



Explainable Deep Fusion for Breast Cancer Classification in Histopathology Images

Sitthichat Meekhwan¹, Buntueng Yana², Ratsada Praphasawat³ and Kunaree Wongrach⁴

ABSTRACT

Breast cancer diagnosis from histopathology images is accurate but labor-intensive and subject to interobserver variability. Deep learning has shown strong performance on digital pathology benchmarks; however, limited transparency, magnification variability, and patient-level data leakage can reduce clinical trust and robustness. This paper proposes an explainable deep fusion framework for magnification-robust breast histopathology classification on the BreakHis dataset. In the experiments, all folds are generated at the patient level, and model selection is performed in two stages: first, single backbones and backbone-pair fusion models are compared without attention; second, attention modules are evaluated on the selected fusion pair. The results identify VGG16–ResNet18 fusion without attention as the final configuration, achieving an accuracy of 0.8890, an F1-score of 0.8888, and an AUC of 0.9326 under mixed-magnification, patient-level 5-fold cross-validation. Although the improvement over the strongest single-backbone baseline is modest and not statistically significant in the fold-level Wilcoxon test, an image-level discordant-case analysis shows that the proposed fusion model corrects more EfficientNetV2 errors than it introduces. To support transparency, Grad-CAM-based case analysis is used to examine corrected, newly misclassified, and persistently misclassified examples, with LIME and SHAP retained as complementary explanation tools when generated for selected cases. These results suggest that feature-level fusion can provide useful gains for some difficult histopathology images while highlighting remaining limitations caused by ambiguous tissue morphology, staining variation, magnification-dependent texture changes, and attention to non-diagnostic regions.

Article information:

Keywords: Breast Cancer, Histopathology, Deep Learning, Feature Fusion, Attention, Explainable AI

Article history:

Received: January 9, 2026

Revised: June 12, 2026

Accepted: August 13, 2026

Published: September 19, 2026

(Online)

DOI: 10.37936/ecti-cit.2026204.265640

1. INTRODUCTION

Globally, breast cancer (BC) remains a leading cause of cancer-related death and one of the most prevalent cancers in women [1]. Recent estimates indicate that 2.3 million new cases and 670,000 deaths from female breast cancer occurred in 2022, with the burden projected to increase substantially by 2050 [2, 3]. Because BC can develop silently in early stages, delayed diagnosis is associated with worse outcomes, indicating the importance of effective early detection

and reliable diagnostic support tools [2, 4].

Across the BC care pathway, multiple imaging modalities are used [5]. While screening is primarily performed using mammography, definitive diagnosis and tumor characterization rely on histopathological examination of H&E-stained tissue, which guides grading and treatment decisions [6]. With the adoption of digital pathology and whole-slide imaging, deep learning (DL) has achieved strong performance on various pathology tasks; however, evidence syntheses still report variability in study design, validation

^{1,2}The authors are with the Department of Electrical Engineering, School of Engineering, University of Phayao, Phayao, 56000, Thailand., Email: mr.sitthichat.n@gmail.com and buntueng.ya@up.ac.th

³The author is with the Department of Pathology, School of Medicine, University of Phayao, Phayao, 56000, Thailand., Email: ratsada.pr@up.ac.th

⁴The author is with the Department of Physics, Faculty of Science, Burapha University, Chonburi, 20131, Thailand., Email: kunaree@buu.ac.th

²Corresponding author: buntueng.ya@up.ac.th

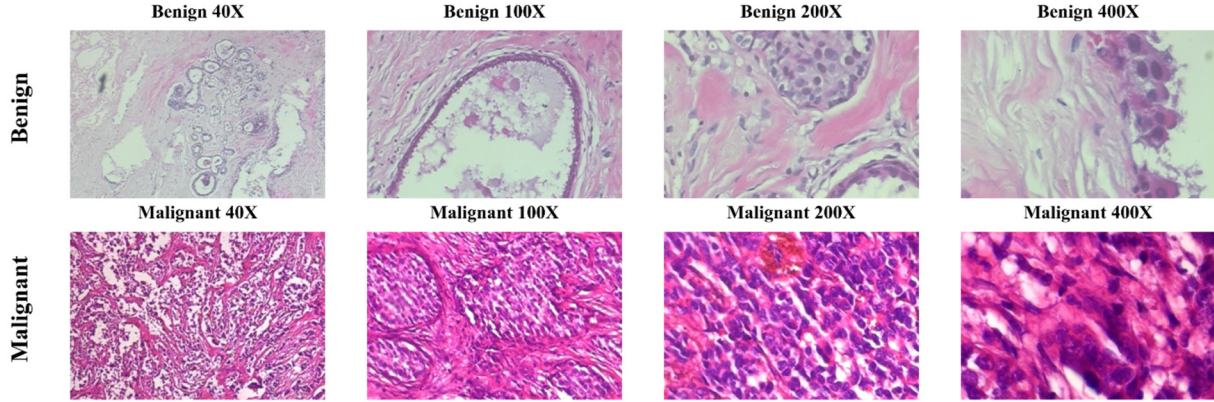


Fig.1: Exemplar BreakHis images at multiple magnification levels (40 $\times$, 100 $\times$, 200 $\times$, and 400 $\times$).

practices, and clinical translation [7].

Despite its central role, histopathology assessment is time-consuming, labor-intensive, and subject to inter-observer variability, particularly for subjective tasks such as grading [8, 9]. Decision-support systems based on DL can improve efficiency and consistency, but many high-performing models operate as black boxes, limiting trust and hindering error analysis in high-stakes settings [10, 11]. Explainable AI (XAI) can mitigate these concerns by providing human-interpretable evidence, and its design should align with clinical workflows and user needs [12–14].

