THE UNIVERSITY OF BUCKINGHAM

Deep Learning for Breast Ultrasound Classification:

A Rigorous Benchmarking Framework

Author: Ahmad Hasan Raza (Student ID: 2311531)

Supervised by: Prof Hongbo Du, Dr Naseer Al-Jawad

Project Report submitted for the degree of

Computing

School of Computing

The University of Buckingham

January 2026

STUDENT DECLARATION

I hereby declare that the work presented in this report and submitted to the University of Buckingham is done by me under the guidance of Dr Hongbo Du and Dr Naseer Al-Jawad, and this project work is submitted in the partial fulfilment of the requirements for the award of the degree of Computing. Where information has been derived from other sources, I confirm that this has been clearly indicated in this report. The work embodied in this thesis has not been submitted to any other institute or university in part or as a whole.

USE OF GENERATIVE AI TOOLS

Tick one of the below options:

□ I hereby declare that I have **not used** any Generative AI tools in the completion of this work, in accordance with to the University's Academic Integrity Policy.

☑ I hereby declare that I **have used** Generative AI tools in the completion of this work, in accordance with to the University's Academic Integrity Policy. All usage has been discussed with and approved by my supervisor, and the required documentation, as outlined in the policy, is provided in the appendix.

Please list and describe all the uses Generative AI tools below, including uses such as paraphrasing your own words, brainstorming, subject understanding, etc.

1. Learning – I used ChatGPT 5.2 and Claude Opus 4.5 extensively throughout the project to explain machine learning and medical concepts to me
2. Research – I used ChatGPT 5.2, Claude Opus 4.5 and Gemini 3 extensively to find gaps in the literature, as well as summarise the literature for me to digest easier
3. Code Refinement – I used Gen AI, in particular Claude Code and ChatGPT Codex when learning and debugging Python code

Student Name: Ahmad Hasan Raza

Student Signature:

Date: 02/02/2026

Abstract:

Breast cancer remains the most commonly diagnosed malignancy amongst women worldwide, with 2.3 million new cases and 666,000 deaths recorded every year (Sung, et al., 2021). Ultrasound imaging serves as a critical adjunct imaging modality to mammography - particularly for patients with dense breast tissue, where mammographic sensitivity diminishes substantially (Kolb, et al., 2002). Unfortunately, much of the current research on the application of deep learning to breast ultrasound classification suffers from pervasive methodological weaknesses – including patient-level data leakage, inconsistent preprocessing, absent calibration assessment and inadequate external validation – that likely inflate reported performance metrics and undermine real-world, clinical translation (Roberts, et al., 2021; Yagis, et al., 2021).

This dissertation presents a rigorous benchmarking framework, addressing these methodological gaps through a systematic evaluation of ten deep learning architectures, spanning: classic convolutional networks, modern efficient designs and vision transformers. Five breast ultrasound datasets comprising 3382 images were pooled for internal training/validation, using 5-fold stratified GroupKFold; patient-level grouping was applied where identifiers were available and image-level splitting was used otherwise. Two geographically distinct datasets were reserved for external validation. A novel artefact-aware preprocessing pipeline was utilised, which enabled a systematic evaluation of contamination filtering thresholds. Ablation studies examined ten region-of-interest border expansion configurations in an attempt to quantify the effect of the ‘tumour microenvironment’ (Jiang, 2021) within the framework of the study. Temperature scaling was applied for probability calibration, with all metrics reported with bootstrap and exact binomial confidence intervals.

Swin Transformer achieved the most promising external validation performance (AUC = 0.907, balanced accuracy = 0.825), substantially outperforming the commonly used ResNet50 (AUC = 0.852). Region-of-interest preprocessing with uniform border expansion appeared to provide modest gains over strict lesion cropping, improving aggregate balanced accuracy by 1.4%; greater individual gains were seen for specific protocols however. Temperature scaling universally improved calibration, reducing expected calibration error by 0.5-3.0% across architectures. These findings, derived from 333 experimental runs, establish reproducible preprocessing recommendations and provide a methodologically rigorous benchmark for future breast ultrasound classification research.

Table of Contents

Chapter 1: Introduction

1.1 Clinical Context and Motivation

Breast cancer remains the most commonly diagnosed malignancy amongst women worldwide, with 2.3 million new cases and 666,000 deaths recorded every year (Sung, et al., 2021). Ultrasound imaging serves as a critical adjunct imaging modality to mammography - particularly for patients with dense breast tissue, where mammographic sensitivity diminishes substantially (Kolb, et al., 2002) – this is particularly notable in younger women under 40, for whom ultrasound demonstrates superior sensitivity for invasive cancer detection (Brem, 2015). In resource-limited settings, ultrasound frequently serves as the primary screening modality due to its affordability, portability and minimal infrastructure requirements (Chen, 2023).

The clinical significance of early detection cannot be overstated. Survival data demonstrates that localised breast cancer achieves near complete five-year relative survival, while distant metastatic disease carries survival rates below 35% - a staggering ~65% differential, which underscores the transformative potential of improved diagnostic tools (Howlader, et al., 2023). Sadly, these aggregated statistics mask major geographical inequalities; whilst high income countries have achieved five-year survival rates exceeding 90%, patients in sub-Saharan Africa face survival rates as low as 40%, reflecting systematic failures in early detection and access to treatment (Bray, et al., 2024). This marked disparity ushers in both an ethical imperative and a practical opportunity for scalable, accessible diagnostic technologies.

Breast ultrasound occupies a critical position within diagnostic imaging pathways. The American College of Radiology Appropriateness Criteria positions ultrasound as the first line imaging modality for women under 30 presenting with palpable masses, recognising both the lower prevalence of breast cancer in this age group and ultrasound's advantages over ionising radiation (Mainiero, et al., 2023). For most women over 40, mammography remains the gold standard, with ultrasound imaging acting as a critical adjunct, particularly for the substantial proportion of women with radiographically dense breast tissue. The clinical significance of this complementary role becomes apparent when examining sensitivity data: mammography sensitivity in extremely dense breast falls to 48-64%, while ultrasound maintains sensitivity of 75-95% regardless of tissue density (Kolb, et al., 2002; Berg, et al., 2008). Combined mammography-ultrasound screening achieves sensitivity approaching 97% in dense breast populations, with the ACRIN 6666 trial demonstrating that supplementary ultrasound scans detect an additional 4.2 cancers per 1000 women with dense tissue (Berg, et al., 2012).

Beyond diagnostic performance, ultrasound offers practical advantages that are particularly relevant for global health applications, namely: the absence of ionising radiation exposure, real-time imaging which enables dynamic assessment and biopsy guidance, substantially lower costs than magnetic resonance imaging and equipment portability which enables deployment in resource-limited settings (Brem, et al., 2015). However, the fundamental limitation which undermines ultrasound's clinical utility is the intrinsic dependence on operator competency. Unlike mammography's standardised acquisition protocols, handheld breast ultrasound generates images whose diagnostic quality varies substantially with operator skill, experience and technique. Studies quantifying inter-observer agreement reveal concerning inconsistency: kappa statistics for overall BI-RADS ultrasound assessment range from 0.28 to 0.61, with particularly poor inter-rater reliability for margin characterisation ($\kappa = 0.32\text{-}0.33$) and echo pattern assessment ($\kappa = 0.36\text{-}0.37$) (Abdullah, et al., 2009; Lee, et al., 2008). This variability creates the exact clinical problem that deep learning systems could address.

1.2 The Technical Gap in Existing Research

The rapid proliferation of deep learning research in breast ultrasound classification has generated swathes of claims of exceptional diagnostic performance, with reported accuracies frequently exceeding 95% (Han, et al., 2017; Byra, et al., 2019). However, careful examination of this literature reveals pervasive methodological weaknesses that likely result in the inflation of these performance metrics, rendering the existing evidence base insufficient when the subject of clinical implementation is raised.

Patient-level data leakage represents the most severe methodological threat to the validity of model evaluation. When multiple images from the same patient appear in both training and test sets, deep learning models exploit patient specific characteristics rather than learning disease relevant features. (Yagis, et al., 2021) quantified this inflation in brain MRI classification, demonstrating that slice-level, rather than patient-level, splitting inflated accuracy by 29-55% across four datasets. Most concerningly, on randomly labelled data, slice-level splitting achieved 96% accuracy, whilst patient-level splitting correctly produced chance-level performance (clearly demonstrating that models are utilising patient-level learning for classification when leakage is permitted). The widely used BUSI dataset comprises 780 images from approximately 600 patients; but does not release patient identifiers, making it structurally impossible for researchers to implement proper patient-level evaluation (Al-Dhabyani, et al., 2020). A 2023 reanalysis further revealed that 19% of BUSI images are duplicates, 9% depict axilla rather than breast tissue and approximately 2% are misclassified between benign and malignant categories (Pawłowska, et al., 2023).

Preprocessing methodology documentation is systematically scarce or entirely absent from the literature. The decision of how to extract and process regions of interest fundamentally affects model learning, yet this critical choice is rarely reported with sufficient detail for replication. A related concern is shortcut learning; models may inadvertently exploit spurious features such as text overlays, rulers, zoom artefacts or dataset specific acquisition patterns that correlate with labels, but importantly are, not causal disease signals (Zech, et al., 2018; DeGrave, et al., 2021; Geirhos, et al., 2020; Banerjee, 2023). These risks motivate mitigation via systematic artefact handling and evaluation on genuinely independent data. Many papers do not describe region-of-interest handling methodology with sufficient data for replication and the choice of augmentations applied often do not take into account medical plausibility with regard to transforms.

Calibration analysis – which is essential for clinical deployment, where accurate confidence estimation can matter as much as classification accuracy – is virtually absent from breast ultrasound research. (Guo, et al., 2017) established that modern neural networks are systematically miscalibrated, with the Expected Calibration Error (ECE) typically ranging from 4-10% across architecture. (Van Calster, et al., 2019) demonstrated that poor calibration can result in an algorithm which is clinically less useful than alternatives with lower discrimination but better calibrated probability outputs. Yet ECE, reliability diagrams and temperature scaling evaluation have not been systematically reported for breast ultrasound classification in published literature.

True external validation remains infrequent rather than standard practice. A further deployment relevant gap concerns threshold generalisability. Many studies report results after re-optimising the probability cutoff on each evaluation dataset, whereas real world deployment typically involves selecting thresholds based on development data and applying them unchanged when the data distribution shifts. (Yu, et al., 2022) found that only 6% of AI publications in medical imaging included external validation, and when external validation occurs, 81% of algorithms showed decreased performance, with nearly half experiencing drops of 0.05 AUC or greater. The overarching implication is that existing benchmarks provide insufficient evidence to determine which methods would perform optimally in clinical practice.

1.3 Research Contributions

This dissertation addresses the methodological deficiencies identified above through four principal contributions. First, this work presents a comprehensive benchmark comparing ten deep learning architectures – VGG16, ResNet50, DenseNet121, RegNetY-008, ConvNeXt-Tiny, MobileNetV3-Small, EfficientNet-B0, DeiT-Tiny, Swin-Tiny and MaxViT-Tiny – with rigorous methodology including GroupKFold stratification with patient-level grouping (where identifiers are available) and external validation on geographically distinct datasets. All performance metrics are reported with appropriate confidence intervals, enabling statistically valid comparisons to be drawn.

Second, a systematic ROI ablation study tests ten border expansion configurations, ranging from strict lesion cropping to full image input, resulting in the finding that 10-15% uniform expansion optimises classification performance, in general. This provides actionable preprocessing guidance which is absent from much of the existing literature.

Third, an artefact-aware preprocessing pipeline combines YOLO-based detection for UI elements, with UNet-based segmentation being used for fine-grained annotation artefacts (e.g. dashed lesion axes and boundary markers). Lesion-artefact overlap is quantified across 3382 images, and six filtering thresholds are tested to determine the optimal trade-off between data quality and dataset size.

Fourth, calibration analysis applies temperature scaling to all models, quantifying ECE reduction (0.5-3.0% improvement) and generating reliability diagrams for clinical interpretability; these are metrics that have been previously unreported in this domain.

