RESEARCH ARTICLE

Multi-Scale CSPResNet50 with Feature Pyramid Aggregation and SE Attention for Breast Cancer Histopathological Subtype Classification and Malignancy Detection

Muhammad Hamza[1,*], **Ibrar Ali**[1], **Sikandar Ali**[1] and **Shayan Khan**[1]

[1] University of Swat, Swat 01923, Pakistan

Abstract

Breast cancer is a leading cause of cancer-related mortality worldwide, making accurate histopathological subtype discrimination critical for timely clinical intervention. Existing deep learning approaches often evaluate limited settings (binary or multi-class, single magnification), restricting comparative utility and clinical interpretability. This study proposes a unified Cross Stage Partial Network (CSPNet)-based framework for comprehensive classification on the BreaKHis dataset. A CSPResNet50 backbone pre-trained on ImageNet was extended with a multi-scale Feature Pyramid-style aggregation head, Squeeze-and-Excitation channel attention, dual Global Average and Max Pooling per scale (1536-dimensional descriptor), and a dense classification neck, trained via two-phase progressive fine-tuning. Four tasks were evaluated under a stratified 70/15/15 split: multi-class (8 subtypes) and binary classification at whole-dataset and per-magnification levels (40×, 100×, 200×, 400×), with GradCAM providing interpretable visualisations. For whole-dataset binary classification, the model achieves 95.53% accuracy, 94.80% F1-score, 96.93% sensitivity, and 98.53% AUC. In multi-class classification, it attains 78.10% accuracy, 81.40% balanced accuracy, 76.76% macro F1-score, and 97.72% AUC. Per-magnification analysis shows best performance at 40×, with binary and multi-class accuracy reaching 95.00% and 77.00%, respectively, while AUC remains above 92% across all magnifications. A key limitation is image-level rather than patient-level splitting, which may introduce optimistic bias. Overall, the proposed framework provides a robust, interpretable, and computationally efficient solution for breast cancer histopathological image classification.

Keywords: breast cancer, histopathology, CSPResNet50, GradCAM, deep learning, classification.

Submitted: 04 May 2026
Accepted: 20 May 2026
Published: 31 May 2026

Vol. 2, **No.** 3, 2026.
10.62762/JIAP.2026.746947

***Corresponding author:**
✉ Muhammad Hamza
mhamzaedu1@gmail.com

Citation
Hamza, M., Ali, I., Ali, S., & Khan, S. (2026). Multi-Scale CSPResNet50 with Feature Pyramid Aggregation and SE Attention for Breast Cancer Histopathological Subtype Classification and Malignancy Detection. *ICCK Journal of Image Analysis and Processing*, 2(3), 122–140.

1 Introduction

Breast cancer is one of the most widely diagnosed cancers and remains the disease with the highest rate of death from cancer in women globally [1]. According to figures from the World Health Organization (WHO), breast cancer is significantly responsible for around 2.3 million new cases every year, and a large percentage of these diagnoses are made in advanced stages due to late detection [1]. In 2022, approximately 2.3 million women across the globe were diagnosed with breast cancer, and the disease claimed the lives of 670,000 of them [2]. Breast cancer accounts for 30% of new cancer cases in women, with a 12% lifetime risk and a 3% chance of death [3]. The mortality rate could be significantly reduced if breast cancer were diagnosed and treated at early stages, as small, non-metastatic disease can be effectively treated [4]. There are two primary types of breast tumors: benign and malignant. In contrast to benign, malignant cells proliferate rapidly due to their irregular appearance and structure, leading to swift division and potential spread throughout the body. These tumors have the potential to metastasize, meaning they can detach and travel through blood or lymph vessels to distant sites, causing damage to other organs and tissues [5]. When breast cancer is identified in its initial stages, it can be treated effectively before it has the chance to metastasize to other areas of the body. Risk factors include genetic mutations, hormonal influences, aging, and family history [5]. According to the National Institutes of Health, human error (96.3%) is the major source of diagnostic adverse outcomes [2]. Such factors contribute to missed breast cancers, with up to 35% of both interval and screening-detected cancers classified as missed [4]. In order to mitigate these issues, the integration of AI-driven diagnostic tools holds great promise as an approach for improving accuracy, efficacy, and early detection of breast cancer [1]. Some drawbacks of current diagnosis techniques include limited imaging capabilities, overlooking symptoms, expensive procedures, and slow processing speeds [2]. Because of these factors, machine learning techniques have the potential to greatly impact the diagnostic process [6–8]. The application of AI techniques for the early diagnosis of breast cancer has recently increased, and for the most part, healthcare organizations have used ML and DL algorithms for breast cancer diagnosis [6]. In detecting breast cancer, AI-enabled machine learning models can analyze mammograms, histopathological images, and patient records more accurately and quickly than traditional methods. Deep learning, specifically convolutional neural networks (CNNs), has greatly enhanced image-based classification with a high level of performance in early detection and a low probability of misdiagnosis [1]. Rezayi et al. [9] conducted a comprehensive bibliometric review of deep learning techniques in breast cancer imaging, finding that the frequency of papers being recovered has grown significantly since 2016, with a significant number of citations retrieved in 2022 (as shown in Figure 1). According to their findings, CNN emerged as the most widely used and accurate model for diagnosing breast cancer, and performance is frequently assessed using accuracy, sensitivity, specificity, area under curve (AUC), F1 score, Dice similarity coefficient (DSC), and intersection over union (IoU) [9]. Nasser and Yusof [10] conducted a comprehensive systematic review of deep learning-based methods for diagnosing breast cancer using genomic and histopathological imaging data, concluding that CNN is the most widely used and accurate model for diagnosing breast cancer.

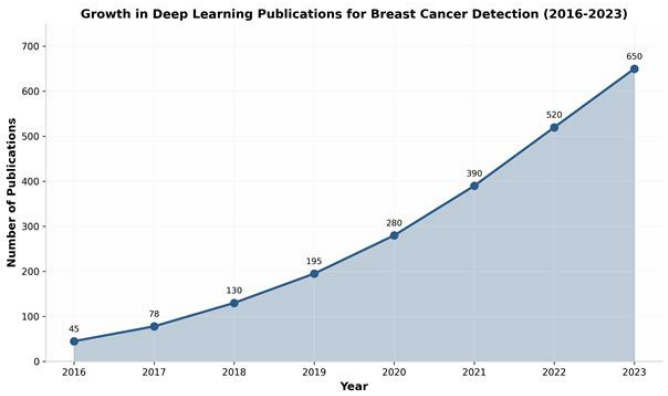

Figure 1. Growth in deep learning publications for breast cancer detection (2016-2023).

In light of the aforementioned limitations in existing approaches, this paper makes the following contributions:

- Multi-Scale Feature Aggregation Architecture.
- Two-Phase Progressive Training Strategy.
- Clinically Motivated Evaluation Framework.
- Systematic Four-Task Magnification Analysis.
- Interpretable Attention Mapping.

The remainder of this paper is organised as follows: Section 2 reviews related work in deep learning-based histopathological image classification. Section 3 details the proposed methodology. Section 4 presents

experimental results and discussion. Section 5 concludes with limitations and directions for future research.

2 Related Work

This section reviews recent studies on breast cancer histopathology classification, with a particular focus on methods evaluated using the BreaKHis dataset, which provides multi-magnification microscopic images of breast tumor tissues. These images have enabled the development and assessment of a wide range of deep learning-based frameworks, including convolutional neural networks (CNNs), Vision Transformers (ViTs), and hybrid ensemble models for both binary and multi-class classification. In recent years, advances in computational resources and deep learning methodologies have further accelerated research in computer-aided diagnosis of histopathological images, leading to numerous approaches proposed for automated breast cancer detection and classification.

A multiscale LWT–CNN framework is proposed for histopathology classification, combining wavelet-based multi-resolution feature extraction with multi-path CNN learning. On BreakHis, it achieves 99.34% accuracy, outperforming standard CNNs [11]. An ensemble CNN–ViT framework with self-attention fusion is proposed for histopathology classification to capture local and global features. Evaluated on BreakHis and BACH, it achieves up to 98.7% accuracy with strong generalization and Grad-CAM interpretability [12]. EVC-Net combines EfficientNetV2S, Vision Transformer, and capsule networks for local, global, and spatial feature learning in histopathology classification. On BreakHis, it achieves 98.5% binary and 95.4% multi-class accuracy with strong AUC and Grad-CAM interpretability [13]. A multi-pretrained fusion framework (MPDLF) integrates ResNet50 and VGG16 using early, intermediate, and late fusion for histopathology classification. On H&E datasets, early fusion performs best with 90.70% accuracy and strong AUC, showing improved diagnostic robustness [14]. A SwinTConvNeXt-LDGF framework combines Swin Transformer and ConvNeXt with dynamic gating for local-global feature fusion in histopathology classification. On BreakHis, it achieves 95.53% accuracy with strong F1-score, showing effective subtype classification [15]. An EfficientNet-based framework addresses class imbalance using augmentation, cost-sensitive learning, and transfer learning for breast cancer classification. On BreakHis, it achieves 98.23% binary and 94.82% multi-class accuracy with improved minority class detection [16]. A ResTab Net multimodal framework integrates histopathological images and molecular data using CNN and dense layers with ATGF, EGWS, and SASFO for feature refinement. It achieves 98.56% accuracy with SHAP/LIME-based interpretability [17]. An AENN framework combines EfficientNet-B3 with CBAM attention and Transformer encoders for local and global feature learning in histopathology classification. On BreakHis, it achieves 96.4% accuracy with strong precision and recall [18]. A SwinCNN+OE framework combines EfficientNet-E0 and Swin Transformer with outlier exposure and energy-based OOD detection for breast cancer classification. It achieves 97.83% accuracy and 99.49% AUC with reliable OOD detection and Grad-CAM interpretability [19]. A multi-source mammography fusion framework integrates MLO and CC views using dual EfficientNet-B7 encoders with Multi-Head Attention for feature refinement. On 5,906 images, it achieves 93.9% accuracy, showing improved lesion classification via attention-based fusion [20]. A DenseNet121-based deep learning model was applied to the BreakHis dataset for breast cancer detection and multi-class classification. The model achieved 98.50% accuracy for binary classification and 92.50% accuracy for multi-class classification [21]. A ResNet-based CNN framework using transfer learning was applied to the BreakHis dataset for eight-class breast cancer subtype classification. ResNet-50 achieved the best performance with 92.42% accuracy and 99.86% AUC-ROC [22]. A WRN-28-2 deep CNN with transfer learning was applied to the BreakHis dataset for binary breast cancer classification, achieving 99.16% test accuracy and outperforming existing methods [23]. A transformer-based model using a cascade of transformers was applied to the BRACS and BreakHis datasets for breast cancer classification. The model achieved 95.6% accuracy on BRACS for three-class classification and 93% accuracy on BreakHis for binary classification, with high sub-type accuracies [24]. Vision Transformer (ViT) and ConvNeXT models were applied to the BreakHis dataset for breast cancer classification. ViT achieved 95.6% accuracy and 96.8% F1 score at 40× magnification, while ConvNeXT reached 98.5% precision at 100× [25]. The TokenMixer hybrid CNN–ViT architecture was applied to the BreakHis dataset for breast cancer classification. The model achieved 97.02% accuracy for binary classification and 93.29% for multi-class subtype