This study concentrates on breast histopathology images and introduces an explainable deep fusion model to improve classification accuracy and decision transparency. The proposed approach is evaluated on the BreakHis [15] benchmark dataset under a magnification-robust setting, targeting stable performance when magnification varies across samples [16–18]. The primary contributions of this research comprise (i) a two-stage framework that benchmarks multiple CNN backbones and then fuses the most effective candidates into a unified model, (ii) an attention-guided fusion strategy that recalibrates and combines complementary feature representations for breast histopathology classification, (iii) an explainability pipeline to help users validate model attention, recognize failure scenarios, and support human-in-the-loop review [12, 13], and (iv) detailed reporting of preprocessing, training configuration, and evaluation metrics to facilitate reproducibility and fair comparison with prior work.

This study also narrows the research gap more explicitly. Many BreakHis-based fusion models report high aggregate accuracy, but their contributions are difficult to assess when patient-level splitting, backbone-pair selection, statistical testing, and image-level error correction are not reported together. Existing fusion designs are therefore insufficient for explaining whether a small numerical gain represents a consistent improvement, a few changed predictions, or possible split-related optimism. To address

this issue, the manuscript emphasizes patient-level evaluation, backbone-pair ablation, attention ablation, statistical testing, discordant-case analysis, and explanation-guided review of difficult images.

2. RELATED WORK

This section summarizes prior work on BC classification, with an emphasis on histopathology image analysis. We first review commonly used public benchmarks for breast histopathology and motivate our choice of datasets. We then discuss deep learning approaches for histopathology classification, including CNN backbones, transformer-based models, and fusion/ensemble strategies. Next, we review magnification variability as a robustness challenge and outline evaluation practices for multi-magnification learning. Finally, we summarize attention mechanisms used in multi-branch fusion models and common XAI techniques that support interpretation and error analysis in medical imaging.

2.1 Breast Histopathology Datasets

Several public datasets support benchmarking in breast histopathology, spanning microscopy-image classification and whole-slide image (WSI) analysis, including BreakHis [15], BACH [19], CAME-LYON17 [20], and TUPAC16 [21].

BACH, released for the ICIAR 2018 challenge, provides a complementary microscopy benchmark for four-class diagnosis. However, its microscopy images are captured at a fixed magnification, and the dataset size is small [19]. CAMELYON17 is primarily a WSI benchmark focused on detecting lymph node metastasis, where models typically learn from patches extracted from multi-resolution slide pyramids rather than from images acquired at different optical magnifications [20]. TUPAC16 also targets WSI-based learning, focusing on tumor proliferation assessment and mitosis-related tasks, and similarly relies on WSI pipelines and patch aggregation [21].

BreakHis is a widely used microscopy benchmark

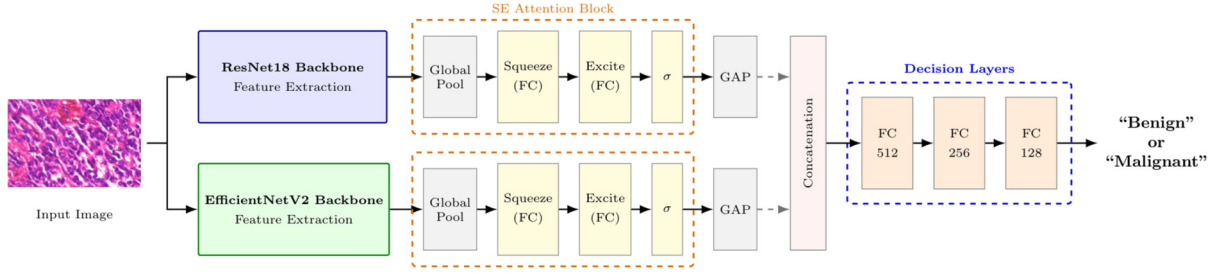


Fig.2: General two-stream fusion architecture used for backbone-pair and attention ablation experiments.

explicitly designed for multi-magnification evaluation. It contains 7,909 H&E-stained RGB biopsy images from 82 patients acquired at four magnifications ($40\times$, $100\times$, $200\times$, and $400\times$), with binary labels (benign/malignant) and subtype annotations [15]. The dataset includes 2,480 benign and 5,429 malignant images, reflecting a moderate class imbalance. Beyond the binary label, BreakHis provides eight histological subtypes: four benign and four malignant [15]. Because the images are acquired at different optical magnifications, the effective pixel size and the appearance of nuclei, glands, and stromal structures vary across $40\times$ – $400\times$. Fig. 1 illustrates these scale-dependent changes, which motivate explicit robustness analysis under magnification variability. A further practical consideration is that multiple images are available per patient, so patient-level splitting is important to avoid leakage when forming cross-validation folds.

In this work, we use BreakHis as our primary benchmark for three concrete reasons: (i) it provides an explicit multi-magnification acquisition setup ($40\times$ – $400\times$), which directly supports evaluating robustness under magnification variability; (ii) it offers a sufficiently large number of labeled microscopy images across benign and malignant cases, enabling stable training and stratified cross-validation; and (iii) it is widely adopted in the literature, facilitating direct and fair comparison with prior microscopy-based approaches [15].

2.2 Breast Histopathology Classification

Breast histopathology classification commonly leverages transfer learning with CNN backbones to balance performance and complexity in data-constrained settings. Across public benchmarks, VGG- and ResNet-based pipelines remain strong baselines, often combined with data augmentation and careful validation [4, 22–24]. More recent CNN families can further improve accuracy when appropriately fine-tuned and regularized [25].