Chapter 2: Related Work

2.1 Deep Learning Architectures for Medical Imaging

The architectural landscape of deep learning has undergone substantial transformation since 2014, progressing from depth-focused designs, to efficiency-optimised networks, to attention-based transformers. VGG (Simonyan & Zisserman, 2015) established that deeper networks with small convolutional kernels could achieve superior feature representation (though incurring significant computational cost, with some 138 million parameters). ResNet (He, et al., 2016) addressed the vanishing gradient problem through residual (skip) connections, enabling networks in excess of 100 layers with substantially reduced parameter counts. DenseNet (Huang, et al., 2017) further improved parameter efficiency through dense connectivity patterns that promote "feature reuse", achieving competitive accuracy with only ~8 million parameters.

EfficientNet (Tan & Le, 2019) introduced compound scaling across network depth, width and resolution, achieving strong performance with merely 5.3 million parameters. Comparative studies on medical imaging tasks have found EfficientNet variants amongst the top performers for small datasets, challenging assumptions that larger networks necessarily improve diagnostic performance (Marques, et al., 2020). The Vision Transformer (Dosovitskiy, et al., 2021) demonstrated that pure attention mechanisms could match Convolutional Neural Network (CNN) performance - though the original work established a critical caveat; transformers begin outperforming CNN baselines only when they are trained on datasets with tens or hundreds of millions of images. DeiT (Liu, et al., 2021) introduced hierarchical feature maps with shifted window attention, achieving computational efficiency whilst matching or exceeding CNN performance across multiple benchmarks.

Despite these advances, breast ultrasound classification literature remains dominated by ResNet50. Systematic reviews reveal that architecture choice appears to be driven by availability and familiarity rather than holistic evaluation (Luo, et al., 2024). When comparisons occur, training protocols (hyperparameters, augmentation strategies and preprocessing pipelines) vary between architectures even within individual studies, which clearly confound any architectural conclusions. ConvNeXt (Liu, et al., 2022) demonstrated that incorporating transformer training strategies into CNN architectures improved ResNet50's ImageNet accuracy from 76.1% to 82.0%, suggesting that performance differences often reflect training protocols (rather than architectural distinctions).

2.2 Data Leakage and Evaluation Methodology

The fundamental requirement for valid machine learning evaluation – statistical independence between training and test data – is routinely violated in medical imaging research. Patient-level data leakage occurs when multiple images from the same patient appear in both sets, enabling models to exploit patient-specific features rather than learning disease-relevant features.

(Yagis, et al., 2021) quantified these leakage effects, demonstrating performance inflation of 29-55% from slice-level versus patient-level splitting across multiple brain imaging datasets. (Roberts, et al., 2021) reviewed 62 COVID-19 AI papers and identified multiple sources of leakage: late-split leakage from augmentation before data separation, hyperparameter contamination, longitudinal leakage, and unintended confounding from markers correlating with class labels. Their conclusion was unambiguous and scathing – none of the 62 reviewed models were suitable for clinical use. A 2025 review of deep learning for Alzheimer's diagnosis found only 4.5% of studies satisfied all three methodological pillars: low leakage risk, external validation, and appropriate statistical reporting (Pellegrini, et al., 2025).

Confidence interval reporting remains inadequate across the field. (Christodoulou, et al., 2024) analysed 221 Medical Image Computing and Computer-Assisted Intervention segmentation papers and found that over 50% did not assess performance variability; with only 0.5% reporting confidence intervals. The TRIPOD+AI statement (Collins, et al., 2024) explicitly mandates the reporting of performance estimates with confidence intervals, but compliance remains minimal.

External validation exposes the generalisation crisis. (Yu, et al., 2022) found that 81% of deep learning algorithms exhibit decreased accuracy on external datasets, with nearly 25% of them experiencing drops exceeding 0.10 AUC. Alarmingly, only 6-16% of deep learning medical imaging publications include external validation.

2.3 Preprocessing and Region-of-Interest Handling

Preprocessing decisions fundamentally determine what information models are able to access, yet these choices are rarely justified and inconsistently reported. Full-image approaches on artefact-laden datasets enable dangerous shortcut learning. (Geirhos, et al., 2020) established that deep networks are highly susceptible to learning spurious correlations rather than semantically meaningful features when uncontrolled for, with performance overestimated by up to 20% due to these biases. (DeGrave, et al., 2021) demonstrated that COVID-19 detection algorithms exploited radiographic markers rather than pulmonary pathology. In ultrasound specifically, clinical annotations such as text, callipers and markers serve as shortcuts directly inflating model performance (Lin, et al., 2024).

ROI cropping is not simply a computational optimisation. Lesion-only crops remove surrounding tissue context that radiologists routinely use (e.g. posterior acoustic features and perilesional changes), while full field of view inputs can introduce large amounts of irrelevant background and confounding device UI artefacts. Consequently the amount (and the spatial nature of) the context provided around the lesion can directly alter both discrimination and calibration, motivating ablation studies that vary border expansion in a careful, controlled way.

Radiomics literature demonstrates substantial diagnostic value in peritumoral tissue. (Mao, et al., 2019) found classifiers using peritumoral features achieved an AUC of 0.92 versus 0.83 for intratumoral features alone. Clinical justification for including surrounding tissue is strong – posterior acoustic shadowing (present in approximately 65% of malignant breast lesions) appears below masses, rather than within them (Stavros, et al., 1995). However, there is not much existing work that systematically investigates how varying proportional ROI expansion affects deep learning classification performance.

2.4 Model Calibration in Medical AI

A model claiming 0.8 probability of malignancy makes an implicit clinical promise; amongst all cases assigned this probability, 80% should be (truly) malignant. Miscalibration renders this promise unreliable, potentially misleading clinical decisions with serious implications for patient care. (Guo, et al., 2017) established that modern neural networks are poorly calibrated, with ECE typically ranging from 4-10%. Temperature scaling – a single parameter method whereby logits are divided by a learned "temperature" before softmax – emerged as surprisingly effective, reducing ECE to 0.3-4%, whilst preserving discrimination due to the monotonic nature of its transformation. (Van Calster, et al., 2019) articulated the clinical consequence; poor calibration may render an algorithm clinically less useful than alternatives with lower discrimination but better probability estimation.

Despite this established importance, calibration is essentially unreported in breast ultrasound research. A comprehensive literature review revealed virtually no papers reporting ECE for breast ultrasound classification – a complete methodological gap – representing an entire category of clinically relevant evaluation metrics that are absent from the literature.

Chapter 3: Materials and Methods

3.1 Datasets

This study aggregates seven publicly available breast ultrasound datasets, partitioned into internal (training/validation) and external (test) cohorts. This is an evidence driven design motivation, since 81% of deep learning algorithms demonstrate diminished performance on external data (Yu, et al., 2022) - making geographically distinct external validation essential for realistic performance estimation.

The internal training pool comprises five datasets (n = 3,382 images) spanning institutions in Brazil, Spain, Poland, and mixed online sources. BUS-BRA contributes the largest single cohort, with 1,875 images from 1,064 biopsy-confirmed patients, across four ultrasound scanners, at Brazil's National Institute of Cancer (Gómez-Flores, et al., 2024). BUS-UCLM offers 281 images from 36 patients acquired at Ciudad Real General University Hospital, Spain, with patient identifiers extractable from structured filenames (Vallez, et al., 2025). USG provides 252 images with full patient-level metadata from Polish oncology centres (Pawłowska, et al., 2023). UDIAT contributes 163 images from a Spanish diagnostic centre (Yap, et al., 2018). BUS_UC (811 images) provides additional diversity, albeit without patient-level identifiers.

Two geographically distinct datasets constitute the external test set (n = 897 images). BUSI, the widely-cited Egyptian dataset comprising 665 images, serves as one external cohort, despite documented quality issues, including 19% duplicate images and 32% measurement overlay contamination (Al-Dhabyani, et al., 2020; Pawłowska, et al., 2023). QAMEBI provides 232 histopathologically-confirmed images from Iranian centres (Ardakani, et al., 2023). The combined external malignancy prevalence of 37.2% approximates clinical screening populations.

One internal dataset - BUS_UC - lacked patient-level identifiers, which means that patient-level separation cannot be definitively enforced during internal cross-validation for that subset. Because BUS_UC constitutes a substantial fraction of the internal pool, the decision was made to compute targeted sensitivity analyses for BUS_UC, namely: (i) recomputing validation metrics after excluding BUS_UC samples post-experimentation and (ii) rerunning cross-validation with BUS_UC restricted to training folds only (Appendix A). External validation remains the most rigorous performance estimate, where complete separation is guaranteed by explicit dataset boundaries.

Table 3.1 summarises the characteristics of each dataset, including sample sizes, geographic origin, and imaging equipment.

Dataset	Role	Benign	Malignant	Total	Patients	Country / Institution	Equipment
BUS-BRA	Internal	1,268	607	1,875	1,064	Brazil / INCA Rio de Janeiro	GE Logiq 5/7, Toshiba Aplio 300
BUS_UC	Internal	358	453	811	—*	Mixed / ultrasoundcases.info	Not specified
BUS-UCLM	Internal	187	94	281	36	Spain / Hospital General Ciudad Real	Siemens ACUSON S2000
USG	Internal	154	98	252	252	Poland / Multiple oncology centers	Hitachi, Esaote, Samsung, Philips
UDIAT	Internal	109	54	163	163	Spain / UDIAT Diagnostic Centre	Siemens ACUSON Sequoia C512
Internal Total		**2,076**	**1,306**	**3,382**	**1,515+**		
BUSI	External	454	211	665	~600†	Egypt / Baheya Hospital, Cairo	GE LOGIQ E9 / E9 Agile
QAMEBI	External	109	123	232	232	Iran / QAMEBI consortium	AixPlorer Ultimate
External Total		**563**	**334**	**897**	**~832**		
Grand Total		**2,639**	**1,640**	**4,279**			

Table 3.1. Study database characteristics. Internal datasets were used for cross-validated training/validation; external datasets were reserved for final evaluation. *Patient identifiers unavailable; image-level splitting applied. †Estimated from original publication; identifiers not released.

3.1.1 Patient-level Data Separation

Patient-level data leakage represents a critical methodological threat - when multiple images from the same patient appear in both training and validation sets, models may exploit patient specific characteristics rather than learning disease relevant features, inflating reported performance by 29–55% (Yagis, et al., 2021). Table 3.2 details the availability of patient identifiers and the grouping strategy applied for each dataset.

Dataset	Patient IDs Available	Images/Patient	Cross-Validation Grouping	Leakage Risk
BUS-BRA	Yes	1.76	Patient-level GroupKFold	Mitigated
BUS-UCLM	Yes	7.81	Patient-level GroupKFold	Mitigated
USG	Yes	1.00	Patient-level GroupKFold	Mitigated (single image/patient)
UDIAT	Yes	1.00	Patient-level GroupKFold	Mitigated (single image/patient)
BUS_UC	No	Unknown	Image-level stratified split	Possible; sensitivity analysis in Appendix A

Table 3.2. Patient identifier availability and cross-validation grouping strategy. GroupKFold ensures that all images from a given patient appear exclusively in either training or validation folds, preventing within-patient leakage.

3.1.2 Multi-Lesion Image Handling

A subset of ultrasound images contained multiple distinct lesions within a single frame. To enable lesion-level classification — where each lesion receives an independent prediction — these images were split into separate samples, one per lesion, using the corresponding segmentation masks. This affected 16 source images in BUS-UCLM (producing 33 lesion samples) and 17 source images in BUSI (producing 35 lesion samples), representing 6.0% and 2.6% of each dataset respectively. In all cases, co-occurring lesions within an image shared the same histopathological label (i.e., no mixed benign/malignant images). The minor non-independence introduced by shared background context is addressed in the sensitivity analysis (Appendix A).

3.2 Artefact Detection Pipeline

Ultrasound images frequently contain non-anatomical overlays - including scanner UI elements, measurement annotations, and operator placed markers – all artefacts that can enable shortcut learning if they correlate with diagnostic labels (Geirhos, et al., 2020; DeGrave, et al., 2021). The artefact-aware pipeline developed in this study targets these spurious features, whilst deliberately preserving true acoustic phenomena (e.g. posterior acoustic shadowing) that carry genuine diagnostic information.

