Article

Enhancing AI-Driven Diagnosis of Invasive Ductal Carcinoma with Morphologically Guided and Interpretable Deep Learning

Suphakon Jarujunawong and Paramate Horkaew * (ID)

School of Computer Engineering, Institute of Engineering, Suranaree University of Technology,
Nakhon Ratchasima 30000, Thailand; suphakon.jarujunawong@gmail.com
* Correspondence: phorkaew@sut.ac.th

Abstract: Artificial intelligence is increasingly shaping the landscape of computer-aided diagnosis of breast cancer. Despite incrementally improved accuracy, pathologist supervision remains essential for verified interpretation. While prior research focused on devising deep model architecture, this study examines the pivotal role of multi-band visual-enhanced features in invasive ductal carcinoma classification using whole slide imaging. Our results showed that orientation invariant filters achieved an accuracy of 0.8125, F1-score of 0.8134, and AUC of 0.8761, while preserving cellular arrangement and tissue morphology. By utilizing spatial relationships across varying extents, the proposed fusion strategy aligns with pathological interpretation principles. While integrating Gabor wavelet responses into ResNet-50 enhanced feature association, the comparative analysis emphasized the benefits of weighted morphological fusion, further strengthening diagnostic performance. These insights underscore the crucial role of informative filters in advancing DL schemes for breast cancer screening. Future research incorporating diverse, multi-center datasets could further validate the approach and broaden its diagnostic applications.

Keywords: computer-aided diagnosis; deep learning; explainable AI; Gabor wavelet; whole slide imaging

1. Introduction

According to GLOBOCAN 2022, nearly 20 million new cancer cases were diagnosed, and 9.7 million fatalities were reported [1]. The article also estimated that 20% of people develop cancer in their lifetime. Lung and female breast cancers were among the most frequently reported cases, accounting for 12.4% and 11.6% of all cancer patients that year, respectively. If the current trend continues, an estimated 35 million cases will be identified by 2050. Cao et al. [2] analyzed the cancer profiles based on the reports published during 2020 to 2022. They observed a notable increase in thyroid cancers, whilst cases and fatality rates of both stomach and esophageal cancers showed a decline. Although cancer remains a leading health concern, advancements in treatment and prevention efforts play a crucial role in reducing its impacts. However, their Mortality-to-Incidence Ratio (MIR) vary notably across countries with differing socioeconomic statuses.

The surveys highlight that breast cancer (BCa) is a most fatal type, severely impacting women's health worldwide. Invasive ductal carcinoma (IDC) is the most common type of BCa, accounting for about 80% of all breast cancers [3,4]. IDC both affects patients' quality of life and contributes to higher MIR. Once IDC originated, its extensive genetic and molecular diversity facilitates rapid spread into adjacent breast tissues [5]. The intra-tumor heterogeneity observed in IDC complicates disease progression, posing challenges

for therapeutic strategies and patient outcomes. Current research prioritizes the development of technologies for diagnostics and interventions. Effective interdisciplinary studies, such as genomic profiling, computerized subtype prediction, and immunomodulatory intervention [4–7], could drive management strategies and enhance treatment responses.

In addition to the above interdisciplinary efforts, artificial intelligence (AI) has also recently played a pivotal role in computer-aided diagnosis (CAD) [8–10]. Computer vision for modern medicine has been driven by Machine Learning (ML) and Deep Learning (DL) models, enabling precise identification and characterization of biomarkers and pathological features. A convolutional neural network (CNN) is an Artificial Neural Network (ANN) that has been widely adopted in recent CAD research, due to their accessible and versatile software libraries [11,12]. Prominent CNN architectures, e.g., DenseNet [13], VGG [14], and ResNet [15], have proven effective in identifying abnormality and disease classification in medical images. By extracting deep features and associating them with known pathological patterns, these decision-support models can assist physicians in reducing diagnostic errors and optimizing treatment outcomes [16–18]. Building on these advancements, AI in digital health is set to address socioeconomic gaps that continue to hinder cancer patient management [2].

Whole slide imaging (WSI) is recognized as the current gold standard for screening and serves as crucial component in digital pathology. It digitizes glass tissue slides at high resolution, preserving rich diagnostic information, enabling comprehensive coverage and detailed interpretation of pathological images [19]. When utilized in CAD, WSI is often divided into smaller patches for detailed analysis [4,20,21]. Although this modality has demonstrated significant value in cancer screening, its exceptionally large data size (e.g., $100 \text{ K} \times 100 \text{ K}$ pixels, corresponding to $0.25 \times 0.25 \text{ } \mu\text{m}^2$) presents challenges, requiring extensive archiving capability and considerable computational resources. In an AI-based system [22,23], even problems involving moderately sized patches require hours of training and demand several gigabytes of Graphical Processing Unit (GPU) memory. Depending on the model complexity and the dataset being evaluated, accuracy ranges from about 75 to 95%. Therefore, care must be taken when deploying AI models in practical clinical settings to ensure patient safety. This process involves rigorous validation and calibration against established diagnostic standards and ensuring compliance with regulatory guidelines [24]. Importantly, AI models should be utilized solely as assistive tools, with their 'decisions' thoroughly reviewed under pathologist supervision.

Due to the abovementioned variability in accuracy and requirement for clinical supervision, this study primarily emphasizes the importance of enhanced multi-band visual features in improving pattern association performance. The remaining sections of this paper are organized as follows: Section 2 provides an overview of previous research in the field of AI-driven cancer screening utilizing WSI. Section 3 describes the data and the proposed methodology. Sections 4 and 5 present the experimental results and discussion, respectively. Section 6 gives the concluding remarks, as well as the current limitations and future work.

2. Related Works

A key limitation in WSI analysis is inconsistencies present within acquired data [25]. Without appropriate image correction, visual inspection of digital pathology is prone to significant errors. Similarly, this variability can compromise the accuracy of computerized classification models. Hematoxylin and Eosin (H&E) staining is a standard method used to highlight cellular and tissue structures in WSI. However, variations in H&E staining protocol across laboratories affect the consistency and reliability of the results [26]. The effect becomes particularly pronounced when a data-dependent AI-model is applied to

diversely varied images. To address this issue, Murtaza et al. [25] preprocessed the WSI using stain normalization by color transfer [27], prior to geometrical augmentation. An ensemble tree-based DL model was developed to classify different pathologies of breast tumors (BrT). The model weights were transferred from an AlexNet [28], with newly trained Fully Connected (FC) layers. While very high accuracy was achieved, its main limitation was reliance on several hierarchical models for indirect classification.

Another subsequent study examined the advantage of image pre-processing on DL-based classification of breast cancer images [4]. In that study, histopathological images were processed using various filters, e.g., details, contour, and edge enhancements, along with histogram-based equalization. The classification process used VGG-16, VGG-19, MobileNet, ResNet-50, and DenseNet-121, with VGG-16 and VGG-19 exhibiting superior performance. However, an Analysis of Variance (ANOVA) revealed that the choice of filters was not statistically significant. In other words, with an optimal model architecture, determining features can be learned effectively, even without trivial enhancement. In fact, the process typically worked in reverse. CNN was utilized to enhance image quality, such as via fusion [29]. This suggests that salient features can be discovered and enhanced concurrently, as the learning progress.

The Explainable AI (XAI) paradigm has attracted considerable attention from the CAD research community [30,31]. In a recent study, an ensemble-based transfer learning approach was taken to rate dementia levels in Alzheimer's Disease (AD) [18]. The framework employed pretrained VGG-16, VGG-19, DenseNet-169, and DenseNet-201 for analyzing Magnetic Resonance Images (MRI). The classifications were ensembled to improve accuracy. The model was equipped with saliency and gradient weighted Class Activation Mapping (MAP) to highlight the neural regions influencing the decisions. The maps intuitively captured the pathological Regions of Interest (ROIs).

