Comparative Performance Analysis of Deep Learning Techniques for Automated Breast Cancer Classification from Medical Imaging

CHRISTIAN KUTABUNA KUBAKISA[1], WITESYAVWIRWA VIANNEY KAMBALE[3], ISAAC LUKUSA KAYEMBE[1], MAHMOUD HAMED[2], DIVIN KAYEYE KABEYA[1], KYANDOGHERE KYAMAKYA[1,2]

[1]Faculté Polytechnique, Université de Kinshasa (UNIKIN), Kinshasa, DEMOCRATIC REPUBLIC OF THE CONGO

[2]Institute for Smart Systems Technologies, Universität Klagenfurt, AUSTRIA

[3]Faculty of Information and Communication Technology, Tshwane University of Technology, Pretoria, SOUTH AFRICA

Abstract: Early and reliable breast cancer diagnosis remains a public-health priority, yet routine screening still suffers from inter-reader variability, heavy workload, and uneven resource availability. To move beyond generic claims of "deep learning works," this paper reports a transparent, like-for-like comparison of three families of models under a single protocol and fixed pre-processing: a custom Convolutional Neural Network (CNN), MobileNetV3, and a Vision Transformer (ViT-L16). We evaluate on two complementary benchmarks—MIAS mammography (n=322 images) and BreakHis histopathology (n=7,909 tiles)—covering both binary (benign vs. malignant/abnormal) and multiclass settings. All models are trained on 224×224 inputs with harmonized augmentation; MIAS splits are image-level, whereas BreakHis splits are patient-level to avoid leakage. On MIAS, MobileNetV3 yields the strongest results with 93.25% accuracy for binary detection and 92.00% for multiclass classification, surpassing both the bespoke CNN and ViT-L16 under identical conditions. On BreakHis, ViT-L16 achieves the best validation performance (91.62% binary; 90.28% multiclass), which is consistent with the advantage of long-range self-attention for texture-rich histology. A side-by-side reading against recent literature shows our scores to be competitive and, in several cases, state-of-the-art under comparable validation protocols. Beyond headline numbers, we discuss deployment trade-offs: MobileNetV3 offers a favorable accuracy–efficiency balance for embedded or resource-constrained settings, while ViT-L16 is better suited to centralized, high-capacity pipelines. Taken together, the results argue for context-aware model selection and provide a reproducible baseline (unified splits, inputs, and training recipe) to facilitate independent verification and clinical translation.

Key-Words: Breast Cancer Detection, Medical Imaging, Deep Learning, MobileNet, Vision Transformer, CNN, MIAS, BreakHis, Binary Classification, Multiclass Classification, Clinical Deployment

Received: May 8, 2025. Revised: July 17, 2025. Accepted: August 23, 2025. Published: March 2, 2026.

1 Introduction

1.1 Background and Motivation

Breast cancer remains one of the most prevalent and life-threatening forms of cancer worldwide, representing a leading cause of cancer-related mortality among women, [1]. According to the World Health Organization (WHO), over 2.3 million new cases were diagnosed in 2020, and the incidence continues to rise, particularly in low- and middle-income countries, [2]. Beyond these headline figures, the clinical consensus is clear: earlier detection consistently shifts stage at diagnosis, improves survival prospects, and reduces treatment burden.

Medical imaging modalities such as mammography, ultrasound, magnetic resonance imaging (MRI), and histopathological slides form the cornerstone of breast cancer screening and diagnosis. In practice, however, these workflows hinge on expert interpretation under significant time pressure, and many health systems struggle with shortages or uneven distribution of radiologists and pathologists. Moreover, diagnostic errors and inter-observer variability remain persistent concerns, especially in high-volume or resource-constrained environments, [3]. Against a backdrop of ever-increasing imaging throughput, these factors jointly motivate the development of diagnostic support systems that are both robust and scalable.

In recent years, deep learning (DL), particularly

convolutional neural networks (CNNs), has emerged as a transformative tool in medical image analysis. CNNs are capable of automatically learning hierarchical features from raw image data, thereby reducing reliance on hand-engineered features and domain-specific preprocessing, [4], [5]. A growing body of evidence further shows that, on curated cohorts and well-defined tasks, DL systems can approach the performance of experienced clinicians, [6]. This progress has catalyzed end-to-end pipelines for detection, segmentation, and even risk stratification, moving beyond proof-of-concept toward operational pilots.

Despite this progress, one critical challenge remains: selecting an optimal model architecture for real-world deployment. The landscape of deep learning encompasses a diverse array of architectures, each with distinct trade-offs in terms of accuracy, interpretability, parameter size, training cost, and computational efficiency. For example, while state-of-the-art models like Vision Transformers (ViTs) offer global attention mechanisms and perform well on large-scale, high-resolution tasks, they are often computationally intensive, [7]. In contrast, lightweight models such as MobileNetV3, **howard2019mobilenetv3**, are optimized for speed and efficiency on mobile or embedded devices but may sacrifice some performance. Custom CNNs tailored to the data domain and clinical objectives remain attractive; however, they require careful architectural choices and substantial tuning to generalize reliably.

Taken together, these considerations underscore the value of head-to-head comparisons under a single, transparent protocol. Comparative studies that systematically analyze multiple architectures under a unified experimental setup are essential for evidence-based model selection. In safety-critical clinical contexts, such evaluations must weigh practical constraints—such as computational limits, data availability, and latency—alongside headline accuracy, thereby aligning methodological rigor with the realities of deployment.

1.2 Problem Statement and Research Questions

Automating breast cancer detection and subtype classification from medical images raises two intertwined challenges: achieving reliable early detection and integrating advances in imaging and AI into deployable clinical tools. The first is fundamentally methodological (robust accuracy under real-world variability); the second is operational (making models usable when expertise and computing power are limited).

Although numerous deep learning models have been proposed for breast cancer classification, comparative studies that put *heterogeneous* architectures on equal footing remain scarce—especially when designs differ in depth, parameter count, and hardware appetite. In clinical practice, particularly in settings such as rural clinics or mobile screening units, the choice of model must be aligned not only with performance indicators but also with deployment feasibility. Consequently, headline accuracy alone is insufficient; memory footprint, latency, and calibration also matter when workflows are time- and resource-constrained.

This study aims to fill this gap by evaluating three distinct deep learning architectures for automated breast cancer classification from medical images: (i) a lightweight MobileNetV3 optimized for efficiency, a fully developed custom CNN, and a large-scale Vision Transformer (ViT-L16). All models are trained and assessed under a unified protocol using two complementary datasets: the Mammographic Image Analysis Society (MIAS) dataset, which poses challenges due to low resolution and a limited sample size, and the BreakHis dataset, which introduces variability in histopathological magnifications and morphological diversity. Together, these corpora encompass radiographic and microscopic modalities, allowing for a modality-agnostic comparison.

The key research questions addressed in this paper are:

- **Q1:** Among CNNs, mobile-efficient CNNs, and transformer-based models, which architecture families are most effective for automated breast cancer detection and classification across modalities?
- **Q2:** How do diagnostic metrics (e.g., accuracy, F1-score) differ across the custom CNN, MobileNetV3, and ViT-L16 under an identical training/validation pipeline?
- **Q3:** What are the practical trade-offs between computational efficiency (parameters, throughput) and diagnostic precision (calibrated performance, recall for malignant cases)?
- **Q4:** Which model offers the best balance of accuracy, robustness, and deployability for real-world, resource-constrained environments (e.g., mobile screening units or district hospitals)?

1.3 Contributions of this Paper

This work presents a comparative, modality-spanning evaluation of three complementary deep learning families for breast cancer imaging, accompanied by a rigorously controlled pipeline that emphasizes both

accuracy and deployability. The main contributions are:

1. **Performance under a unified protocol.** Using identical preprocessing, splits, and evaluation, the models achieve strong and consistent results on mammography (MIAS) and histopathology (BreakHis). On MIAS, MobileNetV3 reached 93.25% accuracy for binary detection and 92.00% for multiclass classification, while ViT-L16 achieved 92.00% in multiclass classification. On BreakHis, ViT-L16 achieved 91.62% for binary detection and 90.28% for multiclass classification, clearly surpassing the custom CNN in both settings.
2. **Compute–accuracy trade-offs made explicit.** We quantify the practical trade-offs between accuracy and resource usage, showing that MobileNetV3 delivers a favorable balance of accuracy and computational cost, which makes it suitable for real-time or energy-constrained deployments without sacrificing diagnostic reliability.
3. **Complementary model roles.** By combining three distinct architectures —custom CNN, MobileNetV3, and ViT-L16— we delineate how each family contributes: the transformer captures long-range patterns in histopathology, the mobile CNN offers efficient inference for point-of-care scenarios, and the shallow CNN provides a transparent baseline for ablation and error analysis.
4. **Reproducible data pipeline.** The study adopts an end-to-end workflow with standardized resizing/normalization, label harmonization, class rebalancing via clinically safe geometric augmentation, and compressed caching, with fixed random seeds and archived splits to support repeatable experiments and downstream benchmarking.
5. **Clinical integration outlook.** We translate experimental findings into deployment guidance: MobileNetV3 for portable or embedded settings, and ViT-L16 for centralized, high-throughput pipelines. This mapping clarifies which models are best aligned with screening workflows that prioritize recall, versus confirmatory settings that prioritize calibration and precision.

The remainder of the paper is organized as follows. Section 2 surveys prior work in breast cancer imaging. Section 3 details datasets, preprocessing, and methodology. Section 4 presents the model architectures, and Section 5 describes the experimental setup. Section 6 reports the empirical results, followed by comparative analysis and discussion in Section 7. Finally, Section 8 offers conclusions and outlines future research directions.

2 Related Work

A clear view of how breast cancer diagnosis has evolved helps position the present contribution within the broader literature. This section first reviews conventional diagnostic pathways and their well-documented constraints, then summarizes classical machine learning approaches, and finally highlights recent advances in deep learning. We conclude with a comparative analysis of performance and clinical feasibility, which delineates the specific gaps our study addresses.

2.1 Traditional Breast Cancer Diagnostic Methods

Prior to the adoption of data□driven techniques, breast cancer detection primarily relied on a combination of clinical examination, imaging modalities, and histological assessment. While mammography remains the reference screening tool, its sensitivity declines markedly in dense breasts, which can delay diagnosis and increase the likelihood of false negatives, [8], [9]. Ultrasound helps distinguish cystic from solid lesions but is vulnerable to operator dependence and elevated false□positive rates (often $> 15\%$), [10]. Magnetic Resonance Imaging (MRI), although highly sensitive ($> 90\%$), is constrained by cost and availability for routine workflows, [8].

Biopsy procedures—including Fine Needle Aspiration (FNA), Core Needle Biopsy, and surgical biopsy—remain essential for pathological confirmation. FNA offers a less invasive option but typically yields lower diagnostic precision than Core Needle Biopsy, which achieves accuracies above 90%, [9]. Table 8 (Appendix) summarizes the strengths and shortcomings of these modalities. Despite their clinical value, time delays, inter-operator variability, and resource constraints underscore the need for AI-assisted solutions.