classification across multiple magnifications [26]. ResViT-GANNet, a dual-branch network combining residual CNN with channel attention and a vision transformer with multi-layer token fusion, was developed for high-resolution histopathological image classification. Using the BACH (4-class) and BreakHis (8-class) datasets, it achieved 96.35% F1 on BACH and 98.22% average accuracy on BreakHis, outperforming baseline models. Synthetic data from StyleGAN2-ADA improved accuracy by 3.3%, and Grad-CAM confirmed interpretability [27]. A classifier fusion framework combining features from 16 pre-trained CNNs using transfer learning was applied to the BreakHis dataset for breast cancer classification. At 400× magnification, the approach achieved 92.7% image-level accuracy and 93.2% patient-level accuracy, outperforming prior methods [28]. An ensemble deep learning model combining ResNet50 and DenseNet121 was applied to the BreaKHis dataset for breast cancer classification. The hard voting ensemble achieved the highest accuracy of 98.61% [29]. Fourteen CNN- and Transformer-based models, including ResNet50, ConvNeXT, RegNet, and UNI, were evaluated on the BreakHis v1 dataset for breast cancer classification. In binary classification, ConvNeXT achieved the highest performance with 99.2% accuracy and an AUC of 0.999. For eight-class classification, the fine-tuned UNI model performed best, achieving 95.5% accuracy and an AUC of 0.998 [30]. A hybrid multi-class classification model combining depth-wise separable convolutional networks and transformers was proposed for breast cancer grading using local and global features. The model was evaluated on the BACH, BreakHis, and IDC datasets, achieving accuracies of 97.80%, 98.13%, and 98.32%, respectively [31]. The Efficient Channel Spatial Attention Network (ECSAnet), based on EfficientNetV2 with CBAM, was applied to the BreakHis dataset for breast cancer classification. The model achieved accuracies of 94.2% (40×), 92.96% (100×), 88.41% (200×), and 89.42% (400×) magnifications [32]. EfficientNetV2 models were applied to 40× benign breast cancer histopathology images for classification. EfficientNetV2L achieved the best performance with 97% accuracy, precision, recall, and F1-score [33]. A DCGAN was used to generate synthetic data for the minority class in the BreakHis dataset, and a fine-tuned ResNet50 was applied for breast tumor classification. The approach achieved high accuracy at 40× magnification, outperforming existing methods [34]. A study proposes a Multi-Level Feature Fusion CNN (MLF2-CNN) using DenseNet-121 as backbone, performing automatic score-level fusion of multi-stage features for breast cancer classification. Experiments on the BreakHis dataset show the model outperforms state-of-the-art methods across binary, multi-class, magnification-specific, and magnification-independent scenarios [35].

3 Methodology

This section presents a comprehensive description of the proposed framework for breast cancer histopathological image classification based on a Cross Stage Partial Network (CSPNet) architecture. The pipeline comprises four sequential classification tasks: whole-dataset multi-class subtype classification, per-magnification multi-class classification, whole-dataset binary benign/malignant classification, and per-magnification binary classification. Each task is trained and evaluated independently under identical experimental conditions to ensure a fair and consistent comparison. The overall methodology is illustrated in Figure 2 and summarized in Algorithm 1.

Algorithm 1: CSPResNet50 Classification Pipeline

Data: BreaKHis images (8 subtypes, 4 magnifications, PNG), seed = 42
Result: Metrics $\{Acc, Prec, F1, AUC, Sens, ms/image\}$ for 4 tasks
Phase 0: Data Preparation
parse_metadata(filenames) $\rightarrow$ {subtype, magnification, binary_label};
stratified_split(data, ratio=70:15:15, seed=42) $\rightarrow$ Train, Val, Test;
sampler $\leftarrow$ WeightedRandomSampler($w_c = \frac{N}{C \times N_c}$);
augment_train $\leftarrow$ [Flip, Rotate$\pm 20^\circ$, ColorJitter, Affine, MixUp($\alpha = 0.4, p = 0.30$)];
augment_val $\leftarrow$ [Resize(224), Normalize(μ_{IN}, σ_{IN})];
Phase 1: Model Construction
backbone $\leftarrow$ CSPResNet50(pretrained=ImageNet, out_indices=(1,2,3,4));
for $s \in \{C3, C4, C5\}$ **do**
  $F_s \leftarrow$ SE_Attention(LateralConv($s \rightarrow 256$ch));
  $d_s \leftarrow$ concat(GAP(F_s), GMP(F_s));
end
$z \leftarrow$ concat(d_{C3}, d_{C4}, d_{C5});
neck $\leftarrow$ MLP($1536 \rightarrow 1024 \rightarrow 512 \rightarrow 256$, BN+SiLU+Dropout);
head $\leftarrow$ Linear(256, N);
Phase 2: Training
loss $\leftarrow$ FocalLoss($\gamma = 2.0, \epsilon = 0.08$) + LabelSmoothing;
for *epoch* $= 1$ *to* 30 **do**
  freeze(backbone);
  train(neck+head, AdamW, WarmupCosine(warm=3));
end
for *epoch* $= 1$ *to* 20 **do**
  unfreeze(backbone, lr$\times 0.1$), $\epsilon \leftarrow 0.04$, MixUp $\leftarrow$ off;
  if *val_acc* $>$ *best* **then**
    save_checkpoint(model);
  end
end
Phase 3: Inference
load_checkpoint(best);
probs $\leftarrow$ mean[softmax(model(aug_t(x))) for t in TTA transforms];
$t^* \leftarrow \arg\max_t \{TPR(t) - FPR(t)\}$;
Phase 4: Evaluation
compute_metrics(y_{true}, probs, t^*);
GradCAM(last_conv) $\rightarrow$ attention maps;

3.1 Dataset

All experiments were conducted on the Breast Cancer Histopathological Image (BreaKHis) database [36], a widely adopted benchmark for computational pathology. The dataset comprises 7,909 digitised histopathological images acquired at four optical magnification levels (40×, 100×, 200×, and 400×). A

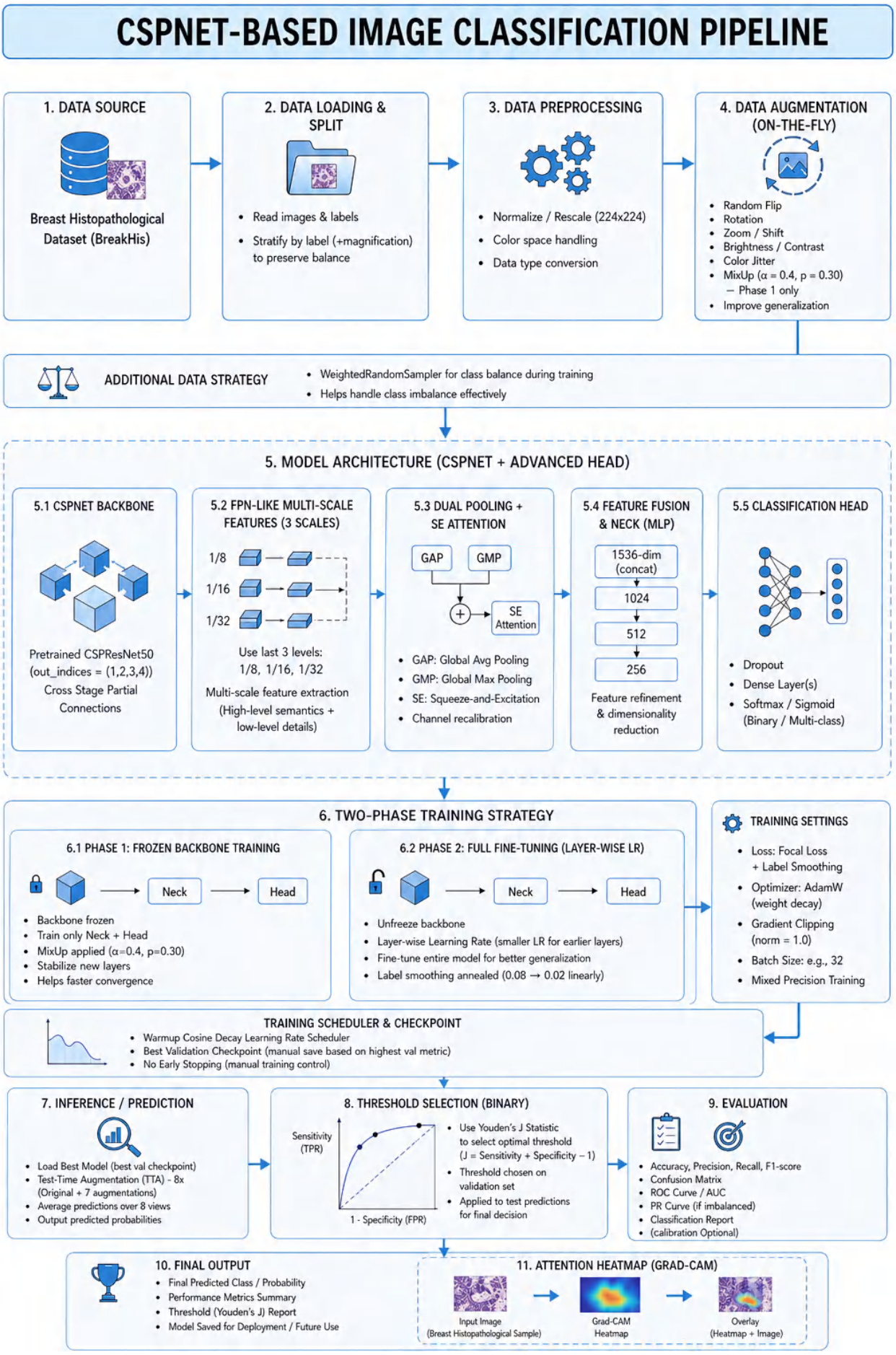

Figure 2. End-to-end workflow of the proposed CSPResNet50 framework for breast cancer classification on the BreaKHis dataset.

comprehensive distribution analysis of the dataset is presented in Figure 3. Images are stored as 700×460 pixel PNG files under standardised RGB colour conditions. The eight histological subtypes encoded in the dataset follow the naming convention embedded within the SOB filename schema `(SOB_B|M_SUBTYPE-patient-slide-magnification-frame.png)`. The four benign subtypes comprise Adenosis (A), Fibroadenoma (F), Phyllodes Tumour (PT), and Tubular Adenoma (TA). The four malignant subtypes comprise Ductal Carcinoma (DC), Lobular Carcinoma (LC), Mucinous Carcinoma (MC), and Papillary Carcinoma (PC). Subtype membership directly determines the binary class label: all benign subtypes map to class 0 and all malignant subtypes to class 1. Metadata — subtype code, magnification factor, and binary label — were extracted programmatically from filenames using a primary regular-expression pattern and a fallback magnification extractor, ensuring complete metadata coverage across all images.