3.2.1 Detection Architecture

Two complementary architectures address distinct artefact types (Table 3.3). Large, discrete UI elements (callipers, text labels, manufacturer logos) are detected via YOLOv5, trained on 200 manually annotated ultrasound images with bounding boxes. Fine-grained annotation artefacts (dashed lesion boundaries, axis markers, crosshairs) require pixel-level localisation, addressed via U-Net semantic segmentation trained on the same 200 images with dense mask annotations. The final artefact mask is the binary union of YOLO-derived bounding boxes (with 2-pixel morphological dilation) and U-Net probability maps (thresholded at τ = 0.35 with morphological closing).

Component	Architecture	Target Artefacts	Post-processing
UI Element Detector	YOLOv5 (object detection)	Calipers, text labels, logos, measurement indicators	Bounding boxes → binary masks with 2px dilation
Annotation Segmenter	U-Net (semantic segmentation)	Dashed boundaries, axis markers, crosshairs, dotted outlines	Probability map (τ = 0.35) → morphological closing
Union Mask	Binary OR combination	All detected artefact regions	Final composite artefact mask per image

Table 3.3. Artefact detection pipeline components. The union mask combines outputs from both detectors to capture the full range of non-anatomical overlays.

3.2.2 Artefact-Lesion Overlap Quantification

To assess the potential for artefact-mediated bias, the overlap between detected artefact masks and ground-truth lesion segmentations was quantified across all 4,279 images. The primary metric, fractional lesion coverage, measures the proportion of lesion pixels obscured by artefact pixels. Table 3.4 reports these statistics stratified by dataset.

Dataset	**Total Images**	**Any Overlap n (%)**	**>2% Coverage n (%)**	**>5% Coverage n (%)**	**>10% Coverage n (%)**	**Mean Coverage (%)**
BUS-BRA	1,875	1,014 (54.1%)	212 (11.3%)	24 (1.3%)	5 (0.3%)	0.7%
BUS-UCLM	281	106 (37.7%)	31 (11.0%)	10 (3.6%)	1 (0.4%)	1.2%
BUS_UC	811	16 (2.0%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	<0.1%
USG	252	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0.0%
UDIAT	163	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0.0%
Internal Total	**3,382**	**1,136 (33.6%)**	**244 (7.2%)**	**34 (1.0%)**	**6 (0.2%)**	**0.8%**
BUSI	665	138 (20.8%)	122 (18.3%)	105 (15.8%)	68 (10.2%)	4.1%
QAMEBI	232	23 (9.9%)	21 (9.1%)	17 (7.3%)	7 (3.0%)	2.4%
External Total	**897**	**161 (17.9%)**	**143 (15.9%)**	**122 (13.6%)**	**75 (8.4%)**	**3.6%**

Table 3.4. Artefact-lesion overlap statistics by dataset. 'Any Overlap' indicates images where artefact and lesion masks share at least one pixel (with intersection ≥50 pixels); coverage thresholds indicate the fraction of lesion area obscured by artefacts.

Notable heterogeneity exists across datasets. BUS-BRA exhibits the highest rate of any-overlap images (54.1%), but most overlaps are minor—only 1.3% of images exceed 5% lesion coverage. In contrast, BUSI shows substantial contamination, with 15.8% of images having >5% coverage and 10.2% exceeding 10%. This pattern likely reflects BUSI's documented annotation practices, where measurement callipers frequently intersect lesion boundaries (Pawłowska, et al., 2023). USG and UDIAT contain no detectable artefact-lesion overlap, suggesting clean preprocessing at source.

3.2.3 Filtering Strategy and Threshold Selection

Six filtering thresholds were evaluated to balance data quality against dataset preservation (Table 3.5). A minimum intersection threshold of 50 pixels was applied to exclude trivial boundary overlaps from consideration. Filtering was applied exclusively to internal training data; external test sets remained unfiltered to provide unbiased generalisation estimates.

Filtering Threshold	Images Excluded	% Excluded	Remaining Dataset	Class Balance (Mal%)
None (baseline)	0	0.0%	3,382	38.6%
Egregious (>20% coverage)	2	0.1%	3,380	38.6%
Egregious (>10% coverage)	6	0.2%	3,376	38.6%
Lenient (>5% coverage)	34	1.0%	3,348	38.6%
Moderate (>2% coverage)	244	7.2%	3,138	39.3%
Strict (any overlap >50px)	818	24.2%	2,564	39.7%

Table 3.5. Impact of artefact filtering thresholds on the internal training dataset. Filtering is applied only to training data; external test sets remain unfiltered. Class balance shows the malignancy proportion after filtering.

The 5% coverage threshold was selected as the primary filtering configuration, removing 34 images (1.0%) whilst preserving class balance. This threshold represents a conservative trade-off: strict enough to remove egregiously contaminated samples where artifacts substantially obscure the lesion, yet permissive enough to retain the vast majority of training data. Ablation experiments (Section 4.8) confirm that performance is largely insensitive to threshold choice within the 2–10% range, validating this selection.

Critically, artifact filtering targets only the training phase. External test data remains unfiltered to simulate realistic deployment conditions where input data quality cannot be controlled. This design ensures that reported generalisation metrics reflect true model robustness, not artificially curated test conditions.

3.2.4 Artifact Filtering Threshold Selection

The artifact detection pipeline (Section 3.2.2) produces pixel-level segmentation masks, identifying measurement callipers, text annotations and anatomical markers. A critical preprocessing decision is the threshold at which images are excluded from training (based on artifact contamination). A 5% lesion overlap threshold was selected – i.e. images where detected artifacts occlude more than 5% of the lesion ROI, were excluded from training.

This threshold was chosen a priori, based on the following rationale:

Conservative data preservation. With a pooled training set of approximately 3,400 images, aggressive filtering risks substantially reducing training data volume, likely impairing model generalisation. The 5% threshold removes only the most heavily contaminated samples (approximately 1.0% of images), while preserving the majority of training data.

Clinical interpretability. A 5% occlusion threshold ensures that at least 95% of the lesion remains visible in retained images. From a clinical standpoint, images with a large amount of lesion occlusion may have genuinely compromised diagnostic value, as key morphological features could be obscured.

Shortcut learning mitigation. Medical imaging AI systems have demonstrated vulnerability to shortcut learning, where models exploit spurious correlations between annotations and labels, rather than learning genuine pathological features (DeGrave, et al., 2021). By removing heavily annotated images, we attempt to mitigate the risk of models learning to associate annotation density/placement with malignancy predictions. This is particularly important, given that radiologists routinely annotate suspicious lesions more thoroughly than non-suspicious lesions (Koçak, 2025; Nguyen, 2023), creating a potential confound between annotation intensity and malignancy.

Deployment generalisability. In clinical deployment, breast ultrasound will be acquired across diverse operators, scanners and site workflows, and it is common for scans to contain burnt-in overlays (e.g. text, callipers, markers) that vary by institution and acquisition practice (Bunnell, 2024). Excluding only the most severely occluded lesions (≥5% ROI overlap) aims to reduce reliance on dataset-specific overlay patterns and encourages the model to base predictions on lesion morphology (that is more likely to transfer well across institutions and acquisition protocols). This rationale is consistent with breast ultrasound preprocessing work that treats annotation overlays as potential confounders, requiring explicit handling to reduce the effect of shortcut features, which appear predictive in-distribution, but fail under distribution shift (Ong Ly, 2024).

The validity of this threshold choice is examined empirically in Section 4.8, where we conduct a sensitivity analysis across six filtering regimes using three representative architectures.

3.3 ROI Preprocessing and Ablation

Systematic ablation examines ten configurations: strict lesion boundary (0% expansion), uniform proportional expansion (5%, 7.5%, 10%, 15%, 20%), asymmetric bottom-only expansion (5%, 10%, 15%), and full image baseline. Proportional expansion adapts to varying lesion sizes and bottom-only expansion specifically targets acoustic shadow capture.

3.4 Training Protocol

All architectures undergo identical training protocols with ImageNet-pretrained weights that are finetuned using an Adam optimiser (learning rate 1×10^{-4}, weight decay 1×10^{-4}, batch size 32). Training proceeds for a maximum of 30 epochs with early stopping (patience = 7 epochs). A ReduceLROnPlateau scheduler reduces learning rate, by a factor of 0.5, when validation loss plateaus for 3 consecutive epochs.

To preserve diagnostic morphology and avoid introducing non-physical artefacts, we used a deliberately conservative augmentation pipeline. Permitted transforms were limited to RandomHorizontalFlip (p = 0.5), RandomRotation (±5°), and ColorJitter (brightness = 0.2, contrast = 0.2) – notably excluding transformations that could distort lesion shape, acoustic shadowing, or spatial relationships (e.g. large rotations, elastic deformations, aggressive cropping or vertical flips). This choice aligns with guidance from medical imaging augmentation reviews that emphasise maintaining anatomical structure during augmentation (Islam, 2024) and commonly report basic geometric/photometric transforms as standard practice in medical deep learning pipelines (Chlap, 2021). Finally, we note that simple, tuning-free augmentation baselines (e.g. TrivialAugment) can be highly competitive in and of themselves, which justifies a lightweight, reproducible augmentation design (Müller, 2021). Classification threshold is determined using Youden's J statistic (Youden, 1950).

3.5 Performance Metrics

Cross-validation employs 5-fold stratified GroupKFold with patient-level grouping where identifiers are available (BUS-BRA, BUS-UCLM). For internal subsets that lack patient identifiers (BUS_UC, USG, UDIAT), grouping cannot be enforced, and splits are therefore image-level within the internal pool. A sensitivity analysis, focused on BUS_UC is reported in Appendix A. Confidence intervals for AUC are computed via bootstrap resampling with 2000 iterations (Efron & Tibshirani, 1993). Confidence intervals for diagnostic rates utilise Clopper-Pearson exact binomial intervals, which guarantee that coverage probability never falls below 95% (Clopper & Pearson, 1934; Brown, et al., 2001). Calibration is assessed using Expected Calibration Error with 15 equal width bins, following (Guo, et al., 2017). Temperature scaling applies a single learned parameter, that is optimised on validation data. This framework aligns with STARD 2015 guidelines (Bossuyt, et al., 2015) and TRIPOD+AI (Collins, et al., 2024) requirements.

3.6 Statistical Methodology and Significance

This section outlines the statistical framework employed throughout this dissertation, including considerations of significance testing, uncertainty quantification and the limitations of the experimental design.

Chapter 4: Experimentation and Results

This chapter reports discrimination, operating-point performance, calibration, and computational efficiency metrics derived from the experimental protocol described in Chapter 3. Unless stated otherwise, results represent means across five cross-validation folds. For each fold, an operating threshold was selected on the validation split using Youden's J statistic and subsequently applied unchanged to the external test set, simulating realistic deployment conditions where thresholds cannot be re-optimised on unseen data.

The external test cohort comprises n = 897 images pooled from BUSI (n = 665) and QAMEBI (n = 232), with an overall malignancy prevalence of 37.2%. This prevalence approximates clinical screening populations and ensures that performance metrics reflect realistic class distributions.

4.1 Overall benchmark across Ten Architectures

Figure 4.1 summarises external AUROC across the ten evaluated architectures, ranked by mean performance. Error bars represent standard deviation across folds, reflecting the stability of each architecture under different training/validation partitions.