2.2 Classical Machine Learning Approaches

With the widespread digitization of breast imaging, classical machine learning (ML) methods were among the first to automate diagnostic decision support. Techniques such as Support Vector Machines (SVMs), Decision Trees, k-Nearest Neighbors (k-NN), and Bayesian classifiers have been reported to exhibit moderate performance when coupled with carefully designed, task-specific

feature extractors based on texture and morphology. For example, [11], reported about 82% accuracy on DDSM with an SVM, while Bayesian models achieved AUC values up to 0.85 for ultrasound classification, [9].

Despite these advances, the main bottleneck of classical ML was its dependence on handcrafted features and its limited scalability to rich, high-resolution imaging data. Decision Trees often overfit small cohorts, and k-NN incurs heavy computational costs at inference. Although shallow Artificial Neural Networks (ANNs) could reach accuracies up to 87%, [12], their limited depth constrains hierarchical abstraction and hinders transferability across datasets. Table 9 (Appendix) outlines the principal strengths and caveats of these classical approaches.

2.3 Deep Learning in Breast Cancer Imaging

The advent of deep learning fundamentally shifted breast image analysis by enabling end-to-end, hierarchical feature learning directly from pixels. Early studies repurposed general-purpose CNNs such as AlexNet and VGGNet.The authors in [12], demonstrated an AlexNet-style CNN for mass classification, reporting roughly 84% accuracy. With greater depth and capacity, VGGNet achieved an AUC of 0.89 on INbreast, [13].

Subsequent architectures emphasized stability and feature reuse. ResNet, via residual connections, enabled deeper optimization and reached an AUC of 0.93 on CBIS-DDSM, [14]. DenseNet further improved gradient flow and feature reuse, with studies reporting around 91–92% accuracy on MIAS and CBIS-DDSM, [15], [16]. For segmentation, U-Net has become the de facto standard, consistently yielding a Dice score greater than 0.87 for lesion delineation, [17].

In parallel, resource-aware models explored the accuracy–efficiency frontier. EfficientNet-B4 delivered AUC $\approx$ 0.94 with reduced memory footprint, [18]. Mobile-first families (e.g., MobileNet) have proven attractive for embedded use, attaining $\sim$88–90% accuracy on mammography while lowering compute and energy demands, [19], [20]. Table 10 (Appendix) summarizes representative deep models, their reported performance, and practical trade-offs relevant to clinical deployment.

2.4 Comparative Insights and Clinical Impact

Taken together, contemporary deep learning models have materially advanced breast imaging practice, delivering higher sensitivity and specificity than conventional workflows. Architectures that emphasize feature reuse, such as DenseNet, demonstrate consistent gains in small-lesion recognition, with reports exceeding 91% accuracy, [16]. For delineation tasks, U-Net has become a reference design, routinely achieving Dice > 0.87, [17]. On the efficiency front, compound scaling in EfficientNet has enabled strong accuracy (AUC ≈ 0.94) with lower computational cost, [18]. Meanwhile, mobile-oriented families such as MobileNetV2/V3 deliver competitive accuracy ($\sim$89–93%) while remaining suitable for embedded or point-of-care deployments, [19], [20].

Important practical caveats persist. High-capacity backbones (e.g., ResNet, DenseNet) can be memory- and time-intensive, which complicates their use in constrained clinics or edge devices. Conversely, lightweight models may trade a few points of peak accuracy for speed and energy savings. Segmentation pipelines (e.g., U-Net) are further dependent on pixel-level annotations that are costly to curate, and lean segmentation variants (e.g., LightSegNet) can be vulnerable to acquisition noise or motion artifacts, [21].

These observations suggest models that strike a balance between accuracy, computational efficiency, interpretability, and deployability. Such a balance is especially critical in low-resource settings, where inference latency, energy budget, and maintainability are binding constraints. In this spirit, the present study focuses on an optimized mobile backbone (MobileNetV3) and a transformer-based alternative (ViT-L16), probing how each trades off robustness and cost to enable scalable, trustworthy breast cancer diagnosis across diverse care environments. A concise overview of clinical implications by architecture is provided in Table 11 (Appendix).

3 Dataset and Methodology

This section presents the datasets used in this study, followed by the unified preprocessing strategy and the methodological framework adopted for training and evaluating the three deep learning models under comparison.

3.1 Datasets

To comprehensively assess the performance of deep learning architectures in breast cancer classification, we consider two publicly available datasets representative of different imaging modalities: mammography (MIAS) and histopathology (BreakHis).

3.1.1 MIAS Dataset

The Mammographic Image Analysis Society (MIAS) dataset is a widely used benchmark for

mammography-based breast cancer research, [22]. It was developed in collaboration with the UK National Breast Screening Programme to support the development of computer-aided diagnostic systems.

Origin and Acquisition Protocol The database contains 322 digitized grayscale mammographic images of 161 patients, including left and right mediolateral oblique (MLO) views. The images were acquired with a spatial resolution of 50 microns per pixel, an 8-bit depth, and an optical density range of 0 to 3.2, and stored in the Portable GrayMap (PGM) format. Each image is annotated with metadata, including tissue density, lesion type, severity, and coordinates of the abnormality.

Class Selection and Encoding The MIAS catalog includes several lesion phenotypes (e.g., spiculated masses, architectural distortion, asymmetry). For this study, we selected three clinically representative classes:

- **Normal (NORM)** – Images without reported abnormality.
- **Calcification (CALC)** – Microcalcifications that may signal early lesions.
- **Circumscribed Mass (CIRC)** – Well-delimited masses, potentially benign or malignant.

These classes were assigned integer labels (0: Normal, 1: CALC, 2: CIRC) to standardize supervision and streamline the training pipeline.

Class Distribution and Imbalance Among the 255 retained samples, 207 correspond to normal tissue, while only 25 and 23 belong to CALC and CIRC, respectively. This pronounced skew (Figure 1) motivated targeted rebalancing via augmentation, as detailed in Section 3.2.

Justification We selected MIAS because it pairs accessible, well-documented mammograms with rich annotations that are directly actionable for supervised learning. Its moderate size supports rapid iteration and ablation while preserving enough heterogeneity to benchmark classification pipelines under controlled conditions.

3.1.2 BreakHis Dataset

The Breast Cancer Histopathological Image (BreakHis) collection complements our setting by providing a microscopic viewpoint in addition to mammography. Introduced by [23], it has become a reference corpus for benchmarking histopathology classifiers.

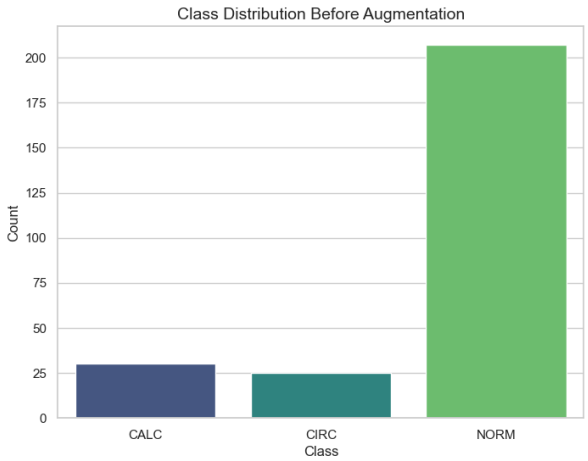

Fig. 1: Original class distribution in MIAS dataset before data augmentation

Origin and Acquisition Protocol BreakHis comprises 7,909 colour micrographs from 82 patients. Specimens were stained with Hematoxylin and Eosin (H&E) and digitised through an optical microscope at four magnifications ($40\times$, $100\times$, $200\times$, $400\times$). All images have a resolution of 700×460 pixels and are stored in PNG format, which limits acquisition variability beyond the intended magnification factors, [23].

Class Selection and Encoding The dataset groups diagnoses into two superclasses—**benign** (adenosis, fibroadenoma, phyllodes tumour, tubular adenoma) and **malignant** (ductal, lobular, mucinous, papillary carcinomas). In line with the clinical questions addressed in this paper, we work with two labelling schemes:

- A *binary* task, encoded as benign $= 0$ and malignant $= 1$; and
- A *three-class* task retaining **Benign** (all benign subtypes pooled) and the two malignant subtypes with the most material in practice—**Ductal carcinoma (DUC)** and **Mucinous carcinoma (MUC)**.

This mirrors the MIAS protocol (binary normal vs. abnormal; three classes) while respecting BreakHis subtype availability.

Class Distribution and Imbalance The raw corpus is skewed towards malignant tissue (2,480 benign vs. 5,429 malignant images). In the three-class variant, Benign dominates, whereas DUC and MUC remain comparatively under-represented. The global benign/malignant split is shown in Figure 2.

To reduce bias, we combine magnification-aware sampling with targeted augmentation (Section 3.2) and enforce patient-level, magnification-stratified partitions so that images from the same case never appear across different splits.

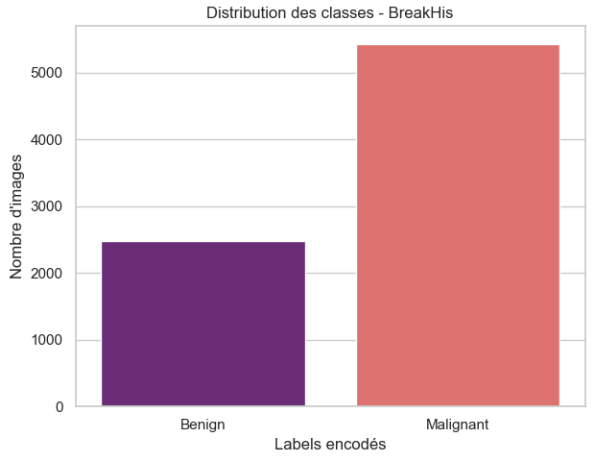

Fig. 2: BreakHis: benign vs. malignant distribution prior to rebalancing

Justification and Relevance Compared with MIAS, BreakHis stresses different visual cues—texture, nuclei morphology, and stain variability—rather than radiographic contrast. Assessing the same architectures (custom CNN, MobileNetV3, ViT-L16) on both binary and three-class BreakHis settings, therefore, probes cross-modal generalization and subtype separability (DUC, MUC) against a pooled benign class. Paired with MIAS, this results in a multimodal, multi-task testbed that is closer to real clinical heterogeneity and more informative about deployability.

3.2 Data Preprocessing and Augmentation

3.2.1 Data Preprocessing

We applied a single, carefully controlled preprocessing workflow to MIAS and BreakHis, ensuring that architectural comparisons were methodologically fair. Although the two corpora differ in modality—grayscale mammography for MIAS and RGB histopathology for BreakHis—the same sequence of transformations was enforced to the extent compatible with each domain.

