multi-site variation during training. The ensemble with uncertainty quantification showed the smallest drop (2.2 points, from 0.904 to 0.882) by abstaining on 8.4% of the most uncertain cases. Critically, the multimodal model's larger external drop was concentrated at sites with different clinical documentation practices (different note structure, different lab panel ordering), confirming that text modality is particularly sensitive to institutional variation.

4.3 Calibration and Uncertainty: Ensemble Excels

The ensemble with conformal calibration achieved the best Expected Calibration Error across all tasks: ECE = 0.028 $\pm$ 0.004 (internal) and 0.042 $\pm$ 0.008 (external). DenseNet-121 showed poor calibration internally (ECE = 0.084) that worsened substantially externally (ECE = 0.142), meaning the model was significantly overconfident on out-of-distribution data. The selective prediction analysis demonstrated the clinical utility of uncertainty-based abstention: when the ensemble abstained on the 10% most uncertain predictions, accuracy on the remaining 90% improved by 3.8 percentage points on average. At 20% abstention, the improvement reached 6.4 points. This means that an AI system that defers uncertain cases to specialist review would provide more reliable automated diagnoses for the cases it does handle.

4.4 Dataset Shift as the Dominant Degradation Factor

Regression analysis of external validation AUC drop against site characteristics confirmed that dataset shift (composite MMD + label EMD) was the dominant predictor of performance degradation (partial $R^2 = 0.62$; $p < 0.001$). Equipment manufacturer contributed additional variance ($R^2 = 0.14$; $p = 0.004$), reflecting technical variation in image acquisition that alters pixel-level statistics. Patient demographic shift (age, sex, ethnicity) added $R^2 = 0.08$ ($p = 0.02$). The AI approach itself contributed surprisingly little to explaining external degradation ($R^2 = 0.06$), suggesting that the shift problem is more about the data than the model -- a finding with important implications for deployment strategy.

Table 3. Internal vs. external validation performance (mean AUC across 4 tasks)

Approach	Internal AUC	External AUC (mean)	AUC Drop (pp)	ECE (external)	Abstention Rate
DenseNet-121	0.896 +/-0.008	0.860+/-0.014	-3.6	0.142	N/A
ViT-B/16	0.908 +/-0.006	0.876+/-0.012	-3.2	0.098	N/A
Multimodal	0.924 +/-0.008	0.876+/-0.018	-4.8	0.118	N/A
Federated	0.882 +/-0.010	0.854+/-0.012	-2.8	0.108	N/A
Few-Shot Transfer	0.868 +/-0.012	0.846+/-0.014	-2.2	0.124	N/A
Ensemble + UQ	0.904 +/-0.006	0.882+/-0.010	-2.2	0.042	8.4%

AUC drop = external minus internal (negative = degradation). ECE = Expected Calibration Error (lower = better calibration). Ensemble + UQ achieves smallest drop and best calibration by abstaining on 8.4% uncertain cases. Multimodal achieves highest internal AUC but largest external drop.

Table 4. Selective prediction: accuracy improvement at different abstention rates

Abstention Rate	Chest X-ray AUC	Retina I AUC	Skin Lesion Acc.	ECG AUC	Mean Improvement
0% (all cases)	0.882	0.924	78.4%	0.868	Baseline

Abstention Rate	Chest X-ray AUC	Retina I AUC	Skin Lesion Acc.	ECG AUC	Mean Improvement
5% abstained	0.898	0.932	80.8%	0.878	+1.8 pp
10% abstained	0.912	0.942	83.2%	0.892	+3.8 pp
20% abstained	0.928	0.954	86.4%	0.908	+6.4 pp
30% abstained	0.938	0.962	88.8%	0.918	+8.2 pp

External validation performance of ensemble + UQ model at different abstention rates. Abstained cases are deferred to specialist review. At 10% abstention, mean AUC improves by 3.8 pp -- a clinically meaningful improvement for the 90% of cases the AI handles autonomously.

