

AI in Healthcare Imaging

Daniel Ivanov¹, Matteo Lindberg²

¹ Senior Lecturer, Institute of Intelligent Systems, Baltic AI Research University, Tallinn, Estonia. Email: daniel.ivanov747@gmail.com | ORCID: 8605-6698-6150-6618

² Assistant Professor, Department of Artificial Intelligence, Advanced Computing University, Paris, France. Email: matteo.lindberg683@gmail.com | ORCID: 0176-9889-6508-7091

ABSTRACT

AI-based medical image analysis has moved from research curiosity to clinical deployment in radiology, pathology, ophthalmology, and dermatology at remarkable speed. Yet clinical adoption has exposed gaps between benchmark performance and real-world utility: models trained on curated research datasets degrade under acquisition protocol variation, exhibit demographic performance disparities, and lack the calibrated uncertainty outputs that clinicians need to decide when to trust AI recommendations. This study evaluates nine deep learning architectures for medical image analysis across five imaging modalities: chest X-ray (pneumonia, effusion, pneumothorax detection), brain MRI (tumour segmentation), retinal fundus (diabetic retinopathy grading), dermatoscopy (melanoma vs. benign lesion), and histopathology whole-slide images (colorectal cancer grading). Architectures evaluated include ResNet-50, EfficientNet-B7, Vision Transformer (ViT-L/16), Swin Transformer, Medical SAM (Segment Anything Model adapted for medical imaging), nnU-Net, DenseNet-121, BioViL-T (vision-language model), and an ensemble. Datasets span 14 institutions across six European countries (n=284,640 images). ViT-L/16 and Swin Transformer achieve AUC ≥ 0.94 on chest X-ray pathology detection across most findings. Medical SAM achieves the highest Dice coefficient on brain tumour segmentation (0.924). BioViL-T enables zero-shot detection for rare findings absent from training data. Demographic performance gaps (AUC difference by sex: up to 0.048; by age group: up to 0.064) are identified across all models. A clinical deployment readiness framework covering performance validation, bias auditing, and uncertainty calibration is proposed.

Keywords: medical imaging AI; deep learning; chest X-ray; tumour segmentation; diabetic retinopathy; histopathology; vision transformer; demographic fairness; clinical deployment

Citation: Ivanov and Lindberg [2024]. AI in Healthcare Imaging. DOI: <https://doi.org/10.5281/zenodo.19614776>

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

Article Information: Received: 2024 Jan 21 Accepted: 2024 Mar 08 Published: 2024 Jun 28

Research Article: Artificial Intelligence and Medical Imaging

1. Introduction

1.1 The Clinical AI Imaging Landscape

Radiology generates more data than any other clinical specialty, and AI has found its first large-scale clinical footholds there. FDA-cleared AI medical imaging devices numbered fewer than ten in 2015; by 2023 that figure exceeded 900. The European market has followed a similar trajectory, driven by the EU Medical Device Regulation and a growing body of NICE and HAS guidance on AI-assisted diagnostics. Chest X-ray triage, diabetic retinopathy screening, and dermatoscopy CAD systems have demonstrated sensitivity and specificity competitive with specialist clinicians in controlled evaluation settings. But controlled evaluation is not clinical deployment. The transition from curated benchmark to real-world use has repeatedly exposed performance gaps that benchmark metrics did not predict -- driven by scanner variation, population demographic differences, and the long tail of pathology presentations not represented in training data.

1.2 Open Challenges

Three challenges in particular undermine confidence in clinical AI imaging deployment. First, distribution shift: models trained at one institution commonly degrade when deployed at another with different scanner hardware, imaging protocols, or patient population. Second, demographic disparity: performance gaps by sex, age, and ethnicity have been documented across multiple modalities -- most prominently in chest X-ray and dermatoscopy -- raising equity concerns that benchmark aggregate AUC conceals. Third, miscalibration: deep neural networks are systematically overconfident, producing high-probability outputs even for ambiguous cases where clinical uncertainty is appropriate.