Beyond single-backbone classifiers, ensembles and fusion strategies are frequently adopted to improve robustness by integrating complementary representations. Ensembles typically aggregate predictions from multiple models, whereas fusion combines in-

formation at the feature or logit level within a single architecture [24, 25]. Multi-branch and multi-scale designs fuse features across depths or receptive fields to capture both local texture and broader tissue context [26, 27]. Hybrid CNN–transformer approaches have also been explored to combine CNN-derived local cues with transformer-derived global context, motivating representation-level fusion as a practical design choice for histopathology classification [28]. Motivated by these trends, we benchmark strong CNN backbones and then adopt a two-stream fusion model to exploit complementary feature spaces while maintaining a lightweight and reproducible pipeline.

2.3 Image Magnification Variability

A key challenge in microscopy images is magnification variability. Tissue structures can appear substantially different across magnifications, altering the relative importance of local texture versus broader architectural patterns. As a result, magnification-dependent evaluation may overestimate performance if models rely on scale-specific cues that do not generalize. Early work explicitly studied cross-magnification generalization and motivated magnification-aware evaluation protocols for breast histology [16]. Subsequent studies proposed magnification-invariant architectures and transfer strategies to improve robustness under magnification shifts [17, 18, 29, 30].

In this paper, we focus on a magnification-robust setting in which training and testing contain mixed magnification levels, aiming for stable performance under heterogeneous magnification conditions. A strict magnification-robust evaluation, where training and testing magnifications differ, is a natural extension and is left for future work.

2.4 Attention Mechanisms in Fusion Models

Attention mechanisms are widely used in fusion models to emphasize informative features before merging streams. This is particularly relevant in breast histopathology, where non-diagnostic regions, stain variation, or background can dominate the field of view. Channel attention is a common and lightweight choice: squeeze-and-excitation

(SE) blocks learn channel-wise weights from global context and can be inserted with minimal overhead [31]. Spatial attention is often paired with channel attention to highlight discriminative regions; CBAM sequentially applies channel and spatial attention and is frequently used in medical imaging pipelines [32]. Non-local/self-attention blocks capture long-range dependencies that standard convolutions may miss, supporting broader tissue-context modeling [33]. Attention-enhanced fusion has been reported to improve breast histopathology classification by strengthening representation integration, e.g., via combined channel and spatial attention in fusion settings [34]. Related work also explores soft-attention designs with an explicit emphasis on interpretability alongside performance [35]. Motivated by these findings, we study attention within a fusion architecture and select SE attention as a practical accuracy-complexity trade-off in our final model.

2.5 Model Transparency

XAI is increasingly used to make deep learning decisions more inspectable in medical imaging, including histopathology, where models may otherwise rely on spurious cues. This consideration is amplified in fusion models, where multiple representation streams and attention components can complicate error analysis. In this work, we employ Grad-CAM [36], LIME [37], and SHAP [38].

Grad-CAM produces class-discriminative localization maps by using the gradients of the target class with respect to the final convolutional feature maps, offering a coarse but intuitive heatmap for CNN-based models [36]. LIME explains individual predictions by perturbing superpixels and fitting a local surrogate model, which can help reveal whether a prediction is driven by meaningful tissue regions or artifacts [37]. SHAP provides additive feature attributions grounded in game-theoretic principles, enabling more consistent comparisons across models and supporting attribution-based error analysis in benchmarking studies [38]. These tools collectively support transparency, qualitative verification, and failure-case analysis in our breast histopathology classification pipeline.

3. METHODOLOGY

This section presents the proposed fusion framework for magnification-robust breast histopathology classification and the experimental protocol used for evaluation. We first describe the backbone baselines and then introduce a fusion model augmented with attention. We conclude with implementation details, cross-validation, and evaluation criteria.

3.1 Proposed Model Architecture

The proposed framework is organized as a two-stage fusion-selection pipeline. In Stage 1, three single-backbone baselines, VGG16, ResNet18, and EfficientNetV2, are trained, and three two-stream fusion pairs are evaluated: VGG16–ResNet18, VGG16–EfficientNetV2, and ResNet18–EfficientNetV2. This stage is performed without attention so that the contribution of backbone pairing can be isolated. In Stage 2, the best fusion pair from Stage 1 is further evaluated with different attention modules. This design provides a systematic basis for selecting the final backbone pair rather than assuming that ResNet18–EfficientNetV2 is automatically the best choice.

An overview of the general two-stream fusion architecture is shown in Fig. 2. Given an input image, two CNN backbones extract feature maps in parallel. The feature maps are pooled, concatenated, and passed to a fully connected classifier for benign/malignant prediction. The final configuration uses the VGG16–ResNet18 pair because it provides the highest mean F1-score in the patient-level 5-fold experiments while also having lower inference time than the other fusion pairs.

Given an input RGB image $x \in R^{H \times W \times 3}$, where H and W denote the image height and width, a backbone $f(\cdot)$ produces a deep feature map $\mathbf{F} = f(x) \in R^{C \times h \times w}$. Here, C is the number of feature channels, and h and w are the spatial dimensions of the feature map. We aggregate spatial information using global average pooling (GAP) to obtain a compact embedding $\mathbf{z} \in R^C$:

$$z_c = \frac{1}{hw} \sum_{i=1}^h \sum_{j=1}^w F_c(i, j), \quad c = 1, \dots, C, \quad (1)$$

where $F_c(i, j)$ denotes the activation at the spatial location (i, j) in the channel c .