The main conclusions of Murcia-Gómez et al.'s work [4] raised an important question: What pre-processing strategies can then improve the learning mechanism, particularly for histopathological cancer image classification? Despite focusing on a different pathology, Mahmud et al. [18] also presented results that led us to hypothesize that enhancing local texture appearance may effectively improve model performance, without relying on pixel-wise treatments [4]. This approach was explored across scientific disciplines. Contextual heuristic edge discovery and visual texture analysis were applied in barley grain grading using CNN [32]. Following grain ROI segmentation, a selected Gabor wavelet sub-band served as the input tensor. The accuracy surpassed baseline performance by over 10%, compared to ResNet-50, DenseNet-121, and VGGNet-19. A similar strategy has been used in contemporary research across multiple fields [33–36]. However, their protocols appear to be restricted to isolated objects, making them unsuitable for WSI.

Motivated by multi-focal image fusion proposed by Liu et al. [29], and deep learning of explicit textual features proposed by Embeyale et al. [32] and others, this paper examines the benefits of multi-band feature fusion in enhancing deep learning outcomes. Based on our survey [16,17,21], ResNet-50 has consistently demonstrated superior performance in histopathological studies. It was employed as the primary architecture. The ResNet-50 model with optimal configuration will be evaluated alongside other leading DL models, i.e., VGG-16, VGG-19, and DenseNet-121.

3. Materials and Methods

The following subsections describe the involved dataset, data management, multi-band image fusion, and model architectures and their development, hyperparameter settings, and evaluations.

3.1. Dataset

To ensure reproducibility and encourage research extension, this study utilized a BCa histopathology image collection acquired from Kaggle [37,38]. The updated dataset comprised 162 WSI specimens scanned at $40\times$ magnification. It contains 78,911 and 198,637 images labeled as normal (IDC$-$) and pathological (IDC+) classes, respectively. Each image was stored in Portable Network Graphic (PNG) format with a resolution of 50×50 pixels. Information such as class, patient identification (ID), and cropped coordinates were encoded in the file name.

3.2. Data Management and Data Folds

The dataset contained 277,548 images in total, with class imbalance of approximately 1:2.5. These files were organized into 278 patient-specific folders, each containing a different number of images from either class. To achieve a balanced distribution, up to 32 images were randomly selected per patient per class. Consequently, a total of 17,792 images were expected, i.e., 32 (images) $\times$ 278 (patients) $\times$ 2 (classes). There were 3 and 16 patients with fewer than 32 images in the IDC$-$ (21.00 ± 7.00) and IDC+ (25.06 ± 6.06) classes, respectively. Therefore, only 17,648 images were available for analysis, consisting of 8863 IDC$-$ and 8785 IDC+ images.

Of the 17,648 images, they were split into training, validation, and testing folds at a 64:16:20 ratio, ensuring equal representation of both classes. No synthetic data was generated to balance the dataset.

3.3. Multi-Band Visual Feature Extraction and Fusion

As shown in the literature, while neither convolutional filters nor histogram-based enhancements have significantly improved model accuracy, Gabor wavelet has demonstrated promising potential. In this study, the Gabor wavelet was extended to images containing non-isolated objects. Instead of applying the wavelet to a pre-segmented ROIs [32], it was fused with the original image using simple weighted linear combinations.

3.3.1. Gabor Wavelets

Rather than characterizing geometrical objects, a pathologist usually analyzes a slide by interpreting histological patterns, cellular arrangement, and tissue morphology to diagnose abnormalities. Even a small patch of histopathologic image exhibits very complex structural details and textural patterns. These properties can be captured by Gabor responses [32,39], defined in Equation (1).

$$G(x, y; \sigma, \theta, \phi) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(i\left(2\pi\phi x' + \psi\right)\right) \tag{1}$$

where σ is the standard deviation of the Gaussian envelope, capturing the scale; θ is the orientation of the envelope, capturing the direction; ϕ is the frequency of the sinusoid, capturing the spatial variation; ψ is the phase offset; and γ is the spatial aspect ratio. However, ψ was set to 0 and γ was set to 1, as they contribute less to recognition. Additionally, ϕ depended on σ, ensuring that the envelope contained only one octave of the wave. Finally, the local coordinate conversion terms are defined by $x' = x\cos(\theta) + y\sin(\theta)$ and $y' = -x\sin(\theta) + y\cos(\theta)$. A kernel ($G$) was prescribed by scale (σ) and orientation (θ).

3.3.2. Orientation Invariant Features

In traditional ML-based strategies, image responses to Gabor wavelets with varying kernel parameters are typically stacked into a Gabor jet [40]. Their statistical descriptors were then extracted to form feature vectors. Alternatively, real-value Gabor jets were used

in the first layer of a CNN [41]. These approaches were either applied to objects with fixed poses, neglecting scale and orientation variabilities, or relied on highly complex neurons to implicitly learn such variations.

This study analyzed histopathological images, digitized from glass slides at a known magnification. Stained tissue samples with similar textural structures but differing in poses can be represented by totally different responses. Instead of extracting individual wavelets, this study used orientation invariant Gabor responses, i.e., max and mean Gabors, as defined in Equations (2a) and (2b), respectively.

$$r_{max}(\mathbf{p}; \sigma) = \max_{\theta}(G_{\sigma, \theta} * I)(\mathbf{p}) \tag{2a}$$

$$r_{mean}(\mathbf{p}; \sigma) = \underset{\theta}{mean}(G_{\sigma, \theta} * I)(\mathbf{p}) \tag{2b}$$

An image patch (I) was split into RGB planes. Each plane was processed using Gabor kernels at a specified scale (σ), with orientations set at $0°$, $45°$, $90°$, and $135°$. The maximum response was derived from impulse responses, whose intensity I at a coinciding coordinate $\mathbf{p} = [x \ y]^{\mathrm{T}}$ represented the highest value across the entire spectra. Similarly, the mean response was obtained by averaging intensities across this range. The resulting planes (r_{R}, r_{G}, r_{B}) were then recomposed (R) with respect to the corresponding response type.

Figure 1A illustrates the responses of a single channel to filters with four distinct orientations, while Figure 1B illustrates the true-color responses to maximum and mean orientation invariant filters.
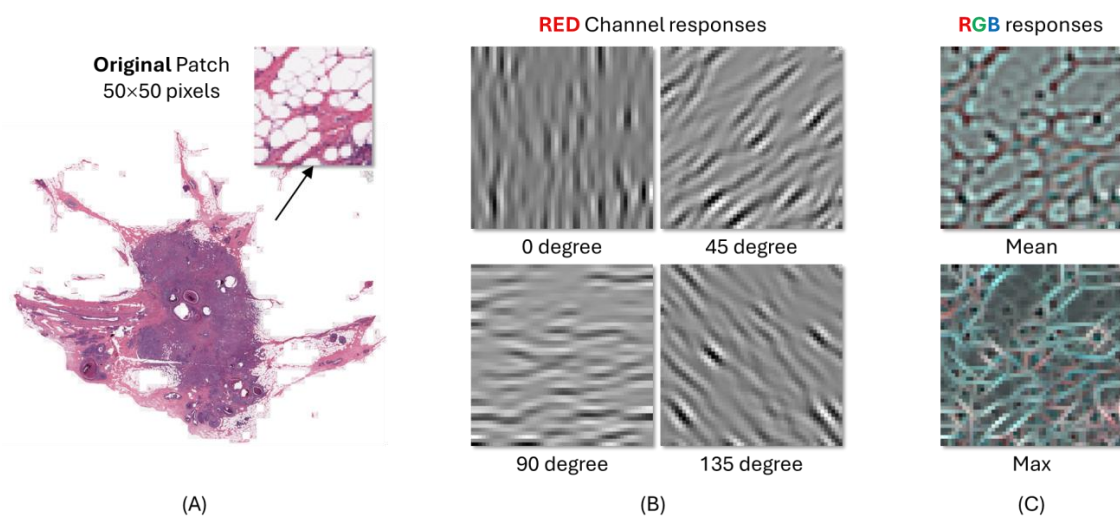
Figure 1. WSI involved in this study, showing an original image patch of size 50×50 pixels (**A**), a set of Gabor responses with $\lambda = 4$ at $\theta = 0, 45, 90$, and 135 degrees (**B**), and the corresponding orientation invariant response R_{Mean} and R_{Max} (**C**).