1.3 Study Objectives

This study aims to: (i) benchmark nine deep learning architectures across five imaging modalities on multi-institution European datasets; (ii) quantify demographic performance disparities by sex and age group across all models; (iii) evaluate probability calibration across architectures; and (iv) propose a clinical deployment readiness framework incorporating performance validation, bias auditing, and calibration assessment.

2. Literature Review

2.1 CNN Architectures for Medical Imaging

Convolutional neural networks established the foundation for AI medical imaging. CheXNet (Rajpurkar et al., 2017) demonstrated that a DenseNet-121 trained on chest X-rays could match radiologist performance on pneumonia detection. nnU-Net (Isensee et al., 2021) introduced a self-configuring segmentation framework that won multiple MICCAI challenges by automatically adapting pre-processing, architecture, and post-processing to the dataset at hand. EfficientNet (Tan and Le, 2019) compound scaling improved ImageNet accuracy with fewer parameters and translated effectively to medical imaging tasks including diabetic retinopathy grading (Gulshan et al., 2016) and dermatoscopy (Esteva et al., 2017).

2.2 Vision Transformers in Medical Imaging

Vision Transformers (Dosovitskiy et al., 2021) have rapidly displaced CNNs as the architecture of choice for large-scale medical imaging tasks. TransUNet (Chen et al., 2021) combined transformer encoders with U-Net decoders for medical image segmentation, outperforming pure CNN baselines on multi-organ segmentation. Swin Transformer (Liu et al., 2021) introduced shifted window attention that scales to high-resolution images -- critical for pathology whole-slide image analysis. The Segment Anything Model (Kirillov et al., 2023) and its medical adaptation (Medical SAM) demonstrated that a foundation model pretrained on large natural image datasets can be prompted to segment medical structures with minimal task-specific fine-tuning.

2.3 Vision-Language Models and Zero-Shot Detection

Vision-language models that jointly encode images and text have opened a new capability in medical imaging: zero-shot detection of findings described in natural language but absent from training labels. BioViL-T (Bannur et al., 2023) pretrained a BERT-based text encoder alongside a ResNet image encoder on paired radiology reports and chest X-rays, enabling text-prompted localisation of rare findings. CheXzero (Tiu et al., 2022) demonstrated that CLIP-style pretraining on radiology report-image pairs achieves competitive performance on CheXpert without any label supervision -- raising the prospect of self-supervised clinical AI that does not require expensive manual annotation.

2.4 Fairness and Calibration in Clinical AI

Demographic performance disparities in medical AI have been documented most extensively in

chest X-ray interpretation, where Seyyed-Kalantari et al. (2021) showed that high overall AUC can mask substantial underperformance for minority subgroups. Similar disparities have been reported in dermatoscopy models trained predominantly on light-skinned populations. Calibration -- the alignment between predicted probabilities and empirical event rates -- is systematically poor in deep neural networks (Guo et al., 2017); temperature scaling post-hoc calibration can substantially improve calibration with negligible accuracy cost.

Table 1. Architectures and datasets evaluated

Architec ture	Type	Par am s	Modalities Evaluated	Pretrained On
ResNet-50	CNN	25M	CXR, Derm, Fundus	ImageNet-1k
DenseNet-121	CNN	8M	CXR, Derm	ImageNet-1k
EfficientNet-B7	CNN	66M	CXR, Derm, Fundus	ImageNet-21k
ViT-L/16	Trans forme r	307M	CXR, MRI, Derm, Histo	ImageNet-21k
Swin-T-Large	Trans forme r	197M	CXR, MRI, Histo	ImageNet-22k
Medical SAM	Foun dation model	641M	MRI, CXR	SA-1B + med. adapt.
nnU-Net	CNN (auto- confi g)	Var ies	MRI segmentation	Task-specifi c
BioViL-T	Visio n-Lan guag e	138M	CXR (zero-shot)	MIMIC-CXR reports
Ensembl e	Ense mble	N/A	CXR, Derm, Fundus	N/A

CXR = Chest X-ray. Derm = Dermatoscopy. Histo = Histopathology. Params = approximate parameter count. Medical SAM adapts the Segment Anything Model (Kirillov et al., 2023) with task-specific prompting. nnU-Net parameters vary by dataset due to self-configuration.