In the fusion setting, two backbones generate feature maps $\mathbf{F}_1 = f_1(x)$ and $\mathbf{F}_2 = f_2(x)$. For the no-attention fusion model, global average pooling is applied to both streams to obtain embeddings $\mathbf{z}_1 \in R^{C_1}$ and $\mathbf{z}_2 \in R^{C_2}$, which are concatenated as $\mathbf{z} = [\mathbf{z}_1; \mathbf{z}_2]$. The concatenated representation is then passed through a classifier head with dimensions $512 \rightarrow 512 \rightarrow 512 \rightarrow 2$. For attention ablation, an attention module is inserted before pooling to refine each stream, yielding $\tilde{\mathbf{F}}_1$ and $\tilde{\mathbf{F}}_2$. The same classifier head is then used so that changes in performance can be attributed to the attention mechanism rather than to a different classifier.

We evaluate self-attention, SE attention, cross-attention, and CBAM. The attention ablation is designed to determine whether an attention module improves the selected fusion pair under the patient-level protocol. The final model is selected according to the patient-level cross-validation results, and the attention ablation is reported in Section 4.3

Table 1: *Experimental parameters used in this study.*

Parameter	Value
Dataset	BreakHis
Task	Binary classification: benign vs. malignant
Image size	224×224 pixels
Input channels	RGB
Cross-validation	5-fold cross-validation
Split strategy	Patient-level split
Magnification	setting mixed magnification
Magnifications	$40\times$, $100\times$, $200\times$, and $400\times$
Pretrained weights	ImageNet pretrained weights
Normalization	ImageNet mean and standard deviation
Batch size	32
Optimizer	Adam/AdamW, depending on the model implementation
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Label smoothing	0.05
Maximum epochs	40
Early stopping patience	10 epochs
Classifier head	$512 \rightarrow 512 \rightarrow 512 \rightarrow 2$
Random seed	42

3.2 Experimental Setup

All models were implemented in PyTorch with CUDA 12.8 acceleration and trained on a workstation equipped with an AMD Ryzen 7 3700X CPU, 40 GB RAM, and an NVIDIA GeForce RTX 3080 GPU. All experiments use a fixed patient-level protocol for the baseline, fusion, and attention-ablation models. The key settings are summarized in Table 1.

We adopt a mixed-magnification setting in which images from all available magnifications are included in both the training and test sets. This protocol evaluates performance under heterogeneous magnification conditions rather than strict cross-magnification generalization. All images from the same patient are assigned to only one fold before training and validation subsets are constructed, which prevents patient-level leakage. After fold generation, the patient identifiers are checked to verify that no patient appears in more than one fold. Results are reported as the mean performance across folds. All images are resized to 224×224 pixels. Data augmentation is applied only to the training split to reduce overfitting. Augmented images are not saved as additional dataset samples; instead, augmentation is applied inside the training data loader so that validation and test images remain unchanged.

3.3 Evaluation Metrics

We evaluate classification performance using accuracy, precision, sensitivity, specificity, and F1-score,

defined in Eqs. (2)-(6). In addition, we report the ROC curve and the area under the ROC curve (AUC). The ROC curve is obtained by sweeping the decision threshold and plotting the true positive rate against the false positive rate, while the AUC summarizes this curve as a threshold-independent measure. All metrics are computed from the confusion-matrix terms: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \quad (6)$$

To support a more balanced interpretation of small performance differences, we also conduct paired statistical and image-level analyses. Fold-level differences are evaluated using the Wilcoxon signed-rank test. In addition, paired image-level predictions from EfficientNetV2 and the proposed fusion model are summarized into four discordant-case groups: both models correct, both models wrong, EfficientNetV2 wrong but the proposed model correct, and EfficientNetV2 correct but the proposed model wrong. McNemar's test is then applied to the discordant counts to assess whether the proposed model corrects more baseline errors than it introduces.

4. RESULTS AND DISCUSSION

This section presents the patient-level results. The analysis focuses on four questions: (i) how the proposed model compares with internally implemented baselines and relevant BreakHis studies, (ii) why the final backbone pair was selected, (iii) whether attention improves the selected fusion pair, and (iv) which image groups are corrected or newly misclassified by the fusion model.

4.1 Baseline and Fusion Comparison

Table 2 compares three single-backbone baselines with three fusion models trained without attention under the patient-level 5-fold cross-validation protocol. The VGG16-ResNet18 fusion model achieves the highest mean accuracy of 0.8890 and the highest mean F1-score of 0.8888. Among the single-backbone baselines, VGG16 gives the highest accuracy at 0.8712, followed closely by EfficientNetV2 at 0.8703. The improvement of the final fusion model

Table 2: Comparison of base models and fusion models without attention under patient-level validation.

Model	Accuracy	Precision	Sensitivity	Specificity	F1-score
VGG16	0.8712	0.8736	0.8712	0.8305	0.8709
ResNet18	0.8572	0.8647	0.8572	0.7854	0.8564
EfficientNetV2	0.8703	0.8762	0.8703	0.8081	0.8698
VGG16–ResNet18	0.8890	0.8917	0.8890	0.8475	0.8888
VGG16–EfficientNetV2	0.8659	0.8706	0.8659	0.8097	0.8655
ResNet18–EfficientNetV2	0.8635	0.8690	0.8635	0.8025	0.8630

over the strongest single-backbone baseline is therefore modest, but it is consistent in the mean accuracy, precision, sensitivity, specificity, and F1-score.

The absolute performance is lower than several values reported in previous BreakHis-based studies because the experimental protocols differ substantially. In particular, the patient-level protocol used in this study is stricter than image-level splitting. This distinction is important for BreakHis because the dataset contains multiple images from the same patient. If images from the same patient are split across training and test folds, visually related samples may appear in both subsets, which can lead to optimistic performance estimates. In the present protocol, all images from the same patient are assigned to only one fold, so the reported performance represents a more conservative estimate of generalization

to unseen patients.

Fig. 3 summarizes the training behavior of the baseline and fusion models. The curves are used to verify that the reported mean performance is not caused by a single unstable training run. Therefore, the figure should be interpreted together with the patient-level split and the fold-level statistics rather than as evidence of a large performance margin.