Image Resizing and Normalization All images were resized to 224×224 pixels to match the input geometry of MobileNetV3 and ViT-L16 and to permit the direct use of pretrained weights. For MIAS (grayscale), intensities were first scaled to $[0, 1]$, then standardised with dataset-level mean and standard deviation (z-score). For BreakHis (RGB), each channel was normalised with the ImageNet statistics.

$$\mu = [0.485, 0.456, 0.406], \qquad \sigma = [0.229, 0.224, 0.225],$$

This aligns colour distributions with those seen during pretraining and stabilises transfer learning.

Format Conversion and Compression MIAS images originally supplied as PGM were converted to PNG for broad library compatibility; BreakHis images (PNG) were used as provided. To reduce I/O overhead and maintain a faithful record of the processed tensors, image arrays, and metadata were serialized into NumPy `.npz` files. This compressed format both speeds up training-time loading and preserves a reproducible snapshot of the preprocessing stage.

Label Harmonisation and Encoding A consistent labelling scheme was adopted across datasets:

- **MIAS:** three diagnostically meaningful classes—`NORMAL (0)`, `CALC (1)`, `CIRC (2)`—with accompanying CSVs that retain tissue type, severity, and lesion geometry.
- **BreakHis (binary):** the eight diagnostic subtypes were collapsed into `BENIGN (0)` vs. `MALIGNANT (1)` to reflect the screening decision.
- **BreakHis (three-class):** We used `Benign/Normal (0)`, `Ductal carcinoma (1)`, and `Mucinous carcinoma (2)`, a choice guided by prevalence and clinical salience, mirroring the three-class MIAS setting.

Data Splitting Each dataset was partitioned into training/validation splits of 80% and 20%, respectively. For BreakHis, splits were *patient-level* and *magnification-aware* to prevent any leakage of slides from the same case across folds. Splits were generated with PyTorch's `random_split` and stored to ensure identical reuse across experiments.

Loader Integration Custom PyTorch `Dataset` classes handled CSV parsing, lazy decompression of `.npz` arrays, resizing/normalisation/tensor conversion, and (when enabled) the retrieval of augmented variants (Section 3.2.2). These classes plug into the standard `DataLoader` for efficient mini-batching.

Reproducibility Measures To support verifiability, we fixed random seeds (NumPy, PyTorch), verified file integrity prior to compression (hash checks), and archived all scripts plus mapping files. These controls are essential in medical AI, where deployability depends on traceable and repeatable pipelines.

3.2.2 Data Augmentation

Because both MIAS and BreakHis are relatively small and markedly imbalanced, augmentation was used not merely as a regularizer, but as a core design choice to improve generalization and curb overfitting, while maintaining the clinical fidelity of the images to their acquisition processes.

Oversampling to Correct Class Imbalance An exploratory count highlighted severe skew in both corpora. In MIAS, *NORMAL* images (207) far exceeded *CALC* (25) and *CIRC* (23) cases (Figure 1). BreakHis exhibited a different but equally problematic pattern: malignant slides outnumbered benign ones (5,429 vs. 2,480), and the gap persisted within each magnification tier (40×, 100×, 200×, 400×). To neutralise these effects, we adopted class-wise oversampling via geometric variants of minority samples. For MIAS, *CALC* and *CIRC* were augmented until parity with *NORMAL* (207 per class; Figure 3). For BreakHis, balancing was performed *per magnification level* to avoid cross-scale bias, yielding comparable benign/malignant counts at 40×, 100×, 200×, and 400× (Figure 4). This produced training sets that are both balanced and representative.

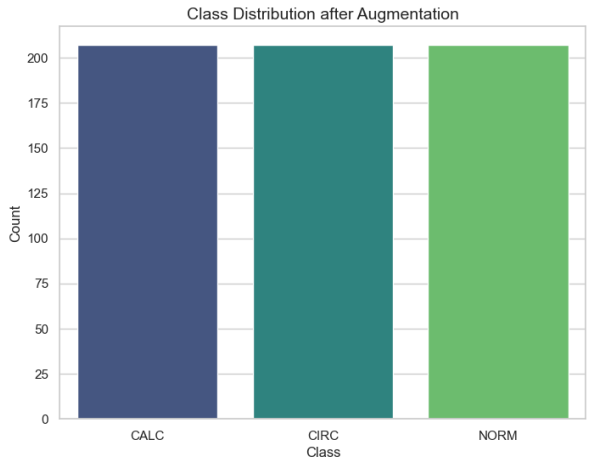

Fig. 3: Balanced class distribution in MIAS after augmentation

Transformation Palette We restricted ourselves to lightweight geometric operations that are both computationally cheap and clinically plausible:

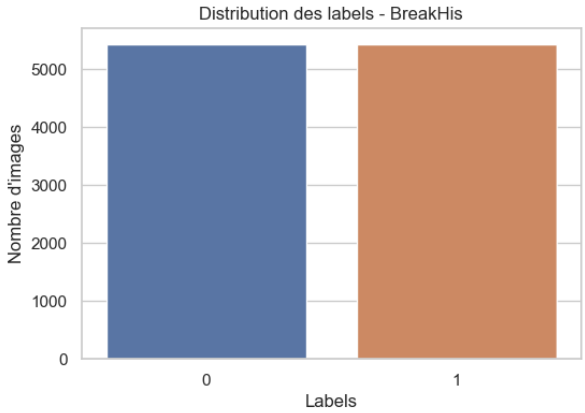

Fig. 4: Balanced class distribution in BreakHis after augmentation

- **Small rotations** (up to ±15°) to reflect positioning variability in mammography and slide orientation in microscopy;
- **Horizontal/vertical flips** to model left–right symmetry and benign orientation changes;
- **Minor translations** (up to 10% of width/height) to increase robustness to shifts during acquisition.

Transforms were applied stochastically at training time using `torchvision.transforms`, so that successive epochs saw distinct but realistic views of the same anatomy.

Clinical Rationale Augmentations were chosen with domain constraints in mind. For MIAS, geometry may change slightly between exposures, whereas intrinsic lesion appearance should not; geometric-only transforms preserve that balance. For BreakHis, H&E colour cues are diagnostic; we therefore *avoided* colour jittering and histogram equalisation. Likewise, aggressive schemes (CutMix, MixUp, elastic warps) were excluded to prevent the fabrication of non-physiological textures or the erasure of subtle cues.

Integration and Reproducibility All augmented tensors were stored in compressed NumPy `.npz` archives and indexed in updated CSVs (e.g., `mias_augmented.csv`, `breakhis_augmented.csv`). The pipeline is fully scriptable: given the same seeds and configuration, the augmented sets can be regenerated bit-for-bit, enabling fair re-runs and external benchmarking.

Summary In short, augmentation served two purposes: restoring class balance and injecting

realistic variability without compromising clinical plausibility. This yielded training distributions that were better matched to deployment conditions, allowing for a fair head-to-head evaluation of MobileNetV3, ViT-L16, and the custom CNN in the subsequent experiments.

4 Model Descriptions

We compare three neural architectures representative of distinct design philosophies: a bespoke Convolutional Neural Network (CNN), the mobile–friendly MobileNetV3, and a transformer-based Vision Transformer (ViT L16). For each, we first recall the guiding principles and then specify the configuration actually used in our experiments.

4.1 Custom CNN Architecture

4.1.1 General Overview

CNNs remain a dependable baseline in medical image analysis because local connectivity and weight sharing promote hierarchical feature formation. In breast imaging, these properties help capture texture, edge patterns, and lesion morphology that are relevant to diagnosis, [4], [12]. The main trade-off, relative to attention-based models, is a more limited ability to encode long-range dependencies or subtle global context.

4.1.2 Implementation in This Study

Design Rationale Our custom network (*NCNN*) ingests RGB images and deliberately keeps the backbone compact to stabilise training and contain memory usage. The layout is conventional: two convolution–pooling blocks extract spatial features, followed by three fully connected (FC) layers that map features to class logits. We adopt 5×5 kernels with modest channel counts to remain sensitive to texture while avoiding unnecessary computation. Because most parameters concentrate in the first FC layer, we apply dropout at that stage to curb overfitting.

Layer Specification Inputs of shape $[3, 224, 224]$ are processed as follows:

- **Conv1**: $3 \to 6$ channels, 5×5 kernels; output $(6, 220, 220)$; ReLU.
- **MaxPool1**: 2×2, stride 2; output $(6, 110, 110)$.
- **Conv2**: $6 \to 16$ channels, 5×5 kernels; output $(16, 106, 106)$; ReLU.
- **MaxPool2**: 2×2, stride 2; output $(16, 53, 53)$.
- **Flatten**: $(16, 53, 53) \to (44{,}944)$.
- **FC1**: $44{,}944 \to 120$; ReLU; dropout $p=0.3$.
- **FC2**: $120 \to 84$; ReLU.
- **FC3 (output)**: $84 \to C$ logits with $C=3$ (multiclass) or $C=2$ (binary).

The head outputs raw logits; cross-entropy computes the softmax internally. As a visual summary, the overall topology of NCNN is depicted in Figure 5.

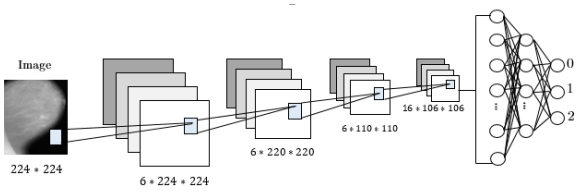

Fig. 5: S chematic of the NCNN used for breast cancer classification

A layer-by-layer summary with tensor shapes and parameter counts is provided in Table 1.

Table 1. NCNN layer dimensions and parameter counts (multiclass configuration, $C=3$)

Layer	Input Shape	Output Shape	Parameters
Conv1	(3, 224, 224)	(6, 220, 220)	456
MaxPool1	(6, 220, 220)	(6, 110, 110)	0
Conv2	(6, 110, 110)	(16, 106, 106)	2,416
MaxPool2	(16, 106, 106)	(16, 53, 53)	0
Flatten	(16, 53, 53)	(44,944,)	0
FC1	(44,944,)	(120,)	5,393,400
FC2	(120,)	(84,)	10,164
Output (FC3)	(84,)	(3,)	255
Total	-	-	**5,406,691**

Training Setup All experiments used PyTorch with cross-entropy loss and the Adam optimiser. Images were resized to 224×224. Early stopping was triggered by validation loss, with a maximum of 100 epochs. The classifier head used $C=2$ for binary detection and $C=3$ for multiclass tasks (MIAS: `NORMAL`/`CALC`/`CIRC`; BreakHis: `Benign`/`DUC`/`MUC`).

4.2 MobileNetV3

4.2.1 General Overview

MobileNetV3, [20], is a family of convolutional networks engineered for low-latency, low-power vision tasks. Its design couples depthwise–separable convolutions with inverted residual blocks (linear bottlenecks) and channel–wise recalibration via Squeeze-and-Excitation (SE), while employing efficient nonlinearities (ReLU6 / h-swish). In practice, this combination delivers a favourable speed–accuracy trade-off, making MobileNetV3 an appealing backbone for point-of-care or embedded clinical scenarios.