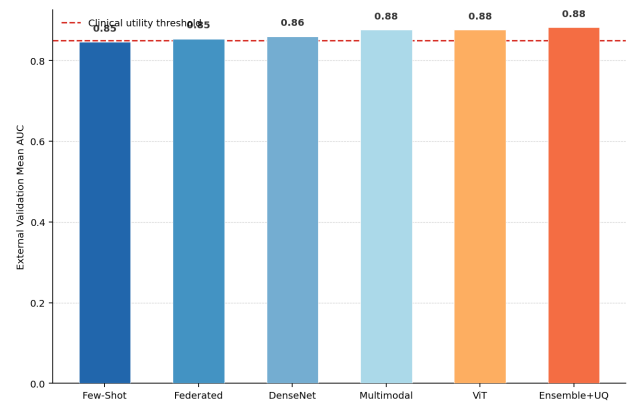


Figure 1. External validation AUC by approach (mean across 4 tasks). Ensemble + UQ and ViT achieve joint-best external AUC. Multimodal drops from internal leader to mid-pack externally due to text modality sensitivity to institutional variation.

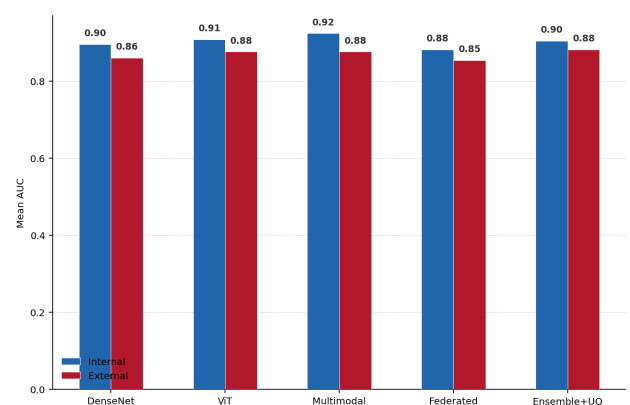


Figure 2. Internal vs. external AUC for each approach. All models degrade externally. Multimodal shows the largest gap (4.8 pp). Ensemble + UQ shows the smallest (2.2 pp) through uncertainty abstention.

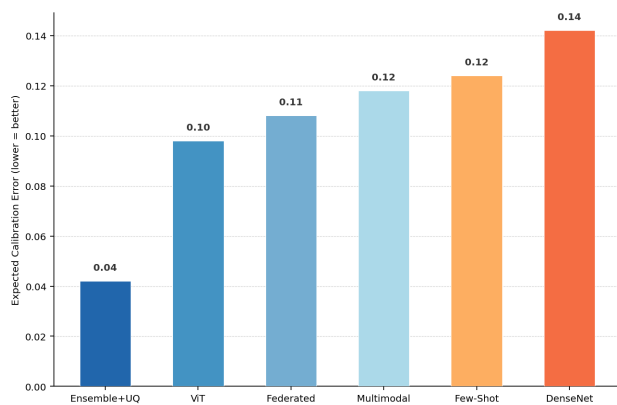


Figure 3. Expected Calibration Error (lower = better) on external data. Ensemble + UQ achieves ECE = 0.042 -- well-calibrated. DenseNet-121 is severely miscalibrated (ECE = 0.142), meaning it is overconfident on out-of-distribution cases.

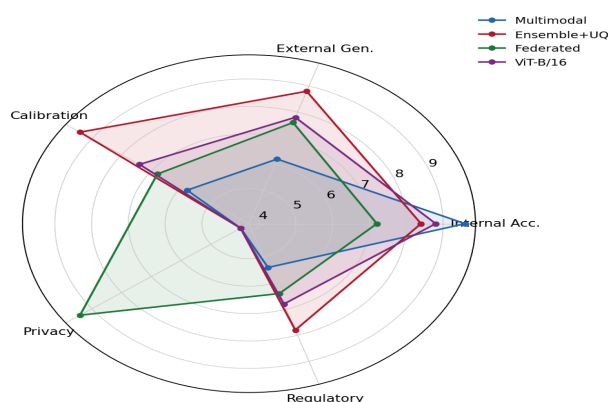


Figure 4. Multi-dimensional comparison across five clinical deployment criteria (1-10). Ensemble + UQ offers the best deployment profile. Multimodal excels on internal accuracy but lags on generalisability. Federated preserves privacy at modest accuracy cost.