Table 3 provides a contextual comparison with selected BreakHis-based studies. Bardou et al. [22] reported high CNN-based binary classification accuracy across magnification settings; Gour et al. [23] reported 0.9252 accuracy and 0.9345 F1-score using residual-learning CNN features with an SVM classifier, and Taheri et al. [18] reported strong performance using a multi-level feature-fusion framework under binary/multi-class and magnification-specific

Table 3: Numeric and methodological comparison with related BreakHis studies.

Parameter	Bardou et al. [22]	Gour et al. [23]	Taheri et al. [18]	Proposed
Best Reported Accuracy	0.9833	0.9252	0.9750	0.8890
F1-score	N/R	0.9345	0.9798	0.8888
AUC	N/R	N/R	N/R	0.9326
Patient-Level Split	$\triangle$	$\checkmark$	$\checkmark$	$\checkmark$
Mixed-Magnification	$\times$	$\times$	$\checkmark$	$\checkmark$
Multi-Model Design	$\checkmark$	$\times$	$\checkmark$	$\checkmark$
Fusion Ablation	$\times$	$\times$	$\triangle$	$\checkmark$
Attention Ablation	$\times$	$\times$	$\times$	$\checkmark$
XAI	$\times$	$\times$	$\times$	$\checkmark$
Statistical Test	$\times$	$\times$	$\checkmark$	$\checkmark$
Discordant-Case Analysis	$\times$	$\times$	$\times$	$\checkmark$

Note: $\checkmark$: reported; $\times$: not reported; $\triangle$: partially reported or the protocol differs. N/R: not reported.

Table 4: Backbone-pair ablation under patient-level 5-fold cross-validation.

Fusion pair	Accuracy	F1-score	AUC	Parameters (M)	Inference time (ms)
VGG16–ResNet18	0.8890	0.8888	0.9326	26.58	1.15
VGG16–EfficientNetV2	0.8659	0.8655	0.9260	35.97	2.50
ResNet18–EfficientNetV2	0.8635	0.8630	0.9274	32.44	2.85

Table 5: Attention ablation for the selected fusion pair.

Attention module	Accuracy	Precision	Sensitivity	Specificity	F1 score
No attention	0.8890	0.8917	0.8890	0.8475	0.8888
SE attention	0.8625	0.8701	0.8626	0.7908	86.18
Self-attention	0.8662	0.8715	0.8662	0.8064	86.57
Cross-attention	0.8573	0.8607	0.8573	0.8092	85.70
CBAM	0.8601	0.8336	0.8601	0.8108	85.98

Table 6: Per-magnification performance of the final proposed model.

Magnification	Accuracy	Precision	Sensitivity	Specificity	F1-score
40×	0.8869	0.8913	0.8869	0.8336	0.8865
100×	0.9161	0.9162	0.9161	0.9130	0.9161
200×	0.9133	0.9141	0.9133	0.8914	0.9133
400×	0.8851	0.8866	0.8851	0.8547	0.8850

settings. These reported values are higher than the performance obtained in this work. However, direct comparison based only on accuracy or F1-score should be interpreted cautiously because the studies differ in data-splitting strategy, magnification setting, model-selection procedure, and validation analysis.

Accordingly, this study does not claim superiority over all previous BreakHis studies. Instead, the contribution is a transparent and reproducible evaluation of deep feature fusion under a patient-level mixed-magnification setting. Although the proposed model obtains lower absolute accuracy than several prior reports, this experiment provides additional evidence that is often missing in aggregate performance comparisons, including backbone-pair ablation, attention ablation, statistical testing, image-level discordant-case analysis, and explainability-guided error discussion. This more conservative interpretation is important because it shows not only whether the fusion model improves the mean metrics, but also where the model corrects baseline errors, where it introduces new errors, and why the performance gain should be interpreted cautiously.

4.2 Backbone-Pair Ablation

Table 4 reports the backbone-pair ablation used to select the final fusion model. VGG16–ResNet18 obtains the highest accuracy, F1-score, and AUC among the three tested fusion pairs. It also has fewer parameters and lower inference time than the VGG16–EfficientNetV2 and ResNet18–EfficientNetV2 pairs. These results explain why the final model uses VGG16–ResNet18 instead of the ResNet18–EfficientNetV2 pair. The VGG16 stream may preserve texture-rich and edge-oriented representations, whereas the ResNet18 stream adds residual features that support more stable optimization. In contrast, the EfficientNetV2 fusion pairs do not improve the patient-level mean metrics in this setting, suggesting that adding a stronger single backbone does not automatically yield a better fusion model when the dataset is small and patient-level constraints reduce the effective training diversity.

4.3 Attention Ablation

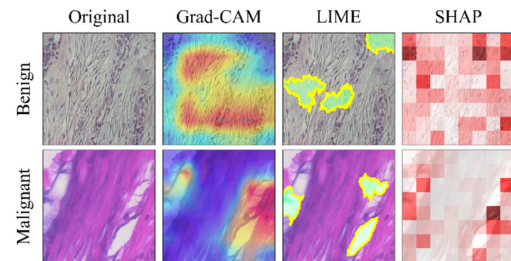
Table 5 evaluates four attention mechanisms on the selected VGG16–ResNet18 pair. The results show that attention does not improve this fusion pair under the patient-level protocol. The no-attention model

EfficientNetV2	Correct	9003 (82.9%)	447 (4.1%)
	Wrong	650 (6.0%)	758 (7.0%)
		Correct	Wrong
		Proposed Model	

Fig.3: Contingency matrix showing correct/wrong prediction distribution between EfficientNetV2 and the proposed model.

remains the best configuration, with 0.8890 accuracy and 0.8888 F1-score. Self-attention gives the strongest performance among the attention variants, but it remains below the no-attention fusion model.