In this work, we adopt the *Large* variant, initialised on ImageNet and adapted to our two settings (binary detection and three-class classification). The generic ImageNet head is removed and replaced by a compact task-specific classifier, allowing the network to retain its pretrained feature extractor while specialising the final mapping to our label spaces.

4.2.2 Implementation Details

Initialisation and Transfer Setup We instantiate `torchvision.models.mobilenet_v3_large` with pre-trained weights. During fine-tuning, the feature extractor remains intact, and only the task head is newly learned. Optionally, early layers can be frozen for the first epochs to stabilize optimization on small medical datasets before being unfrozen for joint fine-tuning. This staged protocol reduces the risk of overfitting while preserving domain-agnostic representations learned on ImageNet.

Classifier Adaptation The original classifier (`nn.Linear(1280, 1000)`) is bypassed using `nn.Identity()` so that the network exposes the 1280-D global embedding. We then add a single linear layer,

$$\texttt{fc\_out} = \texttt{nn.Linear}(1280,\ \texttt{num\_classes}),$$

with `num_classes` set to 2 (binary) or 3 (multiclass). Training uses `CrossEntropyLoss`, which internally applies the softmax to the logits; at inference, probabilities can be recovered with an explicit softmax if needed.

Forward Path (Data Flow) The end-to-end flow is:

1. An RGB input of size $3\times224\times224$ is processed by the MobileNetV3 feature stack (stem convolution followed by inverted residual blocks with SE and ReLU6/h-swish).

2. Global average pooling collapses spatial dimensions to a 1280-D feature vector.

3. The ImageNet head is skipped via an identity layer.

4. The custom `fc_out` projects the embedding to C logits ($C \in \{2, 3\}$).

Figure 6 illustrates the adapted MobileNetV3_Large pipeline used in our experiments.

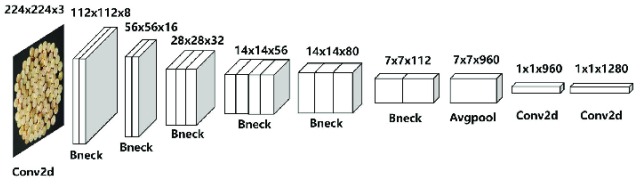

Fig. 6: Schematic of the adapted MobileNetV3_Large used in this study. The pretrained trunk is retained; a lightweight task-specific classifier replaces the ImageNet head

Why MobileNetV3 Here? Compared with conventional CNNs, MobileNetV3 achieves competitive accuracy with substantially fewer parameters ($\sim$5.4 M for the Large variant) and reduced multiply–accumulate operations, owing to depthwise separable kernels and linear bottlenecks. This efficiency is valuable for: (i) rapid experimentation on modest GPUs/CPUs, (ii) prospective deployment on embedded devices, and (iii) lowering energy and memory footprints in clinical IT environments.

Training Notes We fine-tune with Adam (cross-entropy objective), data normalization consistent with ImageNet statistics (for RGB inputs), and moderate regularization (weight decay, dropout in the classifier). Early stopping is based on validation loss; cosine or step-wise learning-rate schedules can further stabilise convergence. In our experiments, this configuration consistently yielded strong validation performance, especially when paired with the balanced augmentation protocol described in Section 3.2.

4.3 Vision Transformer (ViT L16)

4.3.1 General Overview

The Vision Transformer (ViT), [7], departs from a convolutional design by applying a pure transformer stack to image patches. An input image is partitioned into fixed-size tiles, and each tile is linearly embedded; the resulting token sequence is then processed with multi-head self-attention, which enables direct, long-range interactions across the entire field of view. This global receptive field, coupled with positional encodings to retain spatial layout, makes ViT well-suited to recognising subtle, spatially dispersed cues.

In this study, we employ the *Large* configuration with 16×16 patch size (ViT-L/16), initialised on ImageNet and adapted to our binary and three-class tasks. The original ImageNet classifier is removed and replaced with a compact task head, allowing the

pretrained transformer trunk to act as a high-capacity feature extractor tailored to our label spaces.

Key architectural components are:

- a patch embedding layer that maps a $224\times224\times3$ image to $14\times14 = 196$ tokens (each of dimension 1024 for ViT-L);
- learned positional embeddings added to every token;
- a stack of encoder blocks (multi-head self-attention + MLP) with residual connections and LayerNorm;
- a `[CLS]` token prepended to the sequence, whose final representation is used for classification.

4.3.2 Implementation Details

Pretrained Initialisation We load ViT-L/16 weights that are pre-trained on ImageNet-1K via `timm` (or HuggingFace). This transfer initialisation provides strong, generic visual primitives—especially useful given the relatively modest size of medical datasets—thereby shortening warm-up and improving optimisation stability.

Head Replacement and Fine-Tuning To tailor the network to our tasks:

1. the ImageNet head (`Linear(1024, 1000)`) is replaced by `nn.Identity()` so the model exposes the final `[CLS]` embedding;
2. we append a new linear classifier, `fc_out` = `nn.Linear`$(1024,$ `num_classes`$)$, with `num_classes` $\in \{2, 3\}$;
3. training uses `CrossEntropyLoss` on logits; probabilities at inference are obtained via softmax when needed.

A staged schedule (briefly freezing lower blocks before full unfreezing) can be used to curb overfitting at the beginning of fine-tuning; in practice, both staged and end-to-end regimes yielded stable convergence under our augmentation protocol.

Forward Path The computation proceeds as follows:

1. **Patch embedding**: non-overlapping 16×16 patches are flattened and projected to 1024-D vectors;
2. **Tokenisation & positions**: a `[CLS]` token is prepended; positional embeddings are added;
3. **Encoder stack**: the token sequence passes through the transformer blocks (MHSA + MLP) with residual connections and LayerNorm;
4. **Readout**: the terminal `[CLS]` representation is extracted;
5. **Task head**: `fc_out` maps the `[CLS]` vector to class logits.

Advantages and Applicability ViT's attention mechanism affords image-wide context from the outset, contrasting with the locality of small convolutional kernels. For medical imaging, where discriminative evidence may be faint and spatially distributed, this global coupling is advantageous. Although ViT-L has a higher capacity than lightweight CNNs, transfer initialisation and moderate regularisation (weight decay, dropout in the head, and early stopping) keep training costs manageable. In our experiments, the adapted ViT-L/16 consistently provided strong validation results on both MIAS and BreakHis, complementing the efficiency of MobileNetV3 and the transparency of the custom CNN.

A side-by-side synopsis of the three architectures and their computational profiles is provided in Table 12 (Appendix).

5 Experimental Setup

5.1 Experiment Parameters

All models were trained on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU, an AMD Ryzen 9 7900X CPU, and 64 GB of RAM. Automatic mixed precision (AMP) was enabled to accelerate training and reduce the memory footprint. Unless otherwise noted, each experiment was repeated three times with distinct random seeds to account for training stochasticity; we report the mean performance across runs.

To ensure a fair comparison, we used a unified training configuration and identical data splits across models (cf. Section 3). Unless stated otherwise (e.g., ablation runs for NCNN up to 100 epochs), the final model selection adopted the following settings:

- **Loss:** Cross-entropy.
- **Optimizer:** Adam, initial learning rate 3×10^{-4}, cosine decay schedule.
- **Epoch budget:** 40 epochs for MIAS; 30 epochs for BreakHis.
- **Batch size:** 32 (both datasets).
- **Early stopping:** Patience of 5 epochs on validation loss.

For transfer learning models (MobileNetV3, ViT-L/16), we replaced the ImageNet head with a task-specific linear layer and fine-tuned end-to-end after a brief warm-up of the classifier head.

5.2 Evaluation Metrics

We evaluated all configurations using standard metrics derived from the confusion matrix, including accuracy, precision, recall (also known as sensitivity), and F1-score, for multiclass experiments (e.g., *NORMAL*, *CALC*, *CIRC* on MIAS; *Benign*, *DUC*, *MUC* on BreakHis), per-class precision/recall/F1 were computed in a one-vs-all manner and averaged using the *macro* scheme, giving equal weight to each class—an important choice in imbalanced settings.

5.2.1 Accuracy (Acc)

Overall proportion of correct predictions:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (1)$$

5.2.2 Precision (Prec)

Fraction of predicted positives that are truly positive:

$$\text{Prec} = \frac{TP}{TP + FP}. \quad (2)$$

5.2.3 Sensitivity / Recall (Rec)

Fraction of actual positives correctly identified:

$$\text{Rec} = \frac{TP}{TP + FN}. \quad (3)$$

5.2.4 F1-Score

Harmonic mean of precision and recall:

$$\text{F1} = \frac{2 \times \text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}}. \quad (4)$$

5.2.5 Confusion Matrix

The confusion matrix tabulates true vs. predicted labels and helps identify class-specific error patterns.

For completeness, the generic structure of a confusion matrix used throughout this study is reported in Table 2.

Table 2. Structure of the confusion matrix predictions

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Positive	FP	TN

5.2.6 ROC Curve and AUC

Receiver Operating Characteristic (ROC) curves plot true-positive rate against false-positive rate across thresholds; the Area Under the Curve (AUC) provides a threshold-independent summary (0 to 1). When relevant and available, we report AUC alongside the primary metrics above.

6 Experimental Results

6.1 Experimental Results on MIAS Dataset

We report the performance of the three architectures—Custom CNN, MobileNetV3, and Vision Transformer (ViT L16)—on the MIAS dataset. Two tasks were considered to mirror clinical use cases: (i) binary detection (normal vs. abnormal), and (ii) three-way classification among *Normal*, *CIRC*, and *CALC*. Unless stated otherwise, metrics are computed on held-out test sets and include accuracy, sensitivity, precision, and F1-score.

6.1.1 Case Study 1: Binary Classification (Detection)

The detection task collapses all lesions (benign or malignant) into a single *abnormal* category. The numerical results are summarized in Table 3.

Table 3: Performance on Binary Classification Task (MIAS Dataset)

Model	Accuracy	Sensitivity	Precision	F1-Score
Custom CNN	91.50%	91.48%	91.48%	91.48%
MobileNetV3	**93.25%**	**93.13%**	**93.40%**	**93.22%**
ViT L16	89.75%	89.80%	89.80%	89.80%

Overall, MobileNetV3 attains the best balance across metrics (accuracy, sensitivity, precision, and F1), indicating a consistent operating point on this binary decision boundary. ViT L16 remains competitive but trails MobileNetV3 by a few percentage points, while the custom CNN delivers solid performance given its simplicity. The confusion matrices in Figure 7, Figure 8 and Figure 9 illustrate the residual error patterns for each model.