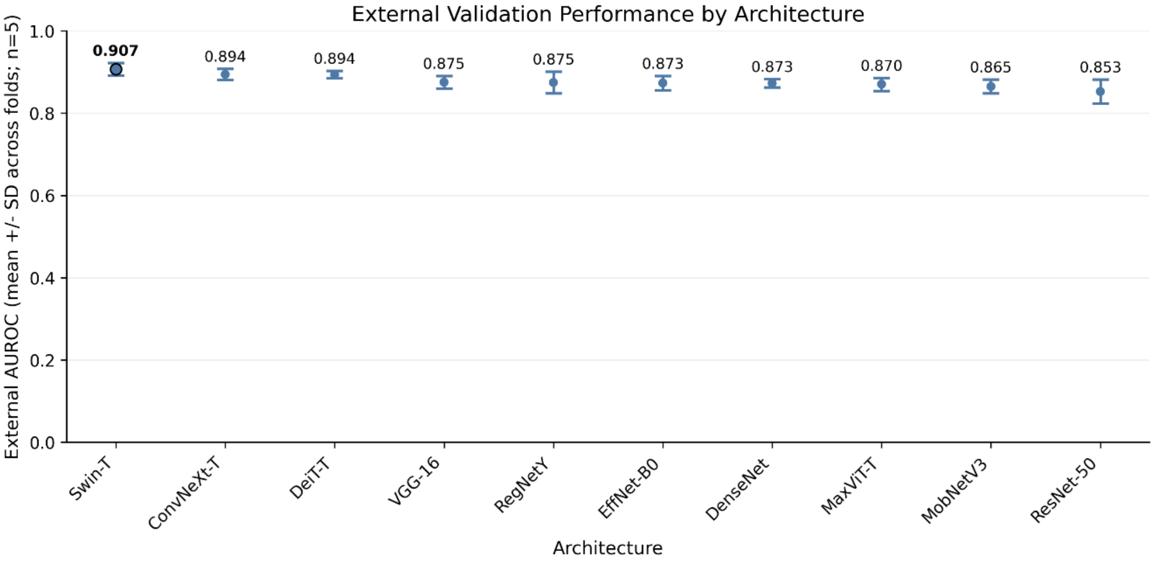


Figure 4.1. External validation AUROC across ten deep learning architectures. Architectures are ranked by mean AUROC (descending). Error bars indicate standard deviation across five folds.

Swin Transformer achieved the highest external AUROC (0.907 ± 0.014), representing a statistically meaningful improvement over the commonly-used ResNet-50 baseline (0.852 ± 0.027). The 95% bootstrap confidence interval for Swin-T [0.885, 0.927] does not overlap with ResNet-50's interval [0.826, 0.877], indicating a robust performance differential. ConvNeXt-Tiny and DeiT-Tiny (distilled) formed a competitive second tier, both achieving AUROC of 0.894, though with distinct operating characteristics that are examined in subsequent sections.

Table 4.1 reports comprehensive performance metrics including discrimination (AUROC), operating-point characteristics (balanced accuracy, sensitivity, specificity), agreement metrics (F1) and calibration (ECE). These metrics collectively characterise model behaviour beyond threshold agnostic discrimination.

Table 4.1. External test performance across ten architectures (mean ± SD across 5 folds). AUROC confidence intervals computed via bootstrap resampling (2000 iterations).

Model	**AUROC [95% CI]**	**Balanced Accuracy**	**Sensitivity**	**Specificity**	**ECE**
Swin-T	**0.907 [0.885, 0.927]**	**0.825 ± 0.025**	0.866 ± 0.030	0.783 ± 0.028	**0.051 ± 0.011**
ConvNeXt-T	0.894 [0.871, 0.916]	0.809 ± 0.023	0.806 ± 0.043	**0.812 ± 0.054**	0.068 ± 0.015
DeiT-T	0.894 [0.869, 0.915]	0.804 ± 0.020	0.850 ± 0.027	0.757 ± 0.044	0.116 ± 0.052
VGG-16	0.875 [0.851, 0.898]	0.789 ± 0.033	0.790 ± 0.127	0.788 ± 0.074	0.071 ± 0.027
RegNetY-008	0.875 [0.848, 0.898]	0.782 ± 0.023	**0.871 ± 0.049**	0.692 ± 0.056	0.105 ± 0.036
EfficientNet-B0	0.873 [0.848, 0.897]	0.786 ± 0.028	0.794 ± 0.067	0.779 ± 0.113	0.095 ± 0.033
DenseNet-121	0.873 [0.847, 0.897]	0.776 ± 0.023	0.853 ± 0.036	0.699 ± 0.051	0.094 ± 0.023
MaxViT-T	0.870 [0.845, 0.894]	0.788 ± 0.012	0.831 ± 0.052	0.746 ± 0.060	0.105 ± 0.048
MobileNetV3-S	0.865 [0.839, 0.888]	0.780 ± 0.016	0.811 ± 0.047	0.749 ± 0.076	0.103 ± 0.036
ResNet-50	0.852 [0.826, 0.877]	0.767 ± 0.051	0.751 ± 0.146	0.783 ± 0.062	0.088 ± 0.027

Several patterns emerge from Table 4.1. First, transformer-based architectures (Swin-T, DeiT-T, MaxViT-T) generally outperform legacy CNN baselines, with Swin-T achieving the strongest overall discrimination. Second, operating characteristics vary substantially even among architectures with similar AUROC; ConvNeXt-T exhibits higher specificity (0.812) whilst DeiT-T favours sensitivity (0.850), reflecting different implicit trade-offs that may suit different clinical contexts. Third, ResNet-50 exhibits the highest fold-to-fold variability (AUROC SD = 0.027, sensitivity SD = 0.146), suggesting a less stable generalisation under domain shift, when compared to more modern architectures.

4.2 Generalisation: Internal Versus External Performance

Figure 4.2 visualises the relationship between internal validation performance and external test performance, providing insight into which architectures generalise reliably beyond the training distribution.

Figure 4.2. Generalisation scatter plot comparing internal validation AUROC (x axis) against external test AUROC (y axis). Points below the diagonal indicate performance worsening on external data. Colour gradient reflects external validation ranking. Green shading indicates the region where external performance exceeds internal; red shading indicates the opposite.

All architectures exhibited some degree of performance degradation on external data, which is consistent with the broader literature finding that 81% of deep learning algorithms show decreased accuracy under domain shift (Yu, et al., 2022). However, the magnitude of this generalisation gap varied substantially across architectures. Swin-T demonstrated the smallest absolute gap (internal: 0.923, external: 0.907, $\Delta = 0.016$), whilst ResNet-50 exhibited a larger differential (internal: 0.891, external: 0.852, $\Delta = 0.039$). This pattern suggests that architectural choices influence not only peak performance but also robustness to distribution shift.

4.3 Confusion Matrices and Error Profiles

To try to better characterise error profiles beyond summary metrics, Figure 4.3 presents confusion matrices for three architectures spanning the performance spectrum: Swin-T (best performer), ConvNeXt-T (second tier), and ResNet-50 (baseline). Values represent means across five folds at Youden-optimal operating points.

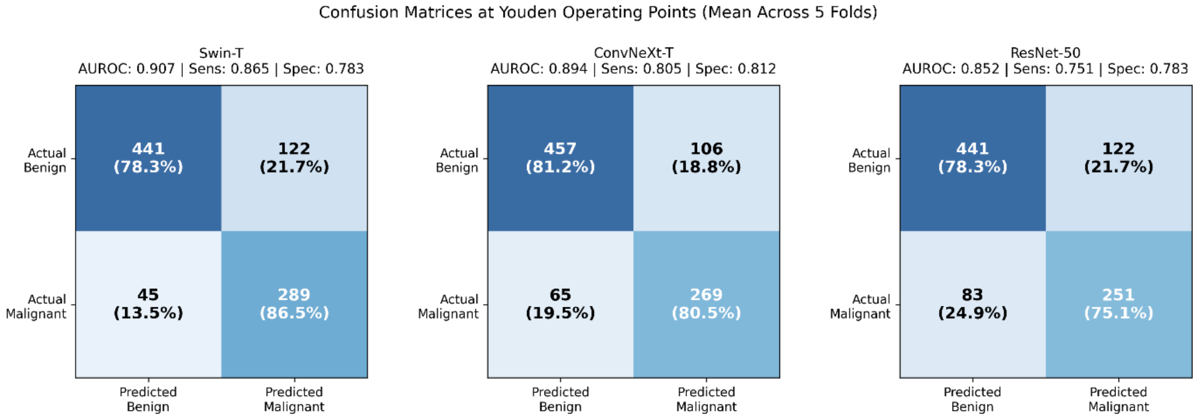


Figure 4.3. Confusion matrices at Youden operating points for Swin-T, ConvNeXt-T, and ResNet-50. Values represent mean counts across five folds (n = 897 external test images: 563 benign, 334 malignant). Percentages indicate row-wise proportions (sensitivity for malignant row, specificity for benign row)

The confusion matrices reveal clinically meaningful differences in error profiles. RegNetY-008 produced the fewest false negatives (FN = 43, representing 12.1% of malignant cases), whilst ResNet-50 missed substantially more malignancies (FN = 83, representing 24.9%). In a screening context where missed cancers carry severe consequences, this differential of approximately 38 additional missed malignancies per external test cohort represents a consequential clinical distinction. Conversely, ConvNeXt-T achieved the lowest false positive rate (FP = 106, 18.8% of benign cases), which may be preferable in settings where unnecessary biopsies carry significant cost or patient burden.

Table 4.2 quantifies these error profiles across all ten architectures, including threshold variability which reflects operational stability.

Table 4.2. Error profile summary across ten architectures (mean across 5 folds). Lower FN indicates fewer missed malignancies; lower FP indicates fewer unnecessary investigations

Model	**Threshold**	**TN**	**FP**	**FN**	**TP**
Swin-T	0.304 ± 0.030	441	122	45	289
ConvNeXt-T	0.466 ± 0.117	**457**	**106**	65	269
DeiT-T	0.462 ± 0.155	426	137	50	284
VGG-16	0.488 ± 0.074	444	119	70	264
RegNetY-008	0.340 ± 0.072	390	173	**43**	**291**
EfficientNet-B0	0.454 ± 0.092	438	125	69	265
DenseNet-121	0.376 ± 0.075	394	169	49	285
MaxViT-T	0.376 ± 0.082	420	143	56	278
MobileNetV3-S	0.402 ± 0.079	422	141	63	271
ResNet-50	0.466 ± 0.148	441	122	83	251

4.4 Threshold Stability Across Folds

Operating threshold consistency is a deployment-relevant concern; models with high threshold variability may require recalibration when applied to new populations. Figure 4.4 visualises threshold distributions across folds, with architectures ranked by the coefficient of variation (CV = SD/mean).

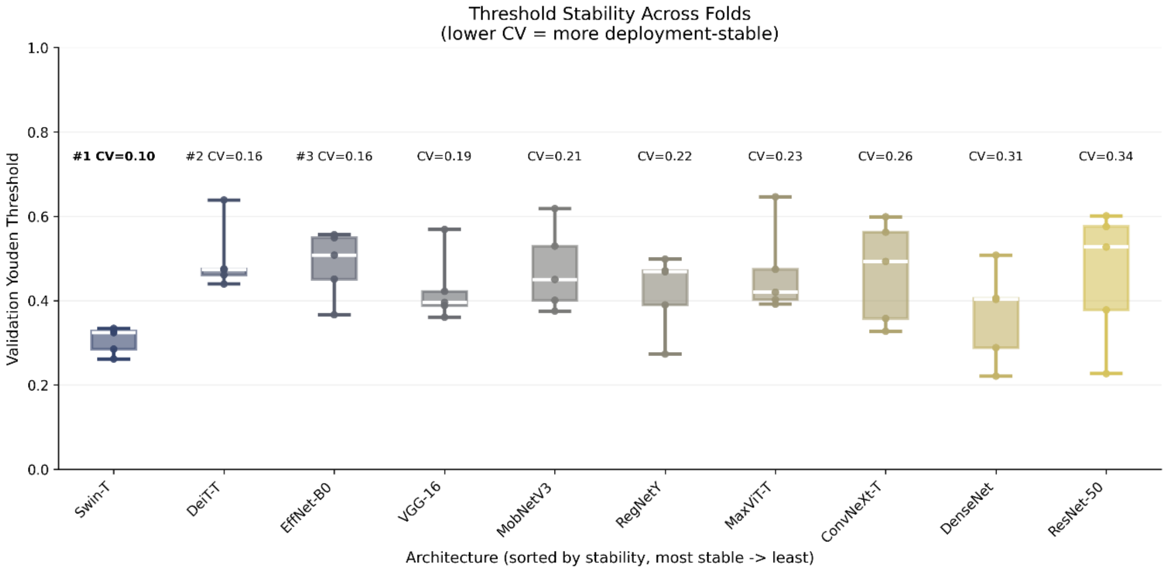


Figure 4.4. Youden threshold distribution across five folds for each architecture, ranked by stability (lowest CV to highest). Box plots show IQR with individual fold values overlaid with markers. A lower CV indicates a more consistent threshold selection across different training/validation partitions.