The decline caused by attention modules may occur because the attention weights suppress subtle texture cues that are useful for BreakHis classification, particularly under mixed magnification. This result also indicates that the fusion benefit in this setting comes primarily from combining complementary backbone representations, not from adding an attention block. Therefore, the final model is defined as VGG16–ResNet18 feature fusion without attention.

**Fig.4:** Representative explanation examples for corrected and misclassified cases.

4.4 Per-Magnification Analysis

Although the model is trained and evaluated in a mixed-magnification setting, Table 6 summarizes the performance of the final proposed model at each magnification level. The best results are obtained at 100×

texture and high-detail nuclear texture can introduce different classification difficulties.

The magnification-specific pattern holds both clinical and technical significance. At magnifications of 100× and 200×, tissue organization and cellular morphology are discernible at a practical scale. However, at 40×, nuclear details may be less distinct, and at 400×, the image reveals highly localized texture patterns that are more susceptible to variations in staining and field-of-view selection. These findings support the implementation of per-magnification reporting alongside pooled mixed-magnification metrics.

4.5 Statistical Testing Analysis

Table 7 reports fold-level Wilcoxon signed-rank tests. The proposed model shows positive mean differences in accuracy and F1-score compared to EfficientNetV2; however, the p-values exceed 0.05. As a result, this study does not assert a statistically significant fold-level improvement. Given that only five folds are analyzed, the Wilcoxon test is regarded as supporting evidence rather than the sole criterion for model comparison.

Table 7: Wilcoxon signed-rank test across five folds.

Comparison	Metric	Mean difference	p-value
EfficientNetV2 vs. Proposed	Accuracy	+0.0133	0.3125
EfficientNetV2 vs. Proposed	F1-score	+0.0134	0.3125
EfficientNetV2 vs. Proposed	AUC	-0.0214	0.4375
No attention vs. Best attention	Accuracy	+0.0129	1.0000
No attention vs. Best attention	F1-score	+0.0116	1.0000
No attention vs. Best attention	AUC	+0.0117	1.0000

To clarify whether the observed metric differences correspond to meaningful changes in individual predictions, Table 8 summarizes the paired image-level predictions between EfficientNetV2 and the proposed fusion model, while the corresponding correct/wrong prediction distribution is visualized as a contingency matrix in Fig. 4. The proposed model corrects 650 images that EfficientNetV2 misclassifies, whereas EfficientNetV2 correctly classifies 447 images that the proposed model misclassifies. Thus, the proposed model corrects 203 more baseline errors than it introduces. McNemar’s test applied to these discordant counts yields ($\chi^2 = 37.20$) with continuity correction and ($p = 1.07 \times 10^{-9}$), indicating that the image-level discordance between the two models is not symmetric.

This image analysis supports a more precise interpretation than a single mean accuracy value. The proposed fusion model does not simply improve every

case; rather, it shifts the decision boundary so that some EfficientNetV2 errors are corrected and some new errors are introduced. The net positive discordance suggests that fusion is useful for a subset of difficult images, while the introduced-error group should be examined carefully before clinical translation.

4.6 Explainability and Error-Case Discussion

To ensure a balanced explainability analysis, the interpretation of model explanations is associated with representative prediction cases rather than focusing solely on correctly classified examples. Representative Grad-CAM, LIME, and SHAP explanations for selected histopathology images are presented in Fig. 5. These complementary explanation methods provide different views of the image regions contributing to the model predictions: Grad-CAM highlights class-discriminative regions in the convolutional feature maps, LIME identifies influential local image regions through perturbation-based analysis, and SHAP provides attribution-based information regarding the contribution of image regions to the prediction. Together, these visualizations support qualitative examination of whether the model focuses on potentially relevant tissue structures or on non-diagnostic regions and artifacts.

The corrected cases are expected to include images where the fusion model captures complementary cues from tissue architecture and local texture that are less prominent in a single EfficientNetV2 representation. Conversely, introduced errors may occur when one stream emphasizes non-diagnostic regions, background, staining artifacts, or magnification-specific texture that is not class-discriminative. Persistently wrong cases are particularly important because they may contain ambiguous morphology, atypical staining, limited tumor tissue, or regions where benign and malignant patterns are visually similar. These cases should be prioritized for pathologist review and for future quantitative explanation stability analysis.

Table 8: Discordant-case summary between EfficientNetV2 and the proposed model.

Case group	No. of images
Both models are correct.	9003
Both models are wrong.	758
EfficientNetV2 is wrong, the proposed model is correct.	650
EfficientNetV2 is correct, the proposed model is wrong.	447

Overall, the results offer a cautious interpretation of the proposed method. The final VGG16-ResNet18 fusion model achieves the highest mean performance among the configurations tested. However, the improvement is modest and not statistically significant according to the fold-level Wilcoxon test. The primary evidence supporting the model includes a com-

bination of patient-level evaluation, backbone-pair ablation, image-level discordant-case analysis, and discussions guided by explainability regarding errors.

5. CONCLUSION

This study presents an explainable deep fusion framework for breast histopathology classification on the BreakHis dataset. All models were evaluated using patient-level 5-fold cross-validation under a mixed-magnification setting. The final VGG16–ResNet18 fusion model without attention achieved 0.8890 accuracy, 0.8888 F1-score, and 0.9326 AUC, outperforming the tested single-backbone baselines and other fusion pairs in this evaluation.