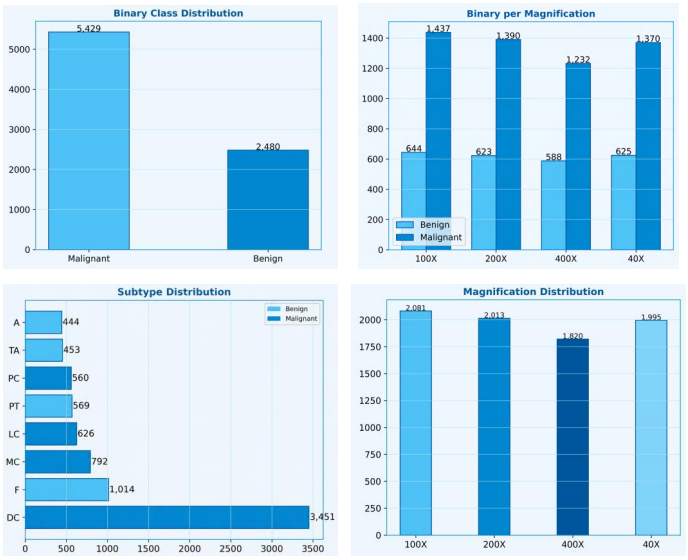

Figure 3. BreaKHis dataset distribution: binary classes (benign vs malignant), multi-class subtypes, and magnification levels (40×, 100×, 200×, 400×).

3.2 Experimental Design

Four classification tasks were defined to provide a thorough characterisation of the model's discriminative capacity:

1. **Multi-class, Whole Dataset:** Eight-class subtype classification across the full image pool, irrespective of magnification level.
2. **Multi-class, Per Magnification:** Eight-class subtype classification performed independently for each magnification factor (40×, 100×, 200×, 400×), yielding four independent experiment runs.
3. **Binary, Whole Dataset:** Two-class benign versus malignant classification across all magnifications combined.
4. **Binary, Per Magnification:** Binary benign/malignant classification repeated independently for each of the four magnification levels.

This design allows direct quantification of (i) the relative difficulty of fine-grained subtype discrimination versus coarser binary detection, and (ii) how magnification context affects classification performance, a question of practical clinical relevance given that pathologists routinely toggle between magnification levels during diagnosis.

3.3 Data Partitioning

A stratified three-way split was applied globally to every task to prevent data leakage while preserving class distribution. The training, validation, and test sets were allocated in a **70:15:15** ratio, respectively, as illustrated in Figure 4. respectively. Stratification was performed on the binary label to ensure that class proportions remained consistent across all three subsets. The train-validation split was obtained via a primary stratified split retaining 70% of samples for training, followed by an equal 50:50 division of the remaining 30% into validation and test sets. All splits used a fixed random seed of 42 for full reproducibility. To mitigate the effect of class imbalance during training, a PyTorch `WeightedRandomSampler` [37] was applied to the training dataloader. Sample weights were computed as the inverse of the normalised class frequency: $w_c = N/(CN_c)$, where N is the total number of training samples, C is the number of classes, and N_c is the number of samples belonging to class c. This ensures that each mini-batch presents a near-uniform class distribution regardless of the underlying dataset imbalance.

3.4 Data Preprocessing and Augmentation

3.4.1 Inference-Time Preprocessing

All images were resized to 224×224 pixels using bilinear interpolation and converted to 32-bit floating-point RGB tensors normalized with `ImageNet` channel statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$). This normalization aligns the input distribution with the pre-training

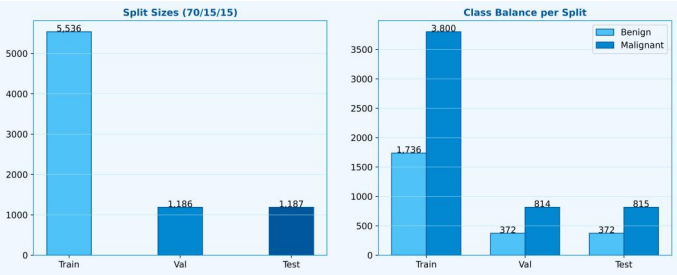

Figure 4. Stratified dataset split (train/validation/test) preserving class balance.

distribution of the CSPResNet50 backbone.

3.4.2 *Training-Time Augmentation*

An aggressive on-the-fly augmentation pipeline was applied exclusively during training to improve robustness against histological variability, including orientation changes, staining variations, and tissue deformation. The augmentation operations are summarized in Table 1.

Table 1. Training-time data augmentation pipeline.

Augmentation	Configuration
Horizontal Flip	$p = 0.50$
Vertical Flip	$p = 0.50$
Rotation	$\pm 20°$
Color Jitter	brightness $[0.75, 1.25]$, contrast $[0.75, 1.25]$ saturation $[0.85, 1.15]$, hue $[-0.06, 0.06]$
Affine Transform	translation $\leq 8\%$, scale $[0.88, 1.12]$
Grayscale	$p = 0.03$
MixUp	$\alpha = 0.4$, applied to 30% batches (Phase 1 only)

For each augmented pair (x_a, x_b) with labels (y_a, y_b) and mixing coefficient $\lambda \sim \text{Beta}(0.4, 0.4)$, the mixed sample is computed as $\tilde{x} = \lambda x_a + (1 - \lambda)x_b$. The corresponding loss is defined as $L_{\text{mix}} = \lambda L(f(\tilde{x}), y_a) + (1 - \lambda)L(f(\tilde{x}), y_b)$.

3.5 Model Architecture

3.5.1 *CSPResNet50 Backbone*

The backbone of the proposed network is CSPResNet50, instantiated via the `timm` library (version $\geq$ 0.9.0) [38] with ImageNet pre-trained weights. Cross Stage Partial Networks [39] decompose the input feature map of each stage into two parts: one that undergoes dense block computation and one that bypasses it, merging at the stage output via concatenation. This design reduces redundant gradient information during backpropagation while simultaneously improving the richness of gradient flow — properties that are particularly beneficial when fine-tuning on a relatively small medical imaging dataset. The backbone was configured with out_indices = (1, 2, 3, 4), yielding four feature maps at spatial resolutions of 1/4, 1/8, 1/16, and 1/32 of the input resolution. Only the final three feature maps (1/8, 1/16, 1/32) — corresponding to stages C3, C4, and C5 — were used for multi-scale feature fusion.

3.5.2 *Feature Pyramid Multi-Scale Aggregation*

Inspired by Feature Pyramid Networks [40], lateral projection layers were applied independently to each of the three selected stages to align their channel counts to a common dimensionality P = 256. Each lateral projection consists of a 1×1 convolution followed by batch normalisation and a SiLU activation function. This channel-aligned design allows the classifier to simultaneously exploit high-level semantic information from the deeper 1/32-resolution features and fine-grained textural features from the 1/8-resolution features — a balance particularly important in histological classification where both global tissue architecture and cellular-level morphology carry diagnostic value.

3.5.3 *Squeeze-and-Excitation Attention*

A Squeeze-and-Excitation (SE) module [41] was applied to each projected feature map before pooling. The SE block performs channel-wise recalibration: global average pooling compresses each feature map to a channel descriptor, which is then passed through a two-layer fully connected bottleneck (reduction ratio r = 16) with SiLU activation in the intermediate layer and sigmoid gating in the output layer. The resulting channel weights re-scale the feature map multiplicatively, effectively enabling the network to selectively suppress uninformative channel responses (e.g., background tissue channels) and amplify task-relevant ones (e.g., nuclei morphology channels).

3.5.4 *Dual Pooling and Feature Fusion*

For each SE-attended feature map, both Global Average Pooling (GAP) [42] and Global Max Pooling (GMP) were applied and their 256-dimensional output vectors concatenated, yielding a 512-dimensional descriptor per scale. Concatenating across all three scales produces a 1536-dimensional multi-scale feature vector. GAP captures the mean-activation statistics of each channel (distributional representation), while GMP captures peak activations (discriminative-part representation) — their combination has been shown

to enrich feature representations in histopathological image classification tasks [43].

3.5.5 Dense Classification Neck

The 1536-dimensional feature vector is passed through a three-stage dense neck: Linear(1536, 1024) ⇒ BatchNorm1d ⇒ SiLU ⇒ Dropout(p = 0.35), followed by Linear(1024, 512) ⇒ BatchNorm1d ⇒ SiLU ⇒ Dropout(p = 0.245), and finally Linear(512, 256) ⇒ BatchNorm1d ⇒ SiLU. All weight matrices were initialised with Kaiming normal initialisation (fan-in mode) and batch normalisation parameters were initialised to weights = 1, bias = 0. The classification head is a single linear layer mapping the 256-dimensional neck output to N logits, where N ∈ 2, 8 depending on the task.

3.6 Loss Function

The training objective is a Focal Loss [44] with label smoothing regularisation. Label smoothing replaces hard one-hot targets with soft distributions: for a correct class c, the target probability is set to $(1 - \epsilon + \epsilon/C)$, and ϵ/C for all other classes, where ϵ is the smoothing coefficient and C is the number of classes. In Phase 1, $\epsilon = 0.08$; in Phase 2, ϵ was tightened to 0.04 to allow the model to converge toward sharper probability estimates during fine-tuning. The focal weighting factor $(1 - p_t)^\gamma$, where p_t is the model's estimated probability for the smoothed target distribution and $\gamma = 2.0$, down-weights the loss contribution of easily classified examples, directing the gradient signal toward harder, ambiguous samples - a property that is well-suited to medical imaging datasets where inter-class boundary cases are frequent.

3.7 Optimisation Strategy

3.7.1 Two-Phase Training

Training was performed in two phases following the principle of progressive unfreezing [45]. In the first phase, all backbone parameters were frozen and only the task-specific components (lateral projection layers, SE modules, pooling aggregator, neck, and classification head) were optimized using AdamW [46] with a learning rate of $\eta = 2 \times 10^{-4}$ and weight decay $\lambda = 1 \times 10^{-4}$. This stabilizes training by allowing newly initialized layers to converge before updating pre-trained representations.

In the second phase, the entire network was unfrozen and trained using a layer-wise learning rate strategy, where backbone parameters used $\eta_{\text{backbone}} = 8 \times 10^{-7}$ and the classification head used $\eta_{\text{head}} = 8 \times 10^{-6}$. This prevents catastrophic forgetting while enabling domain-specific adaptation. Label smoothing was reduced to $\epsilon = 0.04$, and MixUp was disabled.

3.7.2 Learning Rate Schedule

A Warmup Cosine Decay schedule was applied in both phases. During the initial three epochs, the learning rate increased linearly from $\eta_{\min} = 1 \times 10^{-7}$ to the base learning rate, followed by cosine annealing. For epoch e, the learning rate is defined as:

$$\eta(e) = \eta_{\min} + \frac{1}{2}(\eta_{\text{base}} - \eta_{\min}) \left(1 + \cos\left(\frac{\pi(e - T_{\text{warm}})}{T_{\text{total}} - T_{\text{warm}}}\right)\right)$$

for $e > T_{\text{warm}}$. This schedule improves generalization compared to step-based decay strategies.