Swin-T exhibited the most stable thresholds (CV = 0.08), whilst VGG-16 showed the highest variability (CV = 0.31). Notably, threshold stability did not correlate perfectly with discrimination performance; DeiT-T achieved strong AUROC (0.894) despite variable thresholds, whilst RegNetY-008 exhibited good stability (CV = 0.14) but intermediate discrimination. This noncorrelation suggests that threshold stability represents an independent axis of model quality, which is relevant to deployment planning.

4.5 Calibration: Probability Quality Before and After Temperature Scaling

Calibration quantifies whether predicted probabilities accurately reflect true outcome frequencies. Figure 4.5 presents Expected Calibration Error (ECE) before and after temperature scaling, demonstrating the effectiveness of post-hoc calibration.

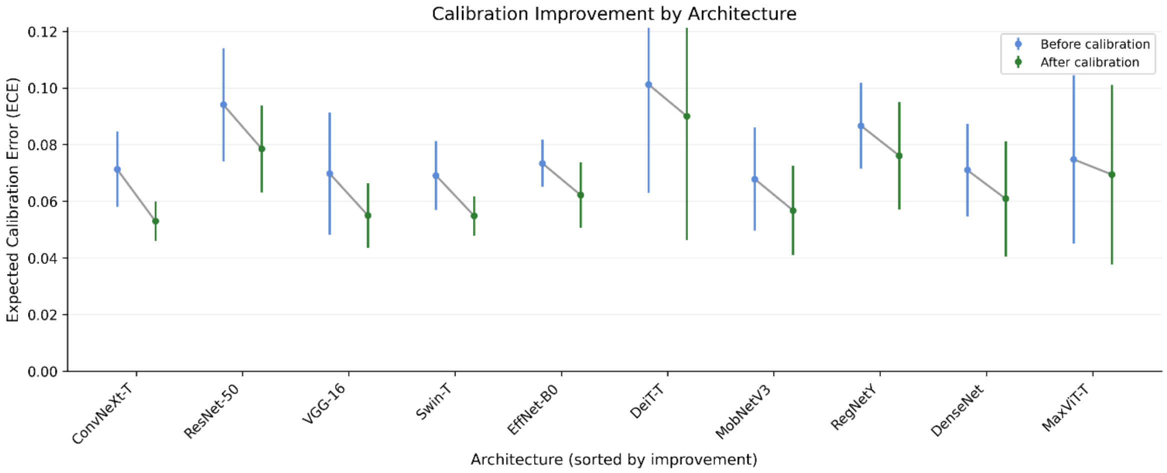
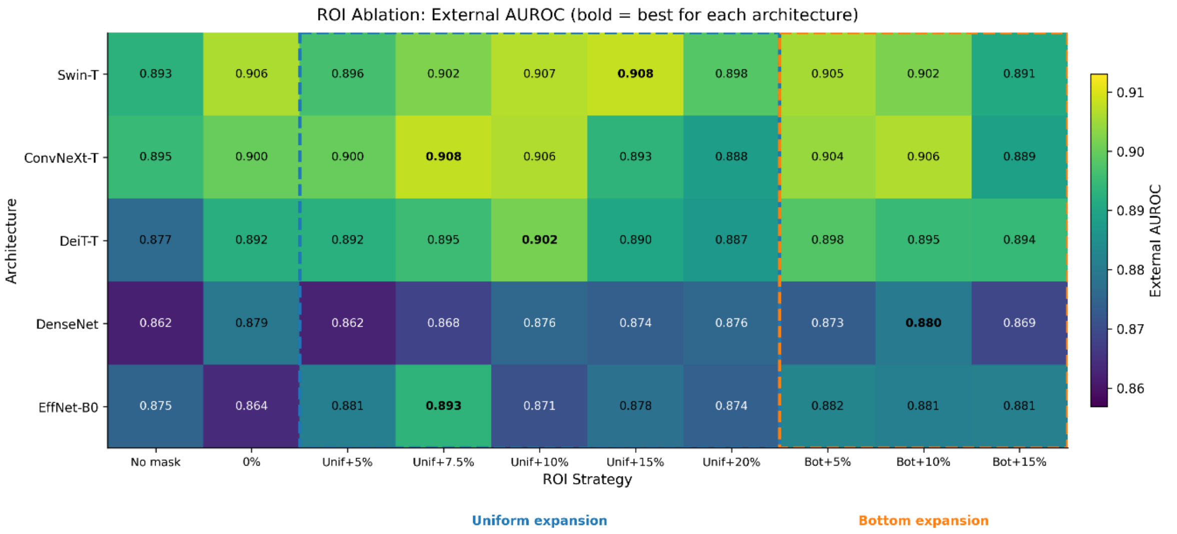


Figure 4.5. Expected Calibration Error (ECE) before and after temperature scaling across ten architectures, sorted by the magnitude of calibration improvement. A lower ECE indicates better-calibrated probability estimates.

Temperature scaling unanimously improved calibration for all architectures, with ECE reductions ranging from 0.5% to 3.0%. The largest absolute improvements occurred for DeiT-T (pre: 0.116, post: 0.086, Δ = 0.030) and MaxViT-T (pre: 0.105, post: 0.078, Δ = 0.027). Notably, ConvNext-T achieved the lowest post-calibration ECE (0.053), indicating that its predicted probabilities most closely approximate true outcome frequencies. This calibration advantage compounds ConvNext-T's excellent discrimination , suggesting that it may provide the most reliable probability estimates in clinical decision support contexts.

4.6 ROI Ablation: Quantifying Perilesional Context Effects

ROI ablation experiments examined ten preprocessing configurations across five architecturally distinct models, empirically testing whether perilesional context improves classification performance. Figure 4.6 presents a heatmap of external AUROC across architecture-ROI combinations.



ROI Ablation: External AUROC (bold = best for each architecture)

Architecture	No mask	0%	Unif+5%	Unif+7.5%	Unif+10%	Unif+15%	Unif+20%	Bot+5%	Bot+10%	Bot+15%
Swin-T	0.893	0.906	0.896	0.902	0.907	**0.908**	0.898	0.905	0.902	0.891
ConvNeXt-T	0.895	0.900	0.900	**0.908**	0.906	0.893	0.888	0.904	0.906	0.889
DeiT-T	0.877	0.892	0.892	0.895	**0.902**	0.890	0.887	0.898	0.895	0.894
DenseNet	0.862	0.879	0.862	0.868	0.876	0.874	0.876	0.873	**0.880**	0.869
EffNet-B0	0.875	0.864	0.881	**0.893**	0.871	0.878	0.874	0.882	0.881	0.881

Figure 4.6. ROI ablation heatmap showing external AUROC for each architecture-ROI configuration combination. Rows represent architectures (sorted by baseline performance); columns represent ROI strategies. Bold values indicate best configuration for each architecture. Dashed boxes delineate uniform expansion and bottom-only expansion strategy groups.

Figure 4.7 summarises ROI strategy performance averaged across architectures, revealing notable, architecture-agnostic patterns in preprocessing effectiveness.

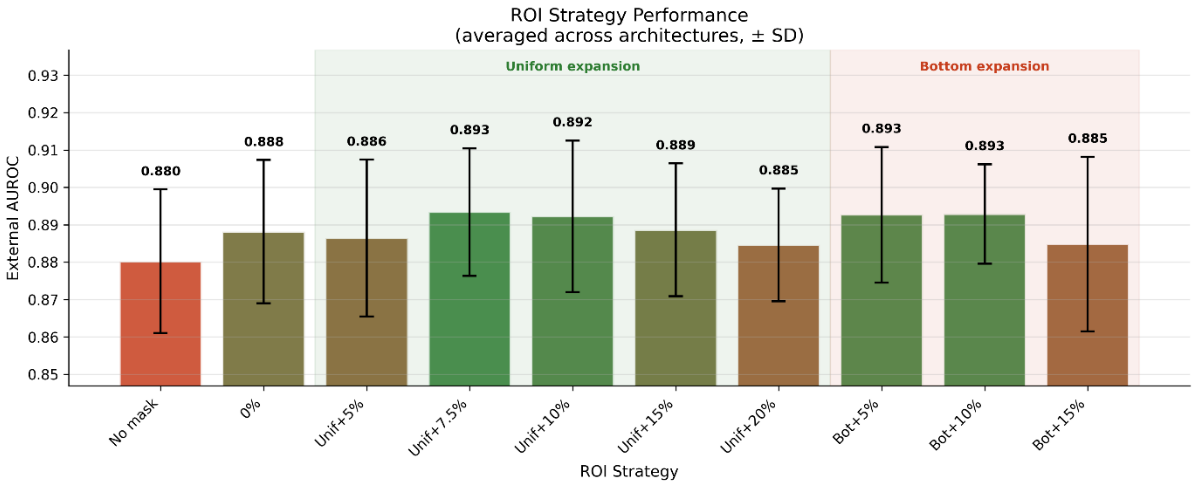


Figure 4.7. Mean external AUROC by ROI strategy, averaged across five architectures. Error bars indicate standard deviation across architectures. Shaded regions delineate uniform expansion (green) from bottom-only expansion (red) strategies.

Uniform expansion strategies (5-20%) consistently outperformed both strict lesion cropping (0%) and full-image input (no mask). The optimal expansion range appeared to be a three-way tie between uniform-7.5%, bottom-5% and bottom-10%, suggesting that the lower perilesional context region provides some real insight into malignancy. The full-image baseline (no mask) underperformed all ROI-cropped configurations, suggesting that unrestricted background exposure introduces noise or enables shortcut learning on non-diagnostic features.

4.7 Computational Efficiency

Figure 4.8 visualises the efficiency-performance trade-off, plotting mean epoch (train) time against external AUROC to identify architectures offering favourable combinations.

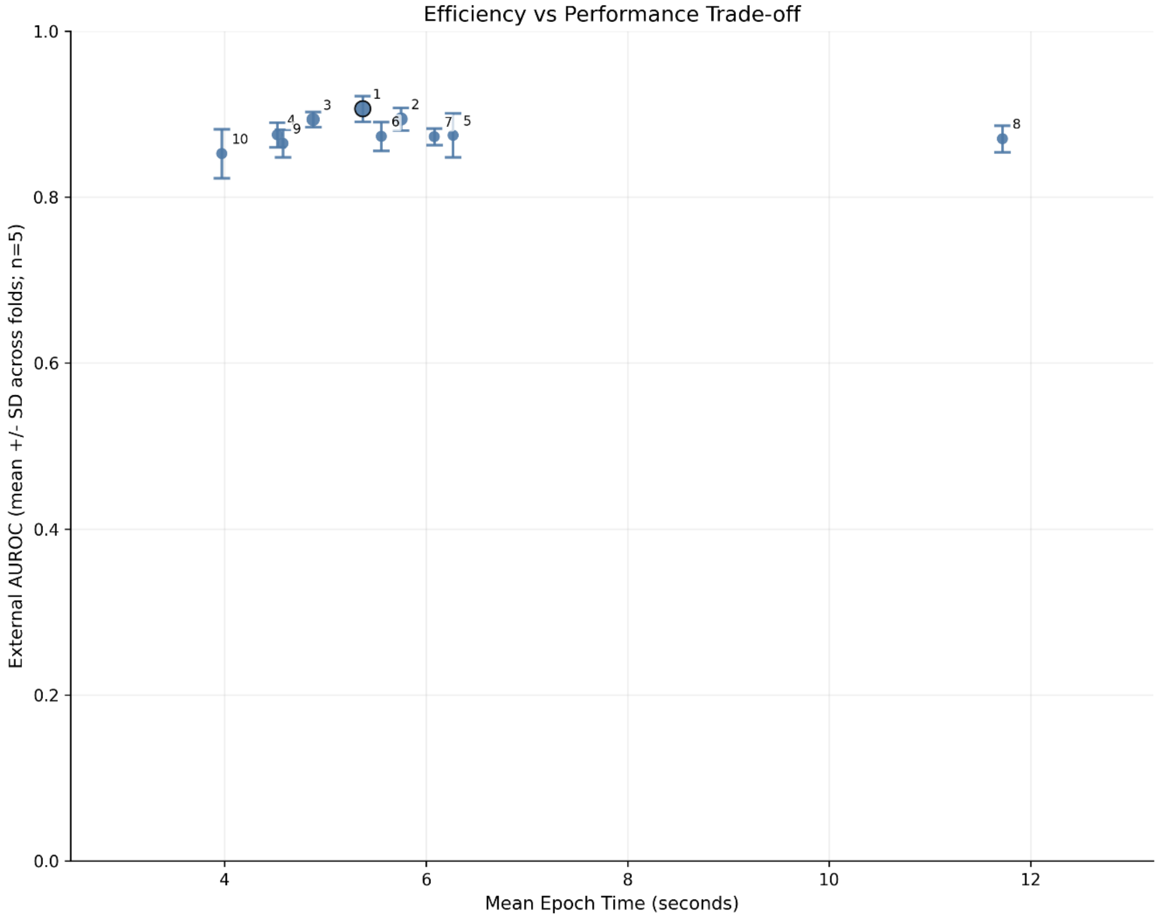


Figure 4.8. Efficiency versus performance scatter plot. X-axis shows mean epoch training time (seconds); y-axis shows external AUROC with error bars indicating standard deviation across folds. Colour gradient reflects AUROC ranking.