3.3.3. Linear Fusion

Gabor responses effectively capture spatial relationships in textural patterns, particularly in defining boundaries and their extensions [42]. This, in turn, enhances structural arrangement and morphology. However, sinusoidal convolution reduces the prominence of localized chromatic distribution within cells. To address this issue, the input tensor (J) was defined by a linear combination of the original image I and its Gabor response with a specific scale and orientation, as expressed by Equation (3).

$$J(x, y) = \alpha I(x, y) + (1 - \alpha)R(x, y) \tag{3}$$

where I and R are the original image and its response, respectively; $\alpha \in [0, 1]$ is the fusion weight.

3.4. Deep Learning Models

In the experiments, ResNet-50 was chosen as the primary model. It was compared against the VGG-16, VGG-19, and DenseNet-121 models. Their detailed descriptions are provided as follows.

3.4.1. Model Architectures

VGG utilizes 3×3 convolutional kernels in each layer [13]. It relies on these stacked small filters to effectively capture intricate image details. Increasing the model's depth improves the accuracy of image recognition [4]. In this study, VGG-16 and VGG-19 were implemented. The VGG-16 architecture is depicted in Figure 2A.
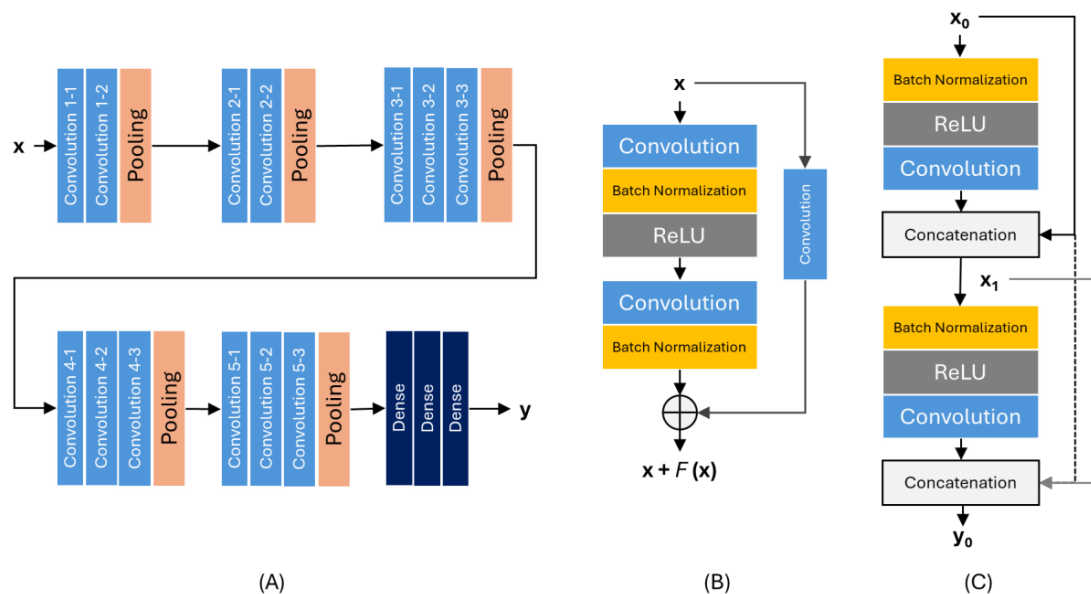
Figure 2. The utilized DL models, i.e., a full VGG network (VGG-16) (**A**), a residue module of ResNet (**B**), and a 2-layer example of DenseNet (**C**).

ResNet was proposed to address the diminishing gradient issue in a deep CNN [15]. As its depth increases, gradients in certain layers of the model become weaker during its updates. Residual learning was introduced by computing the difference between the input value (x) and its function (F(x)). Its main component is depicted in Figure 2B. These shortcut connections allow layers to be bypassed, maintaining efficiency without adding more complexity. In this study ResNet-50 was implemented.

As the name suggests, DenseNet emphasizes dense connections, involving inputs not only from the immediate previous layers but from all preceding ones [14]. Figure 2C illustrates the concept of dense connections within a shallow DenseNet. These connections allow cascading information and its gradient from the initial layer to the deepest one, ensuring seamless learning propagation across layers. Unlike ResNet, it addresses the vanishing gradient issue by reusing features from preceding layers. In this study, DenseNet-121 was implemented.

3.4.2. Hyperparameter Settings

Both Binary Cross-Entropy (BCE) and Categorical Cross-Entropy (CCE) were considered as loss functions, depending on the neural model, to measure differences between the actual and predicted classes. Batch size refers to the number of samples processed in a single pass of the learning process. A smaller batch size of 32 was specified to reduce overfitting [43,44]. The learning rate (LR), which controls convergence of gradient updates,

was varied between 1.0×10^{-7} and 1.0×10^{-5}. A dropout rate of 50% was applied to promote generalization by randomly deactivating half of the neurons during training.

3.4.3. Hardware Environment

The experiments involving data preparation, image processing, and deep model training were conducted on a notebook Personal Computer (PC) running Windows™ 11 Operating System (OS). The PC was equipped with Intel™ Core i9-13900HX Central Processing Unit (CPU), NVIDIA RTX4080 GPU, and 32 GB RAM. The algorithms were implemented in Python version 3.11.2, using Visual Studio (VS) Code version 1.101 as the development environment.

3.5. Evaluation Metrics

In the experiments, various networks with differing configurations will be evaluated and compared using standard metrics, namely accuracy, precision, recall, specificity, and F1-score.

For a given patch, a model classifies it as either invasive ductal carcinoma (IDC+) or normal tissue (IDC−), representing the positive (P) and negative (N) classes, respectively. True Positive (TP) and True Negative (TN) refer to cases where the model correctly classified a sample as IDC+ and IDC−, respectively. False Positive (FP) and False Negative (FN) occur when the model misclassified an IDC− sample as IDC+, and vice versa. The counts of these occurrences in the test set were utilized to construct a confusion matrix and calculate the evaluation metrics, as outlined in Equations (4a)–(4e). Accuracy measures the ratio of correctly classified samples relative to the total cases.

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4a}$$

Focusing on correct IDC+ classification, precision and recall measure the ratio of correctly classified positive samples to the total positive samples and to the predicted positive samples, respectively. In contrast, specificity measures the ratio of correctly classified IDC− samples relative to the total number of actual samples in this class.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4b}$$

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4c}$$

$$Specificity = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4d}$$

The F1-score is calculated as the harmonic mean of precision and recall, offering a balanced metric that evaluates their trade-off.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4e}$$

In addition, Receiver Operating Characteristic (ROC) curve plots the model's performance across various classification thresholds. It shows the trade-off between TP rate (recall) and FP rate (1—precision). The Area Under Curve (AUC) is derived from ROC. AUC values closer to 1 signify that the model can reliably differentiate between IDC+ and IDC− classes, whereas those near 0.5 reflect performance comparable to random guessing. Values below 0.5 suggest systematically incorrect classification or potential label reversal, indicating the model's failure to learn any meaningful pattern.