The results also show that attention modules do not improve the selected fusion pair under the patient-level protocol. Therefore, the final model is simpler than the attention-based variants evaluated in this study. This finding is important because it prevents overclaiming and indicates that the main benefit comes from combining complementary backbone representations rather than from attention. The fold-level Wilcoxon signed-rank test does not show a statistically significant improvement over EfficientNetV2, so the performance gain should be interpreted cautiously. However, the image-level discordant-case analysis shows that the proposed fusion model corrects more EfficientNetV2 errors than it introduces. This supports the usefulness of fusion for a subset of difficult images while also revealing remaining limitations. Misclassified and discordant cases are associated with ambiguous morphology, staining variation, magnification-dependent texture changes, background artifacts, and attention to non-diagnostic regions.

Future work will extend this evaluation in three directions. First, strict cross-magnification testing will be performed to measure generalization when training and testing magnifications differ. Second, external validation on additional breast histopathology datasets will be used to assess dataset transferability. Third, explanation quality will be evaluated more systematically using quantitative XAI stability metrics and expert-in-the-loop review.

ACKNOWLEDGEMENT

This research was supported by the University of Phayao and the Thailand Science Research and Innovation Fund (Fundamental Fund 2026, Grant No. 2372). This work was also supported by the School of Medicine under Grant No. MD68-03.

AUTHOR CONTRIBUTIONS

Conceptualization, Sitthichat Meekhwan, Buntueng Yana, Ratsada Praphasawat, and Kunaree Wongrach; methodology, Sitthichat Meekhwan, Buntueng

Yana, Ratsada Praphasawat, and Kunaree Wongrach; software, Sitthichat Meekhwan; data curation, Sitthichat Meekhwan; writing—original draft preparation, Sitthichat Meekhwan and Buntueng Yana; writing—review and editing, Sitthichat Meekhwan, Buntueng Yana, Ratsada Praphasawat, and Kunaree Wongrach; supervision, Buntueng Yana, Ratsada Praphasawat, and Kunaree Wongrach; funding acquisition, Ratsada Praphasawat. All authors have read and agreed to the published version of the manuscript.

References

- [1] K. Damkliang, T. Wongsirichot, C. Khongrak, and P. Suwannarat, “An optimal deep learning approach to BCA tissue detection using case studies,” *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, vol. 17, no. 1, pp. 60–72, Feb. 2023.
- [2] J. Kim *et al.*, “Global patterns and trends in breast cancer incidence and mortality across 185 countries,” *Nature Medicine*, vol. 31, no. 4, pp. 1154–1162, 2025.
- [3] F. Bray *et al.*, “Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, 2024.
- [4] S. K. Singh and K. S. Patnaik, “Patho-AI: A perceptive breast cancer identification and classification using deep learning methods integrated with explainable AI,” *SN Computer Science*, vol. 6, no. 6, Art. no. 619, 2025.
- [5] J. S. Ahn *et al.*, “Artificial intelligence in breast cancer diagnosis and personalized medicine,” *Journal of Breast Cancer*, vol. 26, no. 5, pp. 405–435, 2023.
- [6] A. Ibrahim, P. Gamble, R. Jaroensri, M. M. Abdelsamea, C. H. Mermel, P.-H. C. Chen and E. A. Rakha, “Artificial intelligence in digital breast pathology: Techniques and applications,” *Breast*, vol. 49, pp. 267–273, 2020.
- [7] C. McGenity *et al.*, “Artificial intelligence in digital pathology: A systematic review and meta-analysis of diagnostic test accuracy,” *npj Digital Medicine*, vol. 7, no. 1, Art. no. 114, 2024.
- [8] P. S. Ginter *et al.*, “Histologic grading of breast carcinoma: A multi-institution study of interobserver variation using virtual microscopy,” *Modern Pathology*, vol. 34, no. 4, pp. 701–709, 2021.
- [9] S. Datwani, H. Khan, M. K. K. Niazi, A. V. Parwani, and Z. Li, “Artificial intelligence in breast pathology: Overview and recent updates,” *Human Pathology*, vol. 162, Art. no. 105819, 2025.
- [10] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