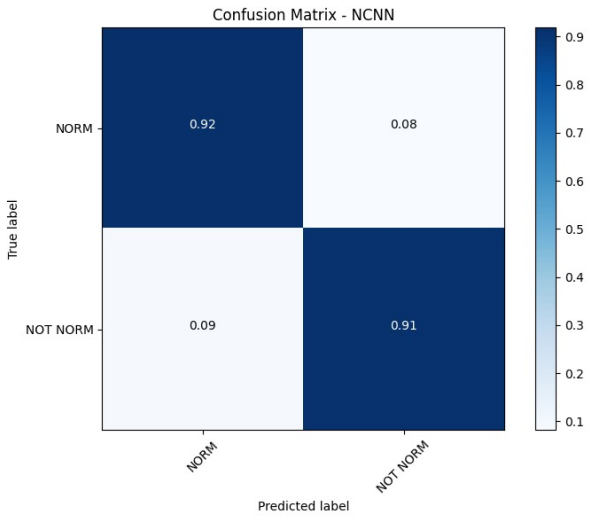

Fig. 7: Confusion matrix of the Custom CNN model (MIAS—detection)

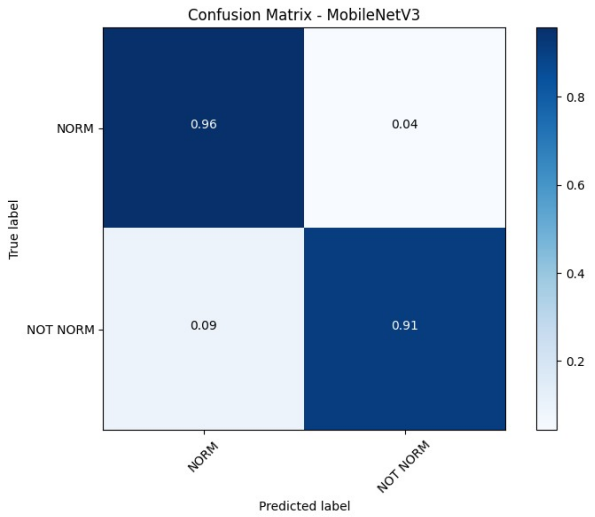

Fig. 8: Confusion matrix of the MobileNetV3 model (MIAS—detection)

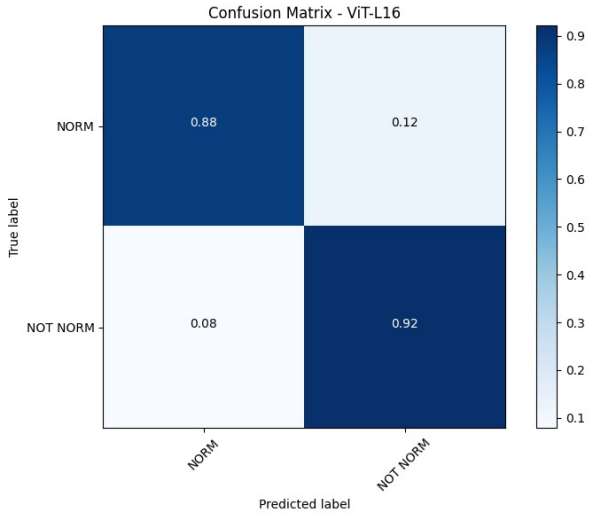

Fig. 9: Confusion matrix of the ViT L16 model (MIAS—detection)

6.1.2 Case Study 2: Multiclass Classification

We next evaluate the three-way classification among *Normal*, *CALC*, and *CIRC*. Table 4 reports the aggregate metrics.

Table 4. Performance on Multiclass Classification Task (MIAS Dataset)

Model	Accuracy	Precision	Recall	F1-Score
Custom CNN	75.20%	74.94%	76.38%	74.67%
MobileNetV3	**92.00%**	**92.00%**	**92.00%**	**92.00%**
ViT L16	**92.00%**	91.87%	92.57%	92.04%

On this more granular task, MobileNetV3 and ViT L16 both reach ≈92% accuracy with closely aligned precision/recall/F1, suggesting that pre-trained, high-capacity backbones transfer well to MIAS despite the limited sample size. The custom CNN exhibits a notable drop relative to the binary setting, which is consistent with the additional inter-class separation required among *CALC* and *CIRC*. Detailed confusion matrices are provided in Figure 10, Figure 11 and Figure 12.

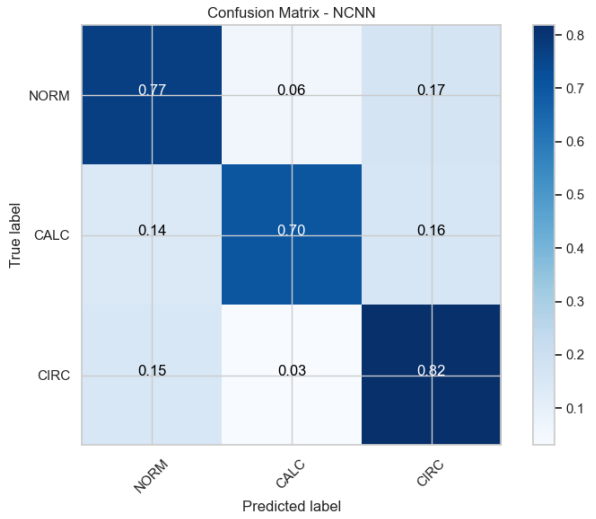

Fig. 10: Confusion matrix of the Custom CNN model (MIAS—multiclass)

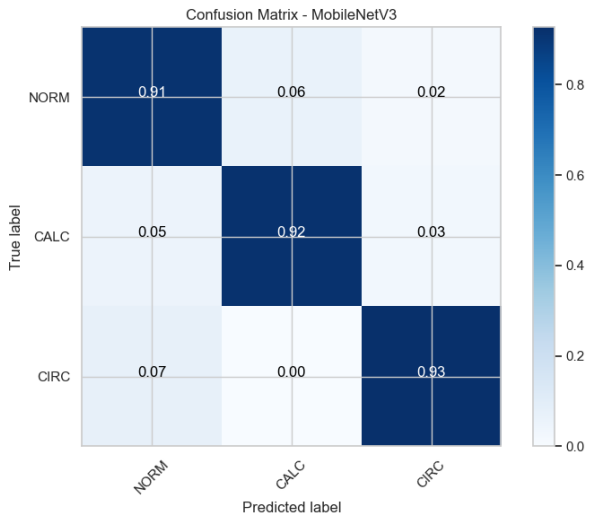

Fig. 11: Confusion matrix of the MobileNetV3 model (MIAS—multiclass)

6.2 Experimental Results on BreakHis Dataset

We now evaluate the three architectures—Custom CNN, MobileNetV3, and ViT L16—on the BreakHis

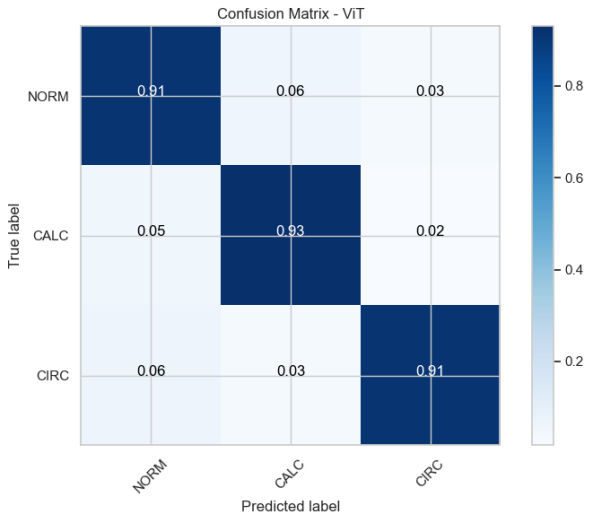

Fig. 12: Confusion matrix of the ViT L16 model (MIAS—multiclass)

dataset. As for MIAS, two tasks are considered: (i) binary detection (benign vs. malignant) and (ii) three-class classification. All metrics (accuracy, precision, recall, F1-score) are reported on the validation split.

6.2.1 Case Study 1: Binary Classification (Detection)

The detection task distinguishes between benign and malignant tissue. Table 5 summarizes the performance for each model.

Table 5. Performance on Binary Classification Task (BreakHis Dataset)

Model	Accuracy	Precision	Recall	F1-Score
Custom CNN	84.53%	84.61%	84.53%	84.52%
MobileNetV3	89.78%	89.94%	89.78%	89.77%
ViT L16	**91.62%**	**91.62%**	**91.62%**	**91.62%**

ViT L16 achieves the highest scores across all indicators (91.62% accuracy and F1), indicating that global self-attention is effective for histopathological texture patterns. MobileNetV3 follows closely (notably 89.94% precision), while the custom CNN trails the two pre-trained backbones. Confusion matrices are shown in Figure 13, Figure 14 and Figure 15.

6.2.2 Case Study 2: Multiclass Classification

For the multiclass setting, models distinguish among *Normal/Benign*, *Ductal Carcinoma (DUC)*, and *Mucinous Carcinoma (MUC)*. Table 6 reports the validation results.

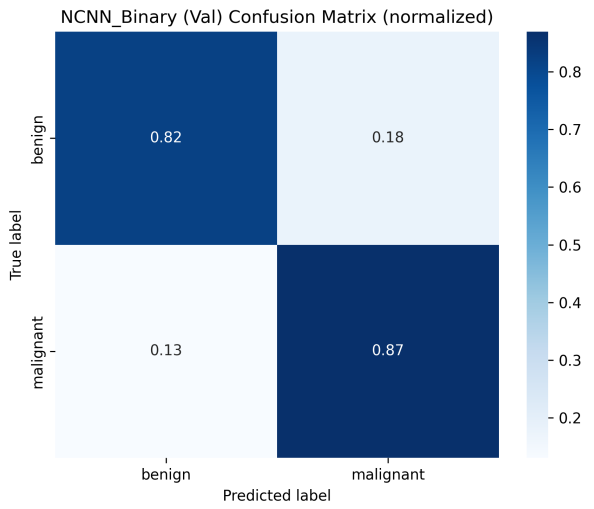

Fig. 13: C onfusion matrix of the Custom CNN model (BreakHis—detection)

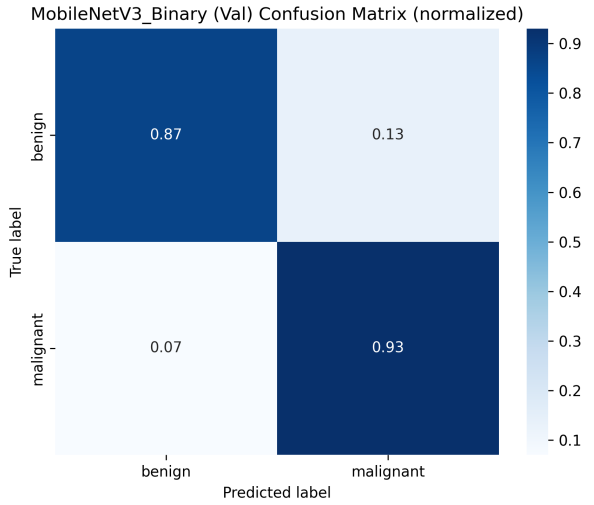

Fig. 14: Confusion matrix of the MobileNetV3 model (BreakHis—detection)