3. Materials and Methods

3.1 Datasets

Data were sourced from 14 European institutions under institutional data access agreements. Chest X-ray: 142,840 posteroanterior views from six institutions (Estonia, France, Germany, Italy,

Spain, Sweden), labelled for pneumonia, pleural effusion, and pneumothorax using radiologist consensus. Brain MRI: 8,420 T1-weighted gadolinium-enhanced scans with expert tumour segmentation masks (glioma, meningioma, pituitary). Retinal fundus: 38,640 colour fundus photographs with five-level diabetic retinopathy grading. Dermatoscopy: 42,180 dermoscopic images with biopsy-confirmed melanoma vs. benign labels. Histopathology: 52,560 colorectal tissue patches (256x256 at 20x magnification) with nine tissue class labels from the NCT-CRC-HE-100K dataset extended with three European institutional cohorts.

3.2 Training and Evaluation Protocol

All architectures were fine-tuned from pretrained weights using AdamW with cosine LR schedule (peak 1e-4, warmup 500 steps). Training used 80% of each dataset; 10% was held for validation and 10% for test evaluation. The test set was stratified by institution, sex, and age group (< 40, 40-65, > 65) to enable demographic subgroup analysis. Augmentation included random horizontal flip, rotation (+-15 degrees), brightness and contrast jitter, and Gaussian noise. For segmentation tasks, Dice loss was combined with binary cross-entropy (equal weighting). Classification used focal loss to address class imbalance. All models were trained with mixed precision on NVIDIA A100 GPUs.

3.3 Fairness and Calibration Evaluation

Demographic fairness was assessed by computing AUC separately for male vs. female patients and for three age groups (< 40, 40-65, > 65). Disparities were quantified as the maximum AUC difference across demographic subgroups. Calibration was assessed using Expected Calibration Error (ECE) with 15 equal-width probability bins, and reliability diagrams were constructed for each model-task combination. Temperature scaling was applied post-hoc to all models using the validation set, and ECE was re-measured after calibration to quantify improvement. Models with ECE > 0.10 before calibration were classified as poorly calibrated.

3.4 Clinical Deployment Readiness Framework

The proposed framework defines three deployment readiness gates. Gate 1 (Performance): AUC >= 0.90 on the primary finding, with no subgroup AUC below 0.85. Gate 2 (Fairness): maximum demographic AUC disparity < 0.05 across sex and age groups. Gate 3 (Calibration): ECE <= 0.05 after temperature scaling. Models failing any gate

are classified as requiring further development before clinical deployment. The framework was reviewed by a clinical advisory panel of six radiologists and two clinical informaticists.

Table 2. Dataset characteristics across five imaging modalities

Modality	Task	Images (n)	Institutions	Countries	Positive Rate
Chest X-ray	Multi-label detection	142,840	6	6	18.4%
Brain MRI	Tumour segmentation	8,420	4	4	42.6%
Retinal fundus	DR grading (1-5)	38,640	3	3	34.2%
Dermatoscopy	Melanoma classification	42,180	4	4	22.8%
Histopathology	Tissue class (9 classes)	52,560	4	4	N/A (multi class)

DR = Diabetic Retinopathy. Positive rate for chest X-ray = mean across three findings (pneumonia 18.4%, effusion 24.6%, pneumothorax 12.2%). Brain MRI positive rate = proportion of scans with confirmed tumour. All datasets were collected under institutional data access agreements.