- [11] R. Shad, J. P. Cunningham, E. A. Ashley, C. P. Langlotz and W. Hiesinger, "Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging," *Nature Machine Intelligence*, vol. 3, no. 11, pp. 929–935, 2021.
- [12] M. Plass *et al.*, "Explainability and causability in digital pathology," *The Journal of Pathology: Clinical Research*, vol. 9, no. 4, pp. 251–260, 2023.
- [13] N. Bienefeld *et al.*, "Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals," *npj Digital Medicine*, vol. 6, no. 1, Art. no. 94, 2023.
- [14] A. Nicolson, E. Bradburn, Y. Gal, A. T. Papa-georgiou and J. A. Noble, "The human factor in explainable artificial intelligence: Clinician variability in trust, reliance, and performance," *npj Digital Medicine*, vol. 8, no. 1, Art. no. 658, 2025.
- [15] F. A. Spanhol, L. S. Oliveira, C. Petitjean and L. Heutte, "A Dataset for Breast Cancer Histopathological Image Classification," in *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, July 2016.
- [16] N. Bayramoglu, J. Kannala and J. Heikkilä, "Deep learning for magnification independent breast cancer histopathology image classification," *2016 23rd International Conference on Pattern Recognition (ICPR)*, Cancun, Mexico, pp. 2440–2445, 2016.
- [17] S. Alkassar, B. A. Jebur, M. A. M. Abdullah, J. H. Al-Khalidy and J. A. Chambers, "Going deeper: Magnification-invariant approach for breast cancer classification using histopathological images," *IET Computer Vision*, vol. 15, no. 2, pp. 151–164, 2021.
- [18] S. Taheri *et al.*, "A comprehensive study on classification of breast cancer histopathological images: Binary versus multi-category and magnification-specific versus magnification independent," in *IEEE Access*, vol. 12, pp. 50431–50443, 2024.
- [19] G. Aresta *et al.*, "BACH: Grand challenge on breast cancer histology images," *Medical Image Analysis*, vol. 56, pp. 122–139, 2019.
- [20] P. Bándi *et al.*, "From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 550–560, 2019.
- [21] M. Veta *et al.*, "Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge," *Medical Image Analysis*, vol. 54, pp. 111–121, 2019.
- [22] D. Bardou, K. Zhang and S. M. Ahmad, "Classification of breast cancer based on histology images using convolutional neural networks," in *IEEE Access*, vol. 6, pp. 24680–24693, 2018.
- [23] M. Gour, S. Jain and T. Sunil Kumar, "Residual learning based CNN for breast cancer histopathological image classification," *International Journal of Imaging Systems and Technology*, vol. 30, no. 3, pp. 621–635, 2020.
- [24] Y. Zheng *et al.*, "Application of transfer learning and ensemble learning in image-level classification for breast histopathology," *Intelligent Medicine*, vol. 3, no. 2, pp. 115–128, 2023.
- [25] M. Azmoodeh-Kalati, H. Shabani, M. S. Maghareh, Z. Barzegar and R. Lashgari, "Leveraging an ensemble of EfficientNetV1 and EfficientNetV2 models for classification and interpretation of breast cancer histopathology images," *Scientific Reports*, vol. 15, no. 1, Art. no. 21541, 2025.
- [26] S. Chattopadhyay, A. Dey, P. K. Singh, D. Oliva, E. Cuevas and R. Sarkar, "MTRRE-Net: A deep learning model for detection of breast cancer from histopathological images," *Computers in Biology and Medicine*, vol. 150, Art. no. 106155, 2022.
- [27] M. M. Rahman, M. S. I. Khan and H. M. H. Babu, "BreastMultiNet: A multi-scale feature fusion method using deep neural network to detect breast cancer," *Array*, vol. 16, Art. no. 100256, 2022.
- [28] M. I. Rahman, "Fusion of vision transformer and convolutional neural network for explainable and efficient histopathological image classification in cyber-physical healthcare systems," *Journal of Transformative Technologies and Sustainable Development Publishing model*, vol. 9, no. 1, Art. no. 8, 2025.
- [29] S. Boumaraf, X. Liu, Z. Zheng, X. Ma, and C. Ferkous, "A new transfer learning based approach to magnification dependent and independent classification of breast cancer in histopathological images," *Biomedical Signal Processing and Control*, vol. 63, Art. no. 102192, 2021.
- [30] V. Kumari and R. Ghosh, "A magnification-independent method for breast cancer classification using transfer learning," *Healthcare Analytics*, vol. 3, Art. no. 100207, 2023.
- [31] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-Excitation Networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 1 Aug. 2020.
- [32] S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, "CBAM: Convolutional block attention module," in *Computer Vision—ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, pp. 3–19, 2018.
- [33] X. Wang, R. Girshick, A. Gupta and K. He, "Non-local Neural Networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern*

Recognition, Salt Lake City, UT, USA, pp. 7794–7803, 2018.

- [34] R. Maurya, N. N. Pandey, M. K. Dutta and M. Karnati, “FCCS-Net: Breast cancer classification using multi-level fully convolutional channel and spatial attention-based transfer learning approach,” *Biomedical Signal Processing and Control*, vol. 94, Art. no. 106258, 2024.
- [35] J. Peta and S. Koppu, “Explainable soft attentive EfficientNet for breast cancer classification in histopathological images,” *Biomedical Signal Processing and Control*, vol. 90, Art. no. 105828, 2024.
- [36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 618–626, 2017.
- [37] M. T. Ribeiro, S. Singh and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, pp. 1135–1144, 2016.
- [38] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 4765–4774, 2017.



Sitthichat Meekhwan received the B.Eng. degree in electrical engineering from the University of Phayao, Phayao, Thailand, in 2019. He is currently pursuing a master’s degree at the Faculty of Engineering, while also working in the energy sector specializing in Industrial IoT (IIoT), web development, and AI. His research interests encompass machine learning, deep learning, robotics, LoRa-based IoT networks, and IC design.

sign.



Buntueng Yana is a Lecturer in the Department of Electrical Engineering, Faculty of Engineering, University of Phayao, Thailand, where he has served since 2009. He received his PhD in Information Science and Technology from Osaka University, Japan (2019), and his M.Eng. and B.Eng. in Electrical Engineering from Chiang Mai University, Thailand (2009 and 2006). His expertise includes electronic circuit and embedded-system design, digital communications, wireless networking, artificial intelligence, data analytics, and satellite image processing. His research focuses on robotics, machine learning, and automatic control systems.



Ratsada Praphasawat received the B.Sc. degree in Microbiology from Naresuan University, Thailand, in 2010 and the M.Sc. degree in Pathobiology from Mahidol University, Thailand, in 2013. She received her Ph.D. degree in Medical Science from Oita University, Japan, in 2020. She is currently an Associate Professor at the School of Medicine, University of Phayao, Thailand. Her research interests focus on cancer biology, the anti-cancer mechanisms of natural products, and the application of machine learning in medical diagnostics.



Kunaree Wongrach received B.Sc. degree in Physics from Chiang Mai University, Thailand, in 2007, and M.Sc. and Dr.rer.nat. in Physics from Heinrich-Heine University Duesseldorf, Germany, in 2011 and 2015, respectively. She is currently a lecturer at the Department of Physics, Burapha University, Thailand, where she began her career. Her research interests include fusion technology, renewable energy, and the application of machine learning.

the application of machine learning.