4. Results

4.1. Optimal Learning Rate

The learning rate is an important hyperparameter in DL. It controls how much neural weights are adjusted during training. A higher RL accelerates convergence but may lead to oscillations, preventing the validation metrics from stabilizing. Conversely, a lower LR improves convergence stability but increases the risk of the model becoming trapped in local minima. This experiment identified the optimal RL, based not only on final prediction accuracy but also on an analysis of convergence characteristics. Figure 3A plots the accuracy of ResNet-50 trained with original BCa histopathology images (I), with initial weights transferred from ImageNet, versus the LR $\in$ [0.0001, 0.0000001]. While the RL of 1.0×10^{-4} gave the most accurate predictions, the corresponding loss values exhibited significant oscillations, as shown in Figure 3B. The RL of 5.0×10^{-6} gave the best compromise, giving relatively high accuracy, while maintaining more stable convergence, as illustrated in Figure 3C. The LR of 5.0×10^{-6} was therefore selected for subsequent experiments.
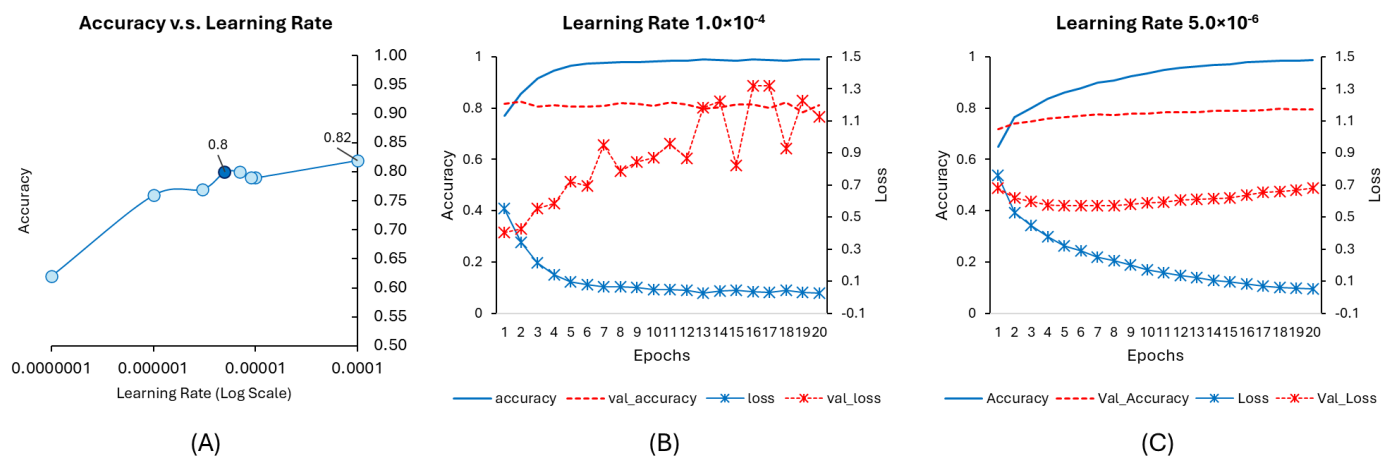
Figure 3. The impact of the learning rate (LR) on model performance, showing the accuracy at varying RL (**A**) and the convergence behaviors with LR = 1.0×10^{-4} (**B**) and 5.0×10^{-6} (**C**).

4.2. Additional Dense Layer

A typical ResNet-50 consists of a bottleneck block consisting of 2048 feature maps. In the proposed architecture, a dense layer with 1024 neurons was inserted before the SoftMax layer. Its role was to capture complex features, while generalizing abstract representation present in the images. Similar experiments were conducted to find the optimal dropout ratios within the range of 10% to 50%. Both original images and their orientation invariant Gabor responses were analyzed. It was found that a dropout ratio of 50% consistently yielded the most accurate predictions, regardless of response types.

4.3. Response Fusion Types and Loss Functions

The next experiment aims to determine the optimal combinations of the original image (I) with the maximum (R_{Max}) and mean (R_{Mean}) responses. The Gabor scale (λ) of 4 pixels was specified. Figure 4A presents the comparison of their accuracies, using the modified ResNet-50, at different fusion weights (α). It was found that the combination with R_{Mean} consistently outperformed its counterpart. Therefore, it was used in the subsequent experiments. Similarly, the appropriate loss function was determined. Figure 4B compares the accuracies using Categorical Cross-Entropy (CCE) and Binary Cross-Entropy (BCE) at varying weights (α).
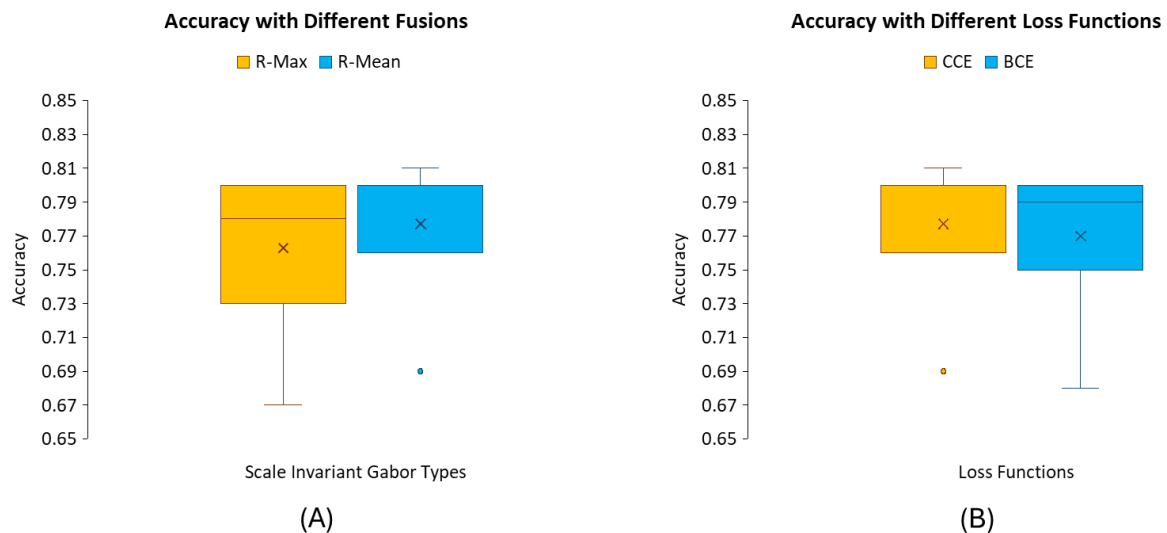
Figure 4. The Box-Wisker plots of the accuracy with different fusion strategies (**A**) and with different loss functions (**B**), at varying weights (α).

4.4. Optimal Weights and Scales

In ML, features extracted from responses across possible scale ranges were typically analyzed. However, adding more feature planes would make a DL model prohibitively complex. To address this limitation, a single most appropriate scale was identified. Figure 5A,B compares the accuracies with scale α set to 4, 8, and 16, using CCE and BCE, respectively. The performance varies depending on the configuration and loss function. For CCE, the optimal setup was $\lambda = 4$ and $\alpha = 0.90$, yielding the best results. In contrast, for BCE, the highest performance was achieved with $\lambda = 16$ and $\alpha = 0.90$. As a result, these models will be thoroughly analyzed and benchmarked against state-of-the-art models to provide a detailed performance comparison.
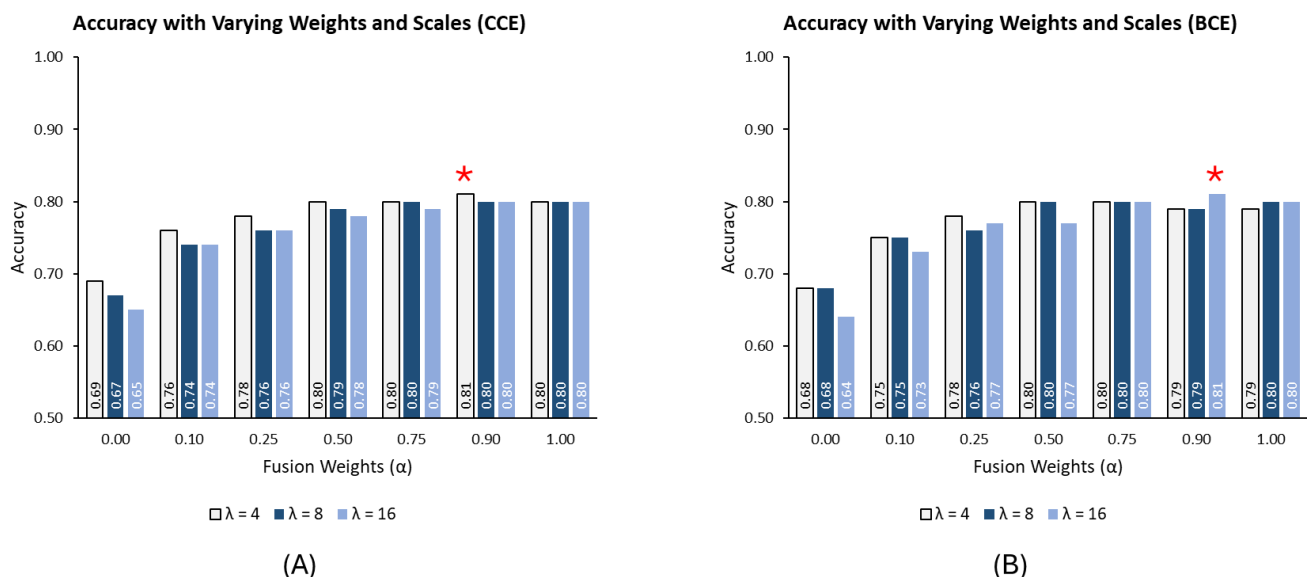