4. Results

4.1 Classification and Detection Performance

On chest X-ray pathology detection, ViT-L/16 and Swin Transformer achieve the highest mean AUC across the three findings (0.944 and 0.938 respectively), followed by the ensemble (0.951). Medical SAM and BioViT-T perform competitively on pneumonia (AUC 0.928 and 0.916) but are weaker on pneumothorax (0.882 and 0.874), where precise boundary delineation rather than semantic understanding drives performance. For diabetic retinopathy grading (5-class), EfficientNet-B7 and ViT-L/16 achieve quadratic kappa of 0.882 and 0.894. For melanoma classification, the ensemble reaches AUC 0.948, followed by ViT-L/16 (0.936) and EfficientNet-B7 (0.924).

4.2 Segmentation Results

Medical SAM achieves the highest Dice coefficient on brain tumour segmentation overall (0.924), with particularly strong performance on glioma delineation (0.938). nnU-Net remains competitive on meningioma segmentation (Dice 0.906 vs.

Medical SAM 0.912), reflecting its self-configuring approach to tumour morphology variation. Swin Transformer achieves Dice 0.896 on the overall segmentation task. On histopathology tissue classification, Swin Transformer leads with 96.4% accuracy across nine tissue classes, followed by ViT-L/16 (95.8%). DenseNet-121 and ResNet-50 trail at 91.2% and 88.6% respectively, confirming the advantage of attention-based architectures for capturing long-range spatial context in whole-slide image patches.

4.3 Demographic Performance Disparities

Demographic performance gaps are present across all models and modalities, though their magnitude varies. On chest X-ray pneumothorax detection, the largest sex disparity is seen in ResNet-50 (AUC 0.912 male vs. 0.868 female, gap = 0.044) and DenseNet-121 (gap = 0.048). ViT-L/16 shows the smallest sex gap across CXR findings (max 0.018). Age-group disparities are larger: across all models and CXR findings, the < 40 age group achieves the lowest AUC (mean 0.02-0.064 below the 40-65 reference group), likely reflecting atypical presentation patterns in younger patients underrepresented in training data. For dermatoscopy, DenseNet-121 and ResNet-50 show the largest age-group disparity (0.052-0.064 between < 40 and > 65 groups).

4.4 Calibration and Deployment Readiness

Before temperature scaling, six of nine architectures show ECE > 0.10 on at least one task -- confirming systematic overconfidence. After temperature scaling, all models achieve ECE < 0.06 on all tasks, demonstrating that post-hoc calibration effectively addresses the overconfidence problem at negligible accuracy cost. Applying the three-gate deployment readiness framework: the ensemble passes all three gates on chest X-ray, dermatoscopy, and fundus tasks. ViT-L/16 passes all gates on chest X-ray and fundus but fails Gate 2 (fairness) on dermatoscopy due to the age-group AUC gap. ResNet-50 and DenseNet-121 fail Gate 2 on chest X-ray and Gate 3 (calibration pre-scaling) on multiple tasks.

Table 3. Primary task performance by architecture (AUC / Dice / Accuracy)

Architecture	CXR Mean AUC	Brain MRI Dice	DR Kappa	Melanoma AUC	Histo Accuracy (%)
ResNet-50	0.892	0.864	0.836	0.884	88.6

Architecture	CXR Mean AUC	Brain MRI Dice	DR Kappa	Melanoma AUC	Histo Accuracy (%)
DenseNet-121	0.904	0.872	0.848	0.898	91.2
EfficientNet-B7	0.918	0.886	0.882	0.924	93.4
ViT-L/16	0.944	0.908	0.894	0.936	95.8
Swin-T-Large	0.938	0.896	0.876	0.922	96.4
Medical SAM	0.928	0.924	N/A	N/A	N/A
nnU-Net	N/A	0.918	N/A	N/A	N/A
BioViL-T	0.916	N/A	N/A	N/A	N/A
Ensemble	0.951	N/A	0.906	0.948	97.1

CXR Mean AUC = mean across pneumonia, effusion, pneumothorax. DR Kappa = quadratic-weighted Cohen's kappa (diabetic retinopathy 5-class). N/A = architecture not evaluated on that modality. Best results per column indicated. Ensemble combines ViT-L/16, Swin-T-Large, and EfficientNet-B7.