Table 6. Performance on Multiclass Classification Task (BreakHis Dataset)

Model	Accuracy	Precision	Recall	F1-Score
Custom CNN	79.11%	62.05%	59.64%	57.75%
MobileNetV3	85.52%	78.13%	78.99%	78.36%
ViT L16	**90.28%**	**86.64%**	**79.23%**	**82.05%**

ViT L16 again leads (90.28% accuracy), with a clear margin in precision (86.64%). MobileNetV3 remains competitive (85.52% accuracy), whereas the custom CNN shows larger drops on recall/F1, indicating difficulties in separating the malignant subtypes from the aggregated benign class. The corresponding confusion matrices in Figure 16, Figure 17 and Figure 18 provide a more granular view

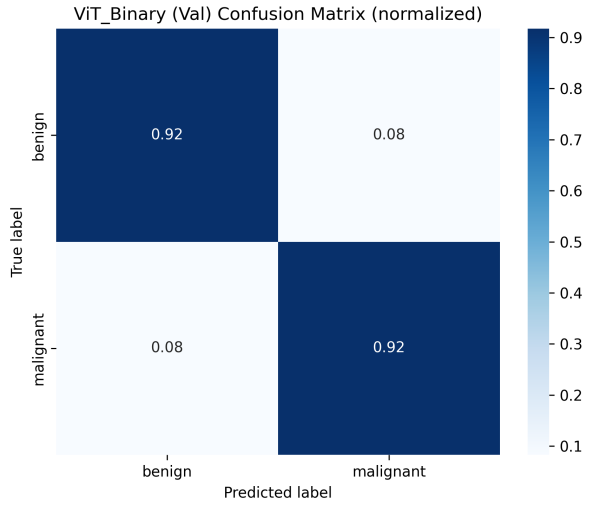

Fig. 15: Confusion matrix of the ViT L16 model (BreakHis—detection)

of per-class errors.

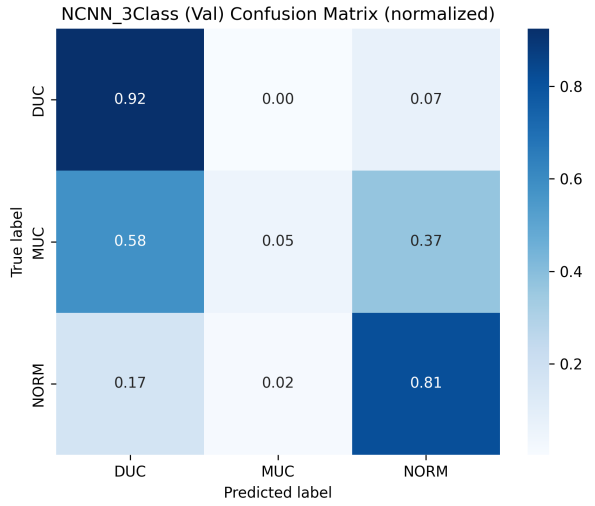

Fig. 16: Confusion matrix of the Custom CNN model (BreakHis—multiclass)

6.3 Analysis on MIAS Dataset

6.3.1 Intra-Model Comparison

Across the three architectures, performance is consistently higher on the binary detection task than on the multiclass setting. This pattern is expected, since binary discrimination imposes a simpler decision boundary and requires less fine-grained inter-class separation. The largest drop in accuracy between the two tasks is observed with the custom CNN, which decreases from 91.50% to 75.20%. This gap indicates limited generalization of a shallow CNN when faced with heterogeneous lesion categories

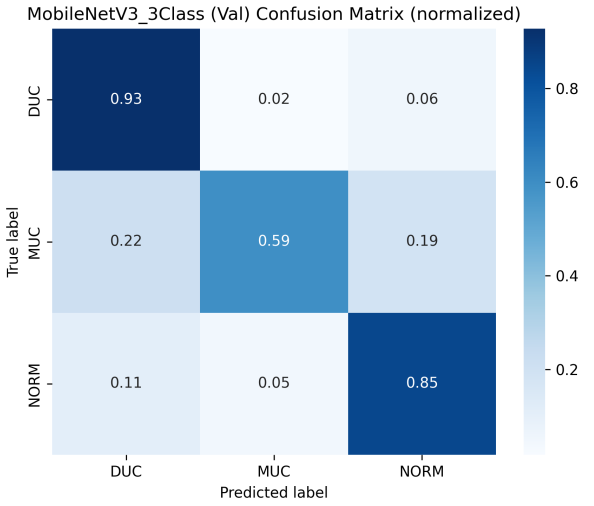

Fig. 17: Confusion matrix of the MobileNetV3 model (BreakHis—multiclass)

such as calcifications (CALC) and circumscribed masses (CIRC).

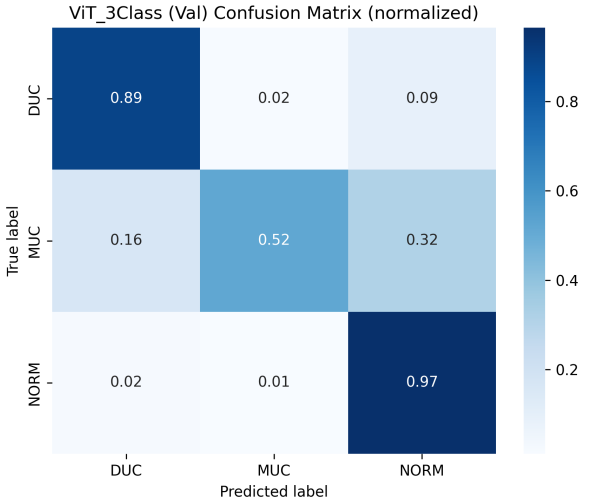

Fig. 18: Confusion matrix of the ViT L16 model (BreakHis—multiclass)

7 Analysis and Comparative Discussion

7.0.1 Inter-Model Comparison

MobileNetV3 achieves the strongest overall results on MIAS, achieving 93.25% accuracy for detection and 92.00% for multiclass classification. These outcomes reinforce the advantage of deeper, pre-trained backbones even when data are relatively scarce. ViT L16 also performs competitively (89.75% in binary; 92.00% in multiclass), suggesting that global context modeling via self-attention is

beneficial in mammography. By contrast, the custom CNN, while serviceable, lags behind, underscoring that handcrafted shallow designs are more sensitive to the dataset's limited size and class diversity unless complemented by stronger regularization or richer features.

7.0.2 Comparison with Related Work (MIAS)

Our findings align with, and in several cases exceed, prior reports on MIAS. A comparative study of AlexNet and GoogLeNet reported 88.24% accuracy and an AUC of 94.65% for GoogLeNet on MIAS, [15]. More recently, DenseNet-169 with a Swish activation achieved 91.72% accuracy on the same dataset, [16]. In comparison, MobileNetV3 achieves 93.25% accuracy in the binary task and 92.00% in multiclass classification, establishing a stronger benchmark under comparable validation protocols. Taken together, these results highlight the effectiveness of transfer learning and point to the promise of transformer-based approaches for mammography analysis.

7.0.3 Clinical Relevance and Deployment Considerations

From a clinical standpoint, the high sensitivity and precision achieved by MobileNetV3 and ViT L16 are desirable in screening pipelines: reducing false negatives is essential to avoid missed cancers, while controlling false positives limits unnecessary biopsies and patient anxiety. Additionally, MobileNetV3's computational efficiency makes it a practical option for on-device inference or telemedicine scenarios, particularly in scenarios where compute resources are constrained.

7.0.4 Limitations and Future Directions

Despite the encouraging results, MIAS remains a small dataset, and even with oversampling and augmentation, generalization may be limited. Broader validation on larger and complementary cohorts (e.g., BreakHis) is warranted to stress-test robustness across modalities and imaging conditions. A further priority is model interpretability: integrating techniques such as Grad-CAM or attention visualization (for ViT) could help elucidate decision drivers and support clinical adoption.

7.1 Analysis on BreakHis Dataset

7.1.1 Intra-Model Comparison

Across the two BreakHis tasks, all three models perform better on binary detection than on the three-class setting (Tables 5 and 6). For the custom CNN, accuracy drops from 84.53% (binary) to 79.11% (multiclass), with a sharp decline in macro F1 from 84.52% to 57.75%. This pattern suggests that the model struggles to capture subtle morphological cues that distinguish benign subtypes from malignant classes, and that minority classes have a significant impact on macro-averaged metrics. MobileNetV3 exhibits a milder degradation (accuracy: 89.78% $\rightarrow$ 85.52%; F1: 89.77% $\rightarrow$ 78.36%), consistent with its stronger feature extraction via depthwise-separable convolutions and Squeeze-and-Excitation. ViT L16 is the most stable: accuracy decreases only modestly (91.62% $rightarrow$ 90.28%), although macro F1 drops to 82.05%. The gap between precision (86.64%) and recall (79.23%) in the multiclass task indicates uneven sensitivity across classes, likely due to residual imbalance and inter-class morphological overlap.

7.1.2 Inter-Model Comparison

On BreakHis, ViT L16 consistently outperforms MobileNetV3, which in turn surpasses the custom CNN, for both tasks. In binary detection, ViT achieves 91.62% accuracy (macro F1: 91.62%), followed by MobileNetV3 at 89.78% (89.77%), and the custom CNN at 84.53% (84.52%). In the multiclass setting, ViT remains ahead (90.28% accuracy; 82.05% macro F1), then MobileNetV3 (85.52%; 78.36%), and the custom CNN (79.11%; 57.75%). These results support the view that transformer-based self-attention is advantageous for histopathology, where diagnostically relevant textures span multiple spatial scales and long-range dependencies. MobileNetV3 remains a strong and efficient baseline, while the shallow CNN is more sensitive to the higher intra-class variability present in BreakHis.

7.1.3 Comparison with Related Work (BreakHis)

Under protocols comparable to ours (image- or patch-wise classification with standard accuracy/F1 metrics), reported performances on BreakHis cover a broad range. Spanhol *et al.* (the original dataset paper) reported baselines around 80–85% for both binary and multiclass classification using classical descriptors with shallow classifiers, [23]. Agarwal *et al.* later achieved $\approx$94.7% accuracy in the binary task using VGG16-based transfer learning, albeit with lower recall ($\approx$80.52%), [24]. More recently, attention-based approaches (e.g., ECSAnet) improved robustness across magnifications, reaching up to $\approx$94.2% at 40$\times$ and 89–93% depending on magnification, [25]. Another study, reported binary accuracies around $\approx$89% and F1-scores near 0.90 across multiple CNNs (Xception, ResNet, InceptionResNet) under settings similar to ours, [26]. Positioned against this backdrop, our ViT L16 results (91.62% binary; 90.28% multiclass)

lie near the upper tier under comparable validation protocols, underscoring the competitiveness of transformer-based pipelines for histopathological analysis.