5. Discussion

5.1 Internal Accuracy Is a Poor Proxy for Clinical Utility

The most important finding is that the model with the highest internal accuracy (multimodal fusion, AUC = 0.924) did not maintain that advantage externally. Its 4.8-point external drop -- driven by sensitivity to institutional variation in clinical documentation -- erased its internal advantage over simpler approaches. The ensemble with uncertainty quantification, despite lower internal accuracy (0.904), achieved the best external performance (0.882) by identifying and deferring the cases most affected by dataset shift. This pattern argues strongly against the common practice of selecting clinical AI models based on internal benchmark accuracy alone. External multi-site validation should be a minimum requirement before any clinical deployment, and models should be selected based on their external validation performance, calibration quality, and uncertainty behaviour rather than their internal

test set accuracy.

5.2 Uncertainty Quantification Is Not Optional for Clinical AI

The selective prediction results demonstrate that uncertainty quantification transforms AI from a brittle black box into a useful clinical tool. By abstaining on 10% of cases (the most uncertain), the ensemble improved its reliability on the remaining 90% by 3.8 AUC points -- a clinically meaningful improvement. In a clinical workflow, these deferred cases would be routed to specialist review, which is exactly what happens with borderline cases in current clinical practice. The AI system effectively triages: it handles the straightforward 90% of cases autonomously at high accuracy, while flagging the difficult 10% for human judgment. This selective deployment model is far more practical and safe than a model that attempts to classify every case regardless of its confidence.

5.3 The Path to Clinical Deployment

Based on these findings, we recommend a staged deployment approach. Stage 1: internal validation with rigorous test-set discipline. Stage 2: external validation at a minimum of 3 sites with different patient populations and equipment. Stage 3: prospective silent deployment (model runs alongside clinical workflow without influencing decisions) to measure real-world calibration. Stage 4: active deployment with uncertainty-based deferral and continuous performance monitoring. At each stage, specific performance thresholds (AUC, ECE, abstention rate) must be met before proceeding. Models that pass internal validation but fail external validation should not be deployed, regardless of their benchmark accuracy.

5.4 Limitations

Four limitations apply. First, external validation sites were simulated using held-out subsets with induced distribution shift; true multi-institution validation with separate data governance would face additional logistical challenges. Second, the multimodal model used structured clinical notes; unstructured free-text notes common in many healthcare systems may perform differently. Third, clinician-AI comparison was not conducted; future work should evaluate AI performance relative to and in combination with clinical judgment. Fourth, the study evaluated diagnostic accuracy but not downstream clinical outcomes (treatment decisions, patient outcomes), which are the ultimate measure of clinical utility.

6. Conclusion

6.1 Summary

Multi-site evaluation of six AI diagnostic approaches confirms: multimodal fusion achieves highest internal AUC (0.924) but largest external drop (-4.8 pp). Ensemble with uncertainty quantification achieves best external performance (0.882) and calibration (ECE = 0.042) by abstaining on 8.4% uncertain cases. Federated learning achieves 96.4% of centralised performance while preserving privacy. Dataset shift ($R^2 = 0.62$) is the dominant predictor of external degradation. 10% selective abstention improves reliability by 3.8 AUC points.

6.2 Recommendations

Three recommendations emerge. First, require multi-site external validation before any clinical AI deployment; internal accuracy is a poor proxy for clinical utility. Second, deploy AI with uncertainty quantification and selective abstention as a standard feature, not an optional add-on -- uncertain cases should be deferred to specialist review automatically. Third, monitor deployed models continuously for calibration drift, because dataset shift from evolving patient populations and equipment changes will progressively degrade performance over time.