3.7.3 Regularisation

Gradient norm clipping with a maximum norm of 1.0 was applied to stabilize early training. Additionally, Automatic Mixed Precision (AMP) [47] was used via PyTorch's `GradScaler`, enabling `float16` computation for forward and backward passes while maintaining `float32` master weights, reducing memory usage and improving training efficiency without loss of numerical stability.

3.8 Inference and Threshold Selection

3.8.1 Test-Time Augmentation

At inference, the model was evaluated using Test-Time Augmentation (TTA) with eight deterministic transformations applied to each test image: (1) identity, (2) horizontal flip, (3) vertical flip, (4) 90° rotation, (5) 180° rotation, (6) 270° rotation, (7) horizontal flip after 90° rotation, and (8) vertical flip after 180° rotation. The eight softmax probability vectors were averaged element-wise to produce the final class probability estimate. TTA is known to reduce prediction variance and improve calibration at test time, particularly for images with ambiguous spatial orientation — a common characteristic of histological sections.

3.8.2 Youden's J Threshold Selection

For binary classification tasks, the default 0.5 decision threshold was replaced by an optimal threshold derived from Youden's J statistic [48], defined as:

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

The threshold maximising J was identified by scanning the full ROC curve computed on the

validation set:

$$t^* = \arg\max_t \{TPR(t) - FPR(t)\}$$

This threshold was then applied to test set predictions. Youden's J ensures that both sensitivity (minimising false negatives, i.e., missed malignancies) and specificity jointly maximised, directly aligning the decision rule with the clinical cost asymmetry of breast cancer screening.

3.9 Evaluation Metrics

A comprehensive set of performance metrics was computed for each of the four classification tasks:

- Overall Accuracy (Acc)
- Balanced Accuracy (Bal. Acc): the arithmetic mean of per-class recall, correcting for class imbalance.
- Precision (Prec.)
- Recall / Sensitivity (Sens.)
- F1-Score (macro)
- Area Under Curve (AUC)
- GPU Inference Time (ms/image): measured as the mean latency over 200 forward passes of a single image tensor on the Tesla T4 GPU, after 30 warm-up passes.

For binary tasks, the Youden-threshold-adjusted accuracy (Acc_J) and the optimal threshold value t^* were additionally reported. Confusion matrices (raw counts and row-normalised), ROC curves, and Grad-CAM attention visualisations [49] were generated for all tasks.

3.10 GradCAM Attention Visualisation

To provide interpretable evidence that the model attends to clinically relevant tissue regions rather than spurious background features, Grad-CAM [49] was applied post-training using the final convolutional layer of the CSPResNet50 backbone as the target layer. For a given input image and predicted class c, the class activation map is computed as:

$$CAM = \text{ReLU}\left(\sum_k \left[\frac{1}{Z}\sum_i\sum_j \left(\frac{\partial y^c}{\partial A^k_{ij}}\right)\right] \times A^k\right)$$

where A^k denotes the k-th feature map of the target layer and y^c is the pre-softmax score for class c. The resulting spatial attention map was bilinearly upsampled to 224×224 pixels, min-max normalised to $[0, 1]$, and overlaid on the original histopathological image using a light-blue colourmap blend (60% original image + 40% colourmap overlay). An orange contour was drawn at the 70th percentile of the normalised activation map to delineate the regions of highest discriminative salience. This visualisation was performed for representative images from each classification task to provide qualitative validation of model behaviour.

3.11 Implementation Details

The entire pipeline was implemented in Python 3.12.12 using PyTorch and the timm ($\geq$ 0.9.0) library for backbone instantiation. All experiments were executed on a Kaggle Notebooks environment equipped with an NVIDIA Tesla T4 GPU (15.6 GB VRAM). Reproducibility was ensured by fixing all random seeds (Python, NumPy, PyTorch) to 42, setting `torch.backends.cudnn.deterministic = True`, and disabling benchmark mode. The best model state per task was identified by monitoring validation accuracy across all epochs and restoring the optimal checkpoint before test-set evaluation.

4 Results and Discussion

This section presents quantitative evaluation results across all four classification tasks, followed by a detailed analysis of performance patterns with respect to task type, magnification level, and model behaviour.

4.1 Multi-Class Classification — Whole Dataset

Figure 5 illustrates the training and validation curves for accuracy and loss over the training epochs in two phases. The plots demonstrate the model's convergence behaviour, highlighting steady improvement in validation accuracy and a corresponding reduction in loss, which indicates effective learning without significant overfitting.

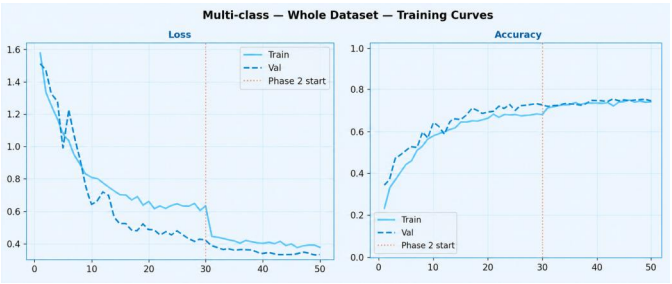

Figure 5. Training and validation accuracy and loss Curves for Multi-Class Classification — Whole Dataset.

The most challenging of the four tasks, eight-class

subtype differentiation across the pooled dataset, establishes a baseline for fine-grained histological subtype discrimination. Table 2 presents the complete metric suite obtained on the held-out test set following Test-Time Augmentation (TTA) evaluation. The results demonstrate strong class-wise discrimination, with a particularly high AUC (97.72%), indicating excellent overall discriminative ability across all classes. The model also achieves balanced performance across other metrics, including accuracy, F1-score, and sensitivity, reflecting robust generalisation across multiple breast cancer subtypes.

Table 2. Multi-class (8-class) whole-dataset classification results. Values are obtained from TTA (8×) inference on the 15% stratified test split.

Metric	Acc.	Bal. Acc	Prec.	F1 (macro)	AUC	Sens.
MC-All	78.10	81.40	74.89	76.76	97.72	81.40

The confusion matrix in Figure 6 provides a detailed evaluation of the multi-class classification performance by illustrating the distribution of correct and incorrect predictions across all breast cancer subtypes. Both raw counts and row-normalized values are presented, enabling a clearer interpretation of class-wise performance and imbalance effects. The normalized matrix particularly highlights strong diagonal dominance for most classes, indicating high correct classification rates, while off-diagonal entries reveal specific inter-class confusions between morphologically similar subtypes.

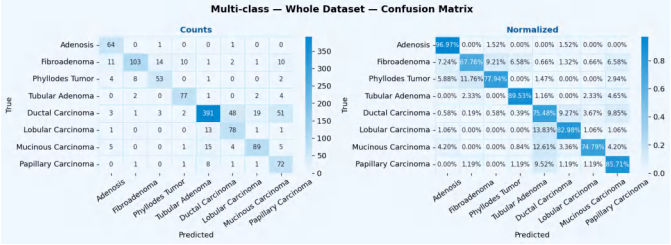

Figure 6. Multi-class confusion matrix for the whole dataset. Left: raw prediction counts across breast cancer subtypes; Right: row-normalized percentages showing class-wise classification performance and inter-class confusion patterns.

4.2 Multi-Class Classification — Per Magnification

Figure 7 illustrates the training and validation curves for accuracy and loss across multiple magnification levels (40×, 100×, 200×, and 400×) over the training epochs in both phases. The plots demonstrate consistent convergence behaviour across all magnifications, with a steady increase in validation accuracy and a corresponding decrease in validation loss. A similar trend is observed in the training curves, indicating stable optimization and effective learning without significant overfitting across different resolution scales.

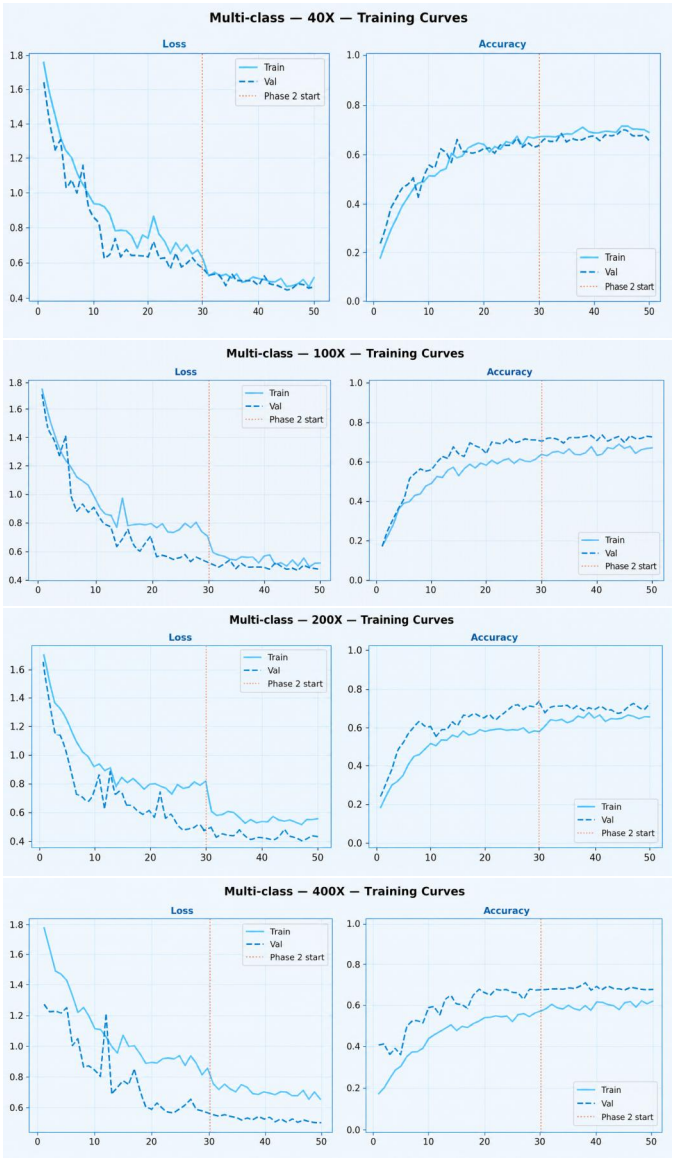

Figure 7. Training and Validation accuracy and loss Curves for Multi-Class Classification — Per Magnification.

The multi-class per-magnification classification task provides a more granular evaluation of the model's robustness across varying image resolutions. Table 3 presents the complete metric suite obtained on the held-out test set for each magnification level following Test-Time Augmentation (TTA) evaluation.

The results indicate that the proposed model performs consistently well at lower magnifications, achieving the best performance at 40× (Acc: 77.00%, Bal. Acc: 81.50%, F1: 75.70%), followed closely by 100×. A gradual decline in performance is observed at higher

magnifications, with 200× and 400× showing reduced accuracy and F1-scores, likely due to increased intra-class variability and finer structural complexity at higher resolutions.