7.1.4 Clinical Relevance and Deployment Considerations

From a workflow perspective, the high recall of ViT in the binary task (91.62%) is desirable for reducing missed malignancies, whereas the multiclass recall (79.23%) indicates residual confusion among morphologically similar subtypes. Threshold tuning and cost-sensitive training can rebalance precision and recall according to clinical priorities, thereby optimizing the trade-off between precision and recall. MobileNetV3 offers a favorable efficiency profile for embedded or on-premise deployment where compute is limited, while ViT L16 provides additional robustness for centralized systems handling large, heterogeneous cohorts.

7.1.5 Limitations and Future Directions (BreakHis)

BreakHis exhibits variability across magnifications and staining, and macro-averaged metrics can be depressed by minority or challenging subclasses. Future work will incorporate stain normalization (e.g., Macenko/Reinhard), magnification-aware sampling, class-balanced or focal losses, and per-class calibration. Methodologically, hybrid or ensemble designs (CNN backbones coupled with transformer heads), multi-scale ViTs, and instance-level attention over tiles may further enhance sensitivity for challenging categories. Extending to slide-level (MIL) settings and adding cross-institutional validation would further strengthen clinical generalizability.

7.2 Cross-Dataset Comparative Analysis: MIAS vs. BreakHis

7.2.1 Summary table of cross-dataset performance

Table 7 summarizes the validation performance of the three evaluated models across both datasets and both tasks. Results are expressed as accuracy (%). The Figure 19 displays the accuracies of the three models on the two datasets (MIAS and BreakHis) for the two tasks (binary vs. multiclass).

Table 7. Cross-dataset accuracy summary (Validation / MIAS and BreakHis)

Model	MIAS (Binary)	MIAS (Multiclass)	BreakHis (Binary)	BreakHis (Multiclass)
Custom CNN (NCNN)	91.50%	75.20%	84.53%	79.11%
MobileNetV3	93.25%	92.00%	89.78%	85.52%
ViT L16	89.75%	92.00%	91.62%	90.28%

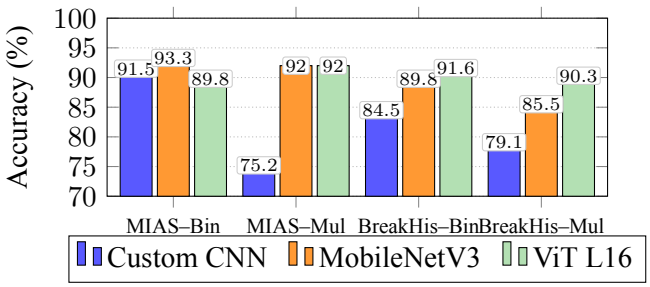

Fig. 19: Cross-dataset accuracy comparison (validation) for the three models

7.2.2 Key observations

- **Consistent task effect**: All models perform better on binary detection than on multiclass classification in most cases (Table 7). This is expected: binary detection requires a simpler decision boundary while multiclass classification demands finer discrimination among visually similar classes.
- **Model ranking varies with dataset characteristics**: On MIAS (grayscale mammograms, relatively small and homogeneous), MobileNetV3 attains the best performance overall (93.25% binary; 92.00% multiclass). On BreakHis (histopathological color images, greater intra-class variability and scale differences), ViT L16 leads (91.62% binary; 90.28% multiclass). MobileNetV3 remains competitive on BreakHis but loses some ground relative to ViT.
- **Shallow vs. deep / global-context models**: The Custom CNN (NCNN) shows the largest degradation when moving from MIAS to BreakHis and from binary to multiclass tasks (e.g., MIAS multiclass 75.20%). This suggests that shallow handcrafted architectures lack the capacity to capture the higher morphological variability and multi-scale texture present in histopathology images.
- **Transformer strengths on diverse data**: ViT's relative stability across datasets (small drop or even improvement in multiclass on MIAS) indicates that self-attention and global context modeling help capture complex spatial dependencies that are characteristic of histopathology and multiclass mammography tasks.

7.2.3 Explanatory factors

Several dataset–specific characteristics help account for the performance gaps observed between MIAS and BreakHis:

- **Imaging modality.** MIAS comprises digitized mammograms (grayscale, macroscopic breast anatomy), whereas BreakHis contains histopathology patches (RGB, microscopic tissue morphology). Cues that are discriminative in one modality need not transfer to the other.
- **Scale and texture.** BreakHis exhibits rich, fine-grained textures across multiple magnifications. Shallow convolutional stacks (e.g., NCNN) can under-aggregate long-range context, while the self-attention in ViT naturally links distant regions, improving discrimination among subtle patterns.
- **Data volume and diversity.** Transfer learning benefits (MobileNetV3, ViT pretrained on ImageNet) are more pronounced as sample diversity increases. On the smaller, less diverse MIAS set, MobileNetV3 offers a favorable bias–variance trade-off; on BreakHis, ViT scales better with heterogeneity.
- **Color information.** BreakHis images carry stain-dependent color signals that pretrained RGB backbones can exploit. Preserving channel interactions is therefore advantageous, particularly for transformer or modern CNN designs.
- **Label granularity and imbalance.** Three-class settings (e.g., CALC vs. CIRC on MIAS; Benign/DUC/MUC on BreakHis) amplify class imbalance and require finer boundaries. Models with limited representational capacity or without class-aware training tend to be disproportionately affected.

7.2.4 Practical implications and recommendations

- **Model choice by deployment context.** For on-device or point-of-care mammography screening, MobileNetV3 delivers a strong accuracy–efficiency compromise. For centralized analysis of histopathology with higher visual diversity, ViT-family models provide greater robustness.
- **Training protocols.** Improve cross-dataset generalization with domain-aware fine-tuning (e.g., magnification-conditioned curricula on BreakHis; stain-preserving preprocessing), strict patient-wise splits, and imbalance-oriented objectives (focal loss, calibrated oversampling).
- **Hybridization and ensembles.** Lightweight CNN backbones augmented with attention modules, or simple ensembles combining MobileNetV3 and ViT outputs, can recover both efficiency and global-context sensitivity and often yield consistent gains in binary and multiclass tasks.
- **Transparency and calibration.** Clinical adoption benefits from visual explanations (Grad-CAM, attention maps) and well-calibrated probabilities (temperature scaling, Dirichlet calibration) to quantify uncertainty—especially when models operate across modalities.

7.2.5 Limitations of the cross-dataset comparison

The synthesis in Table 7 should be interpreted with caution:

- **Protocol heterogeneity.** Variations in splitting strategy (image- vs. patient-level), magnification handling, augmentation, and averaging schemes can shift reported accuracy by several points, even with identical backbones.
- **Preprocessing sensitivity.** Steps such as CLAHE, normalization choices, or stain handling can differentially favor some architectures, complicating head-to-head contrasts.
- **External validity.** Results on curated public datasets are necessary but insufficient for clinical readiness; prospective, multi-center validation remains essential to establish generalizability.

7.2.6 Concluding remark

Taken together, the evidence suggests that there is no universally dominant architecture across modalities. Model selection should be aligned with the imaging domain, computational budget, and clinical question. In practice, our results support MobileNetV3 for efficient mammography triage and ViT-L16 for high-accuracy histopathology, with ensembles and domain-aware fine-tuning offering a pragmatic route toward robust, clinically deployable systems.

8 Conclusion and Future Work

This work compared three complementary deep learning architectures—a custom CNN, MobileNetV3, and a Vision Transformer (ViT-L16)—within a single, reproducible pipeline across two imaging modalities: mammography (MIAS) and histopathology (BreakHis). A harmonized protocol for preprocessing, augmentation, data splits, and evaluation ensured that models were judged under comparable conditions for both binary detection and three-class classification.

Key findings. On *MIAS*, MobileNetV3 delivered the strongest overall performance, attaining 93.25% accuracy for binary detection (normal vs. abnormal) and 92.00% for the three-class task (NORMAL, CALC, CIRC). ViT-L16 matched MobileNetV3 in the multiclass setting (92.00%) and remained competitive for binary detection (89.75%). On *BreakHis*, ViT-L16 was consistently superior: 91.62% accuracy (with precision/recall/F1 all 91.62%) for benign–malignant discrimination and 90.28% accuracy in the three-class configuration (NORMAL, DUC, MUC), with macro precision/recall/F1 of 86.64%/79.23%/82.05%. MobileNetV3 provided strong second-best results (89.78% and 85.52% accuracy in binary and multiclass experiments, respectively), whereas the custom CNN lagged, especially when the task required finer inter-class separation. Taken together, these outcomes suggest a practical division of labor: lightweight CNNs such as MobileNetV3 offer an attractive accuracy–efficiency trade-off for radiographic screening. At the same time, transformer models yield additional gains for texture-rich histopathology.

Methodological and practical contributions. Beyond raw performance, the work contributes (i) a *modality-agnostic* and *reproducible* pipeline (standardized 224×224 inputs, conservative but effective augmentation, patient-level splits for BreakHis, magnification-aware sampling, and seed control), (ii) *evidence-based guidance* for deployment: MobileNetV3 is well-suited to embedded or resource-constrained environments, while ViT L16 is preferable for centralized infrastructures handling heterogeneous, high-variability cohorts, and (iii) a *calibrated positioning* of our results with respect to the literature on MIAS and BreakHis, showing that our best models align with or surpass strong baselines under comparable validation conditions.

Clinical implications. In screening and triage workflows, the high sensitivity and precision observed with MobileNetV3 (MIAS) and ViT L16 (BreakHis) are desirable to reduce missed malignancies and unnecessary callbacks. Nevertheless, the precision–recall asymmetry observed for ViT L16 in BreakHis multiclass (precision 86.64% vs. recall 79.23%) highlights the need for task-specific thresholding and potential cost-sensitive optimization when false negatives carry higher clinical risk.

Limitations. MIAS remains small and imbalanced despite augmentation, which may limit generalization. BreakHis exhibits variability across stains and magnifications; even with magnification-aware sampling, class imbalance and subtype overlap depress the macro-averaged recall/F1-score. Our analysis relied on image/patch-level supervision rather than whole-slide inference and did not include prospective, multi-center external validation. Finally, we report standard metrics but did not quantify calibration (e.g., ECE) or uncertainty, which are critical for safe clinical integration.

Future work. Future research directions include: (i) extending validation to multi-center cohorts with prospective protocols; (ii) incorporating stain normalization and domain adaptation, as well as cost-sensitive/focal losses to improve recall on difficult subtypes; (iii) exploring hybrid CNN–Transformer designs and multi-scale ViTs, along with self-/semi-supervised pretraining to leverage unlabeled data; (iv) moving toward slide-level multiple instance learning for histopathology and view-aware modeling for mammography; (v) reporting clinically meaningful operating points (e.g., sensitivity at fixed specificity), model calibration, and uncertainty quantification; (vi) integrating explainability (Grad-CAM, attention rollouts) into user-facing prototypes; and (vii) optimizing on-device inference paths (e.g., INT8 quantization, mixed precision) to enable real-time, privacy-preserving deployment in resource-constrained settings.