Figure 5. The accuracy achieved by models configured different weights (α) and scales (λ), based on CCE (**A**) and BCE (**B**) loss functions. An asterisk marks the best value within each configuration.

4.5. Model Analysis and Benchmarking

For clarity, the model configured with $\lambda = 4$ and $\alpha = 0.90$, using CCE, as the loss function is referred to as Model A, while the model with $\lambda = 16$ and $\alpha = 0.90$, using BCE, is

designated as Model B. The confusion matrixes given by Models A and B are depicted in Figures 6A and 6B, respectively. Each matrix is utilized to evaluate the performance of the respective model by summarizing its predictions against the actual classes. In the following reports, both models were evaluated across four independent runs, and their performance measures were averaged to ensure reliable interpretations.
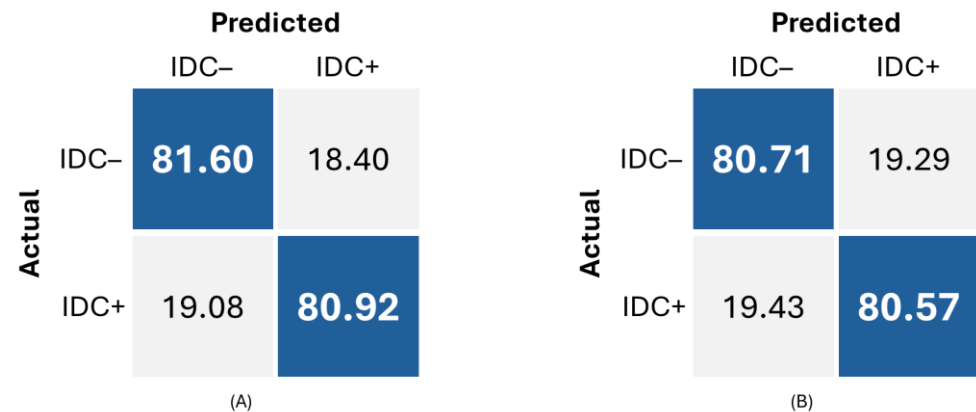
Figure 6. Confusion matrices achieved by Model A ($\lambda = 4$, $\alpha = 0.9$, CCE) (**A**) and Model B ($\lambda = 16$, $\alpha = 0.9$, BCE) (**B**). Numbers are expressed in percentages relative to each respective actual class.

Deeper insights into the model effectiveness in distinguishing between IDC classes are shown in Figure 7A. The graph compares the accuracy, precision, recall, specificity, and F1-score between both models. Their overall accuracies were comparable to two decimal places. However, Model A demonstrated consistently higher on all metrics, especially precision. The ROCs in Figure 7B show that the AUC of Model A was marginally higher than Model B (0.8761 vs. 0.8716). While F1-score and AUC may exhibit similar trends for both models, these metrics assess different aspects of performance. Therefore, care should be taken when interpreting model effectiveness.
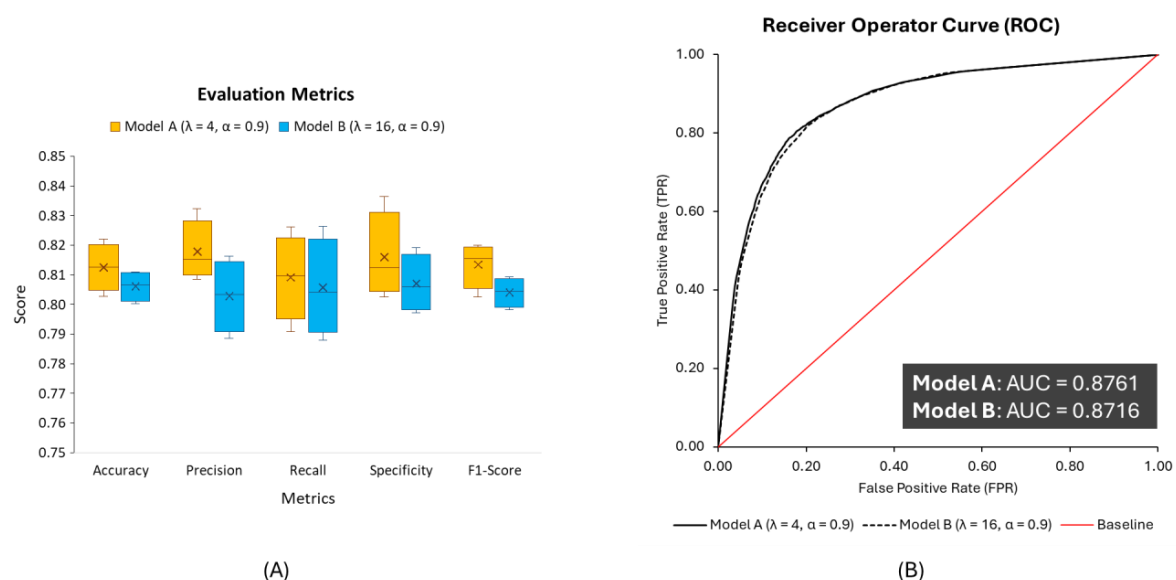
Figure 7. Comparisons of evaluation metrics distributions (**A**) and ROC curve (**B**) between Model A ($\lambda = 4$, $\alpha = 0.9$, CCE) and Model B ($\lambda = 16$, $\alpha = 0.9$, BCE).

Finally, the individual performance metrics of Model A and Model B were then compared with state-of-the-art models, with the same scale (λ), weights (α), and loss functions, and are listed in Table 1 and Table 2, respectively.

Table 1. A comparison of the performance metrics obtained by Model A.

| | Model A: $\lambda = 4$ and $\alpha = 0.90$ (CCE) | | | | |
	VGG-16	VGG-19	ResNet-50	DenseNet-121	p [1]
Accuracy	0.7890 ± 0.0051	0.7896 ± 0.0087	$\mathbf{0.8125 \pm 0.0069}$	0.7921 ± 0.0059	0.003 ***
Precision	0.7938 ± 0.0142	0.8063 ± 0.0069	$\mathbf{0.8179 \pm 0.0089}$	0.7921 ± 0.0110	0.038 **
Recall	0.7770 ± 0.0236	0.7588 ± 0.0335	$\mathbf{0.8092 \pm 0.0126}$	0.7919 ± 0.0112	0.084 *
Specificity	0.8011 ± 0.0218	$\mathbf{0.8198 \pm 0.0171}$	0.8160 ± 0.0126	0.7923 ± 0.0152	0.219
F1-Score	0.7849 ± 0.0058	0.7813 ± 0.0151	$\mathbf{0.8134 \pm 0.0066}$	0.7919 ± 0.0062	0.005 **

[1] *** $p < 0.005$, ** $p < 0.05$, and * $p < 0.1$. Numbers in bold indicate the best-performing models for each measure.