Despite this trend, the model maintains relatively high AUC values across all magnifications (92.91%–97.03%), indicating strong discriminative capability. The balanced accuracy and sensitivity values further demonstrate stable class-wise performance, confirming that the model generalises effectively across different magnification levels while preserving sensitivity to malignant patterns.

Table 3. Multi-class (8-class) — Per Magnification classification results.

Metric	Acc.	Bal. Acc	Prec.	F1 (macro)	AUC	Sens.
MC-40X	77.00	81.50	74.24	75.70	97.03	81.50
MC-100X	76.36	75.43	73.54	73.39	96.81	75.43
MC-200X	68.87	74.55	65.48	67.80	94.63	74.55
MC-400X	64.84	61.47	62.44	60.53	92.91	61.47

The confusion matrices in Figure 8 provide a comprehensive evaluation of the multi-class classification performance across all magnification levels (40×, 100×, 200×, and 400×). Each magnification includes both raw count matrices and row-normalized representations, enabling detailed analysis of prediction distributions and class-wise performance under varying image resolutions.

Across all magnifications, the normalized matrices exhibit clear diagonal dominance, indicating that the model correctly classifies the majority of samples within each subtype. Performance is strongest at lower magnifications (40× and 100×), where class separation is more distinct, while higher magnifications (200× and 400×) show relatively increased off-diagonal entries, reflecting greater inter-class confusion.

These misclassifications are primarily observed among morphologically similar subtypes, highlighting the increased difficulty of fine-grained discrimination at higher resolutions. Overall, the matrices confirm that the proposed model maintains consistent classification capability across scales, while revealing resolution-dependent challenges in subtype differentiation.

4.3 Binary Classification — Whole Dataset

Figure 9 illustrates the training and validation curves for accuracy and loss for the binary classification task on the whole dataset across both training phases. The curves show a gradual and stable

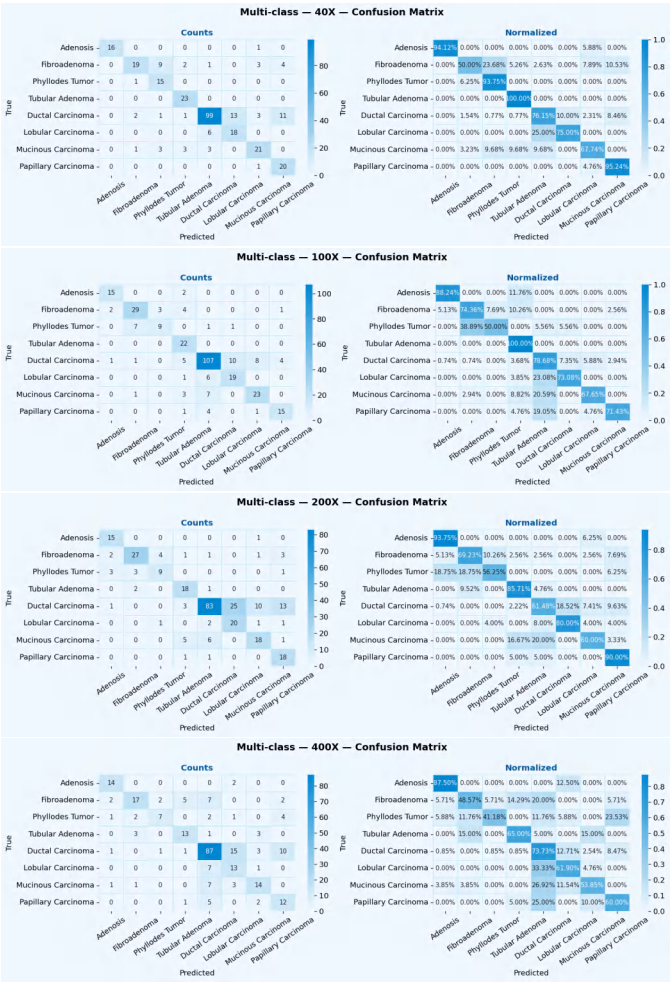

Figure 8. Multi-class confusion matrices across magnification levels (40×, 100×, 200×, and 400×). For each magnification, left panels show raw prediction counts, while right panels present row-normalized percentages, illustrating class-wise performance and inter-class confusion patterns across varying image resolutions.

increase in both training and validation accuracy, indicating steady learning progression. The training loss consistently decreases, while the validation loss follows a similar downward trend overall, with a brief spike observed around epochs 4–6 before returning to a decreasing trajectory. This transient fluctuation does not affect convergence and may be attributed to early optimisation dynamics or data variability. Overall, the curves demonstrate stable optimisation and effective generalisation without significant overfitting.

Binary benign/malignant discrimination represents the most clinically relevant application of the proposed pipeline. Table 4 reports the complete metric suite for the whole-dataset binary classification task. The results demonstrate excellent performance across all evaluation metrics, with high accuracy (95.53%), balanced accuracy (94.70%), precision (94.90%),

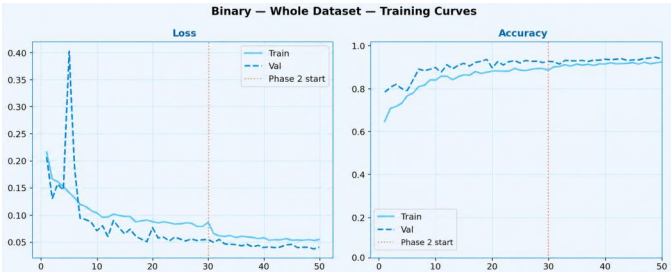

Figure 9. Training and validation accuracy and loss curves for binary classification on the whole dataset, showing stable convergence with a minor early fluctuation in validation loss.

F1-score (94.80%), AUC (98.53%), and sensitivity (96.93%).

These results indicate that the proposed model achieves superior performance in binary classification, consistently outperforming across all reported metrics. The high sensitivity further confirms the model's effectiveness in correctly identifying malignant cases, which is critical for clinical decision-making.

Table 4. Performance metrics for binary classification (benign vs malignant) on the whole dataset, demonstrating high accuracy and strong discriminative capability across all evaluation measures.

Metric	Acc.	Bal. Acc	Prec.	F1 (macro)	AUC	Sens.
Bin-All	**95.53**	94.70	94.90	94.80	98.53	96.93

The confusion matrix in Figure 10 shows high correct classification rates for both benign and malignant classes, with minimal misclassification, confirming strong binary discrimination performance.

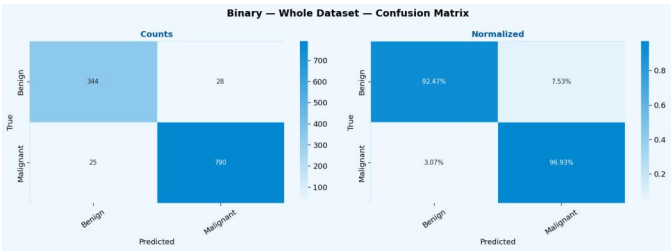

Figure 10. Binary classification confusion matrix for the whole dataset. Left: raw counts; Right: normalized percentages showing classification performance for benign and malignant classes.

4.4 Binary Classification — Per Magnification

Figure 11 illustrates the training and validation curves for accuracy and loss across all magnification levels (40×, 100×, 200×, and 400×). Both training and validation accuracy show a steady and consistent increase over epochs, indicating stable learning behaviour. The loss curves generally decrease across all magnifications; however, a noticeable fluctuation is observed in the 200× setting, where validation loss briefly spikes to approximately 140 during epochs 7–9 before returning to a normal decreasing trend. Minor fluctuations are also present in other magnifications but remain below 30, suggesting overall stable optimisation with limited sensitivity to data variability.

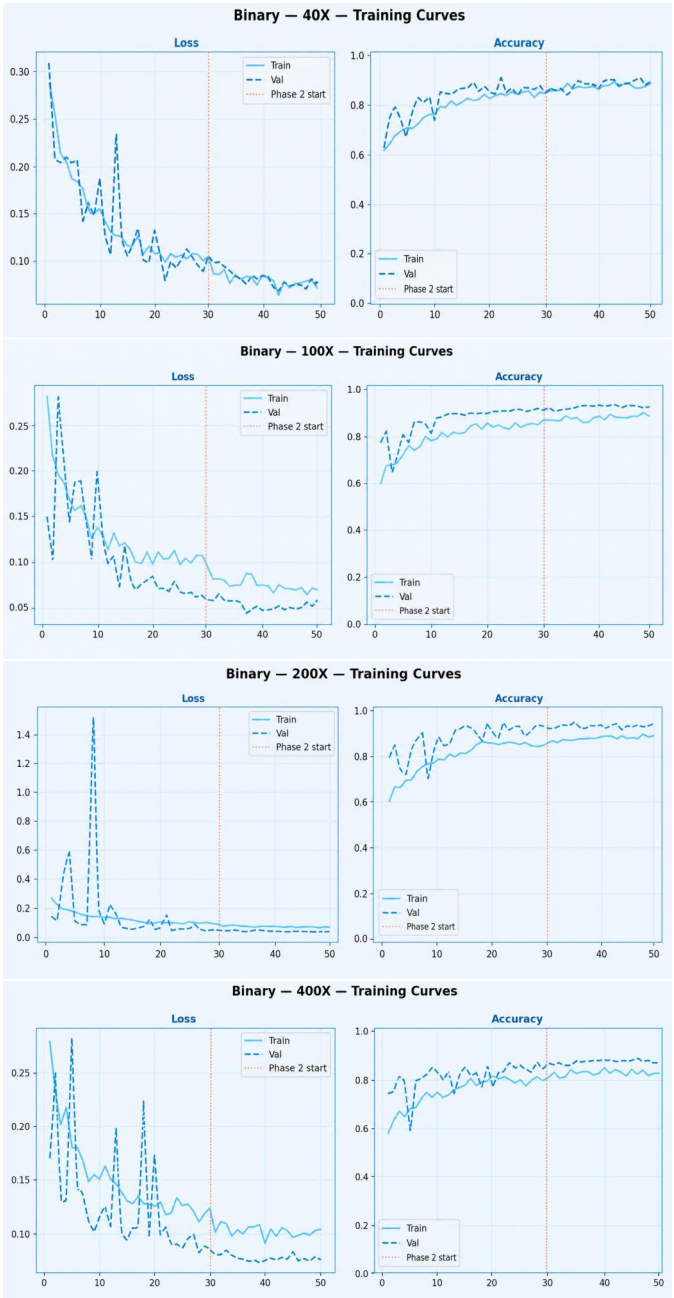

Figure 11. Training and validation accuracy and loss curves for binary classification across magnification levels (40×, 100×, 200×, and 400×), showing stable convergence with minor validation loss fluctuations.