In summary, a carefully engineered, modality-aware pipeline combined with judicious architectural choices yields competitive and deployable performance for breast cancer detection and classification across radiographic and histopathological images. The outlined roadmap targets the remaining gaps to bridge the last mile toward reliable, clinically integrated AI tools.

References:

[1] World Health Organization, "Cancer statistics," *Global Cancer Observatory*, 2023, Accessed: May 5, 2025. [Online]. https://gco.iarc.fr.

[2] World Health Organization, *Breast cancer fact sheet*, 2023. Accessed: May 5, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[3] C. A. Beam, P. M. Layde, and D. C. Sullivan, "Variability in the interpretation of screening mammograms by us radiologists: Findings from a national sample," *Archives of internal medicine*, vol. 156, no. 2, pp. 209–213, 1996. DOI: 10.1001/archinte.1996.00440020119016.

[4] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. DOI: 10.1016/j.media.2017.07.005.

[5] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, no. 1, pp. 221–248, 2017. DOI: 10.1146/annurev-bioeng-071516-044442.

[6] S. M. McKinney et al., "International evaluation of an ai system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020. DOI: 10.1038/s41586-019-1799-6.

[7] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. DOI: 10.48550/arXiv.2010.11929.

[8] D. Saslow et al., "American cancer society guidelines for breast screening with mri as an adjunct to mammography," *CA: A Cancer Journal for Clinicians*, vol. 57, no. 2, pp. 75–89, 2007. DOI: 10.3322/canjclin.57.2.75.

[9] R. A. Vierkant et al., "Mammographic breast density and risk of breast cancer in women with atypical hyperplasia: An observational cohort study from the mayo clinic benign breast disease (bbd) cohort," *BMC Cancer*, vol. 17, no. 1, p. 84, 2017. DOI: 10.1186/s12885-017-3082-2.

[10] L. Jiang, J. Gilbert, H. Langley, R. Moineddin, and P. A. Groome, "Breast cancer detection method, diagnostic interval and use of specialized diagnostic assessment units across ontario, canada," *Health Promotion and Chronic Disease Prevention in Canada*, 2018. DOI: 10.24095/hpcdp.38.10.02.

[11] E. H. Houssein, M. M. Emam, A. A. Ali, and P. N. Suganthan, "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review," *Expert Systems with Applications*, vol. 167, p. 114 161, 2021. DOI: 10.1016/j.eswa.2020.114161.

[12] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 248–257, 2016. DOI: 10.1016/j.cmpb.2015.12.014.

[13] L. Tsochatzidis, L. Costaridou, and I. Pratikakis, "Deep learning for breast cancer diagnosis from mammograms—a comparative study," *Journal of Imaging*, vol. 5, no. 3, p. 37, 2019. DOI: 10.3390/jimaging5030037.

[14] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53 040–53 065, 2019. DOI: 10.1109/ACCESS.2019.2912200.

[15] S. A. Hassan, M. S. Sayed, M. I. Abdalla, and M. A. Rashwan, "Breast cancer masses classification using deep convolutional neural networks and transfer learning," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 30 735–30 768, 2020. DOI: doi.org/10.1007/s11042-020-09518-w

[16] D. M. Bhatt, P. Oza, P. Sharma, and S. Patel, "Deep learning based novel approach for mammogram classification using densenet-169," in *International Conference on Machine Intelligence and Smart Systems*, Springer, 2023, pp. 3–13. DOI: doi.org/10.1007/978-3-031-31723-1_1

[17] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021. DOI: 10.48550/arXiv.2102.04306.

[18] B. Yuan et al., "Comparative analysis of convolutional neural networks and transformer architectures for breast cancer histopathological image classification," *Frontiers in Medicine*, vol. 12, p. 1 606 336, 2025. DOI: 10.3389/fmed.2025.1606336.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520. DOI: 10.1109/CVPR.2018.00474.

[20] A. Howard et al., "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019, pp. 1314–1324. DOI: 10.48550/arXiv.1905.02244.

[21] A. Vinod, P. Guddati, A. K. Panda, and R. K. Tripathy, "A lightweight deep convolutional neural network implemented on fpga and android devices for detection of breast cancer using ultrasound images," *IEEE Access*, 2024. DOI: 10.1109/ACCESS.2024.3506334.

[22] J. Suckling, "The mammographic images analysis society digital mammogram database," in *Exerpta Medica. International Congress Series*, Accessed 2025-08-05, vol. 1069, 1994, pp. 375–378.

[23] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2015. DOI: 10.1109/TBME.2015.2496264.

[24] P. Agarwal, A. Yadav, and P. Mathur, "Breast cancer prediction on breakhis dataset using deep cnn and transfer learning model," in *Data Engineering for Smart Systems: Proceedings of SSIC 2021*, Springer, 2021, pp. 77–88. DOI: 10.1007/978-981-16-2641-8_8.

[25] L. A. Aldakhil, H. F. Alhasson, and S. S. Alharbi, "Attention-based deep learning approach for breast cancer histopathological image multi-classification," *Diagnostics*, vol. 14, no. 13, p. 1402, 2024. DOI: 10.3390/diagnostics14131402.

[26] İ. Sayın et al., "Comparative analysis of deep learning architectures for breast cancer diagnosis using the breakhis dataset," *arXiv preprint arXiv:2309.01007*, 2023. DOI: 10.48550/arXiv.2309.01007.

[27] M. A. Mohammed, B. Al-Khateeb, A. N. Rashid, D. A. Ibrahim, M. K. Abd Ghani, and S. A. Mostafa, "Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images," *Computers & Electrical Engineering*, vol. 70, pp. 871–882, 2018. DOI: 10.1016/j.compeleceng.2018.01.033.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy) The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself No funding was received for conducting this study.

Conflicts of Interest The authors have no conflicts of interest to declare.

APPENDIX

Table 8. Comparative Diagnostic Performance of Conventional Imaging Modalities in Breast Cancer Assessment [8], [9], [10]

Modality	Sensitivity	Specificity	Advantages	Limitations
Mammography	60–85%	90%	Early detection, moderate cost	Low performance in dense breast tissues
Ultrasound	75–80%	89%	Radiation-free, portable	High false-positive rate
MRI	$>$ 90%	85%	High spatial resolution	High cost, limited availability
Biopsy	95–98%	99%	Gold-standard diagnostic accuracy	Invasive procedure, time-consuming

Table 9. Comparative Performance of Classical Machine Learning Methods for Breast Cancer Diagnosis, [27], [11], [12]

Method	Accuracy	AUC	Advantages	Limitations
SVM	82%	0.78	Effective linear separation	Fails to model nonlinear patterns
Bayesian Networks	80%	0.85	Good interpretability	Limited multivariate dependencies
Decision Trees	78%	0.72	Simple and easy to implement	Overfits on small datasets
k-NN	75%	0.68	No training phase needed	High computational cost at inference
ANN	87%	0.82	Hierarchical feature learning	Requires manual feature engineering

Table 10. Representative deep neural network architectures applied to breast cancer imaging, [12], [13], [14], [15], [16], [17], [18], [19], [20]

Architecture	Performance Metric	Advantages	Limitations
AlexNet-inspired CNN	Accuracy: 84%	Foundation of CNNs in medical imaging	Shallow depth, limited generalization
VGGNet-16	AUC: 0.89	Strong feature extraction	High computational cost
ResNet-50	AUC: 0.93	Mitigates vanishing gradients	Demands significant GPU memory
DenseNet-169	Accuracy: 91–92%	Efficient feature reuse, strong sensitivity	Architectural complexity
U-Net	Dice: 0.87–0.90	Accurate lesion segmentation	Requires detailed annotations
EfficientNet-B4	AUC: 0.94	Energy-efficient scaling	Sensitive to artifacts
MobileNetV2/V3	Accuracy: 88–90%	Lightweight, suited for embedded systems	Lower accuracy than deeper CNNs

Table 11. Comparative Performance of Neural Network Architectures in Breast Cancer Imaging, [12], [13], [14], [16], [17], [18], [19], [20], [21]

Architecture	Reported Performance	Clinical Impact	Clinical Limitations
AlexNet	Accuracy: 84%	Foundation of automated diagnosis	Limited depth and generalization
VGGNet-16	AUC: 0.89	High-resolution microcalcification detection	High computational burden
ResNet-50	AUC: 0.93	Reduced false negatives in dense tissues	High memory/computation requirements
DenseNet-169	Accuracy: 91.72%	Enhanced small mass sensitivity	Slow training convergence
U-Net	Dice: 0.87	Accurate lesion segmentation for planning	Requires pixel-level annotations
EfficientNet-B4	AUC: 0.94	Resource-efficient classification	Sensitive to image artifacts
LightSegNet	Dice: 0.91	Real-time mobile segmentation	Vulnerable to noise/artifacts
MobileNetV2	Accuracy: 89%	Optimized for embedded systems	Slightly lower accuracy than deeper CNNs
MobileNetV3	Accuracy: 93%	Strong balance between accuracy and efficiency	Still limited on very complex patterns

Table 12. Comparative Summary of the Three Implemented Architectures

Model	**NCNN (Custom CNN)**	**MobileNetV3 (Large)**	**ViT-L16 (Vision Transformer)**
Architecture Type	Convolutional Neural Network	Lightweight CNN with inverted residuals and SE blocks	Pure Transformer with Multi-Head Self-Attention
Pretraining	None (from scratch)	ImageNet (transfer learning)	ImageNet-1K (transfer learning)
Input Size	$3 \times 224 \times 224$	$3 \times 224 \times 224$	$3 \times 224 \times 224$
Model Depth	2 conv + 3 FC layers	$\sim$16 inverted residual blocks + head	24 Transformer layers
Total Parameters	5.4 million	$\sim$5.4 million (after customization)	$\sim$307 million (ViT-Large)
Backbone Output Size	84 (before logits)	1280	1024 (CLS token embedding)
Final Classifier	FC layer with 3 outputs	Linear(1280, 3)	Linear(1024, 3)
Memory Footprint	Low	Low to Moderate	High
Training Time (relative)	Fast (train from scratch)	Very Fast (few layers fine-tuned)	Moderate to Slow (larger model)
Inference Speed	Fast	Very Fast	Moderate
Suitability for Medical Imaging	High (custom and interpretable)	High (compact and accurate)	Very High (global context modeling)
Key Strengths	Simple and transparent design	Efficient and mobile-friendly	Captures long-range dependencies
Limitations	May lack generalization on complex patterns	Less flexible on fine-grained global features	Computationally intensive, requires more memory