Table 2. A comparison of the performance metrics obtained by Model B.

| | Model B: $\lambda = 16$ and $\alpha = 0.90$ (BCE) | | | | |
	VGG-16	VGG-19	ResNet-50	DenseNet-121	p [1]
Accuracy	0.7882 ± 0.0080	0.7931 ± 0.0065	$\mathbf{0.8062 \pm 0.0045}$	0.7964 ± 0.0049	0.023 **
Precision	0.7837 ± 0.0075	0.7975 ± 0.0137	$\mathbf{0.8029 \pm 0.0107}$	0.7997 ± 0.0110	0.201
Recall	0.7894 ± 0.0209	0.7872 ± 0.0357	$\mathbf{0.8057 \pm 0.0142}$	0.7940 ± 0.0051	0.740
Specificity	0.7871 ± 0.0079	0.7988 ± 0.0253	$\mathbf{0.8071 \pm 0.0085}$	0.7989 ± 0.0134	0.492
F1-Score	0.7863 ± 0.0112	0.7915 ± 0.0131	$\mathbf{0.8041 \pm 0.0043}$	0.7968 ± 0.0035	0.154

[1] ** $p < 0.05$. Numbers in bold indicate the best-performing models for each measure.

The proposed ResNet-50 model outperformed other architectures across all metrics, except for the specificity in Model A. In Model A, accuracy, precision, and F1-score showed statistically significant differences, while recall was notable but less significant, and specificity was not significant. In Model B, although ResNet-50 maintained the best overall performance, significance levels were generally lower compared to Model A, with only accuracy reaching statistical significance.

The CAM of selected examples, one for IDC− and two IDC+ cases, obtained by Models A and B are depicted in Figures 8A and 8B, respectively. They visualize where each model focuses on a WSI when making their prediction.
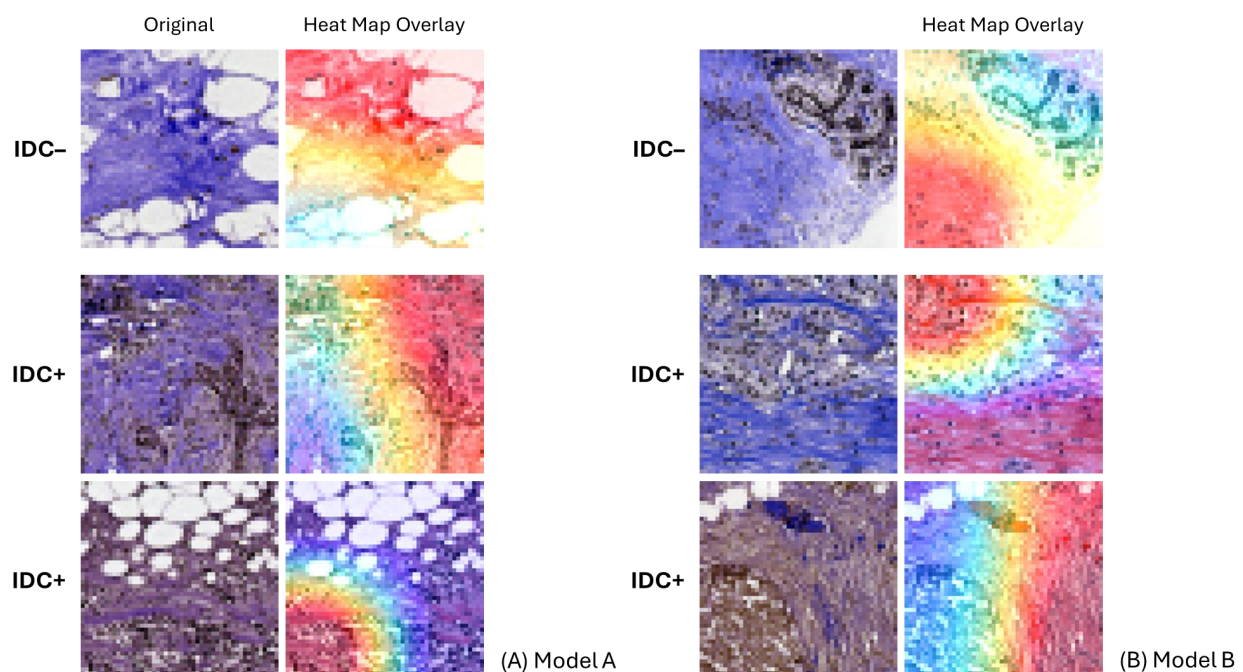
Figure 8. CAM visualization highlighting regions contributing to IDC predictions, made by Model A (**A**) and Model (**B**). Warmer colors (red and yellow) indicate areas of higher relevance to predicted class, while cooler color (blue) represent those with less influence.

5. Discussion

WSI is the gold standard in digital pathology for diagnosing BCa. Several studies have explored AI-based CAD systems designed to classify IDC pathology by using this imaging modality. Recent attempts have been either focusing on feature engineering [37], learning strategies [22,23], or modifying network architecture [4]. The findings reported earlier this year [45] suggested that locally distilling knowledge from dual scale morphology could help enhance lesion identification and pathological annotation. Our research question aims to investigate whether morphologically enhanced filters (i.e., Gabor wavelet) can help improve the classification of IDC, potentially refining diagnostic accuracy.

A ResNet-50 architecture with a modified dense layer was selected for its optimal balance between computation efficiency and prediction performance [16,17,21]. In the above experiments, only basic hyperparameter settings were made. The LR of 5.0×10^{-6} was identified as the most effective in achieving stable convergence for the given dataset. When comparing different fusion types, mean orientation invariant filters (R_{Mean}) demonstrated superior performance over the max filters (R_{Max}), regardless of weights (α). This may be due to R_{Max} responses emphasizing cell boundaries excessively, while suppressing interior patterns, leading to greater degradation of morphological details compared to R_{Mean}. Additionally, R_{Mean} more effectively captured spatial relationships between connecting structures across larger extents, due to accumulated responses in all directions. Rather than implementing formal dual scale distillation [45], we proposed fusing chromatic and morphological patterns via simple linear weighted combinations. The empirical results indicated that $\lambda = 4$ and $\lambda = 16$ were the most suitable scales for CCE and BCE loss function, respectively. It is worth noting that the weights remained 0.90 in both cases, suggesting that classification primarily relied on local chromatic distribution patterns, with additional support from enhanced boundary definitions. This aligns with the way pathologists interpret a histological slide, utilizing not only colors but also cellular arrangement and tissue morphology [24]. While ResNet's hidden neurons already capture the latter feature, augmenting it with Gabor responses clearly demonstrated its added contributions.

This study employed linear combinations to progressively assess the impact of morphological enhancement on ResNet's performance. While effective, adaptive fusion could further improve results, as demonstrated by recent studies on adaptive convolutional kernel updates with Gabor templates [46] and adaptive Gabor filters [47] in image classification and recognition. Their results outline the potential directions for further refinement of the current method.

A comparative analysis of both ResNet variants revealed that Model A demonstrated stronger performance in predicting IDC−, primarily by minimizing falsely predicted IDC+ (FP). This was reflected in its higher precision and specificity. However, both models exhibited similarly lower recall rates, indicating a common tendency to miss some IDC+. In contrast, Model B classified both classes equally well, as evidenced in its more balanced recall and precision. This suggests that Model A, with its local focus (lower λ), was more effective in confirming IDC−, while Model B (higher λ) was better suited to maintaining balance between sensitivity and specificity. The trade-off among these indicators underscores the complementary strengths of multi-band visual scales in diagnostic workflows. In other words, examining the slides at a broader scale can facilitate preliminary, unbiased screening, whereas finer-scale inspection may improve specificity by ruling out false IDC detection. This finding may guide researchers in developing ensemble deep learning [48,49] strategies for computer-aided BCa diagnosing systems.