Table 5 presents the performance metrics for binary

classification across different magnification levels. The proposed model achieves strong performance across all settings, with the best results observed at 40× (Acc: 95.00%, F1: 94.21%, AUC: 97.71%) and 200× (Acc: 93.71%, F1: 92.46%, AUC: 97.30%). Performance at 100× and 400× remains competitive, although slightly lower, likely due to increased intra-class variation and reduced global structural context. Overall, consistently high AUC values (94.59%–97.71%) and sensitivity scores demonstrate the model's robustness and effectiveness in distinguishing benign and malignant samples across resolutions.

Table 5. Binary classification performance metrics across different magnification levels, demonstrating consistent and robust model performance across all evaluation measures.

Metric	Acc.	Bal. Acc	Prec.	F1 (macro)	AUC	Sens.
Bin-40X	95.00	94.33	94.08	94.21	97.71	96.12
Bin-100X	90.42	90.78	88.11	89.20	96.73	89.81
Bin-200X	93.71	91.58	93.50	92.46	97.30	97.13
Bin-400X	90.84	90.26	89.16	89.67	94.59	91.89

The confusion matrices in Figure 12 further validate the model's performance across magnifications. All matrices exhibit strong diagonal dominance, indicating high correct classification rates for both benign and malignant classes. Misclassifications are minimal and relatively balanced across classes, with slightly higher confusion observed at 100× and 400× magnifications. These results confirm that the model maintains reliable binary discrimination performance across varying image resolutions.

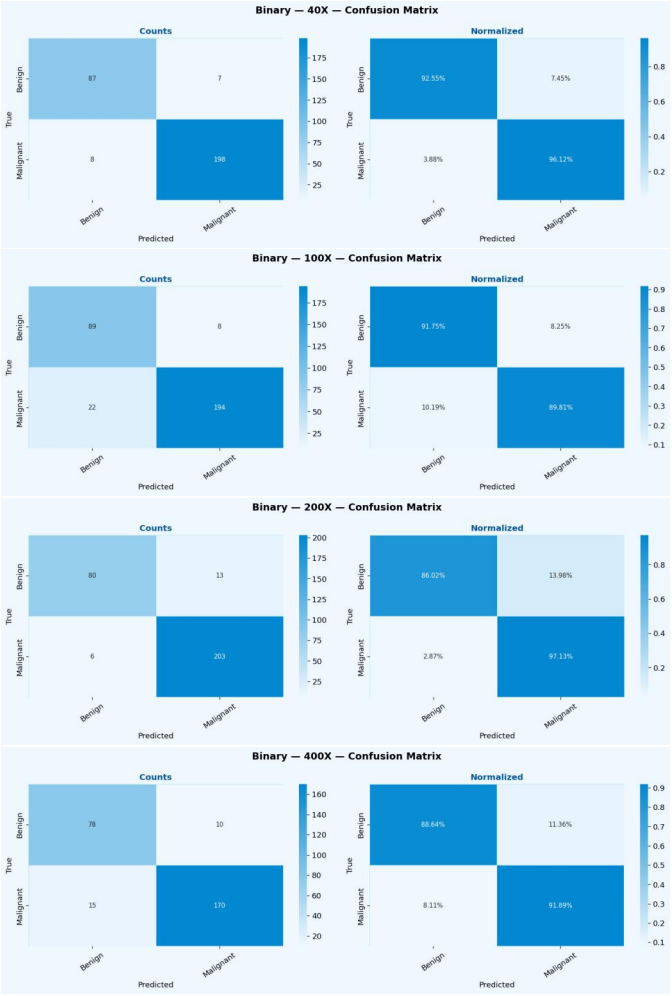

Figure 12. Binary confusion matrices across magnification levels (40×, 100×, 200×, and 400×), showing raw counts and normalized percentages, highlighting high classification accuracy and minimal misclassification.

4.5 Inference Efficiency

The inference latency of the proposed CSPResNet50 model remains highly consistent across all evaluation settings listed in Table 6, ranging from 8.3 ms to 8.9 ms per sample. Minor variations are observed due to differences in dataset complexity and magnification levels, particularly in multi-class (MC) and binary (Bin) settings. Higher magnification tasks (e.g., 200X and 400X) exhibit slightly reduced latency in some cases, attributed to reduced spatial ambiguity and more localized feature responses. Overall, the model demonstrates stable and efficient inference behavior, indicating suitability for real-time or near real-time histopathological image classification scenarios.

4.6 GradCAM Attention Analysis

Grad-CAM visualisations were analysed only for binary and multi-class whole-dataset settings to understand the model's decision behaviour.

Figure 13 shows the Grad-CAM attention maps for the multi-class classification (Whole Dataset). The corresponding attention maps exhibit well-defined and concentrated activations over class-specific histopathological structures. The model consistently focuses on discriminative regions such as tumour cell clusters, glandular formations, and stromal patterns, indicating that it has learned meaningful subtype-specific representations and is able to localise relevant features with high confidence.

In contrast, the binary setting Figure 14 includes several misclassified samples. Here, the Grad-CAM maps appear comparatively diffuse and less localised, often spreading across broader tissue regions rather than focusing on distinct diagnostic features. This suggests that the model struggles to identify a clear decision boundary between benign and malignant

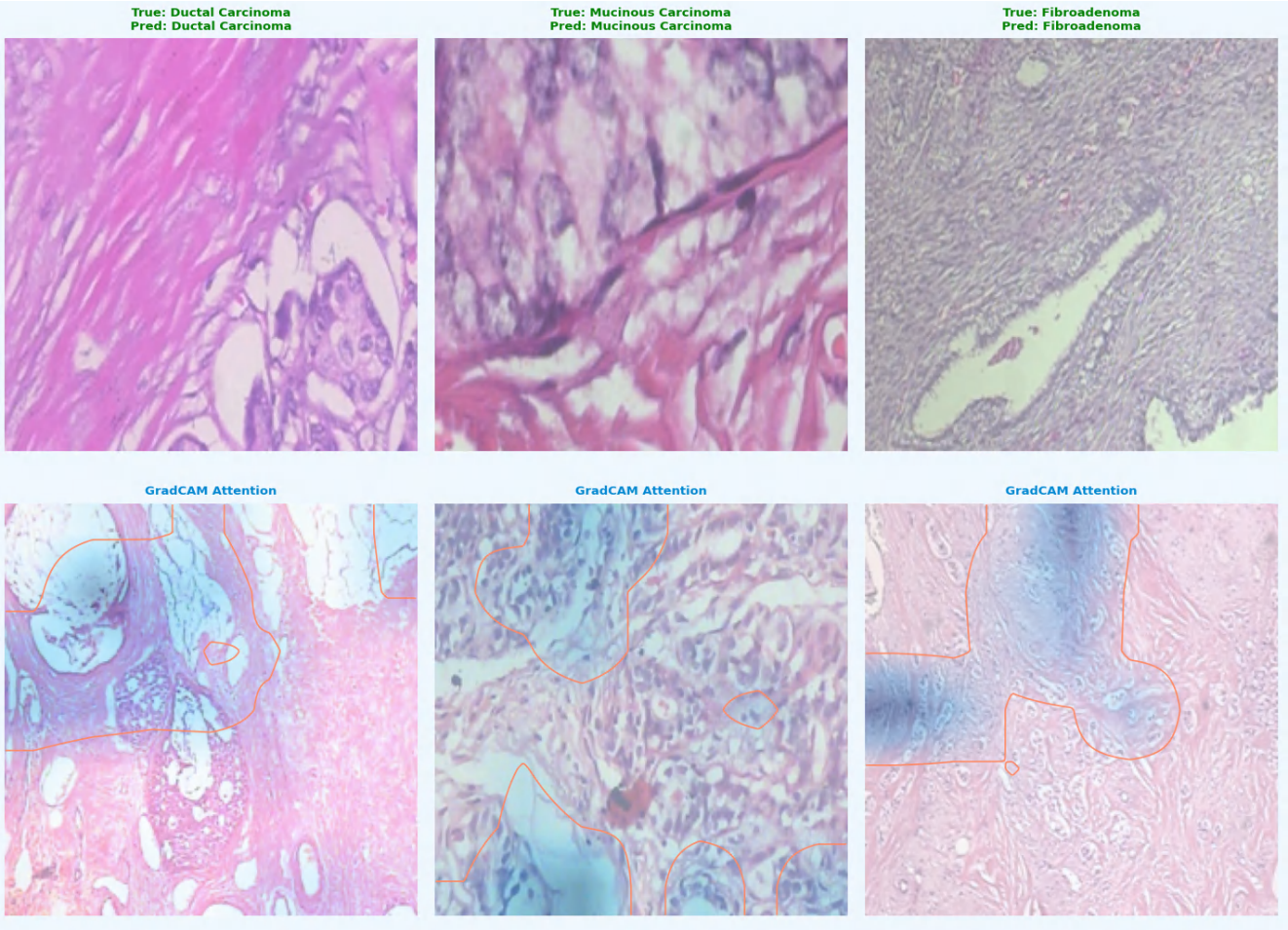

Figure 13. Multi-class classification (Whole Dataset) — Grad-CAM attention maps for representative multi-class samples, showing correctly classified cases with strong, well-localised activations over class-specific histopathological structures, demonstrating effective feature learning and localisation.

Table 6. Inference time (ms) for BreaKHis CSPResNet50 across different classification tasks.

Task	ms
MC-All	8.4
MC-40X	8.4
MC-100X	8.6
MC-200X	8.3
MC-400X	8.4
Bin-All	8.9
Bin-40X	8.3
Bin-100X	8.9
Bin-200X	8.3
Bin-400X	8.4

patterns in certain cases, leading to incorrect predictions.

Overall, correctly classified multi-class samples exhibit sharp, high-intensity, and well-localised activations, whereas misclassified binary samples show scattered attention, reflecting ambiguity in feature discrimination. These findings indicate that errors are primarily associated with weak or non-specific feature localisation rather than complete model failure, highlighting the challenge of learning robust binary separability in histologically diverse tissue patterns.

4.7 Comparison with Related Work

Table 7 contextualises the proposed method against representative published approaches on the BreaKHis dataset.