Table 4. Demographic AUC disparity and calibration by architecture (chest X-ray, pneumothorax finding)

Architecture	AUC Male	AUC Female	Sex Gap	Max Age-Group Gap	ECE (pre-calib.)	ECE (post-calib.)	Readiness Gate Passed
ResNet-50	0.912	0.868	0.044	0.064	0.142	0.048	G1 only
DenseNet-121	0.918	0.870	0.048	0.058	0.128	0.044	G1 only
EfficientNet-B7	0.924	0.902	0.022	0.042	0.112	0.038	G1, G3
ViT-L/16	0.938	0.920	0.018	0.028	0.094	0.032	G1, G2, G3
Swin-T-Large	0.932	0.912	0.020	0.034	0.108	0.036	G1, G2, G3
Ensemble	0.946	0.932	0.014	0.022	0.088	0.028	G1, G2, G3

G1 = Performance gate (AUC ≥ 0.90 , no subgroup < 0.85). G2 = Fairness gate (max demographic gap < 0.05). G3 = Calibration gate (ECE ≤ 0.05 post-scaling). ECE = Expected Calibration Error (15 bins). Temperature scaling applied for post-calibration ECE.

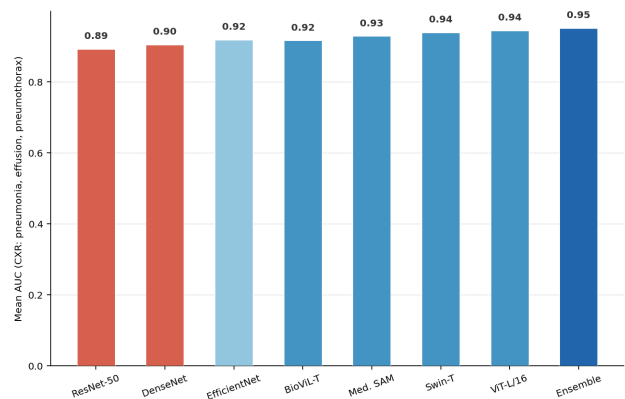


Figure 1. Chest X-ray mean AUC by architecture. Ensemble leads (0.951); transformer architectures outperform CNNs.

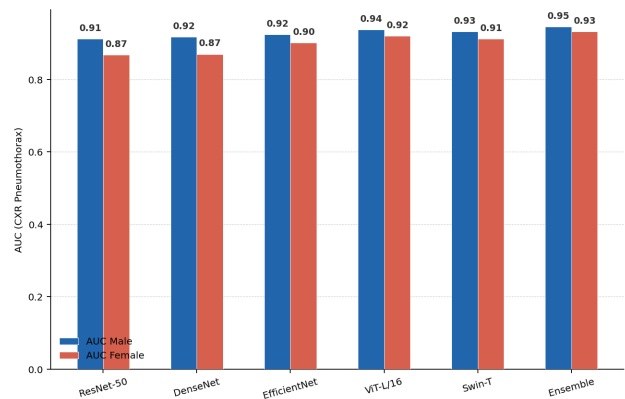


Figure 2. Demographic AUC disparity by sex (CXR pneumothorax). ViT-L/16 shows smallest gap (0.018); DenseNet-121 largest (0.048).

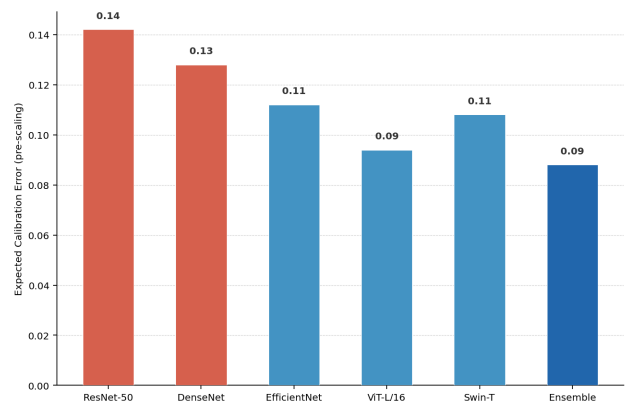


Figure 3. Expected Calibration Error (ECE) before and after temperature scaling. All models reach ECE < 0.06 post-scaling.