The higher F1-score and AUC achieved by Model A indicate its superior overall performance, showing a stronger capacity to differentiate between positive and negative classes across varying thresholds. In contrast, while Model B maintained a balanced trade-

off between TP and FP, this came at the cost of a lower F1-score and AUC. This finding suggests that the choice of loss functions and weight parameters may affect consistency in feature learning. Finetuning model architecture based on the respective fusions is thus recommended. Finally, the benchmarking confirmed previous findings [16,17,21], where ResNet-50 consistently outperformed almost all the other DL models, regardless of multi-band visual fusion strategies.

The CAM heatmaps generated by these models closely aligned with pathologist interpretations, effectively highlighting diagnostically relevant areas. Key characteristic features include irregular, pleomorphic tumor cells actively invading surrounding breast tissue, which are critical indicators of IDC progression [50]. This consistency reinforces the potential of the proposed models to support clinical decision making, particularly in cases where subtle morphological patterns may otherwise be overlooked.

This study utilizes a publicly available dataset to encourage reproducibility, benchmarking, and development in IDC diagnosis. The selection of small patch sizes aligns with the original dataset's intent to validate DL models of varying complexity. While individual patches may not fully capture morphological and anatomical contexts, some studies [20,21,45] have explored distilled classification, where adjacent patches are hierarchically labeled to provide a more comprehensive delineation of pathological lesions. Recognizing the generalization challenges associated with small image datasets, future research will focus on improving applicability by utilizing larger images and incorporating additional morphological and anatomical contexts.

6. Conclusions

AI has played an increasingly active role in CAD of breast cancer. Thus far, existing research has primarily focused on developing sophisticated deep network architectures to address various challenges, achieving certain degrees of incremental accuracy. Despite promising successes, the developed CAD systems should still be supervised by pathologists to ensure valid slide interpretation. In contrast, this study investigated rather the impact of a prominent multi-band visual-enhanced filters, i.e., Gabor wavelets, on the DL-based classification of IDC using WSI modality.

Our findings suggest that images fused by mean orientation invariant filters offer superior performance compared to max filters, as they preserve both cellular arrangement and tissue morphology well. By incorporating spatial relationships across broader extents through accumulated directional responses, the adopted fusion strategy closely aligned with pathological interpretation principles utilized by experts.

Additionally, integrating Gabor filter responses into the input layer of ResNet-50 showed measurable contributions to improving its pattern association. Comparative analysis of its variants underscored the advantages of fusion on a larger scale, which achieved a strong balance between sensitivity and specificity. These results highlight the pivotal role of morphologically enhanced filters in enhancing diagnostic performance and optimizing DL models for BCa screening.

This study assessed ResNet-50, VGG-16 and 19, and DenseNet-121, focusing on well-established architectures to provide a robust baseline for comparison. Our findings may provide a guideline for future research in ensemble DL strategies. While the models offer solid performance analyses, we acknowledge the omission of more recent and lightweight networks. Efficient modern architectures are worth considering for reducing computational complexity.

Additionally, the reliance on a highly curated Kaggle dataset, with limited variability, poses a challenge in terms of generalizability in clinical settings. Specifically, domain shifts (e.g., variations in scanners, staining protocols, and patient populations) may affect model performance when applied to external datasets. Future research should focus on validation

across more diverse, multi-institutional datasets, to improve real-world applicability and extend usability to other pathological conditions.

Abbreviations

The following abbreviations are used in this manuscript:

MIR	Mortality-to-Incidence Ratio
BCa	Breast Cancer
IDC	Invasive Ductal Carcinoma
AI	Artificial Intelligence
CAD	Computer-Aided Diagnosis
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network
ANN	Artificial Neural Network
WSI	Whole Slide Image
GPU	Graphic Processing Unit
H&E	Hematoxylin and Eosin
BrT	Breast Tumor
FC	Fully Connected
ANOVA	Analysis of Variance
AD	Alzheimer's Disease
XAI	Explainable AI
MRI	Magnetic Resonance Image
CAM	Class Activation Mapping
ROI	Region of Interest
PNG	Portable Network Graphic
BCE	Binary Cross-Entropy
CCE	Categorical Cross-Entropy
LR	Learning Rate
PC	Personal Computer
OS	Operating System
CPU	Central Processing Unit
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

ROC   Receiver Operator Characteristic
AUC   Area Under Curve

References

1. Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R.L.; Soerjomataram, I.; Jemal, A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2024**, *74*, 229–263. [CrossRef] [PubMed]

2. Cao, W.; Qin, K.; Li, F.; Chen, W. Comparative study of cancer profiles between 2020 and 2022 using global cancer statistics (GLOBOCAN). *J. Natl. Cancer Cent.* **2024**, *4*, 128–134. [CrossRef] [PubMed]

3. American Cancer Society. Invasive Breast Cancer (IDC/ILC). American Cancer Society. 2021. Available online: https://www.cancer.org/cancer/types/breast-cancer/about/types-of-breast-cancer/invasive-breast-cancer.html (accessed on 17 April 2025).

4. Murcia-Gómez, D.; Rojas-Valenzuela, I.; Valenzuela, O. Impact of image preprocessing methods and deep learning models for classifying histopathological breast cancer images. *Appl. Sci.* **2022**, *12*, 11375. [CrossRef]

5. Fortunato, A.; Mallo, D.; Cisneros, L.; King, L.M.; Khan, A.; Curtis, C.; Ryser, M.D.; Lo, J.Y.; Hall, A.; Marks, J.R.; et al. Evolutionary measures show that recurrence of DCIS is distinct from progression to breast cancer. *Breast Cancer Res.* **2025**, *27*, 43. [CrossRef]

6. Golestan, A.; Tahmasebi, A.; Maghsoodi, N.; Faraji, S.N.; Irajie, C.; Ramezani, A. Unveiling promising breast cancer biomarkers: An integrative approach combining bioinformatics analysis and experimental verification. *BMC Cancer* **2024**, *24*, 155. [CrossRef]

7. Roy, S.; Shanmugam, G.; Rakshit, S.; Pradeep, R.; George, M.; Sarkar, K. Exploring the immunomodulatory potential of Brahmi (*Bacopa monnieri*) in the treatment of invasive ductal carcinoma. *Med. Oncol.* **2024**, *41*, 115. [CrossRef]

8. Mirbabaie, M.; Stieglitz, S.; Frick, N.R.J. Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. *Health Technol.* **2021**, *11*, 693–731. [CrossRef]

9. Kumar, Y.; Koul, A.; Singla, R.; Ijaz, M.F. Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda. *J. Ambient. Intell. Human. Comput.* **2023**, *14*, 8459–8486. [CrossRef] [PubMed]

10. Xu, Y.; Khan, T.M.; Song, Y.; Meijering, E. Edge deep learning in computer vision and medical diagnostics: A comprehensive survey. *Artif. Intell. Rev.* **2025**, *58*, 93. [CrossRef]

11. NVIDIA. CUDA Toolkit (release 11.8). 2022. Available online: https://developer.nvidia.com/cuda-toolkit (accessed on 2 July 2023).

12. Chen, H.; Belash, E.; Liu, Y.; Recheis, M. TensorFlow.NET: Google's TensorFlow full binding in .NET Standard. 2023. Available online: https://github.com/SciSharp/TensorFlow.NET (accessed on 5 February 2025).

13. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015; Computational and Biological Learning Society: Cambridge, UK, 2015; Volume 2015, pp. 1–14.

14. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.

15. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

16. Suganyadevi, S.; Seethalakshmi, V.; Balasamy, K. A review on deep learning in medical image analysis. *Int. J. Multimed. Inf. Retr.* **2022**, *11*, 19–38. [CrossRef]

17. Mall, P.K.; Singh, P.K.; Srivastav, S.; Narayan, V.; Paprzycki, M.; Jaworska, T.; Ganzha, M. A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthc. Anal.* **2023**, *4*, 100216. [CrossRef]

18. Mahmud, T.; Barua, K.; Habiba, S.U.; Sharmen, N.; Hossain, M.S.; Andersson, K. An explainable AI paradigm for Alzheimer's diagnosis using deep transfer learning. *Diagnostics* **2024**, *14*, 345. [CrossRef] [PubMed]

19. Zarella, M.D.; Bowman, D.; Aeffner, F.; Farahani, N.; Xthona, A.; Absar, S.F.; Parwani, A.; Bui, M.; Hartman, D.J. A Practical Guide to Whole Slide Imaging: A White Paper from the Digital Pathology Association. *Arch. Pathol. Lab. Med.* **2019**, *143*, 222–234. [CrossRef]

20. Rodriguez, J.P.M.; Rodriguez, R.; Silva, V.W.K.; Kitamura, F.C.; Corradi, G.C.A.; Bertoletti de Marchi, A.C.; Rieder, R. Artificial intelligence as a tool for diagnosis in digital pathology whole slide images: A systematic review. *J. Pathol. Inform.* **2022**, *13*, 100138. [CrossRef] [PubMed]

21. Fatima, G.; Alhmadi, H.; Ali Mahdi, A.; Hadi, N.; Fedacko, J.; Magomedova, A.; Parvez, S.; Mehdi Raza, A. Transforming diagnostics: A comprehensive review of advances in digital pathology. *Cureus* **2024**, *16*, e71890. [CrossRef]

22. Yang, J.; Chen, H.; Zhao, Y.; Yang, F.; Zhang, Y.; He, L.; Yao, J. ReMix: A General and Efficient Framework for Multiple Instance Learning Based Whole Slide Image Classification. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022. MICCAI 2022*; Lecture Notes in Computer Science; Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2022; Volume 13432.

23. Wang, H.; Luo, L.; Wang, F.; Tong, R.; Chen, Y.W.; Hu, H. Rethinking Multiple Instance Learning for Whole Slide Image Classification: A Bag-Level Classifier is a Good Instance-Level Teacher. *IEEE Trans. Med. Imaging* **2024**, *43*, 3964–3976. [CrossRef] [PubMed]

24. Evans, A.J.; Brown, R.W.; Bui, M.M.; Chlipala, E.A.; Lacchetti, C.; Milner, D.A.; Pantanowitz, L.; Parwani, A.V.; Reid, K.; Riben, M.W.; et al. Validating whole slide imaging systems for diagnostic purposes in pathology: Guideline update from the College of American Pathologists. *Arch. Pathol. Lab. Med.* **2022**, *146*, 440–450. [CrossRef]

25. Murtaza, G.; Abdul Wahab, A.W.; Raza, G.; Shuib, L. A tree based multiclassification of breast tumor histopathology images through deep learning. *Comput. Med. Imaging Graph.* **2021**, *89*, 101870. [CrossRef]

26. Dunn, C.; Brettle, D.; Cockroft, M.; Keating, E.; Revie, C.; Treanor, D. Quantitative assessment of H&E staining for pathology: Development and clinical evaluation of a novel system. *Diagn. Pathol.* **2024**, *19*, 42.

27. Reinhard, E.; Adhikhmin, M.; Gooch, B.; Shirley, P. Color transfer between images. *IEEE Comput. Graph. Appl.* **2001**, *21*, 34–41. [CrossRef]

28. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]

29. Liu, Y.; Chen, X.; Peng, H.; Wang, Z. Multi-focus image fusion with a deep convolutional neural network. *Inf. Fusion.* **2017**, *36*, 191–207. [CrossRef]

30. Biswas, A.A. A comprehensive review of explainable AI for disease diagnosis. *Array* **2024**, *22*, 100345. [CrossRef]

31. Muhammad, D.; Bendechache, M. Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis. *Comput. Struct. Biotechnol. J.* **2024**, *24*, 542–560. [CrossRef]

32. Embeyale, D.; Chen, Y.T.; Assabie, Y. Automatic grading of barley grain for brewery industries using convolutional neural network based on texture features. *J. Agric. Food Res.* **2025**, *20*, 101752. [CrossRef]

33. Serte, S.; Demirel, H. Gabor wavelet-based deep learning for skin lesion classification. *Comput. Biol. Med.* **2019**, *113*, 103423. [CrossRef]

34. Yuan, Y.; Zhang, J.; Wang, Q. Deep Gabor convolution network for person re-identification. *Neurocomputing* **2020**, *378*, 387–398. [CrossRef]

35. Thanh Le, H.; Phung, S.L.; Chapple, P.B.; Bouzerdoum, A.; Ritz, C.H.; Tran, L.C. Deep Gabor Neural Network for Automatic Detection of Mine-Like Objects in Sonar Imagery. *IEEE Access* **2020**, *8*, 94126–94139. [CrossRef]

36. Jaber, A.G.; Muniyandi, R.C.; Usman, O.L.; Singh, H.K.R. A Hybrid Method of Enhancing Accuracy of Facial Recognition System Using Gabor Filter and Stacked Sparse Autoencoders Deep Neural Network. *Appl. Sci.* **2022**, *12*, 11052. [CrossRef]

37. Janowczyk, A.; Madabhushi, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J. Pathol. Inform.* **2016**, *7*, 29. [CrossRef]

38. Mooney, P. (n.d.). Breast Histopathology Images [Data Set]. Kaggle. Available online: https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images (accessed on 2 December 2024).

39. Gabor, D. Theory of communication. *J. Inst. Electr. Eng. —Part. III: Radio. Commun. Eng.* **1946**, *93*, 429–457. [CrossRef]

40. Ghiasi-Shirazi, K. Learning 2D Gabor filters by infinite kernel learning regression. *J. Comput. Math. Data Sci.* **2021**, *1*, 100016. [CrossRef]

41. Alekseev, A.; Bobe, A. GaborNet: Gabor filters with learnable parameters in deep convolutional neural network. In Proceedings of the 2019 International Conference on Engineering and Telecommunication (EnT), Dolgoprudny, Russia, 20–21 November 2019; pp. 1–4.

42. Watt, R.; Ledgeway, T.; Dakin, S.C. Families of models for gabor paths demonstrate the importance of spatial adjacency. *J. Vision.* **2008**, *8*, 23. [CrossRef] [PubMed]

43. Saleh, A.Y.; Chern, L.H. Autism spectrum disorder classification using deep learning. *Int. J. Online Biomed. Eng.* **2021**, *17*, 24603. [CrossRef]

44. Pereira, C.; Guede-Fernández, F.; Vigário, R.; Coelho, P.; Fragata, J.; Londral, A. Image analysis system for early detection of cardiothoracic surgery wound alterations based on artificial intelligence models. *Appl. Sci.* **2023**, *13*, 2120. [CrossRef]

45. Cheng, H.; Liu, X.; Zhang, J.; Dong, X.; Ma, X.; Zhang, Y.; Meng, H.; Chen, X.; Yue, G.; Li, Y.; et al. GLMKD: Joint global and local mutual knowledge distillation for weakly supervised lesion segmentation in histopathology images, *Expert. Syst. Appl.* **2025**, *279*, 127425. [CrossRef]

46. Yuan, Y.; Wang, L.; Zhong, G.; Gao, W.; Jiao, W.; Dong, J.; Shen, B.; Xia, D.; Wei, X. Adaptive Gabor convolutional networks. *Pattern Recogn.* **2022**, *124*, 108495. [CrossRef]

47. Kovač, I.; Marák, P. Finger vein recognition: Utilization of adaptive gabor filters in the enhancement stage combined with SIFT/SURF-based feature extraction. *Signal Image Video Process.* **2023**, *17*, 635–641. [CrossRef]

48. Ganaie, M.A.; Hu, M.; Malik, A.K.; Tanveer, M.; Suganthan, P.N. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105151. [CrossRef]

49.   Yang, Y.; Lv, H.; Chen, N. A Survey on ensemble learning under the era of deep learning. *Artif. Intell. Rev.* **2023**, *56*, 5545–5589. [CrossRef]

50.   Zhao, J.; Lang, R.; Guo, X.; Chen, L.; Gu, F.; Fan, Y.; Fu, X.; Fu, L. Clinicopathologic characteristics of pleomorphic carcinoma of the breast. *Virchows Arch.* **2010**, *456*, 31–37. [CrossRef] [PubMed]