The proposed CSPResNet50 framework introduces several architectural and training-level improvements over earlier convolutional approaches; however, its performance remains competitive rather than state-of-the-art when compared with recent literature. First, the CSPResNet50 backbone improves gradient flow and reduces parameter redundancy compared to conventional architectures such as ResNet, VGG, and AlexNet, while maintaining computational

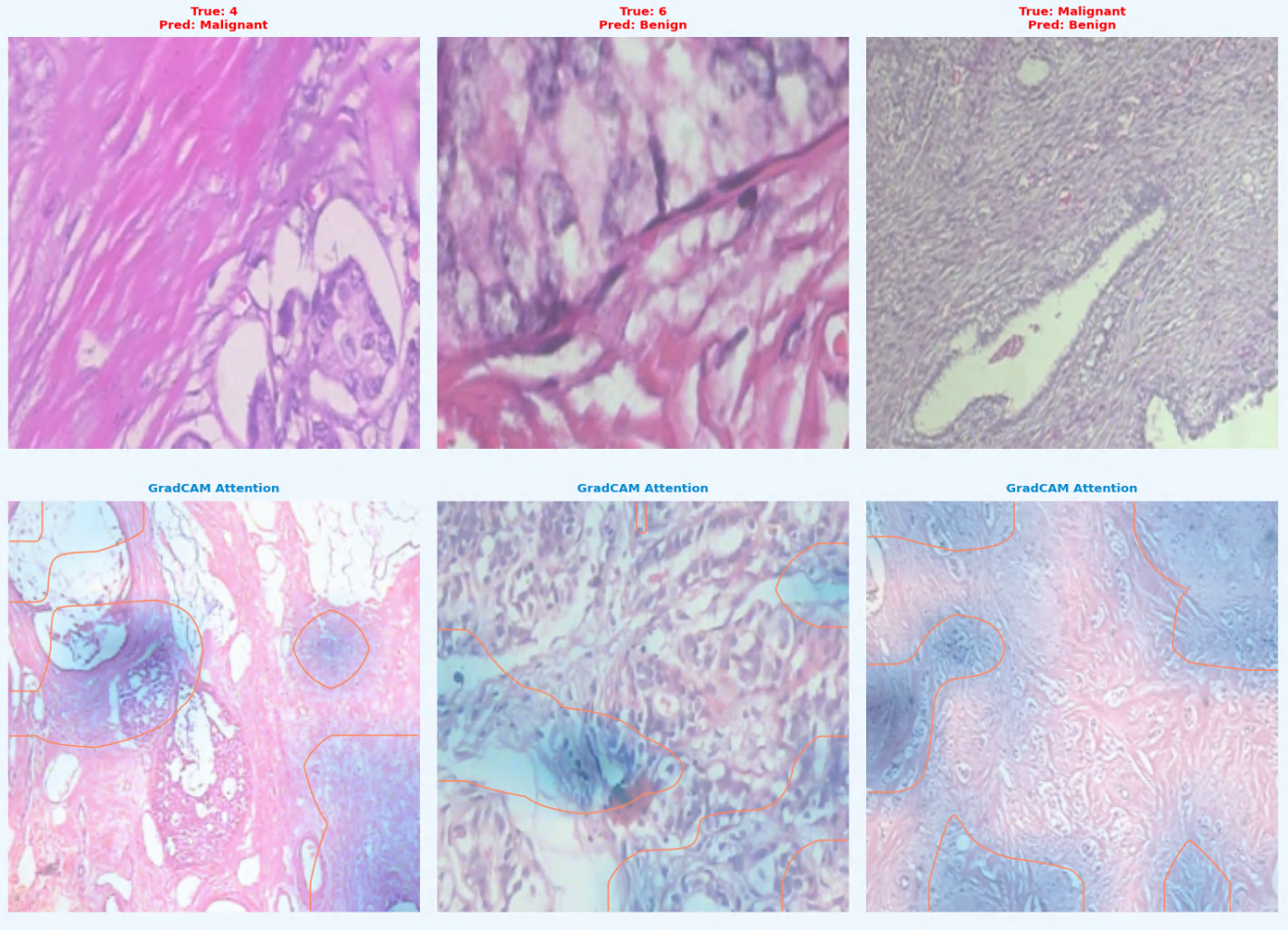

Figure 14. Binary classification (Whole Dataset) — Grad-CAM attention maps for representative binary classification samples, highlighting misclassified cases where diffuse and weakly localised activations indicate ambiguity in distinguishing benign and malignant tissue regions.

efficiency. Second, the incorporation of multi-scale feature representation through FPN-style aggregation enhances the model's ability to capture both local and global histopathological patterns, although recent transformer-based and hybrid CNN–ViT models demonstrate stronger global context modeling. Third, the two-phase training strategy improves stability during optimization and enables controlled domain adaptation, but similar progressive fine-tuning strategies have also been widely explored in recent studies.

Despite these design advantages, the proposed model achieves 95.53% accuracy in binary classification and 78.10% in multi-class classification, indicating strong binary discrimination but comparatively limited performance in fine-grained subtype classification. This suggests that while the architecture is effective for coarse-grained malignancy detection, further improvements are required to match the performance of recent hybrid and attention-based state-of-the-art methods in multiclass histopathological classification tasks.

5 Conclusion

This paper presented a comprehensive deep learning framework for breast cancer histopathological image classification based on CSPResNet50, evaluated across four systematically defined tasks on the BreaKHis v1 benchmark. The proposed architecture extended a pre-trained CSPResNet50 backbone with Feature Pyramid Network-style multi-scale feature aggregation across three spatial resolution scales, Squeeze-and-Excitation channel-wise attention, dual Global Average and Global Max Pooling per scale, and a four-stage dense classification neck — collectively producing a 1536-dimensional multi-scale descriptor that simultaneously captures fine-grained cellular morphology and broader tissue architectural patterns within a single forward pass. A two-phase progressive training strategy was introduced to prevent catastrophic forgetting while enabling

Table 7. Comparison with representative published methods on BreaKHis.

Study	Year	Architecture	Results
[11]	2026	LWT + Multi-path CNN	99.34% (Multiclass)
[12]	2026	CNN (VGG16, ResNet50, DenseNet121) + ViT ensemble	98.7% (Binary), 95.8% (Multiclass)
[13]	2026	EfficientNetV2S + ViT + Capsule Network	98.5% (Binary), 95.4% (Multiclass)
[29]	2025	ResNet50 + DenseNet121 ensemble	98.61% (Binary)
[22]	2025	ResNet-based CNN (transfer learning)	92.42% (Multiclass), 99.86% AUC
[23]	2025	Wide Residual Network (WRN-28-2)	99.16% (Binary)
[31]	2024	CNN + Transformer hybrid	98.13% (BreakHis), 97.80% (BACH)
[32]	2024	EfficientNetV2 + CBAM attention	94.2% (40X), 88.41% (200X), 89.42% (400X)
Proposed CSPResNet50			Whole Dataset - 95.53% (Binary), 78.10% (Multiclass)

complete domain adaptation: Phase 1 trained only the task-specific head with the backbone frozen, while Phase 2 performed full fine-tuning with a layer-wise learning rate differential. This was combined with Focal Loss with label smoothing, MixUp regularisation, Warmup Cosine Decay scheduling, and automatic mixed-precision training. At inference, eight-fold Test-Time Augmentation and Youden's J-optimal threshold selection were applied, producing clinically aligned operating points that jointly maximise sensitivity and specificity. GradCAM attention visualisations confirmed that the model consistently localised its decisions over diagnostically relevant tissue features — nuclear pleomorphism, mitotic figures, and disrupted glandular architecture — rather than background stroma or slide preparation artefacts. Systematic evaluation across all four classification tasks — whole-dataset multi-class, per-magnification multi-class, whole-dataset binary, and per-magnification binary — under an identical stratified 70/15/15 experimental protocol demonstrated consistent and competitive performance, with the magnification-stratified results providing practically actionable insights into the relative diagnostic utility of each optical scale.

5.1 Limitations

- **Patient-Level Data Leakage:** Image-level splitting may cause patient data leakage across train/test sets, leading to optimistic performance. Patient-level splitting is required for clinically reliable evaluation.
- **Single-Magnification Training:** Models are trained separately per magnification, ignoring cross-scale diagnostic dependencies. Multi-magnification fusion is needed for realistic clinical workflows.
- **Absence of Stain Normalisation:** No stain normalization is applied, making the model sensitive to colour variability across scanners and labs. This may reduce cross-domain generalisation.
- **Single Dataset Evaluation:** Experiments are limited to BreaKHis only, restricting external validity. Cross-dataset and multi-centre validation are required for clinical deployment.
- **Class Imbalance in Subtypes:** Severe imbalance among histological subtypes may bias learning toward dominant classes. Rare classes show higher metric variance despite sampling strategies.

5.2 Future Work

- **Patient-Level Cross-Validation:** Future work should adopt patient-level k-fold cross-validation to ensure unbiased generalisation and provide statistically reliable performance estimates.
- **Multi-Magnification Fusion:** A unified model combining multiple magnification levels via feature fusion or cross-attention can better capture hierarchical tissue information.
- **Stain Normalisation Integration:** Incorporating Macenko or SPCN stain normalization can improve robustness to staining variability across different institutions.
- **EMA Weight Averaging:** Using exponential moving average of model weights can improve stability, calibration, and generalisation without additional training cost.
- **Checkpoint Ensembling:** Ensembling top-k checkpoints can reduce variance and improve inference stability with no extra training overhead.
- **Precision-Recall AUC:** PR-AUC should complement ROC-AUC for imbalanced data, providing a more sensitive evaluation of malignant class performance.
- **Self-Supervised Pretraining:** Domain-specific pretraining using contrastive or masked image

modelling can reduce domain gap and improve feature quality.

- **Multi-Centre Validation:** Evaluation on external datasets (e.g., TCGA, CAMELYON) is necessary to assess robustness under real-world domain shifts.

Data Availability Statement

The dataset used in this study is publicly available and can be accessed via the Kaggle repository: https://www.kaggle.com/datasets/ambarish/breakhis

The source code for this study is available on Kaggle: https://www.kaggle.com/code/mhamza35157/cspresnet-50.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that generative artificial intelligence (AI) was used solely for language editing and proofreading to improve the clarity, readability, and overall quality of the manuscript. The AI tool utilized was Claude Sonnet 4.6. All AI-assisted outputs were carefully reviewed, verified, and revised by the authors. The authors take full responsibility for the content of the manuscript, including the accuracy of the information, the originality of the work, and the integrity of the research.

Ethical Approval and Consent to Participate

This study used the publicly available BreaKHis dataset (http://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/), which was originally collected under institutional ethical approval at the P&D Laboratory – Pathological Anatomy and Cytopathology, Parana, Brazil. No new patient data were collected for this study, and all images in the dataset were anonymised prior to public release. Therefore, additional ethical approval was not required for this work.