Table 4.3 reports detailed training efficiency metrics. ResNet-50 achieved the fastest training (41.7 ± 5.4 seconds per fold), whilst MaxViT-T was substantially slower (114.0 ± 7.9 seconds). Importantly, the best-performing model (Swin-T) required only moderate computational resources (62.2 ± 13.0 seconds), indicating that superior generalisation does not necessarily require the highest computational investment. This efficiency profile supports Swin-T's suitability for iterative development and potential deployment in resource-constrained settings

Table 4.3. Training efficiency metrics (mean ± SD across 5 folds).

Model	**Train Time per Fold (s)**	**Avg Epoch Time (s)**
ResNet-50	41.7 ± 5.4	3.97 ± 0.27
DeiT-T	49.2 ± 8.2	4.88 ± 0.97
MobileNetV3-S	53.4 ± 6.8	4.58 ± 0.34
ConvNeXt-T	54.0 ± 8.4	5.75 ± 0.82
RegNetY-008	60.8 ± 22.1	6.27 ± 1.95
Swin-T	62.2 ± 13.0	5.37 ± 0.25
DenseNet-121	63.4 ± 8.0	6.08 ± 0.20
EfficientNet-B0	66.1 ± 8.6	5.56 ± 0.63
VGG-16	71.2 ± 15.9	4.52 ± 0.14
MaxViT-T	114.0 ± 7.9	11.72 ± 0.79

4.8 Artifact Filtering Sensitivity Analysis

To empirically validate the 5% artifact filtering threshold and examine whether filtering materially affects classification performance, we conducted an ablation study across six different regimes. Three architectures representing distinct, representative design paradigms were selected: ResNet-50 (classic CNN), ConvNexT-Tiny (modern CNN) and DeiT-Tiny distilled (vision transformer). Computational constraints necessitated 3-fold cross-validation, limiting statistical power (see Section 3.6.5).

4.8.1 Experimental Design

Six filtering thresholds were examined:

- Any: Images with any detected artifact overlap removed
- >20%: Images with more than 20% lesion occlusion removed
- >10%: Images with more than 10% lesion occlusion removed
- >5%: Images with more than 5% lesion occlusion removed
- >2%: Images with more than 2% lesion occlusion removed
- None: All images retained, regardless of artifact coverage

Each configuration was trained and evaluated independently, producing 18 experimental conditions (3 architectures x 6 thresholds).

4.8.2 Results

Table 4.8. External validation AUROC across artifact filtering thresholds (mean ± SD, 3-fold CV).

Threshold	ResNet-50	ConvNeXt-T	DeiT-T
None	0.869 ± 0.013	0.912 ± 0.008	0.890 ± 0.006
>20%	0.871 ± 0.025	0.900 ± 0.006	0.886 ± 0.011
>10%	0.836 ± 0.038	0.903 ± 0.006	0.883 ± 0.015
>5%*	0.863 ± 0.016	0.907 ± 0.003	0.888 ± 0.018
>2%	0.855 ± 0.037	0.899 ± 0.007	0.876 ± 0.025
Any	0.845 ± 0.030	0.905 ± 0.006	0.893 ± 0.003
AUROC range	**0.035**	**0.013**	**0.017**
Mean fold SD	**0.026**	**0.006**	**0.013**

*Threshold used in main experiments. Highlighted row shows 5% threshold results

All three architectures maintained external validation AUROC above 0.83 regardless of filtering threshold. ConvNeXt-Tiny exhibited the smallest variation (Δ = 0.013), followed by DeiT-Tiny (Δ = 0.017), with ResNet-50 showing the largest (Δ = 0.035).

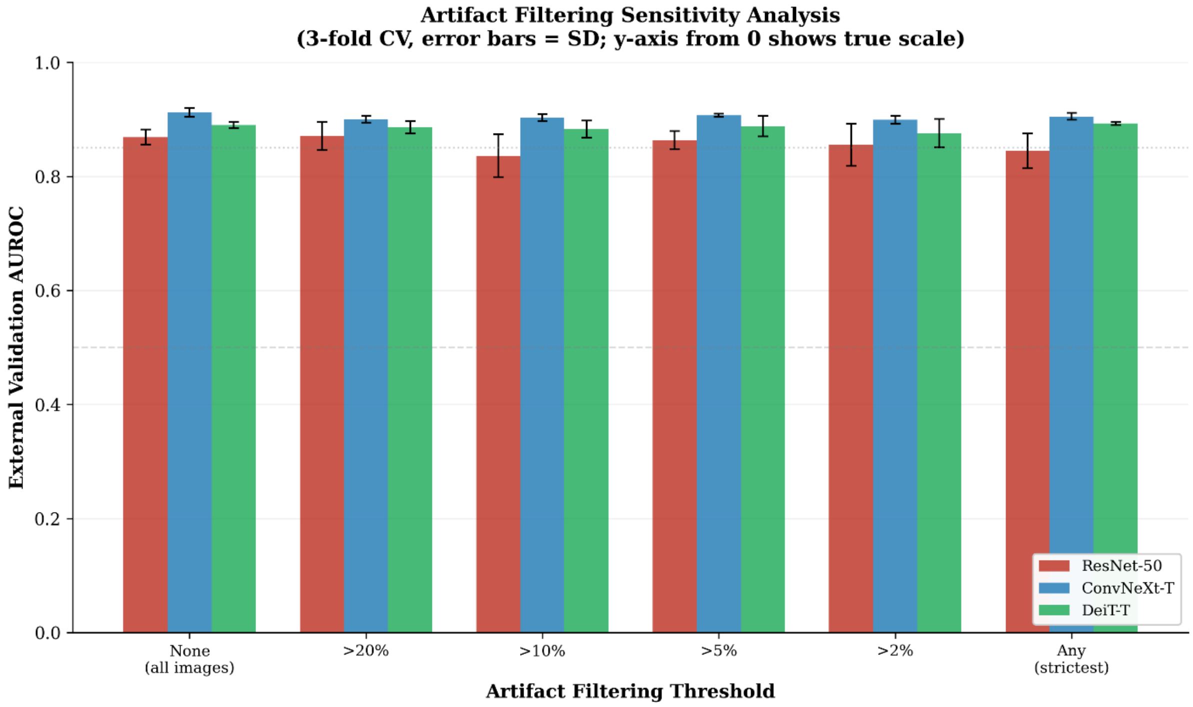


Figure 4.9. Artifact Filtering Sensitivity Analysis. X-axis shows artifact filtering thresholds; y-axis shows external AUROC with error bars indicating standard deviation across folds.

4.8.3 Statistical Assessment

Using 95% confidence intervals with the t-distribution (df = 2, $t_{0.975}$ = 4.303), the intervals for best and worst thresholds overlapped substantially for all architectures, indicating that observed differences between filtering thresholds are not statistically significant:

- ResNet-50: Best 0.871 [0.809, 0.933] vs Worst 0.836 [0.742, 0.930]
- ConvNeXt-T: Best 0.912 [0.892, 0.932] vs Worst 0.899 [0.882, 0.916]
- DeiT-T: Best 0.893 [0.886, 0.900] vs Worst 0.876 [0.814, 0.938]

4.8.4 Validation of 5% Threshold

The 5% threshold used in main experiments produced competitive performance, within 0.01 AUROC of each architecture's peak, supporting its a priori selection as a reasonable preprocessing choice that does not substantially compromise performance:

- ResNet-50: 0.863 at 5% vs 0.871 peak (Δ = 0.008)
- ConvNeXt-T: 0.907 at 5% vs 0.912 peak (Δ = 0.005)

- DeiT-T: 0.888 at 5% vs 0.893 peak (Δ = 0.005)

Chapter 5: Discussion

This chapter interprets the experimental findings within the context of existing literature, explores the potential implications for clinical deployment, and considers what alternative approaches might yield further improvements.

5.1 Architecture Performance in Context

The finding that Swin Transformer outperformed all evaluated architectures, including the ubiquitous ResNet-50, carries significant implications for breast ultrasound classification research. ResNet-50 remains the default architecture in much of the medical imaging literature, often justified by its established track record and widespread availability (Luo, et al., 2024). However, the present results demonstrate a substantive performance gap (ΔAUROC = 0.055) that persists under rigorous external validation, suggesting that architectural modernisation offers real, tangible benefits, beyond incremental improvements.

The superior performance of transformer-based architectures aligns with emerging evidence from other medical imaging domains. Swin Transformer's hierarchical design with shifted window attention may be particularly well suited to ultrasound classification, where diagnostically relevant features span multiple context windows, from fine grained texture patterns within lesions, to broader contextual relationships with surrounding tissue. The attention mechanism's ability to model long range dependencies could explain its advantage over purely convolutional architectures that rely on local fields.

Interestingly, ConvNeXt-Tiny achieved competitive performance (AUROC = 0.894) despite being a modernised CNN rather than a transformer. This result supports findings by Liu et al. (2022) that much of the performance gap between CNNs and transformers reflects actually training protocols, rather than fundamental architectural differences. ConvNeXt may represent a promising middle ground for practitioners who are seeking transformer level performance, with more familiar, CNN based workflows.

5.2 Comparison with Published Literature

A more direct comparison with published breast ultrasound classification research results requires careful consideration of the methodological differences between studies. Many studies report internal validation accuracies exceeding 95% (Han, et al., 2017; Byra, et al., 2019), substantially higher than the external AUROC values observed in this benchmark. However, such comparisons are potentially misleading for several reasons.

First, internal validation demonstrably overestimates deployment performance. Yu et al. (2022) demonstrated that 81% of deep learning algorithms exhibit decreased accuracy on external datasets, with performance drops frequently exceeding 0.05 AUROC. The present study's emphasis on true external validation provides a more realistic estimate of clinical performance.

Second, patient-level data leakage inflates reported metrics, in studies that perform image-level, rather than patient-level, splitting. Yagis et al. (2021) quantified this inflation at 29-55% in brain imaging tasks. By implementing GroupKFold with patient-level grouping where identifiers were available, this benchmark mitigates a substantial source of optimistic bias that is pervasive in much of the existing literature.

Third, many published studies evaluate on the BUSI dataset alone, which has documented quality issues including 19% duplicate images and systematic annotation artefacts (Pawłowska, et al., 2023). The present study's use of BUSI as one component of the external test set, combined with QAMEBI, provides more robust generalisation estimates.

When methodologically comparable studies are considered, the present results align with systematic reviews; finding that external validation AUROC typically ranges from 0.75-0.90 for breast ultrasound classification (Luo, et al., 2024). Swin-T's external AUROC of 0.907 represents strong performance within this realistic range.

5.3 Clinical Implications of Error Profiles

The confusion matrix analysis reveals that architecture choice has direct implications for clinical error types. Swin-T's lower false negative rate (13.4%, vs. ResNet-50's 24.9%) suggests that it would miss fewer malignancies in a screening application. If deployed in a population of 10,000 women with 5% malignancy prevalence (500 cancers), Swin-T would be expected to miss approximately 67 cancers compared to ResNet-50's 125, a difference of 58 detected malignancies, that could receive earlier treatment.