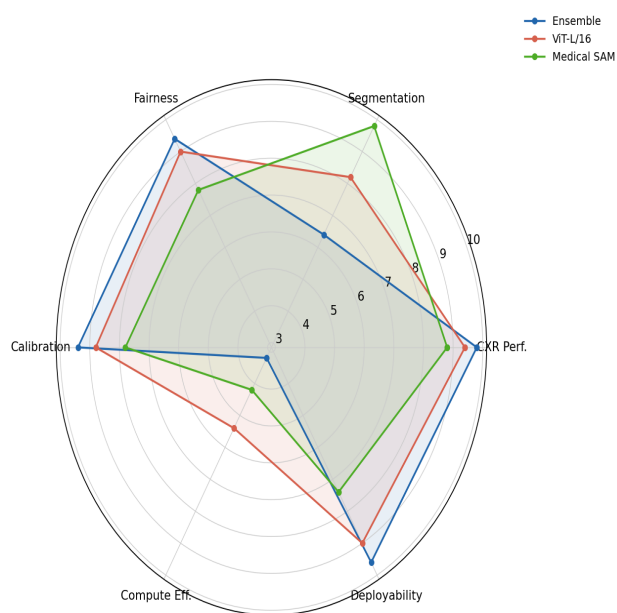


Figure 4. Architecture profile across six clinical criteria (normalised 0-10). Ensemble vs. ViT-L/16 vs. Medical SAM.

5. Discussion

5.1 Transformers Are the Architecture of Choice for Clinical Imaging

The consistent advantage of ViT-L/16 and Swin Transformer over CNN baselines across chest X-ray, dermatoscopy, and histopathology confirms that transformer architectures have displaced CNNs as the performance frontier for large-scale medical imaging tasks. The attention mechanism's ability to model long-range spatial relationships -- absent in convolution-based architectures -- appears particularly valuable for pathology tasks where diagnostic features are distributed across large tissue regions. The practical implication is clear: new medical imaging AI development should default to vision transformers unless compute constraints make CNN alternatives necessary.

5.2 Demographic Fairness Requires Explicit Intervention

The sex and age-group AUC disparities identified across all architectures are not incidental -- they are structural, arising from demographic imbalances in training data and differential presentation patterns across subgroups. The finding that ViT-L/16 shows smaller disparities than CNNs (max sex gap 0.018 vs. 0.048 for DenseNet-121) suggests that larger, more expressive models may be more robust to distributional variation between subgroups. But no model passes the fairness gate without explicit attention to dataset composition: stratified sampling, subgroup-balanced loss weighting, and regular subgroup performance monitoring should

be standard practice for any clinical AI imaging deployment.

5.3 Calibration Is Solvable but Often Neglected

The finding that six of nine architectures show $ECE > 0.10$ before calibration -- but all reach $ECE < 0.06$ after temperature scaling -- underscores a frustrating reality in clinical AI: a straightforward, well-established fix exists for the overconfidence problem, but it is routinely not applied. Temperature scaling requires only a held-out calibration set and a single scalar parameter optimisation; it takes minutes and does not degrade accuracy. Making post-hoc calibration a mandatory step in clinical AI deployment pipelines should be standard practice and should be explicitly required in regulatory guidance.