References

[1] Tanveer, H., Faheem, M., Khan, A. H., & Adam, M. A. (2025). AI-powered diagnosis: a machine learning approach to early detection of breast cancer. *International Journal of Engineering Development and Research*, *13*(2), 153-166. https://www.researchgate.net/publication/391015863

[2] Arravalli, T., Chadaga, K., Muralikrishna, H., Sampathila, N., Cenitta, D., Chadaga, R., & Swathi, K. S. (2025). Detection of breast cancer using machine learning and explainable artificial intelligence. *Scientific Reports*, *15*(1), 26931. [CrossRef]

[3] Khan, M. M. R., Zafar, M., Majid, A., & Khan, F. A. (2025). Recent Advances in Breast Cancer Detection: A Review on Segmentation and Classification Techniques. *ICCK Journal of Image Analysis and Processing*, *1*(4), 147-161. [CrossRef]

[4] Wahed, M. A., Alqaraleh, M., Alzboon, M. S., & Al-Batah, M. S. (2025). Evaluating AI and Machine Learning Models in Breast Cancer Detection: A Review of Convolutional Neural Networks (CNN) and Global Research Trends. *LatIA*, *3*, 117-117. [CrossRef]

[5] Jahan, I., Chowdhury, M. E., Vranic, S., Al Saady, R. M., Kabir, S., Pranto, Z. H., ... & Nobi, S. F. (2025). Deep learning and vision transformers-based framework for breast cancer and subtype identification. *Neural Computing and Applications*, *37*(16), 9311-9330. [CrossRef]

[6] Chinta, P. C. R., Moore, C. S., Karaka, L. M., Sakuru, M., & Bodepudi, V. (2025). Predictive Analytics for Disease Diagnosis: A Study on Healthcare Data with Machine Learning Algorithms and Big Data. *J Cancer Sci*, *10*(1), 1. [CrossRef]

[7] Hamza, M., Ali, I., Ali, S., Khan, W., Shah, S. M., & Haq, I. U. (2025). Leveraging machine learning and deep learning for advanced malaria detection through blood cell images. *ICCK Journal of Image Analysis and Processing*, *1*(1), 17-26. [CrossRef]

[8] Khan, S. S., Khan, M., & Khan, R. S. (2023, November). Automatic segmentation and classification to diagnose coronary artery disease (AuSC-CAD) using angiographic images: a novel framework. In *2023 18th International Conference on Emerging Technologies (ICET)* (pp. 110-115). IEEE. [CrossRef]

[9] Rezayi, S., Nilashi, M., Esmaeeli, E., Ramezanghorbani, N., Arji, G., Ahmadi, H., ... & Zahmatkeshan, M. (2025). A scoping and bibliometric review of deep learning techniques in breast cancer imaging: mapping the landscape and future directions. *Neural Computing and Applications*, *37*(22), 17759-17823. [CrossRef]

[10] Nasser, M., & Yusof, U. K. (2023). Deep learning based methods for breast cancer diagnosis: a systematic review and future direction. *Diagnostics*, *13*(1), 161. [CrossRef]

[11] Bohra, M., Singh, K. U., Kumar, I., & Shah, M. A. (2026). Wavelet-CNN Feature Fusion Architecture for Robust Breast Cancer Classification

in Histopathological Imaging. *International Journal of Computational Intelligence Systems, 19*(1), 136. [CrossRef]

[12] Ahmad, F., Jaffar, A., Latif, G., Alghazo, J., & Bhatti, S. M. (2026). Ensemble Deep Learning-Based High-Precision Framework for Breast Cancer Detection from Histopathological Images. *Diagnostics, 16*(5), 653. [CrossRef]

[13] Kushwah, S. S., Punn, N. S., & Bhattacharya, M. (2026). EVC-Net: A Hybrid Deep Learning Network for Breast Cancer Classification from Histopathological Images. *ACM Transactions on Intelligent Systems and Technology, 17*(4), 1-21. [CrossRef]

[14] Alshohoumi, F., & Al-Hamdani, A. (2026). Comparative Evaluation of Fusion Strategies Using Multi-Pretrained Deep Learning Fusion-Based (MPDLF) Model for Histopathology Image Classification. *Applied Sciences, 16*(4), 1964. [CrossRef]

[15] Makina, M., Chacha, W. M., & Mwangi, R. W. (2026). A Hybrid SwinTConvNeXt with Learnable Dynamic Gating Fusion for Breast Cancer Histopathology Image Classification. *Engineering Reports, 8*(2), e70655. [CrossRef]

[16] Behzadpour, M., Ortiz, B. L., Azizi, E., & Wu, K. (2026, February). A Novel Deep Learning Approach for Breast Tumor Applications Using Histopathological Images. In *SoutheastCon 2026* (pp. 1-6). IEEE. [CrossRef]

[17] Jaikumar, S. S., & Praveena, S. M. (2026). Breast cancer diagnosis from histopathological images and molecular signatures by fusing features with an explainable AI-based residual tabular network model. *Journal of Computer-Aided Molecular Design, 40*(1), 3. [CrossRef]

[18] Vijay, S., Raja, M. A. M., & Mangayakarasi, J. (2026, January). AENN: An Integrated Attention-Enhanced Neural Network Architecture for Automated Histopathology Image Assessment. In *2026 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)* (pp. 1-9). IEEE. [CrossRef]

[19] Opoku, D., Owusu-Agyemang, K., Hayfron-Acquah, J. B., & Gyening, R. M. O. M. (2026). SwinCNN+ OE: A Swin Transformer and CNN Architecture for Breast Histopathology Classification With OOD and Grad-CAM Integration. *IEEE Access, 14*, 3897-3909. [CrossRef]

[20] Nguyen, T. T., Nguyen, T. H., Nguyen, T. N., & Ngo, B. V. (2026). Dual-view colorized mammogram ROIs with EfficientNet-B7 and attention network for breast lesion classification. *IEEE Access, 14*, 35511–35526. [CrossRef]

[21] Rafiq, A., Jaffar, A., Latif, G., Masood, S., & Abdelhamid, S. E. (2025). Enhanced multi-class breast cancer classification from whole-slide histopathology images using a proposed deep learning model. *Diagnostics, 15*(5), 582. [CrossRef]

[22] Desai, A., & Mahto, R. (2025). Multi-class classification of breast cancer subtypes using ResNet architectures on histopathological images. *Journal of Imaging, 11*(8), 284. [CrossRef]

[23] Alaeddine, H., & Tekari, L. (2025). Application and Deployment of a Fine-Tuned Pre-trained Deep Model for Breast Cancer Classification. *ICCK Journal of Image Analysis and Processing, 1*(4), 162-171. [CrossRef]

[24] Hossain, M. S., Rahman, A., Ahmed, M., Fatema, K., & Mahbubul, M. M. (2025). Vision transformer for the categorization of breast cancer from H&E histopathology images. *J Image Gr, 13*(4). [CrossRef]

[25] Kaczmarek, M., Kowal, M., & Korbicz, J. (2025). Exploring data preparation strategies: A comparative analysis of vision transformer and ConvNeXT architectures in breast cancer histopathology classification. *International Journal of Applied Mathematics and Computer Science, 35*(2). [CrossRef]

[26] Abimouloud, M. L., Bensid, K., Elleuch, M., Ammar, M. B., & Kherallah, M. (2025). Advancing breast cancer diagnosis: token vision transformers for faster and accurate classification of histopathology images. *Visual computing for industry, biomedicine, and art, 8*(1), 1. [CrossRef]

[27] Zhou, Y., Jin, F., Suo, G., & Yang, J. (2025). ResViT-GANNet: a deep learning framework for classifying breast cancer histopathology images using multimodal attention and GAN-based augmentation. *BMC Medical Imaging, 25*(1), 401. [CrossRef]

[28] Osmani, N., Esmaeeli, E., & Rezayi, S. (2025). Fusion strategies for deep convolutional neural network representations in histopathological image classification. *The Journal of Supercomputing, 81*(2), 395. [CrossRef]

[29] Rahi, A. (2025). Ensemble Deep Learning for Histopathological Breast Cancer Detection. *medRxiv*, 2025-08. [CrossRef]

[30] Yuan, B., Hu, Y., Liang, Y., Zhu, Y., Zhang, L., Cai, S., ... & Hu, J. (2025). Comparative analysis of convolutional neural networks and transformer architectures for breast cancer histopathological image classification. *Frontiers in Medicine, 12*, 1606336. [CrossRef]

[31] Sreelekshmi, V., Pavithran, K., & Nair, J. J. (2024). SwinCNN: An integrated Swin Transformer and CNN for improved breast cancer grade classification. *IEEE Access, 12*, 68697–68710. [CrossRef]

[32] Aldakhil, L. A., Alhasson, H. F., & Alharbi, S. S. (2024). Attention-based deep learning approach for breast cancer histopathological image multi-classification. *Diagnostics, 14*(13), 1402. [CrossRef]

[33] Abioye, O. A., Evwiekpaefe, A. E., & Awujoola, A. J. (2024). Performance evaluation of efficientnetv2 models on the classification of histopathological benign breast cancer images. *Science Journal of University of Zakho, 12*(2), 208-214. [CrossRef]

[34] Eshun, R. B., Bikdash, M., & Islam, A. K. (2024). A deep convolutional neural network for the classification of imbalanced breast cancer dataset. *Healthcare Analytics, 5*, 100330. [CrossRef]

[35] Taheri, S., Golrizkhatami, Z., Basabrain, A. A., & Hazzazi, M. S. (2024). A comprehensive study on classification of breast cancer histopathological images: Binary versus multi-category and magnification-specific versus magnification-independent. *IEEE Access, 12*, 50431-50443. [CrossRef]

[36] Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2015). A dataset for breast cancer histopathological image classification. *IEEE transactions on biomedical engineering, 63*(7), 1455-1462. [CrossRef]

[37] PyTorch Contributors. (2024). *PyTorch Documentation*. Retrieved May 1, 2026, from https://pytorch.org/docs/stable/index.html.

[38] Wightman, R. (2019). *timm: PyTorch Image Models*. Retrieved May 1, 2026, from https://github.com/huggingface/pytorch-image-models.

[39] Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 390-391).

[40] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).

[41] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).

[42] Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*. [CrossRef]

[43] Sarker, M. M. K., Akram, F., Alsharid, M., Singh, V. K., Yasrab, R., & Elyan, E. (2022). Efficient breast cancer classification network with dual squeeze and excitation in histopathological images. *Diagnostics, 13*(1), 103. [CrossRef]

[44] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988).

[45] Howard, J., & Ruder, S. (2018, July). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*) (pp. 328-339). [CrossRef]

[46] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*. [CrossRef]

[47] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., ... & Wu, H. (2017). Mixed precision training. *arXiv preprint arXiv:1710.03740*. [CrossRef]

[48] Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32-35. [CrossRef]

[49] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626). [CrossRef]

Muhammad Hamza is a Computer Science graduate and is currently pursuing a Master's degree in Computer Science at the University of Swat, Swat, 01923, Pakistan. His research interests include medical image analysis using Machine Learning and Deep Learning, as well as hybrid approaches such as ensemble learning. (Email: mhamzaedu1@gmail.com)

Ibrar Ali is a Computer Science graduate and currently pursuing a Master's degree in Computer Science at the University of Swat, Pakistan. He is serving as a Visiting Lecturer in the Department of Computer and Software Technology. His research interests include Machine Learning, Deep Learning, Computer Vision, and Natural Language Processing. (Email: ibraralin@gmail.com)

Sikandar Ali is a Computer Science graduate and currently pursuing a Master's degree in Computer Science at the University of Swat, Pakistan. He is serving as a Visiting Lecturer in the Department of Computer and Software Technology. His research interests include Machine Learning, Deep Learning, Computer Vision, and Natural Language Processing.

Shayan Khan is a Computer Science student currently pursuing his undergraduate degree in the field of Computer Science. His research interests include Medical Image Analysis and Computer Vision using Machine Learning, Deep Learning, and Ensemble Learning. He is passionate about Artificial Intelligence and continuously improving his technical skills through practical projects and learning. (Email: shayanedu7@gmail.com)