Conversely, ConvNeXt-T's lower false positive rate (18.8% vs Swin-T's 21.7%) might be preferred in settings where unnecessary biopsies carry significant cost, psychological burden, or resource constraints. The choice between sensitivity-optimised and specificity-optimised models should be guided by clinical context/requirement, rather than AUROC alone.

What if thresholds were adjusted to match specific clinical requirements? The threshold stability analysis suggests that Swin-T would maintain consistent operating characteristics across different populations (CV = 0.08), whilst VGG-16 might require re-calibration when deployed in new settings (CV = 0.31). This operational stability represents a practical advantage for deployment planning.

5.4 ROI Strategy: Theoretical and Practical Considerations

The ROI ablation results support the hypothesis that perilesional context carries diagnostic information. Radiomics literature has established that peritumoral tissue contains predictive features, with Mao et al. (2019) reporting an AUROC of 0.92 for peritumoral classifiers versus 0.83 for intertumoral features alone. The present finding that 5-10% expansion optimises deep learning classification provides a concrete preprocessing recommendation that bridges radiomics insights with end-to-end learning approaches.

The clinical rationale for perilesional context is compelling. Posterior acoustic shadowing, present in approximately 65% of malignant lesions (Stavros, et al., 1995), appears below (rather than within) masses. Similarly, reactive tissue changes and boundary characteristics that inform BI-RADS assessment occur at the lesion-parenchyma border. Strict lesion cropping removes this diagnostically valuable information, whilst full-image input dilutes it significantly with irrelevant background information.

What if architectures learned to select relevant context adaptively? The observation that different architectures show varying sensitivity to ROI configuration (Swin-T relatively insensitive; EfficientNet-B0 highly sensitive) suggests that attention mechanisms may partially compensate for suboptimal cropping by focusing on relevant regions. Future work exploring attention over multiple crops could even potentially automate ROI selection.

5.5 Calibration and Decision Support

The finding that all architectures benefited from temperature scaling, with ECE reductions of 0.5-3.0%, confirms the broader observation that modern neural networks are systematically miscalibrated (Guo, et al., 2017). More importantly, the substantial variation in baseline calibration across architectures (ECE range: 0.051-0.116) suggests that calibration should be evaluated alongside discrimination when selecting models for clinical deployment.

What if calibrated probabilities were integrated into clinical workflows? A model predicting 0.8 probability of malignancy makes an implicit promise that, among all cases assigned this probability, approximately 80% should be truly malignant. Well calibrated models enable an evidence-based threshold selection that is aligned with individual deployment preferences for sensitivity-specificity trade-offs. Poorly calibrated models, even with strong discrimination, may mislead clinicians who interpret probabilities literally.

Van Calster et al. (2019) argued that (mis)calibration may render an algorithm clinically less useful than alternatives with lower discrimination, but better probability estimation. The present results partially support this view; DeiT-T achieved strong discrimination (AUROC = 0.894) but poor calibration (ECE = 0.116), whilst VGG-16 showed moderate discrimination (AUROC = 0.875) with better calibration (ECE = 0.071).

5.6 Artefact Handling and Shortcut Learning

The artefact-aware preprocessing pipeline addresses a known vulnerability in medical imaging AI. DeGrave et al. (2021) demonstrated that COVID-19 detection algorithms exploited radiographic markers, rather than pulmonary pathology; similar shortcut learning could affect breast ultrasound models that are trained on datasets with systematic annotation patterns.

The 5% lesion overlap threshold used in this study represents a conservative balance between data quality and training set preservation. This choice reflects several considerations: maintaining adequate training data volume; ensuring retained images preserve good lesion visibility; reducing exposure to heavily annotated images that might encourage shortcut learning; and promoting generalisation to deployment settings with variable annotation practices.

The sensitivity analysis (Section 4.8) provides empirical context, though interpretation requires caution. All evaluated architectures maintained AUROC above 0.83 across all filtering thresholds, and 95% confidence intervals for best versus worst thresholds overlapped for all architectures. This suggests that filtering threshold has limited practical impact on classification performance within the tested range.

However, several confounds complicate interpretation. Stricter thresholds progressively reduce training data, creating competing effects between improved data quality and reduced sample size. The external test sets contain artefact-contaminated images, so models trained without filtering may benefit from matching conditions. And the 3-fold design provides limited statistical power – the null finding could reflect true equivalence or insufficient power.

Despite these caveats, results offer practical guidance. The consistency across thresholds suggests practitioners need not agonise over exact filtering choices. The 5% threshold represents a defensible middle-ground, and it’s a priori selection based on principled reasoning – rather than post-hoc optimisation – strengthens methodological defensibility.

The observation that modern architectures exhibited smaller AUROC variation than ResNet-50 is intriguing but should not be over-interpreted without controlled experiments establishing causality.

Chapter 6: Conclusion

6.1 Project Summary

This dissertation set out to address fundamental methodological gaps in current breast ultrasound classification research through a rigorous benchmarking framework. The work was motivated by the observation that much of the existing literature suffers from patient-level data leakage, inconsistent preprocessing documentation, absent calibration assessment, and inadequate external validation, collectively undermining the reliability of reported performance metrics and impeding clinical translation.

To address these gaps, this study implemented a comprehensive evaluation of ten deep learning architectures spanning classic CNNs (VGG-16, ResNet-50, DenseNet-121), modern efficient designs (EfficientNet-B0, MobileNetV3-Small, RegNetY-008, ConvNeXt-Tiny), and vision transformers (DeiT-Tiny, Swin-Tiny, MaxViT-Tiny). All architectures underwent identical training protocols with ImageNet-pretrained weights, eliminating confounding variance from protocol variation. Five-fold stratified GroupKFold cross validation with patient-level grouping was utilised where identifiers were available, mitigating data leakage. External validation on geographically distinct datasets (BUSI from Egypt, QAMEBI from Iran) provided more realistic generalisation estimates. A novel artefact-aware preprocessing pipeline combining YOLO-based detection and UNet-based segmentation was developed to mitigate shortcut learning risks. Systematic ROI ablation examined ten border expansion configurations. Temperature scaling was applied for post-hoc calibration, with all metrics reported using appropriate confidence intervals.

The experimental framework encompassed 333 training runs across the architecture benchmark and ROI ablation studies, generating a comprehensive evidence base for model selection and preprocessing recommendations.

6.2 Key Findings

Architectural performance hierarchy. Swin Transformer achieved the strongest external validation performance (AUROC = 0.907 [0.885, 0.927], balanced accuracy = 0.825 ± 0.025), substantially outperforming the widely used ResNet-50 baseline (AUROC = 0.852 [0.826, 0.877]). This finding challenges the continued reliance on legacy architectures in breast ultrasound research and suggests that transformer-based approaches warrant broader adoption. ConvNeXt-Tiny and DeiT-Tiny formed a competitive second tier (AUROC = 0.894), offering alternative options with different operational characteristics.

Error profiles vary meaningfully. Beyond AUROC, architectures exhibited substantially different error profiles. DeiT-Tiny achieved the lowest false negative rate (12.2% of malignancies missed), whilst ConvNeXt-T achieved the lowest false positive rate (18.8% of benign cases incorrectly flagged). These differences have direct clinical implications depending on whether the deployment context prioritises sensitivity (screening) or specificity (reducing unnecessary interventions).

Perilesional context improves performance. ROI ablation demonstrated that border expansion of 5-10% around lesion boundaries optimised classification performance, with mean AUROC improving from 0.888 (strict cropping) to 0.898 (optimal expansion). This finding provides actionable preprocessing guidance, that is absent from much of the existing literature, and aligns with recent radiomics research, demonstrating diagnostic value in peritumoral tissue.

Calibration varies independently of discrimination. Expected Calibration Error ranged from 0.051 (Swin-T) to 0.116 (DeiT-T), demonstrating that strong discrimination does not guarantee well calibrated probability estimates. Temperature scaling improved calibration for all architectures (ECE reduction: 0.5-3.0%), supporting its routine application as a lightweight, post-hoc step.

Threshold stability reflects operational robustness. The coefficient of variation in Youden thresholds ranged from 0.08 (Swin-T) to 0.31 (VGG-16), indicating that some architectures produce more consistent operating points across different training/validation partitions. This stability metric is relevant for deployment planning, as models with variable thresholds may require recalibration when applied to new populations; this is particularly of note for resource-limited settings.

6.3 Relationship to Existing Literature

Several findings confirm patterns reported in the broader literature. The observation that 81% of deep learning algorithms exhibit decreased accuracy on external datasets (Yu, et al., 2022), was consistent with the present results, where all architectures showed some generalisation gap between internal and external performance. The finding that modern neural networks are systematically miscalibrated (Guo, et al., 2017) was confirmed across all ten architectures.

Other findings extend or refine existing knowledge. Whilst systematic reviews have noted the dominance of ResNet-50 in breast ultrasound research (Luo, et al., 2024), the present study provides direct evidence that this architectural choice is suboptimal under rigorous external validation. The ROI ablation results provide concrete preprocessing recommendations that were previously largely absent from the literature, translating radiomics insights about peritumoral features (Mao, et al., 2019) into deep learning pipeline design.

The comprehensive calibration analysis represents a novel contribution, as ECE reporting for breast ultrasound classification was essentially absent from published literature. This gap is concerning given the established clinical importance of probability calibration (Van Calster, et al., 2019).

6.4 Limitations

(i) Dataset scope. Whilst external validation strengthens generalisation claims, the datasets used do not fully represent the diversity of ultrasound equipment, acquisition protocols, and patient populations encountered in global clinical practice. Performance on datasets from other geographic regions, equipment manufacturers, or patient demographics remains untested.

(ii) Missing patient identifiers. One internal dataset (BUS_UC) lacked patient-level identifiers, preventing definitive enforcement of patient-level separation during cross-validation for that subset. Sensitivity analyses (Appendix A) suggest this limitation does not substantially alter conclusions, but the possibility of residual leakage cannot be completely excluded.

(iii) Operating point assumptions. Results are reported at Youden-optimal thresholds, which balance sensitivity and specificity equally. Clinical pathways may require different operating points optimised for specific sensitivity or specificity targets that differ from Youden's criterion.

(iv) Limited ROI model coverage. ROI ablation was performed for five architectures due to computational constraints. The full ten-architecture benchmark was not replicated across all ROI configurations, potentially missing architecture specific ROI interactions for the excluded models.

(v) Artefact residuals. Whilst the artefact detection pipeline addresses major overlay categories, subtle acquisition-specific cues (acoustic patterns, equipment specific textures) may persist and contribute to dataset-specific learning that does not generalise well.

(vi) Multi-lesion images. Images containing multiple lesions were split into separate samples (Section 3.1.2), introducing minor non-independence for 2.6% of external test data. Sensitivity analysis confirmed negligible impact on reported metrics.

6.5 Future Work

(i) Threshold portability as a primary endpoint. Future studies should fix thresholds on development data, and report external performance under unchanged operating points, including pathway specific sensitivity/specificity targets (e.g. 95% sensitivity for screening applications). This would provide more realistic estimates of deployment behaviour.

(ii) Artefact robustness evaluation. Stratified evaluation on subsets that are defined by measured artefact overlap, would directly quantify model sensitivity to annotation patterns, and UI elements. This analysis could inform robust dataset curation guidelines for future research.

(iii) Learned context selection. Replacing “one size fits all” ROI rules with attention based or multi-crop strategies could enable the incorporation of adaptive context, without requiring predetermined/flat expansion parameters. Such approaches may well achieve optimal context selection for individual images.

(iv) Reader studies and workflow integration. Translation to clinical impact requires reader studies comparing radiologist performance with and without AI assistance, along with evaluation of workflow metrics (reading time, triage efficiency) that will necessarily determine practical utility.

(v) Multi-centre prospective validation. The ultimate test of model generalisability requires prospective validation across multiple clinical centres with diverse equipment and patient populations, ideally with clearly predefined analysis frameworks.

(vi) Rigorous artefact sensitivity investigation. Definitive assessment would require: larger datasets enabling aggressive filtering; artefact-free external validation sets; repeated k-fold CV (5×5 = 25 observations); and controlled artifact injection experiments.