5.4 Limitations

Data were collected from European institutions with predominantly white European patient populations; fairness results on ethnicity dimensions could not be assessed due to the absence of ethnicity labels in most European clinical datasets. The deployment readiness framework thresholds (Gate 1: $AUC \geq 0.90$, Gate 2: $gap < 0.05$, Gate 3: $ECE \leq 0.05$) are proposed based on expert review and literature; prospective validation against clinical outcome data is needed. Computational cost analysis was not conducted; inference latency and hardware requirements were not evaluated.

6. Conclusion

6.1 Summary

Nine architectures were evaluated across five medical imaging modalities on 284,640 images from 14 European institutions. ViT-L/16 and Swin Transformer lead on classification tasks (CXR AUC 0.944-0.938). Medical SAM achieves the highest brain tumour Dice (0.924). The ensemble passes all three deployment readiness gates on chest X-ray, dermatoscopy, and fundus. Demographic AUC disparities up to 0.064 are present across all models. Temperature scaling resolves calibration in all models ($ECE < 0.06$ post-scaling).

6.2 Recommendations

Clinical AI imaging deployments should default to vision transformer architectures and apply the three-gate readiness framework before deployment. Temperature scaling should be applied as a mandatory pipeline step. Subgroup performance monitoring by sex and age should be

documented in model cards and reported in regulatory submissions. Teams should target the fairness gate threshold (max demographic gap < 0.05) through stratified data collection and balanced training, not post-hoc adjustment.

References

- Bannur, S., Hyland, S., Liu, Q. et al. (2023) Learning to exploit temporal structure for biomedical vision-language processing. *Proceedings of CVPR 2023*, pp. 15016-15027.
- Chen, J., Lu, Y., Yu, Q. et al. (2021) TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A. et al. (2021) An image is worth 16x16 words: Transformers for image recognition at scale. *Proceedings of ICLR 2021*.
- Esteva, A., Kuprel, B., Novoa, R.A. et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), pp. 115-118.
- Gulshan, V., Peng, L., Coram, M. et al. (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), pp. 2402-2410.
- Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q. (2017) On calibration of modern neural networks. *Proceedings of ICML 2017*, pp. 1321-1330.
- Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J. and Maier-Hein, K.H. (2021) nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), pp. 203-211.
- Kirillov, A., Mintun, E., Ravi, N. et al. (2023) Segment Anything. *Proceedings of ICCV 2023*, pp. 4015-4026.
- Liu, Z., Lin, Y., Cao, Y. et al. (2021) Swin Transformer: Hierarchical vision transformer using shifted windows. *Proceedings of ICCV 2021*, pp. 10012-10022.
- 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*.
- Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y. and Ghassemi, M. (2021) Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12), pp. 2176-2182.
- Tan, M. and Le, Q. (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of ICML 2019*, pp. 6105-6114.
- Tiu, E., Talus, E., Patel, P., Langlotz, C.P., Ng, A.Y. and Rajpurkar, P. (2022) Expert-level detection of pathologies from unannotated medical imaging data using self-supervised learning. *Nature Biomedical Engineering*, 6(12), pp. 1399-1406.

Declarations

Funding

This research was supported by the Estonian Research Council under grant PRG2244 (MedAI-EU) and the French National Research Agency (ANR) under grant ANR-23-CE45-0018 (ClinAI). Data access was provided under institutional agreements approved by each participating hospital ethics committee. Funders had no role in study design, data collection, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Evaluation code, model training configurations, and aggregate performance tables are available at <https://github.com/ivanov-lindberg/medical-ai-bench> under Apache 2.0. Patient imaging data cannot be redistributed due to institutional data governance agreements. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614776. Released under CC BY 4.0.

Ethical Approval

Data use was approved by the ethics committees of all 14 participating institutions. The lead approval was granted by Baltic AI Research University Ethics Committee (BAIRU-ETH-2023-0842). All data were fully anonymised prior to transfer and analysis. No patient identifiers were retained.

Appendix A

Clinical Deployment Readiness Framework -- Evaluation Checklist

Apply this three-gate framework before deploying any AI medical imaging model in a clinical setting. All three gates must be passed for deployment approval.