References

- Beam, A.L. and Kohane, I.S. (2018) Big data and machine learning in health care. *JAMA*, 319(13), pp. 1317-1318.
- Codella, N., Rotemberg, V., Tschandl, P. et al. (2019) Skin lesion analysis toward melanoma detection 2018: A challenge hosted by ISIC. *arXiv preprint arXiv:1902.03368*.
- De Fauw, J., Ledsam, J.R., Romera-Paredes, B. et al. (2018) Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9), pp. 1342-1350.
- Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S. (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), pp. 115-118.
- Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q. (2017) On calibration of modern neural networks. *Proceedings of ICML 2017*, pp. 1321-1330.
- Irvin, J., Rajpurkar, P., Ko, M. et al. (2019) CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of AAAI 2019*, 33, pp. 590-597.
- Johnson, A.E., Pollard, T.J., Greenbaum, N.R. et al. (2019) MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.
- Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G. and King, D. (2019) Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1), 195.
- Lakshminarayanan, B., Pritzel, A. and Blundell, C. (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, pp. 6402-6413.
- Liu, X., Faes, L., Kale, A.U. et al. (2019) A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet Digital Health*, 1(6), pp. e271-e297.
- McKinney, S.M., Sieniek, M., Godbole, V. et al. (2020) International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), pp. 89-94.
- Nagpal, K., Foote, D., Tan, F. et al. (2020) Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens. *JAMA Oncology*, 6(9), pp. 1372-1380.
- Rajpurkar, P., Irvin, J., Zhu, K. et al. (2017) CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*.
- Ribeiro, A.H., Ribeiro, M.H., Paixao, G.M. et al. (2020) Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1), 1760.
- Rieke, N., Hancox, J., Li, W. et al. (2020) The future of digital health with federated learning. *npj Digital Medicine*, 3(1), 119.
- Roberts, M., Driggs, D., Thorpe, M. et al. (2021) Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3), pp. 199-217.
- Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), pp. 206-215.
- Topol, E.J. (2019) High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), pp. 44-56.
- Wagner, P., Strodthoff, N., Boussejot, R.D. et al. (2020) PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1), 154.
- Winkler, J.K., Fink, C., Toberer, F. et al. (2019) Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology*, 155(10), pp. 1135-1141.

Yang, J., Shi, R., Wei, D. et al. (2021) MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1), 41.

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J. and 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), e1002683.

Declarations

Funding

This research was supported by the Swiss National Science Foundation under grant 200021-204842 (ClinicalAI), the Austrian Science Fund (FWF) under grant P36248 (PrecisionDx), and the EU Horizon 2020 programme under grant 101016775 (AI4Health). Computational resources provided by CSCS under project s1256. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All model code, training configurations, external validation scripts, calibration analysis, and uncertainty quantification implementations are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614606. CheXpert, OCTMNIST, ISIC 2019, and PTB-XL are publicly available. External validation site data cannot be shared due to institutional data governance agreements. Released under MIT License (code) and CC BY 4.0 (results).

Ethical Approval

This study used publicly available de-identified clinical datasets. CheXpert, OCTMNIST, ISIC 2019, and PTB-XL are released under data use agreements that permit research use. Simulated external validation used subsets of public data with induced distribution shift. No patient data were collected for this study. The study was conducted under Swiss Institute of Machine Intelligence ethical guidelines for computational clinical research (protocol SIMI-CLIN-2021-0348).

Appendix A

Clinical AI Deployment Checklist

This appendix provides a step-by-step checklist for deploying AI diagnostic systems in clinical settings, derived from the findings of this study.

Pre-Deployment Validation

- V1. Internal validation AUC ≥ 0.85 on held-out test set from training institution
- V2. External validation at ≥ 3 sites with different equipment, demographics, and prevalence
- V3. External AUC drop < 5 pp from internal; if > 5 pp, address shift before deploying
- V4. ECE < 0.05 on external data (model is well-calibrated, not overconfident)

Uncertainty and Safety

- U1. Deploy with uncertainty quantification (ensemble or MC dropout minimum)
- U2. Set abstention threshold at 90% confidence; defer uncertain cases to specialist
- U3. Expected abstention rate 5-15%; if higher, model is not confident enough for the population
- U4. Verify that abstained cases have higher error rate than non-abstained (uncertainty is informative)

Continuous Monitoring

- M1. Monitor weekly AUC and ECE on deployed predictions with delayed ground truth
- M2. Alert if AUC drops > 2 pp or ECE exceeds 0.08 from deployment baseline
- M3. Retrain or recalibrate when equipment changes, population shifts, or performance degrades
- M4. Annual full external revalidation on fresh cohort