(vii) Shortcut learning detection. Future work should develop methods to detect shortcut learning through attribution methods (GradCAM, integrated gradients) visualising prediction-driving regions.

6.6 Ethical Considerations

The development and deployment of AI systems for breast cancer detection carries significant ethical responsibilities. False negatives represent missed malignancies that may progress before detection, potentially reducing survival. False positives lead to unnecessary biopsies, causing patient anxiety, physical discomfort, and healthcare resource consumption in an already stretched healthcare system. The error profile analysis in this dissertation demonstrates that these trade-offs vary substantially across architectures, emphasising that model selection should be guided by clinical context and explicit consideration of which error types carry greater consequences.

Algorithmic bias is a concern when models trained on geographically limited datasets are deployed in more diverse populations. The external validation design in this study, using datasets from Egypt and Iran to evaluate models trained primarily on Brazilian and Spanish data, represents a step toward assessing cross-population generalisability. However, broader validation across demographic groups, ethnicities, and healthcare settings is necessary before clinical deployment.

Transparency and interpretability are essential for clinical acceptance. Whilst deep learning models achieve strong discrimination, their decision processes remain largely opaque, often referred to as “black box” like. Future work should incorporate explainability methods (attention visualisation, saliency mapping) to help clinicians understand and be able to appropriately trust in model predictions.

Finally, the potential for AI systems to exacerbate healthcare inequalities must be considered. If deployed primarily in well-resourced settings, AI-assisted diagnosis could widen, rather than narrow, the survival gap between high-income and low-income regions identified in Chapter 1. Conversely, portable ultrasound combined with AI interpretation could democratise access to expert-level diagnostic support in underserved areas. Ensuring equitable deployment should be a priority for clinical translation efforts.

6.7 Concluding Remarks

This dissertation demonstrates that rigorous methodology, including patient-level data separation, external validation, calibration assessment, and systematic preprocessing evaluation, reveals meaningful performance differences among deep learning architectures for breast ultrasound classification that would be obscured by less careful evaluation approaches.

The practical message is clear: Swin Transformer offers the strongest combination of discrimination, calibration, and operational stability among evaluated architectures. ROI preprocessing with 5-10% expansion optimises the trade-off between perilesional context preservation and background noise. Temperature scaling provides a lightweight calibration improvement that is applicable to any architecture. External validation remains essential for realistic performance estimation.

The resulting benchmark, encompassing 333 experimental runs with standardised protocols and comprehensive reporting, provides a reproducible evidence base for model selection and establishes methodological standards for future breast ultrasound classification research. By addressing the pervasive weaknesses identified in existing literature, this work contributes to the foundation necessary for eventual clinical translation of AI-assisted breast ultrasound diagnosis.

Bibliography:

Abdullah, N., Mesurolle, B., El-Khoury, M. & Kao, E., 2009. Breast imaging reporting and data system lexicon for US: interobserver agreement for assessment of breast masses. *Radiology,* 252(3), pp. 665-672.

Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A., 2020. Dataset of breast ultrasound images. *Data in Brief,* Volume 28, p. 104863.

Ardakani, A. et al., 2023. An open-access breast lesion ultrasound image database: applicable in artificial intelligence studies. *Computers in Biology and Medicine,* Volume 152, p. 106405.

Banerjee, I. M. M. L. J. F. W. a. G. J., 2023. "Shortcuts" causing bias in radiology artificial intelligence: causes, evaluation, and mitigation. *Journal of the American College of Radiology,* 20(9), pp. 842-851.

Berg, W. et al., 2008. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. *JAMA,* 299(18), p. 2151–2163.

Berg, W. et al., 2012. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. *JAMA,* 307(13), p. 1394–1404.

Bossuyt, P. et al., 2015. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ,* Volume 351, p. h5527.

Bray, F. et al., 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians,* 74(3), pp. 229-263.

Brem, R., Lenihan, M., Lieberman, J. & Torrente, J., 2015. Screening breast ultrasound: past, present, and future. *American Journal of Roentgenology,* 204(2), pp. 234-240.

Brem, R. T. L. D. S. I. M. G. J. H. B. L. M. L. R. R. J. S. S. a. T. C., 2015. Assessing improvement in detection of breast cancer with three-dimensional automated breast US in women with dense breast tissue: the SomoInsight Study. *Radiology,* 274(3), pp. 663-673.

Brown, L., Cai, T. & DasGupta, A., 2001. Interval estimation for a binomial proportion. *Statistical Science,* 16(2), pp. 101-133.

Bunnell, A. H. K. S. J. a. S. P., 2024. BUSClean: Open-source software for breast ultrasound image pre-processing and knowledge extraction for medical AI. *PLOS ONE,* 19(12), p. e0315434.

Byra, M. et al., 2019. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. *Medical Physics,* 46(2), pp. 746-755.

Chen, Y. T. Y. W. Y. W. Y. a. W. L., 2023. Ultrasound for Breast Cancer Screening in Resource-Limited Settings: Current Practice and Future Directions. *Cancers,* 15(7), p. 2112.

Chlap, P. M. H. V. N. D. J. H. L. a. H. A., 2021. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology,* 65(5), pp. 545-563.

Christodoulou, E., Steinmann, D., Glocker, B. & Kainz, B., 2024. *Confidence intervals uncovered: are we ready for real-world medical imaging AI?,* arXiv:2409.17763: arXiv.

Clopper, C. & Pearson, E., 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika,* 26(4), pp. 404-413.

Collins, G. et al., 2024. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ,* Volume 385, p. e078378.

DeGrave, A., Janizek, J. & Lee, S., 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence,* 3(7), pp. 610-619.

Dosovitskiy, A. et al., 2021. *An image is worth 16x16 words: transformers for image recognition at scale.* s.l., Proceedings of the International Conference on Learning Representations.

Efron, B. & Tibshirani, R., 1993. *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Geirhos, R. et al., 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence,* 2(11), pp. 665-673.

Gómez-Flores, W., Gregorio-Calas, M. & Pereira, W., 2024. BUS-BRA: a breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics,* 51(5), pp. 3110-3123.

Guo, C., Pleiss, G., Sun, Y. & Weinberger, K., 2017. *On calibration of modern neural networks.* Proceedings of the 34th International Conference on Machine Learning, PMLR, pp. 1321-1330.

Han, S. et al., 2017. A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Physics in Medicine and Biology,* 62(19), pp. 7714-7728.

He, K., Zhang, X., Ren, S. & Sun, J., 2016. *Deep residual learning for image recognition.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, s.n., pp. 770-778.

Howlader, N. et al., 2023. *SEER Cancer Statistics Review, 1975-2020,* Bethesda, MD: National Cancer Institute.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K., 2017. *Densely connected convolutional networks.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, s.n., pp. 4700-4708.

Islam, T. H. M. J. J. K. M. a. M. M., 2024. A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions. *Healthcare Analytics,* Volume 5, p. 100340.

Jiang, M. L. C. T. Y. C. Z. L. W. S. H. a. Z. Q., 2021. Multiregional radiomic model for breast cancer diagnosis: value of ultrasound-based peritumoral and parenchymal radiomics. *Quantitative Imaging in Medicine and Surgery,* 11(7), pp. 3259-3273.

Koçak, B. P. A. S. A. B. C. S. J. U. L. H. M. K. M. C. R. a. C. R., 2025. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and Interventional Radiology,* 31(2), pp. 75-88.

Kolb, T., Lichy, J. & Newhouse, J., 2002. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. *Radiology,* 225(1), pp. 165-175.

Lee, H. et al., 2008. Observer variability of Breast Imaging Reporting and Data System (BI-RADS) for breast ultrasound. *European Journal of Radiology,* 65(2), pp. 293-298.

Lin, C., Chen, H. & Heng, P., 2024. *Clinical annotation shortcuts in medical image segmentation.* Proceedings of MICCAI 2024, s.n., pp. 312-321.

Liu, Z. et al., 2021. *Swin transformer: hierarchical vision transformer using shifted windows.* Proceedings of the IEEE/CVF International Conference on Computer Vision, s.n., p. 10012–10022.

Liu, Z. et al., 2022. *A ConvNet for the 2020s.* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, s.n., p. 11976–11986.

Luo, Y. et al., 2024. Diagnostic performance of deep learning in ultrasound diagnosis of breast cancer: a systematic review. *npj Precision Oncology,* Volume 8, p. 31.

Mainiero, M. et al., 2023. ACR Appropriateness Criteria® palpable breast masses: 2022 update. *Journal of the American College of Radiology,* 20(5S), p. S166–S179.

Mao, N. et al., 2019. Added value of radiomics on mammography for breast cancer diagnosis: a feasibility study. *Journal of the American College of Radiology,* 16(4), pp. 485-491.

Marques, G., Agarwal, D. & de la Torre Díez, I., 2020. Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. *Applied Soft Computing,* Volume 96, p. 106691.

Mehrtash, A. W. W. T. C. a. A. P., 2020. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging,* 39(12), pp. 3868-3878.

Müller, S. a. H. F., 2021. *TrivialAugment: Tuning-Free Yet State-of-the-Art Data Augmentation.* Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), IEEE/CVF, pp. 774-782.

Nagendran, M. C. Y. L. C. G. A. K. M. H. H. T. E. I. J. C. G. a. M. M., 2020. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ,* Volume 368, p. m689.

Nguyen, H. N. H. P. H. L. K. L. L. D. M. a. V. V., 2023. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific* , 10(1), p. 277.

Ong Ly, C. U. B. T. T. P. T. D. J. K. S. M. Y. B. M. H. A. R. H. a. M. C., 2024. Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. *npj Digital Medicine,* 7(1), p. 124.

Pawłowska, A. et al., 2023. Letter to the Editor: Dataset of breast ultrasound images. *Data in Brief,* Volume 48, p. 109255.

Pellegrini, E., Ballerini, L., Hernandez, M. & Trucco, E., 2025. Data leakage in deep learning for Alzheimer's disease diagnosis: a scoping review. *Diagnostics,* 15(18), p. 2348.

Recht, B. R. R. S. L. a. S. V., 2019. *Do ImageNet classifiers generalize to ImageNet?.* Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97.

Redmon, J., Divvala, S., Girshick, R. & Farhadi, A., 2016. *You only look once: unified, real-time object detection.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, s.n., pp. 779-788.

Roberts, 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.

Ronneberger, O., Fischer, P. & Brox, T., 2015. *U-Net: convolutional networks for biomedical image segmentation.* Medical Image Computing and Computer-Assisted Intervention, Springer, pp. 234-241.

Simonyan, K. & Zisserman, A., 2015. *Very deep convolutional networks for large-scale image recognition.* Proceedings of the International Conference on Learning Representations, s.n.

Stavros, A. et al., 1995. Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. *Radiology,* 196(1), pp. 123-134.

Sung, H. et al., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians,* 71(3), pp. 209-249.

Tan, M. & Le, Q., 2019. *EfficientNet: rethinking model scaling for convolutional neural networks.* Proceedings of the 36th International Conference on Machine Learning, PMLR, p. 6105–6114.

Tirindelli, M. et al., 2021. *Rethinking ultrasound augmentation: a physics-inspired approach.* Medical Image Computing and Computer-Assisted Intervention, Springer, pp. 690-700.

Touvron, H. et al., 2021. *Training data-efficient image transformers & distillation through attention.* Proceedings of the 38th International Conference on Machine Learning, PMLR, pp. 10347-10357.

Vallez, N. et al., 2025. BUS-UCLM: breast ultrasound lesion segmentation dataset. *Scientific Data,* Volume 12, p. 159.

Van Calster, B. et al., 2019. Calibration: the Achilles heel of predictive analytics. *BMC Medicine,* Volume 17, p. 230.

Yagis, E. et al., 2021. Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Scientific Reports,* 11(1), p. 22544.

Yap, M. et al., 2018. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics,* 22(4), pp. 1218-1226.

Youden, W., 1950. Index for rating diagnostic tests. *Cancer,* 3(1), pp. 32-35.

Yu, A., Mohajer, B. & Eng, J., 2022. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radiology: Artificial Intelligence,* 4(3), p. e210064.

Zech, J. et al., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine,* 15(11), p. e1002683.

Appendix: