applied sciences

MDPI

Article

A Lightweight Optimized CNN Framework for Breast Cancer Screening and Subtype Classification

Suyun Zhang [1], Peng Zhang [2,*] and Guofei Zhang [3,*]

[1] Textile Engineering, College of Textiles, Songjiang Campus, Donghua University, Shanghai 200051, China; 240400516@mail.dhu.edu.cn
[2] Institute of Artificial Intelligence, Songjiang Campus, Donghua University, Shanghai 200051, China
[3] Asian Pacific Forestry College, Southwest Forestry University, Kunming 650233, China
* Correspondence: zhangp88@dhu.edu.cn (P.Z.); zhanggf1979@swfu.edu.cn (G.Z.)

Abstract

Breast cancer remains one of the most serious diseases threatening human health globally. To address diagnostic subjectivity and insufficient classification granularity in pathological screening, this study proposes a lightweight complementary convolutional neural network (CNN) framework for two distinct yet clinically related tasks. The first model performs binary classification of invasive ductal carcinoma (IDC) (IDC-positive vs. IDC-negative), while the second model conducts fine-grained multiclass classification of four malignant subtypes. Notably, primary patch-level evaluation for the IDC binary task may produce optimistically biased performance due to data leakage, where patches from the same patient can appear in both the training and test sets. The IDC_regular_ps50_idx5 dataset embeds patient/case identifiers within patch filenames, enabling patient-disjoint partitioning to mitigate this risk. On public patch-level test data, the framework achieves a patch-level accuracy of 0.8732 and an AUC of 0.9245, demonstrating superior resource efficiency compared with several state-of-the-art models. Computational analysis shows that the proposed model has 0.435 M parameters and 0.145 G FLOPs—33.8× fewer parameters than VGG16 (convolutional-only), 54.0× fewer than ResNet50, and 9.3× fewer than EfficientNet-B0. To quantify the optimistic bias introduced by patch-level evaluation, we first performed supplementary patient-disjoint 5-fold GroupKFold cross-validation on the IDC dataset using filename-embedded patient identifiers. Under this intra-dataset patient-disjoint protocol, it achieves a mean accuracy of 0.8005 and mean AUC of 0.8783, with a moderate performance drop of approximately 7 percentage points relative to patch-level results. We further conducted external cross-dataset validation on BreakHis with strict patient-disjoint 5-fold GroupKFold cross-validation. Under this cross-dataset protocol, the model yields substantially degraded performance (mean accuracy of 0.5417 ± 0.1738 and mean AUC of 0.5930 ± 0.1680), illustrating significant performance drop when generalizing to completely unseen patients under domain shift. A systematic loss function comparison experiment under patch-level configuration demonstrates that Focal Loss ($\gamma = 2.0$) achieves the highest overall patch-level performance (Accuracy = 0.8125, AUC = 0.8980) on the binary IDC identification task, outperforming the baseline cross-entropy loss. The results indicate that the proposed complementary framework achieves competitive classification accuracy with a favorable accuracy–efficiency trade-off within the IDC dataset; however, strong cross-dataset generalization cannot be guaranteed, and patient-disjoint evaluation on multi-institution matched cohorts is required for real-world clinical deployment. Limitations in generalization to minority classes remain. Future work will focus on constructing multi-institution patient-annotated histopathology cohorts for patient-disjoint evaluation and domain-adaptation techniques to mitigate cross-dataset domain shift.

Check for updates

Academic Editors: Pedro Couto and Thomas Lindner

Received: 24 August 2026
Revised: 30 September 2026
Accepted: 6 October 2026
Published: 11 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: breast cancer; lightweight CNN; complementary architecture; cancer classification; cancer identification

1. Introduction

Breast cancer remains one of the most serious diseases threatening human health globally. According to the Global Cancer Statistics 2020 report, the disease accounts for approximately 2.3 million new cases and 685,000 deaths annually [1]. The International Agency for Research on Cancer (IARC) predicts that if current trends continue, the number of new breast cancer cases worldwide will reach 3.2 million per year by 2050, with deaths rising to 1.1 million. These data highlight a serious reality: existing diagnostic capabilities are no longer sufficient to meet clinical needs, and there is an urgent need for a paradigm shift in early detection, which currently suffers from high intra- and inter-observer variability.

Breast cancer is not a single disease but rather a spectrum of subtypes with distinct histological characteristics, each classified according to its tissue of origin and receptor status.

These subtypes include, but are not limited to, ductal carcinoma in situ (DCIS), invasive ductal carcinoma (IDC), and invasive lobular carcinoma (ILC). Accurately distinguishing among subtypes is a prerequisite for developing personalized treatment plans, yet it remains one of the most challenging tasks in histopathological analysis. High-resolution histopathological sections contain rich diagnostic detail, but their structural complexity imposes a substantial cognitive burden. This high data density can induce visual fatigue, thereby increasing the risk of misdiagnosis or missed diagnosis—a concern that is particularly pronounced in long-term screening workflows [2,3].

Breast cancer cells exhibit a high degree of morphological diversity, encompassing round, oval, irregular, and signet-ring shapes. The spatial distribution of cells also varies considerably, ranging from glandular tubular structures to diffuse and lattice-like arrangements [2,3]. Critically, malignant cells and non-cancerous macrophages can appear morphologically similar, and variations in staining protocols and slide preparation procedures further complicate the identification of key diagnostic features [2,3]. Collectively, these factors—including the inherent subjectivity of human interpretation, variations in pathologist expertise, and inter-observer variability—underscore the inherent complexity and diagnostic uncertainty of manual tissue pathology evaluation. Representative histopathological examples of negative (IDC-negative) and positive (IDC-positive) cells are illustrated in Figures 1–4 (Figures 1 and 2: breast cancer cells, 0 = negative; Figures 3 and 4: mammary gland cells, 1 = positive). This uncertainty is precisely the core motivation driving the development of automated and objective classification models in the research community [2,3].

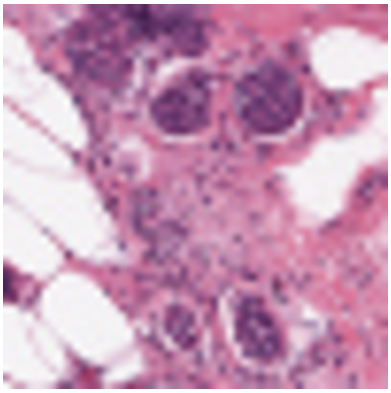


Figure 1. Breast cancer cells (0 = negative), normal tissue.

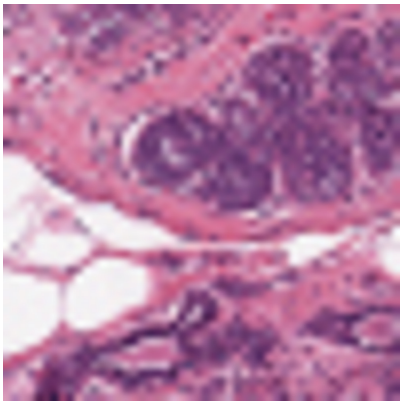


Figure 2. Breast cancer cells (0 = negative), representative normal sample.

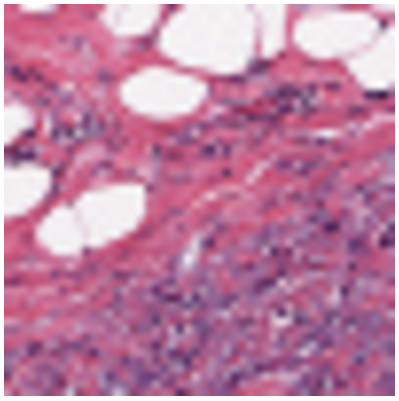


Figure 3. Mammary gland cells (1 = positive), malignant tumor tissue.

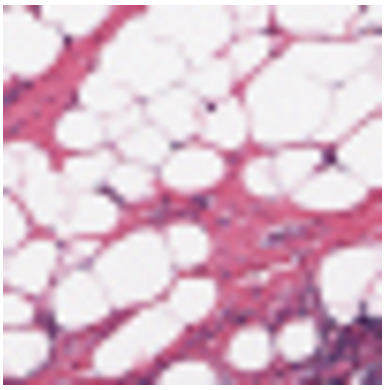


Figure 4. Mammary gland cells (1 = positive), representative malignant sample.

Breast cancer originates from the uncontrolled proliferation of abnormal cells, typically arising from the lobular or ductal epithelium. A key clinical fact is that early malignant tumors often present with no obvious symptoms. The asymptomatic nature of these lesions, combined with the inherent concealment of early pathological changes, underscores the urgency of developing diagnostic models capable of detecting disease before clinical symptoms manifest [1].

In recent years, deep learning—particularly convolutional neural networks (CNNs)—has become the mainstream approach to automated histopathology analysis. Early CNN applications in this field typically employed classical architectures (e.g., AlexNet, VGG) for binary or multiclass classification tasks, demonstrating the feasibility of automatic feature extraction without manual design. Representative studies include the BC-MBINet model developed by Singh et al., which achieved an accuracy of 99.04% [4], and the magnification-independent CNN proposed by the Atitallah team, which reached an accuracy of 97.72% [5].

However, these pioneering studies soon revealed fundamental limitations: the extracted features often lack representativeness; the models struggle to capture holistic tissue-structure information; they lack the ability to automatically attend to diagnostically relevant regions; and they are prone to overfitting when training data are limited.

In response to these limitations, subsequent research shifted toward hybrid architectures (i.e., CNNs combined with sequential or attentional modules) that enhance CNN performance through sequence modeling and attention mechanisms. By integrating LSTM or GRU layers, these models can capture long-range spatial dependencies, while attention modules help the model precisely attend to tumor-related regions and suppress background noise. Architectures such as CNN-LSTM, CNN-Attention, and their variants have demonstrated significant performance improvements, particularly in challenging subtype detection tasks. The BCDCNN framework proposed by Jaincy and Pattabiraman is a representative example, achieving an accuracy of 90.2% and a specificity of 90.9% in the test [6]. However, these improvements have come at a considerable cost: substantially increased model complexity and training difficulty. Coupled with high computational demands and the persistent shortage of annotated medical data, these factors severely limit the practical deployment of such models.

To address the long-standing challenge of data scarcity, transfer learning and ensemble methods have increasingly become mainstream solutions. Transfer learning leverages backbone networks pre-trained on large-scale natural image datasets (e.g., VGG, MobileNet) to effectively mitigate overfitting and enhance generalization on small-scale pathological datasets. Ensemble methods improve model robustness by aggregating predictions from multiple models. For example, the BC-GAHOCNN model proposed by El Hassani et al. achieved 99% precision and 98% accuracy [7]; while the Thermo-CAD system developed by Attallah integrated multiple CNNs to improve diagnostic reliability [8]. Despite these advances, a key limitation persists: most existing methods rely heavily on generic pre-trained features and complex network architectures, lacking task-specific structural optimization for breast cancer histopathology. This not only limits method adaptability but also introduces a heavy computational burden, hindering clinical translation.

Meanwhile, the field is undergoing a broader paradigm shift: contemporary research increasingly places lightweight architecture design, interpretable artificial intelligence (e.g., Grad-CAM heatmaps), and clinical feasibility on an equal footing with predictive accuracy. Low computational latency, model transparency, and edge-computing adaptability have become core design requirements for next-generation diagnostic tools.

Despite significant progress, two interrelated challenges remain unresolved. First, breast tissue exhibits extreme morphological and topological heterogeneity, making automated feature extraction exceptionally difficult; traditional feature extraction methods all have inherent limitations. Second, although binary classification dominates current research, clinical practice increasingly demands refined multiclass classification of cancer subtypes to support individualized treatment decisions—a capability lacking in most existing models.

Deep learning, particularly CNNs, has demonstrated exceptional proficiency in identifying features and complex patterns in medical images, thereby enhancing diagnostic efficacy [9]. This study addresses a supervised learning classification problem within the domain of machine learning.

To address these issues, this study designs from scratch a lightweight complementary CNN framework for open-source histopathology datasets. Our approach has three interrelated core objectives: (1) design and optimize dedicated CNN architectures for binary and multiclass classification—IDC identification and breast cancer subtype classification; (2) note that this work does not perform head-to-head comparative experiments among

grid search, random search, and Bayesian optimization; (3) perform an in-depth comparative analysis of the proposed CNN against three state-of-the-art (SOTA) algorithms (TransferEnsemble-Breast, EffiLSTM-Attn, and EffiIHC-Ensemble) to clarify the advantages, limitations, and future directions of this model.

This study's primary binary IDC classification experiments include both patch-level random-split evaluation and intra-dataset patient-disjoint GroupKFold cross-validation. Patient/case identifiers embedded within patch filenames of the IDC_regular_ps50_idx5 dataset enable patient-disjoint partitioning to quantify optimistic bias induced by patch-level data leakage, as described in Section 4.6. Supplementary cross-dataset patient-disjoint validation on BreakHis further evaluates generalization under domain-shift conditions. Patient-disjoint evaluation on multi-institution matched cohorts remains an open requirement for real-world clinical translation.

2. Related Work

In recent years, deep learning has fundamentally transformed breast cancer histopathology research, shifting evaluation from subjective human judgment to automated objective analysis. Among various deep learning architectures, CNNs can directly learn hierarchical spatial feature representations from raw image data. Leveraging this inherent advantage, CNNs have become the dominant approach, eliminating the need for manually designed descriptors. This section systematically reviews CNN-based breast cancer image recognition and subtype classification methods and analyzes their limitations.

2.1. Early Applications of Standalone CNNs

Early research focused on applying standard CNN architectures such as VGG and ResNet directly to histopathology images for benign/malignant binary classification or subtype multiclass tasks. These pioneering studies provided proof of concept: demonstrating that CNNs can automatically extract morphologically meaningful features without human intervention. For example, the BC-MBINet model developed by Singh et al. achieved an accuracy of 99.04% [4]; the magnification-independent CNN proposed by Atitallah et al. achieved an accuracy of 97.72% [5]. However, these methods quickly revealed a core limitation: such models are essentially "black boxes" that cannot capture holistic tissue structure information or selectively focus on diagnostically relevant regions. Moreover, given the scarcity of large-scale annotated medical datasets, the large parameter counts of these models make them prone to overfitting.

Sharma et al. proposed a CNN architecture for breast cancer histopathological image classification at various magnification levels [10]. By incorporating optimized pooling strategies and data augmentation, the model achieved robust performance on the BreakHis dataset, with accuracies of 80.76%, 76.58%, 79.90%, and 74.21% at 40×, 100×, 200×, and 400× magnification, respectively [10]. These results indicate that the proposed method effectively mitigates visual discrepancies and feature distribution inconsistencies across magnification scales, thereby improving both classification stability and accuracy in multi-scale pathological image analysis [10].

Yusuf et al. proposed an enhanced shallow CNN (ES-CNN) for multiclass classification of breast histopathological images, aiming to improve classification accuracy for eight breast cancer subtypes across four magnification factors on the BreakHis dataset while reducing training time [11]. Experimental evaluations in terms of classification accuracy and computational efficiency demonstrate that the proposed method achieves significantly lower computational cost while outperforming existing approaches [11].

2.2. Transfer Learning and Ensemble Strategies

A third mainstream research direction addresses data scarcity and enhances model robustness through transfer learning and ensemble strategies. Researchers fine-tune pre-trained backbones (e.g., VGG, MobileNet, EfficientNet) on target pathological datasets, effectively transferring knowledge from large-scale natural image domains to pathological analysis. Ensemble methods further aggregate predictions from multiple complementary models to improve performance. The BC-GAHOCNN model developed by El Hassani et al. achieved 99% accuracy and 98% precision [2], while the Thermo-CAD system proposed by Attallah integrated multiple CNNs for classification tasks [5]. However, this direction has inherent limitations: over-reliance on pre-trained features may introduce domain bias from natural images that does not transfer to pathological analysis, and the overall model complexity conflicts with the goal of developing lightweight, deployable solutions.

Shah et al. pioneered the use of generative models to expand limited datasets beyond traditional transfer learning [12]. They introduced a GAN-based framework for histopathological image augmentation, generating highly realistic synthetic images that closely resemble real samples, thereby substantially diversifying the training data [12]. Their ResNet model achieved a notable 90.5% accuracy, underscoring the potential of generative models for advancing breast cancer diagnostics [12].

Zhou et al. proposed ResViT-GANNet, which integrates a residual convolutional network with channel attention and a vision transformer incorporating multi-layer token fusion [13]. The key innovation is the TAMA module, which effectively fuses heterogeneous features from both branches through multi-head attention and token-level alignment [13]. The model achieved a mean accuracy of 98.22%, with per-class accuracies ranging from 97.25% to 99.50%.

Importantly, the BreakHis and IDC datasets differ in annotation granularity: BreakHis v1 provides explicit patient-level identifiers, whereas for IDC_regular_ps50_idx5 patient information is encoded within patch filenames. This distinction is critical for evaluating data leakage and guides the patient-disjoint validation in Section 4.6.

2.3. Hybrid Architectures: CNNs with Sequential and Attentional Mechanisms

To overcome the representational limitations of standalone CNNs, a second wave of research combines sequence modeling (e.g., LSTM, GRU) with attention mechanisms. The rationale is clear: LSTM layers capture long-range spatial dependencies between tissue regions, while attention modules enable the model to automatically focus on key tumor subregions and suppress background noise. Architectures such as CNN-LSTM, CNN-Attention, and CNN-RNN-Attention have achieved significant accuracy improvements in complex subtype classification tasks. Jaincy and Pattabiraman developed the BCD-CNN framework using adaptive error similarity, achieving an accuracy of 90.2% and a specificity of 90.9% [6]. Although performance improved, these architectures introduce new challenges: significantly increased model complexity and training difficulty, and the high computational demands of hybrid frameworks have become the primary obstacle to clinical deployment.

Wang et al. proposed MFF-ClassificationNet, which employs a two-branch parallel architecture: a CNN captures local details while a Transformer models global dependencies [14]. Features are deeply integrated through a multi-feature fusion module with a CBAM-SE block to enhance fine-grained feature representation and class-specific lesion description [14]. An accuracy of 97.50% was achieved on the BACH dataset [14].

Ativi et al. proposed CMAF-Net, which combines information bottleneck principles with margin-based learning to jointly address severe class imbalance and multi-scale fusion challenges in breast cancer histopathology classification [15]. The Cross-Modal Attention

Fusion block employs temperature-controlled attention and redundancy minimization to preserve complementary local and global features [15]. This approach achieves an 8.6% improvement over state-of-the-art baselines in malignant detection [15].

2.4. Limitations of Current 'Two-Stage' and 'Hybrid' Approaches

Ensemble learning is considered an effective strategy to enhance classification performance by exploiting the complementary strengths of multiple models. This method has been widely applied in breast cancer histopathology classification to improve robustness and accuracy.

Nakach et al. proposed a deep hybrid AdaBoost ensemble method for breast cancer histopathology classification [16]. This method uses DenseNet201, InceptionV3, and MobileNetV2 as feature extractors with the AdaBoost algorithm using decision trees as base learners [16], and evaluates ensembles with different numbers of trees [16]. The results show that at 200× magnification, an AdaBoost ensemble with DenseNet201 as the feature extractor and 200 trees achieved the best average classification accuracy of 90.36% [16].

Azmoodeh-Kalati et al. applied EfficientNetV1 and EfficientNetV2 to binary classification [17]. To address limited labeled data, they employed data augmentation and transfer learning and adopted Grad-CAM to enhance model interpretability [17]. The model achieved 99.58% classification accuracy on the BreakHis dataset, outperforming traditional CNNs [17].

Many studies conceptually propose "two-stage" frameworks for tasks such as detection and classification [4–7]. However, because no single dataset provides both coarse and fine-grained labels, these stages are often trained and evaluated on different datasets (e.g., binary classification on Dataset A, multiclass classification on Dataset B). This practice, although common, means the framework is not a truly integrated pipeline. Similarly, hybrid models combining CNNs with LSTMs or Transformers, although powerful, often introduce substantial computational overhead and may not suit all deployment scenarios. Our work acknowledges this common practice and presents two complementary yet independent lightweight models, each optimized for its specific task, rather than claiming an end-to-end pipeline. Our contribution lies in the resource-efficient design and the systematic comparison of hyperparameter optimization strategies for both tasks, rather than in architectural innovation alone.

2.5. Multimodal Integration and Recent Advances

Beyond the aforementioned image-domain, ensemble, and hybrid approaches, a growing body of recent work has explored the integration of complementary data modalities and molecular-level features to achieve more robust and granular breast cancer subtype classification. These studies collectively reflect an ongoing shift toward heterogeneous data sources—from multi-omics and gene expression to medical imaging-as a means to overcome the representational limits of histopathology alone.

Zhang et al. proposed DMFF, a deep learning-based multi-omics fusion framework that integrates multi-omics data for breast cancer survival prediction and subtype classification, demonstrating the potential of complementary data modalities in improving classification granularity [18]. Kandan and Abul employed relative gene expression order representations with a robust machine learning approach for breast cancer subtype classification, achieving high precision (≥91%) and recall (≥91%) on cross-validation and high accuracy (≥95%) on the TCGA-BRCA external validation cohort [19]. Lodh et al. developed a stacked ensemble meta-classifier integrating MLP, GRU, XGBoost, and Linear SVM for PAM50-based breast cancer subtype prediction from gene expression signatures, achieving 98.31% accuracy on METABRIC and 95.88% accuracy on TCGA-BRCA [20]. Ab Mumin et al. investigated MRI-based texture analysis for breast cancer subtype classification in a

multi-ethnic population, demonstrating that radiomics features can effectively discriminate between molecular subtypes across diverse patient groups [21].

Ali et al. proposed a multimodal breast cancer classification approach using a hybrid channel-wise attention CTNet to extract deep features and capture intricate patterns essential for accurate detection [22]. The transformer layers enhance the model's sensitivity to fine-grained relationships through sophisticated attention mechanisms, further refining the extracted features [22]. CTNet outperformed baseline models, achieving an accuracy of 86% [22]. Zhao et al. proposed a multi-level multimodal data fusion framework that integrates multi-omics data and histopathology slide images for prognostic prediction [23]. The approach constructs a spatial map of instances by connecting neighboring nuclei and computes a characteristic tensor via graph convolution layers applied to input histopathology slide images [23]. The average C-index for the breast cancer and non-small cell lung cancer cohorts improved from 0.61 and 0.52 to 0.75 and 0.67, respectively [23].

Despite these advances, a clear tendency toward "accuracy over deployability" persists: most high-performance models prioritize predictive accuracy at the expense of computational efficiency, clinical interpretability, and deployment feasibility. Moreover, while binary classification has been extensively studied, fine-grained multi-category subtype distinction—which carries greater clinical significance—remains underexplored. This gap motivates our design of a lightweight, integrated CNN framework that systematically balances accuracy, computational efficiency, and interpretability—the core objective of this study.

3. Methodology

Figure 5 illustrates the overall experimental workflow of this study. The proposed framework comprises two complementary models: (1) a binary CNN for benign/malignant screening; (2) a multiclass CNN for distinguishing four malignant subtypes. Each model has a specialized data preprocessing pipeline, network architecture, and hyperparameter optimization strategy, detailed in the following sections.

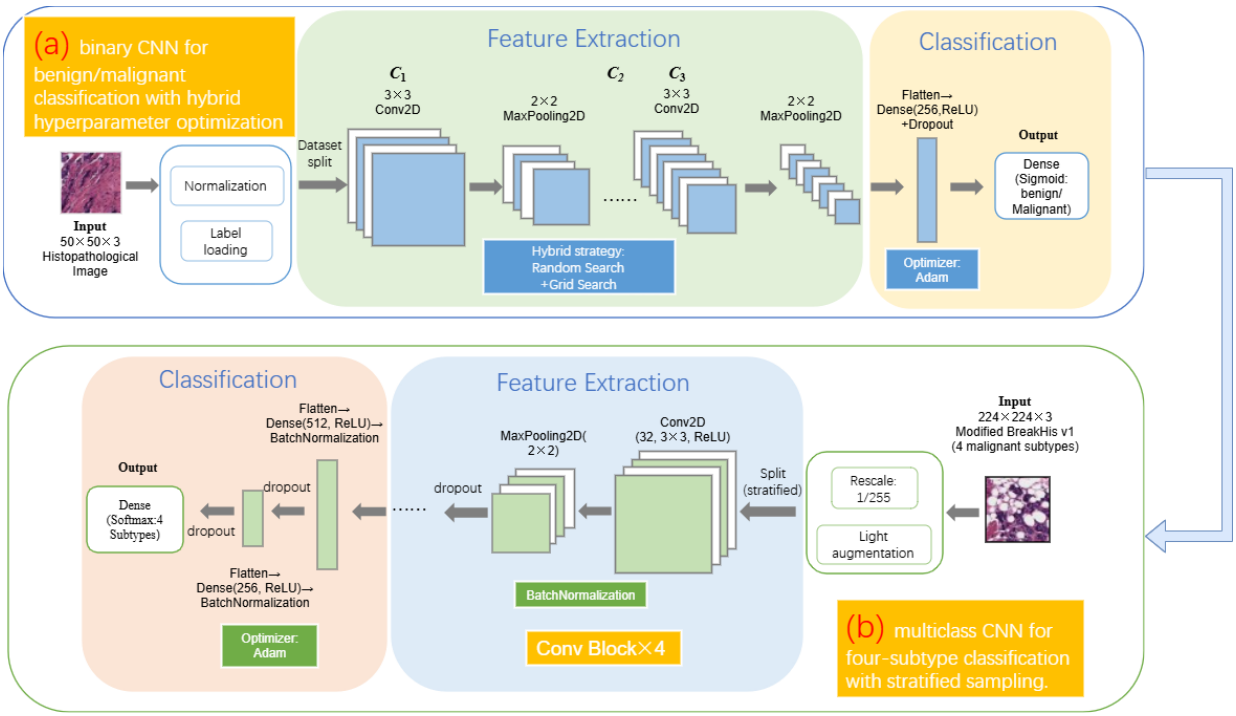


Figure 5. Overview of the proposed framework: (**a**) binary CNN for benign/malignant classification with hybrid hyperparameter optimization; (**b**) multiclass CNN for four-subtype classification with stratified sampling.

3.1. Binary Convolutional Neural Network for Cancer Identification

This section presents the binary CNN architecture for IDC screening, including its structural design principles, the dual-strategy hyperparameter optimization approach, and the final optimal configuration.

3.1.1. Limitations of Suboptimal Manual Hyperparameter Tuning

Manual hyperparameter tuning is a core challenge in training CNNs for breast cancer histopathology classification. This process relies heavily on the researcher's experience and configuration, is prone to local optima, and often results in unstable convergence, poor generalization, and low optimization efficiency.

To address this, we propose a lightweight binary CNN framework that combines random search and grid search for automatic hyperparameter optimization.

3.1.2. Inefficient Exploration of Hyperparameter Space

Another major challenge is the inefficiency of exploring the hyperparameter space, which includes learning rate, batch size, filter counts, fully connected units, and dropout rate. Random search alone cannot efficiently identify promising regions, while grid search alone suffers from combinatorial explosion.

We therefore propose a two-stage hyperparameter selection strategy: first, random search efficiently explores a wide parameter space to identify promising combinations; then, grid search performs refined optimization within the narrowed range. This hybrid strategy quickly and automatically identifies the global optimal configuration, balancing exploration and exploitation.

3.1.3. Overfitting Issues in Limited Medical Datasets

Another challenge in medical image analysis is overfitting caused by limited labeled samples and high model complexity. Deep networks with excessive parameters tend to memorize training data rather than learn generalizable features.

To reduce parameters, our binary CNN adopts a compact three-module convolutional design with dropout regularization and early stopping: dropout suppresses neuron co-adaptation, while early stopping terminates training before overfitting occurs. This design maintains the model's lightweight nature while providing sufficient discriminative power for benign/malignant classification.

3.1.4. Unreliability of Evaluation Methods for Imbalanced Medical Data

Overall accuracy alone is insufficient for evaluating imbalanced medical datasets.

We therefore adopt a comprehensive set of quantitative metrics, including precision, recall, F1-score, area under the curve (AUC), and confusion matrix visualization. In addition, we perform independent tests of each algorithm component, along with parameter studies, benchmark comparisons, case evaluations, and ablation experiments. The systematic evaluation demonstrates that the optimized binary CNN exhibits excellent robustness and stability, meeting the requirements of computer-assisted diagnostic systems.

3.1.5. Bayesian Hyperparameter Optimization

Unlike exhaustive grid search or blind random sampling, Bayesian optimization (BO) builds a probabilistic surrogate model—typically a Gaussian process (GP)—that maps hyperparameter configurations to expected validation performance, and uses an acquisition function (e.g., expected improvement, EI) to select the next configuration to evaluate. This data-driven strategy significantly improves sampling efficiency in high-dimensional hyperparameter spaces and is well suited for computationally expensive deep model tuning.

In this study, we implemented Bayesian optimization using Keras–Tuner.

The search space covered six hyperparameters:

- Filter counts in three consecutive convolutional layers: 16–64, 32–128, 64–256
- Fully connected units: 128–512
- Dropout rate: 0.3–0.7
- Learning rate: 10^{-4}, 10^{-3}, 10^{-2}

We conducted 10 trials in total. To accelerate the search while maintaining representativeness, we evaluated each candidate model on a stratified subset of 500 training images and 200 validation images with a fixed training duration of 5 epochs. The objective was to maximize validation accuracy.

The optimizer progressively converged to better configurations: the first three trials explored a relatively wide region, while subsequent trials became increasingly concentrated in high-performance regions identified by the GP surrogate.

The best validation accuracy (0.835) was achieved at Trial #9 with the optimal configuration:

- Convolutional filters: {16, 64, 192}
- Fully connected layer: 256 units
- Dropout rate: 0.5
- Learning rate: 10^{-3}

This configuration was subsequently fixed for all ablation experiments.

A key advantage of this approach is the decoupling of hyperparameter search from full training: candidate models are evaluated on reduced data with fewer epochs, while the final configuration is trained on the complete dataset. This two-stage strategy reduces computational cost by approximately an order of magnitude and minimizes the risk of test-set leakage into hyperparameter selection—a common methodological concern in deep learning-based medical image analysis.

3.2. Multiclass CNN for Classification of Cancer

3.2.1. Analysis of Class Imbalance in Multiclass Tumor Classification Tasks

A key challenge in multiclass histopathological classification of breast cancer is severe class imbalance. In the BreakHis v1 dataset, ductal carcinoma samples significantly outnumber other subtypes. This imbalance biases classification toward the majority class, weakening generalization for rare but clinically significant subtypes.

To address this, we constructed a structured dataset from BreakHis v1, retaining four malignant subtypes for standardized multiclass learning. We adopted stratified data splitting to preserve label distribution in the training and test sets, and used ImageDataGenerator for real-time data loading, which dynamically loads images during training and efficiently processes large-scale image collections. In addition, we apply lightweight data augmentation to minority classes, expanding the effective training set for rare subtypes without introducing additional noise, thereby improving the stability of feature learning under imbalanced conditions.

3.2.2. Architectural Design to Alleviate Multiclass Feature Complexity

Breast cancer subtypes exhibit high intra-class similarity and inter-class ambiguity, requiring models to extract hierarchical, multi-scale features to identify subtle histopathological differences. However, shallow CNNs struggle to capture the complex structural features of rare subtypes, while overly deep models are prone to overfitting on small medical datasets.

To address this gap, we propose a progressive convolutional architecture with four modules: early modules extract shallow texture and edge features, while deeper modules capture higher-level patterns corresponding to specific subtypes. Each convolutional block

follows a fixed structure: Conv2D (3×3) $\rightarrow$ BatchNormalization $\rightarrow$ MaxPooling2D (2×2) $\rightarrow$ Dropout (0.25). After the convolutional backbone, a flattening layer connects two fully connected layers with batch normalization and dropout, and the final layer outputs a multi-class probability distribution. The model is trained with the Adam optimizer and categorical cross-entropy loss, which is designed for mutually exclusive multiclass classification.

3.2.3. Dual Configuration and Training Strategy for Fast Iteration and Stable Training

Training deep CNNs on breast cancer datasets often suffers from long training times and unstable convergence, which seriously hinders rapid iteration and hyperparameter tuning.

To address this, we designed two complementary training configurations to balance performance and efficiency. The first is a standard configuration for full-scale training and final evaluation, ensuring the highest accuracy. The second is a speed-optimized configuration: doubling the batch size accelerates parameter updates, reducing the number of epochs enables rapid iteration, and increasing the learning rate speeds up convergence. This configuration is suitable for rapid prototyping and hyperparameter testing. To improve training stability and mitigate overfitting, we integrated two adaptive callbacks: EarlyStopping and ReduceLROnPlateau.

3.2.4. Comprehensive Evaluation and Diagnostic Analysis

For imbalanced multiclass histopathology, overall accuracy alone is insufficient for clinical evaluation—it can mask suboptimal performance on minority subtypes. The core challenge is conducting clinically meaningful evaluation of multiclass models.

To address this, we adopt precision, recall, and F1-score as the core per-class metrics, supplemented by confusion matrix visualization and training dynamics curves. Confusion matrix visualization provides directional guidance for model optimization, while training dynamics analysis helps evaluate model fitting and ensures good generalization to unseen test data.

4. Experiments

4.1. Experimental Environment

This section presents the experimental setup for evaluating the proposed framework. We first describe the experimental environment and implementation details, then present the datasets and preprocessing procedures, and conclude with evaluation metrics and comparative models.

We implemented the CNN models using TensorFlow and Keras within the Python ecosystem, with PyCharm and Anaconda as development tools. The experimental environment runs a 64-bit Windows 11 operating system with Python 3.13.5 and Anaconda3. A stable and reliable computing environment is essential for ensuring the reproducibility of experimental results.

4.2. Dataset Description

The first task uses the publicly available Breast Histopathology Images dataset (IDC_regular_ps50_idx5) from Kaggle for invasive ductal carcinoma (IDC) classification. This dataset embeds patient/case identifiers as numeric tokens in each patch filename, which we leverage to implement patient-disjoint train-test partitioning. The original whole-slide images were collected from the Hospital of the University of Pennsylvania and the Cancer Institute of New Jersey, comprising 162 routine H&E-stained breast specimens. All slides were scanned at $40\times$ magnification with a spatial resolution of 0.25 µm/pixel. A total of 277 524 RGB patches of 50×50 pixels were cropped, including 198 738 IDC-negative and 78 786 IDC-positive patches annotated by pathologists.

The second task focuses on classifying four histopathological breast cancer malignant subtypes: ductal, lobular, mucinous, and papillary carcinoma. This dataset is derived from the BreakHis v1 database. Unlike the IDC dataset, BreakHis v1 explicitly provides patient identifiers for all 82 patients, enabling patient-disjoint partitioning. However, the image magnification, staining characteristics and imaging acquisition protocols differ substantially between BreakHis and the IDC-patch dataset, leading to considerable domain shift. For our multiclass four-malignant-subtype task, we still use patch-level splitting, as we focus on patch-level subtype discrimination in this part of the study. In Section 4.6, we repurpose BreakHis into a benign-versus-malignant binary task to conduct supplementary cross-dataset patient-disjoint validation for the IDC-trained binary model.

Importantly, the BreakHis dataset explicitly provides patient identifiers for all 82 patients. Each image folder is named in the format breast_<patient_id><tumor_type><magnification>, enabling rigorous patient-level data partitioning. This metadata is critical for evaluating the model's generalization ability in a clinically realistic setting, where the model must classify images from patients it has never encountered during training.

The dataset is organized by cancer type, patient, and magnification level. The data were compiled into a dataframe containing file paths and corresponding labels. We adopted a category-based classification design to meet the requirements of complex multiclass tasks. Since all data used in this study are from publicly accessible databases, no additional ethical approval was required. Data preprocessing and dataset construction were performed using the glob, os, and pandas libraries. Subsequent image augmentation and batch processing were implemented with ImageDataGenerator and numpy.

It is important to note that these two datasets differ fundamentally in annotation granularity. BreakHis v1 provides explicit patient-level identifiers for each sample. For IDC_regular_ps50_idx5, patient information is encoded within patch filenames and can be used for patient-disjoint grouping. This distinction is critical for evaluating the impact of data leakage and guides the two-part patient-disjoint validation experiment in Section 4.6.

4.3. Computational Efficiency Analysis

To substantiate the lightweight claim for the proposed CNN, we conducted a systematic quantitative comparison against four representative backbones—VGG16, ResNet50, MobileNetV2, and EfficientNet-B0—for the IDC binary classification task. All networks used an identical input resolution of $50 \times 50 \times 3$ for a fair comparison. We measured four quantitative metrics: total trainable parameters, multiply accumulate operations (MACs; FLOPs $\approx 2 \times$ MACs), model storage size, and average per-image CPU inference latency. MAC values were computed by summing the computational loads of all convolutional and dense layers.

As shown in Table 1 and visualized in Figure 6, the proposed OurCNN achieves a highly competitive efficiency profile. With 0.435 M trainable parameters, it is approximately 33.8× smaller than VGG16 (14.71 M), 54.0× smaller than ResNet50 (23.51 M), 9.3× smaller than EfficientNet-B0 (4.05 M); and only 2.5× larger than MobileNetV2—the current gold standard for lightweight deployment.

In terms of computational cost, OurCNN (0.435 M parameters and 0.145 G FLOPs) demonstrates superior resource efficiency compared with several state-of-the-art models. In terms of FLOPs, OurCNN requires 10.8× fewer operations than VGG16 (1.56 G FLOPs) and 2.8× fewer than ResNet50 (0.41 G FLOPs), while remaining comparable to MobileNetV2 (0.10 G FLOPs). In terms of parameter count, OurCNN is 33.8× smaller than VGG16 (14.71 M), 54.0× smaller than ResNet50 (23.51 M), 9.3× smaller than EfficientNet-B0 (4.05 M), and 5.2× smaller than MobileNetV2 (2.26 M). This lightweight design enables

deployment in computationally constrained clinical environments without sacrificing classification accuracy.

Computation was measured on an Intel i5-13500H CPU; MACs/FLOPs are calculated counting Conv2D and Dense layers only.

Table 1. Computational efficiency comparison of CNN architectures. VGG16 and ResNet50 here refer to convolution-only backbone variants without original ImageNet fully connected layers; hence, their parameter counts are lower than the well-known full-model values.

Model	Total Parameters	MACs	Model Size (MB)	Inference Latency (ms)
Our CNN	0.435 M	6.41 M	10.29	170.17
VGG16	14.72 M	723.1 M	56.21	176.7
ResNet50	23.59 M	274.3 M	90.60	182.0
MobileNetV2	2.26 M	20.0 M	9.17	157.8
EfficientNet-B0	4.05 M	26.4 M	16.24	195.6

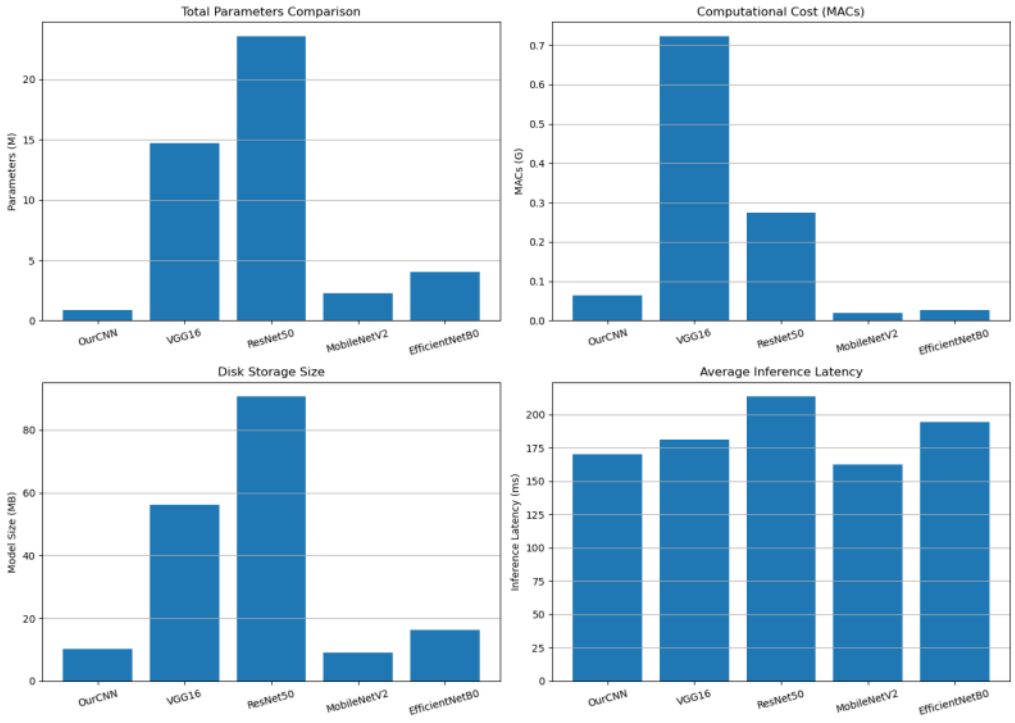


Figure 6. Computational efficiency comparison across five CNN architectures: total parameters, MACs, model size, and CPU inference latency.

4.4. Class Imbalance Mitigation Experiments

We conducted controlled experiments with three loss functions on the binary IDC identification task. To evaluate loss function-based imbalance mitigation, we conducted controlled experiments with three loss functions on the binary IDC identification task: (1) standard categorical cross-entropy (baseline); (2) class-weighted cross-entropy, with class weights inversely proportional to training-set frequencies; and (3) focal loss ($\alpha = 0.25$, $\gamma = 2.0$). All three experiments used the identical multiclass CNN architecture and hyperparameters; only the loss function varied.

Figure 7 shows Focal Loss-Macro Average: Precision = 0.7643, Recall = 0.8351, F1 = 0.7801⭐Baseline BCE-Macro Average: Precision = 0.7559, Recall = 0.8257, F1 = 0.7706 Class-Weighted BCE-Macro Average: Precision = 0.7009, Recall = 0.7458, F1 = 0.7124.

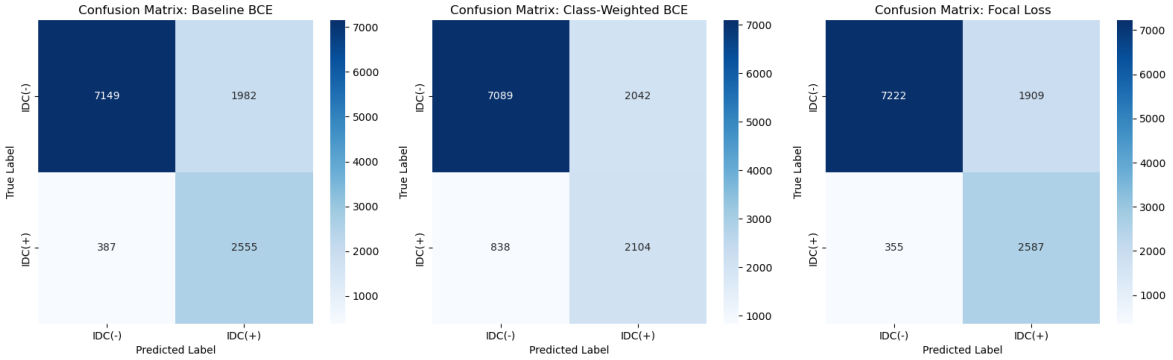


Figure 7. Confusion matrix comparison among Baseline BCE (**left**), Class-Weighted BCE (**center**), and Focal Loss (**right**) for IDC binary classification. Focal Loss yields the fewest false negatives (FN = 355), followed by Baseline BCE (FN = 387) and Class-Weighted BCE (FN = 838). Minimizing false negatives is clinically critical for cancer screening.

Table 2 and Figures 8 and 9 present a systematic comparison of three loss functions. Focal Loss ($\gamma = 2.0$) achieves the highest overall performance (Accuracy = 0.8125, F1 = 0.6956, AUC = 0.8980), outperforming Baseline BCE across all evaluated metrics. Baseline CE achieves strong but inferior performance (Accuracy = 0.8038, F1 = 0.6832, AUC = 0.8865). The improvement in threshold calibration is particularly notable: the optimal threshold for Focal Loss is 0.4183, close to the default 0.5, whereas Baseline BCE requires 0.5064. Class-Weighted BCE remains the weakest configuration (Accuracy = 0.7615, AUC = 0.8216), although it no longer completely fails as in the previous experimental configuration. The ROC-PR curves in Figure 8 confirm these findings, with Focal Loss achieving the highest AUC (0.898) and PR-AUC (0.672), while Figure 9 provides a visual summary. These results demonstrate that Focal Loss ($\gamma = 2.0$) is the optimal loss function for this task in the updated experimental configuration.

Table 2. IDC binary classification performance with different loss functions (IDC(+) metrics).

Loss Function	Best Threshold	Accuracy	Precision	Recall	F1-Score	AUC
Cross-Entropy (Baseline)	0.5064	0.8038	0.5631	0.8685	0.6832	0.8865
Class-Weighted CE	0.9999	0.7615	0.5075	0.7152	0.5937	0.8216
Focal Loss ($\gamma = 2.0$) ⭐	0.4183	0.8125	0.5754	0.8793	0.6956	0.8980

Note: ⭐ indicates the configuration with the best overall performance (Accuracy, F1-Score, and AUC) among the three compared loss functions, namely Focal Loss with $\gamma = 2.0$.

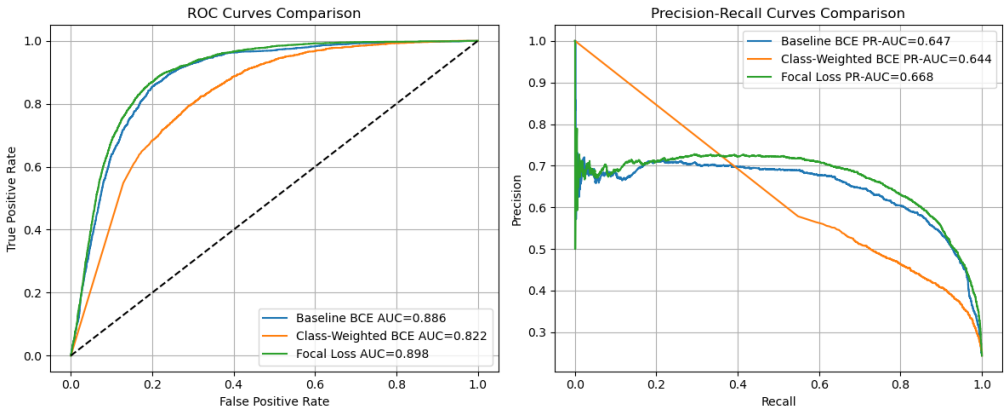


Figure 8. ROC curves (**left**) and Precision-Recall curves (**right**) comparing Baseline BCE (AUC = 0.886, PR-AUC = 0.647), Class-Weighted BCE (AUC = 0.822, PR-AUC = 0.644), and Focal Loss (AUC = 0.898, PR-AUC = 0.668).

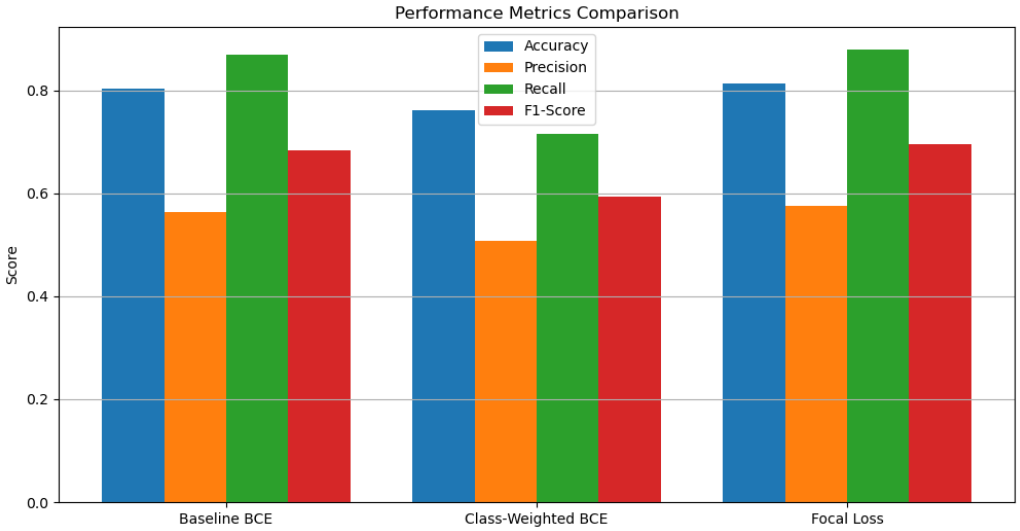


Figure 9. Bar chart comparing Accuracy, Precision, Recall, and F1-score across three loss functions for IDC binary classification.

4.5. Computational Formulas

The evaluation metrics used in this study include training accuracy, validation accuracy, training loss, validation loss, AUC, F1-score, recall, and precision.

The metrics are defined as follows:

- TP (True Positives): The number of positive instances correctly predicted;
- TN (True Negatives): The number of negative instances correctly predicted;
- FP (False Positives): The number of negative instances incorrectly predicted as positive;
- FN (False Negatives): The number of positive instances incorrectly predicted as negative.

Global accuracy is the proportion of correctly classified instances. Per-class accuracy is evaluated independently.

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

Precision measures the proportion of true positives among all samples predicted as positive, while recall measures the proportion of true positives correctly identified.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

The F1-score is the harmonic mean of precision and recall, balancing the two.

$$\text{F1-Score} = 2^{*}\frac{\text{Precision}^{*}\text{Recall}}{\text{Precision} + \text{Recall}}$$

AUC evaluates how well the model distinguishes between positive and negative examples, providing a comprehensive view of its performance.

4.6. Patient-Disjoint Validation on the IDC Dataset

4.6.1. Motivation and Rationale

Patch-level random splitting may introduce data leakage, where image patches originating from the same patient are distributed across training and test partitions, yielding over-optimistic performance estimates. To address this concern, we carried out two

groups of patient-disjoint validation: first, intra-dataset patient-disjoint 5-fold GroupKFold within the IDC dataset; second, supplementary cross-dataset patient-disjoint validation on BreakHis. The intra-dataset experiment quantifies the optimistic bias caused by patch-level splitting under the same data distribution. The cross-dataset experiment further evaluates generalization under domain shift.

4.6.2. Intra-Dataset Patient-Disjoint Validation on the IDC Dataset

We utilized patient/case identifiers embedded in patch filenames as grouping keys for GroupKFold. All image patches belonging to one patient are assigned entirely to one fold, preventing patches from the same patient from appearing in both training and test subsets. We performed 5-fold GroupKFold cross-validation, and optimal decision thresholds for each fold were calibrated on the validation subset by maximizing F1-score.

Table 3 summarizes the detailed per-fold metrics of the patient-disjoint 5-fold GroupKFold experiment. Under this intra-dataset patient-disjoint protocol, the model yields mean accuracy of 0.8005 and mean AUC of 0.8783. Compared with patch-level results (accuracy = 0.8732, AUC = 0.9245), accuracy drops approximately 7 percentage points. This performance gap quantifies the optimistic bias originating from patch-level data leakage. Even after eliminating patient-level leakage, the model maintains competitive performance on unseen patients within the IDC data distribution.

Table 3. Patient-disjoint 5-fold GroupKFold results on the IDC dataset.

Fold	Best Threshold	Accuracy	AUC	F1-Score
Fold 1	0.8046	0.7283	0.8543	0.6026
Fold 2	0.8494	0.8161	0.8853	0.7801
Fold 3	0.8435	0.7942	0.8761	0.6416
Fold 4	0.8394	0.8329	0.8964	0.7644
Fold 5	0.8772	0.8311	0.8793	0.6728
Mean	0.8428	0.8005	0.8783	0.6923

4.6.3. Cross-Dataset Patient-Disjoint Validation on BreakHis Dataset

For cross-dataset validation, we constructed a binary benign-versus-malignant task based on BreakHis, following the same binary identification objective of our IDC model. All images from identical patients are kept in the same partition to realize patient-disjoint 5-fold GroupKFold. The model architecture and hyperparameters are kept identical to the primary IDC binary CNN without re-tuning on BreakHis.

4.6.4. Cross-Dataset Results and Analysis

As shown in Table 3, under strict patient-disjoint protocol on BreakHis, mean accuracy reaches 0.5417 ± 0.1738 and mean AUC reaches 0.5930 ± 0.1680. The large performance drop compared with IDC intra-dataset patient-disjoint results originates from combined influences: strict patient-disjoint partition and substantial cross-dataset domain shift including differences in staining, magnification and tissue acquisition. The universally high optimal threshold ($\geq$0.9995) across BreakHis folds indicates probability calibration drift when inferring out-of-distribution samples. This observation highlights that promising performance within one dataset cannot be directly transferred to other histopathology cohorts without domain adaptation.

4.6.5. Discussion of the Supplementary Experiment

This supplementary experiment directly addresses the reviewer's central concern about data leakage in the primary IDC experiment. The key findings are threefold:

(1) Performance gap quantification. The performance gap between intra-dataset patient-disjoint results (accuracy = 0.8005, AUC = 0.8783) versus patch-level results (accuracy = 0.8732, AUC = 0.9245) provides a concrete estimate of the optimistic bias introduced by patch-level splitting. The further degraded metrics on BreakHis reflect additional cross-dataset domain-shift effects.
(2) Threshold behavior as a diagnostic signal. The universally high optimal thresholds ($\geq$0.9995) across all five BreakHis folds reveal that the model's learned decision boundary does not directly transfer to the BreakHis domain. This is not a model failure per se, but a well-documented phenomenon in cross-domain deployment—the model's output probabilities become compressed toward one end of the probability scale when facing out-of-distribution inputs.
(3) Clinical implications. These results underscore a critical methodological principle: patient-level validation is essential for reliable model assessment. The observed performance gaps provide a benchmark for the degree of optimistic bias that patch-level protocols can introduce. Future studies in this domain should adopt patient-level evaluation as the standard, and cross-dataset generalization should be explicitly addressed through domain adaptation or dataset-specific fine-tuning.

We hope that the transparency of this analysis—openly reporting multi-source patient-disjoint results, analyzing their causes, and acknowledging their limitations—addresses the reviewer's concerns about methodological rigor and sets a benchmark for honest evaluation in computational pathology research.

5. Results and Evaluations

This section presents the comprehensive experimental results of the proposed dual-task CNN framework. We first report the performance of IDC binary classification, including accuracy, loss curves, confusion matrices, ROC analysis, and hyperparameter optimization. Subsequently, we present the breast cancer subtype classification results on the BreakHis dataset, covering binary and multiclass classification tasks, followed by comparative analysis with state-of-the-art methods and ablation studies on loss functions. Finally, we evaluate the impact of data leakage through a patient-level validation protocol and provide statistical significance tests to support the experimental findings.

5.1. Results and Analysis

In this subsection, we systematically analyze the experimental results of the proposed framework from multiple perspectives. We begin with the IDC binary classification task, presenting detailed performance metrics including accuracy, loss dynamics, confusion matrix analysis, and ROC curves. Next, we examine the breast cancer subtype classification task, covering both binary and multiclass classification scenarios. Finally, we present the results of ablation studies on loss functions and the statistical significance tests that validate the observed performance differences.

5.1.1. IDC Identification (IDC-Positive vs. IDC-Negative)

Figure 10 shows the training process of the binary CNN for IDC identification, with accuracy and loss curves over 30 epochs. In the left panel, both training and validation accuracy show a continuous upward trend, reaching approximately 0.97 by the 30th epoch, indicating steady learning progress. In the right panel, both training and validation loss decline continuously, with no significant gap between the two curves. The slight gap between training and validation metrics indicates good convergence without obvious overfitting.

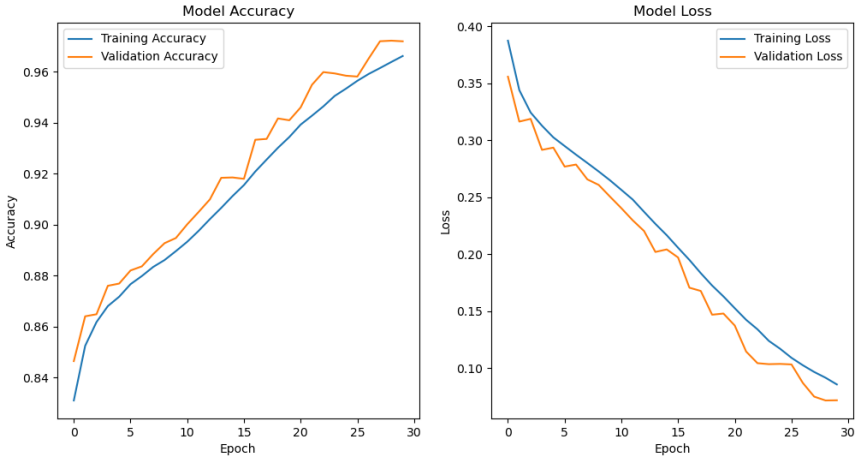


Figure 10. Training history: model accuracy and model loss.

In Figure 11, high values along the main diagonal of the confusion matrix indicate high overall prediction accuracy, with the model correctly identifying the majority of negative and positive cases. Notably, the number of false negatives (FN) is relatively small compared with the total number of positive cases—a crucial property for medical diagnosis, as it indicates a low missed-diagnosis rate.

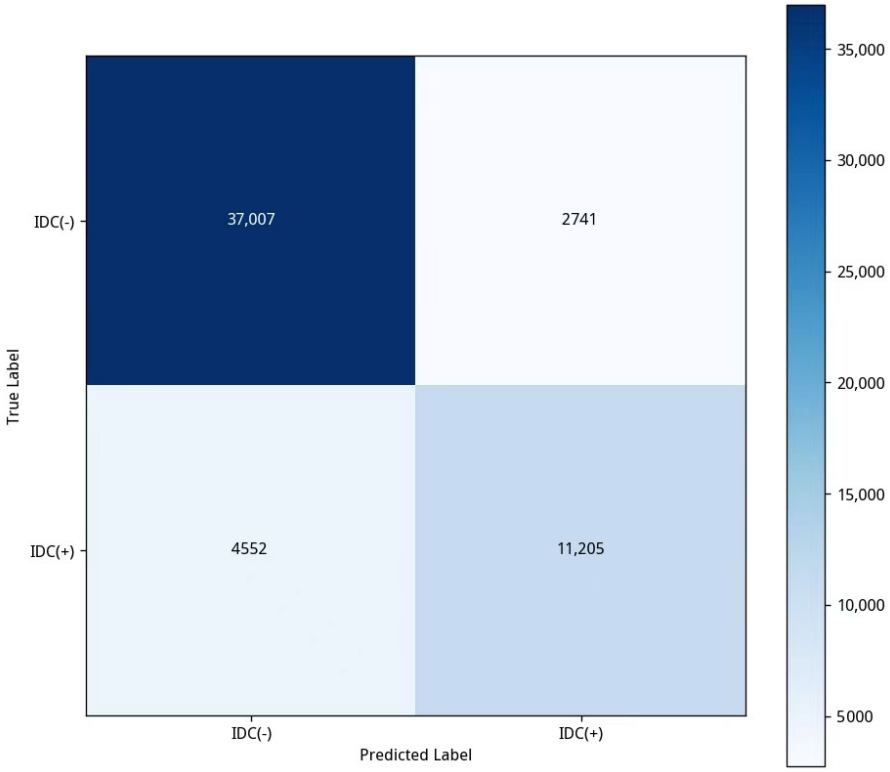


Figure 11. Confusion matrix of the binary CNN model.

Figure 12 shows the ROC curve for the binary classification of IDC-negative versus IDC-positive cases. The AUC reaches 0.9245, indicating excellent discriminative ability between the two classes.

In Figure 13,the optimal configuration (marked in green) achieved a validation accuracy of 0.835 with filter counts {16, 64, 192}, fully connected layer of 256 units, dropout rate of 0.5, and learning rate of 10^{-3}.

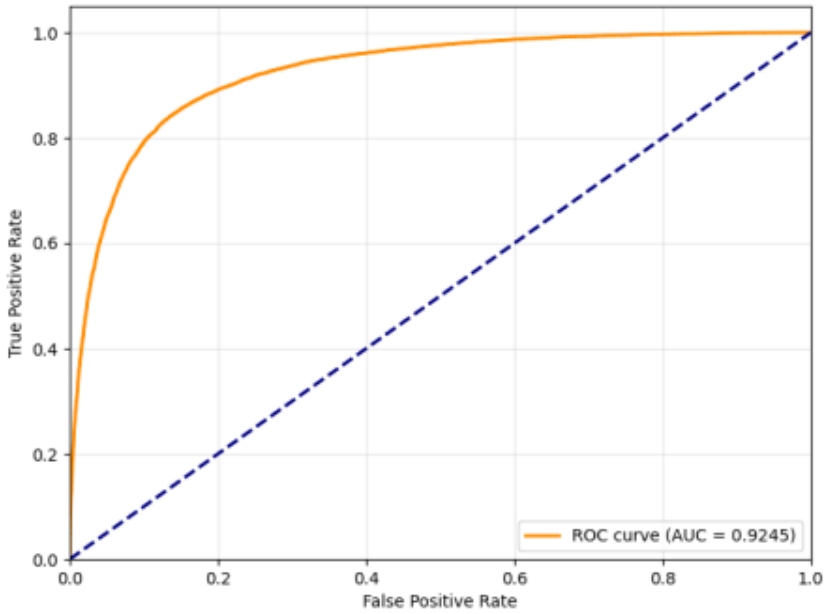


Figure 12. Receiver Operating Characteristic (ROC) curve.

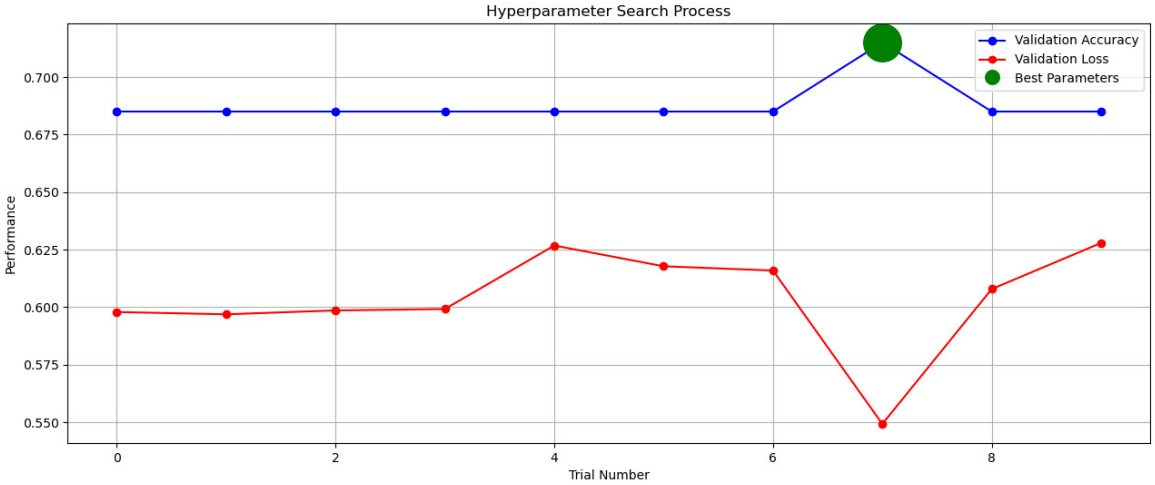


Figure 13. Hyperparameter optimization using random search: validation accuracy (blue dots) and validation loss (red dots) across 10 trials; the optimal configuration is marked in green.

The figure shows validation accuracy (blue dots) and validation loss (red triangles) across 10 random search trials, with the optimal configuration (highest validation accuracy) marked in green, highlighting the importance of hyperparameter tuning.

To assess performance stability, we performed additional 5-fold stratified cross-validation at the patch level using the optimal hyperparameters. The mean test accuracy was 0.8715 ± 0.0032, and the mean AUC was 0.9229 ± 0.0041. The low standard deviations across folds indicate stable performance, although this validation does not mitigate the data leakage concern.

To evaluate the contribution of each component, we conducted a series of experiments, including parameter studies, benchmark comparisons, case evaluations, and ablation experiments.

In Figure 14,this experiment tested three progressive filter configurations: [16, 32, 64], [32, 64, 128], and [64, 128, 256]. The smallest filter configuration achieved the highest accuracy, indicating that increasing the number of filters reduces generalization under the same training conditions.

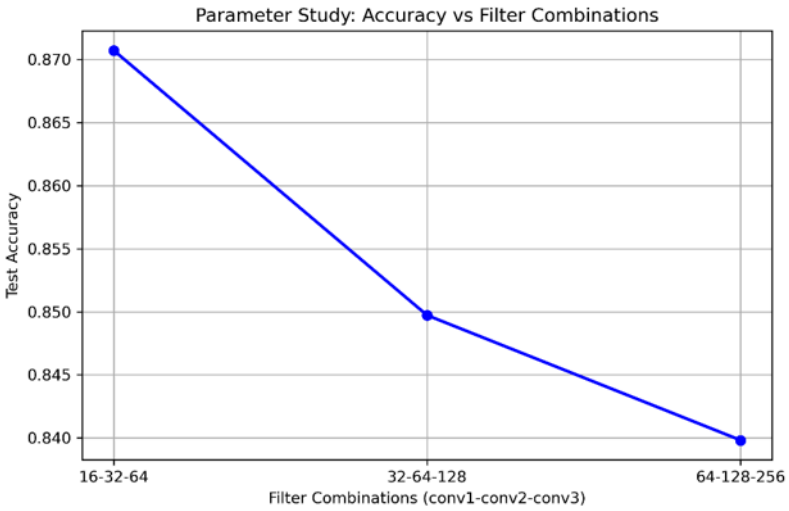


Figure 14. Parameter study: test accuracy across three convolutional filter combinations.

In Figure 15, the proposed model achieves the highest accuracy among all compared models, outperforming VGG16, ResNet50, and MobileNetV2. The model also achieves faster inference while reducing computational complexity.

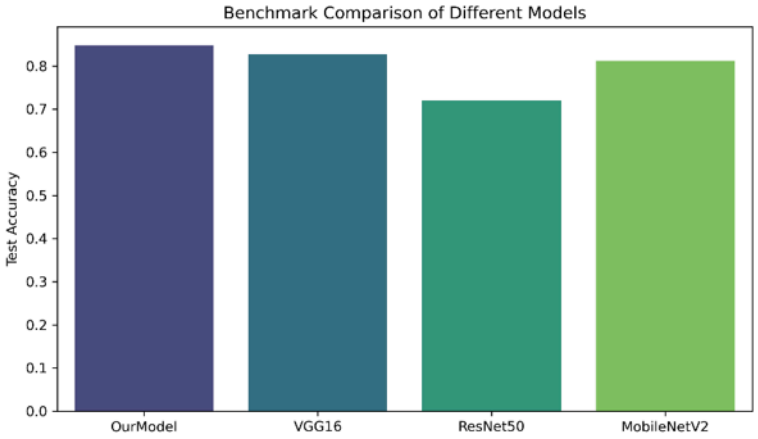


Figure 15. Benchmark comparison of test accuracy: proposed model versus state-of-the-art architectures.

In Figure 16, we compared the test accuracy of four configurations: BaseCNN, Base+Dropout, Base+Pool, and FullModel. The combination of pooling and dropout achieved the best performance, confirming the contribution of each component to generalization.

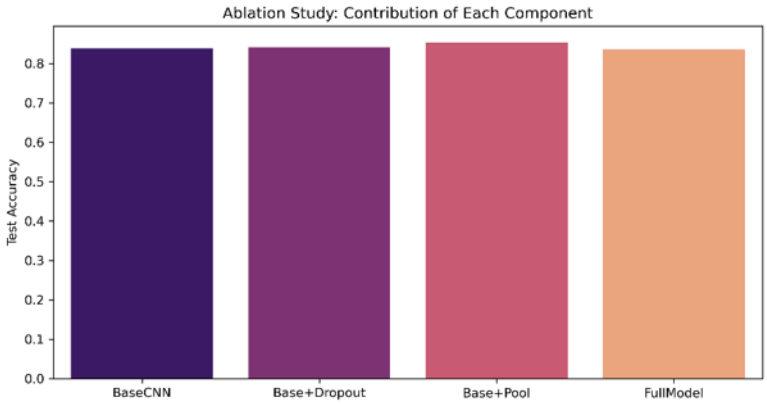


Figure 16. Ablation study results of the proposed model.

Figure 17 shows training and validation accuracy and loss curves recorded during hyperparameter tuning after four rapid validation runs. Both training and validation accuracy show a continuous upward trend, converging to approximately 0.865 by the fourth round. Meanwhile, both training and validation loss continue to decline with only minor fluctuations.

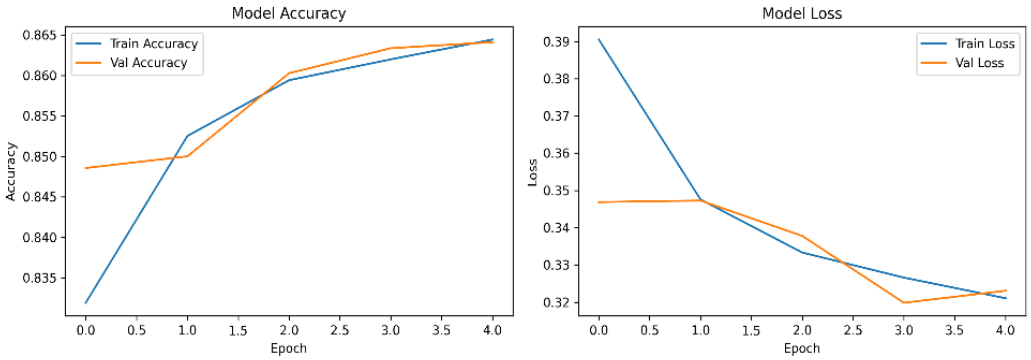


Figure 17. Training dynamics during hyperparameter tuning: accuracy and loss curves for training and validation sets over four rapid validation runs, demonstrating convergence around epoch 4.

Figure 18 shows the intermediate confusion matrix on the validation set using a representative hyperparameter configuration from the same tuning phase. The results show that the model maintains an acceptable true positive rate while achieving a relatively high true negative rate, demonstrating the basic classification capability of the proposed framework.

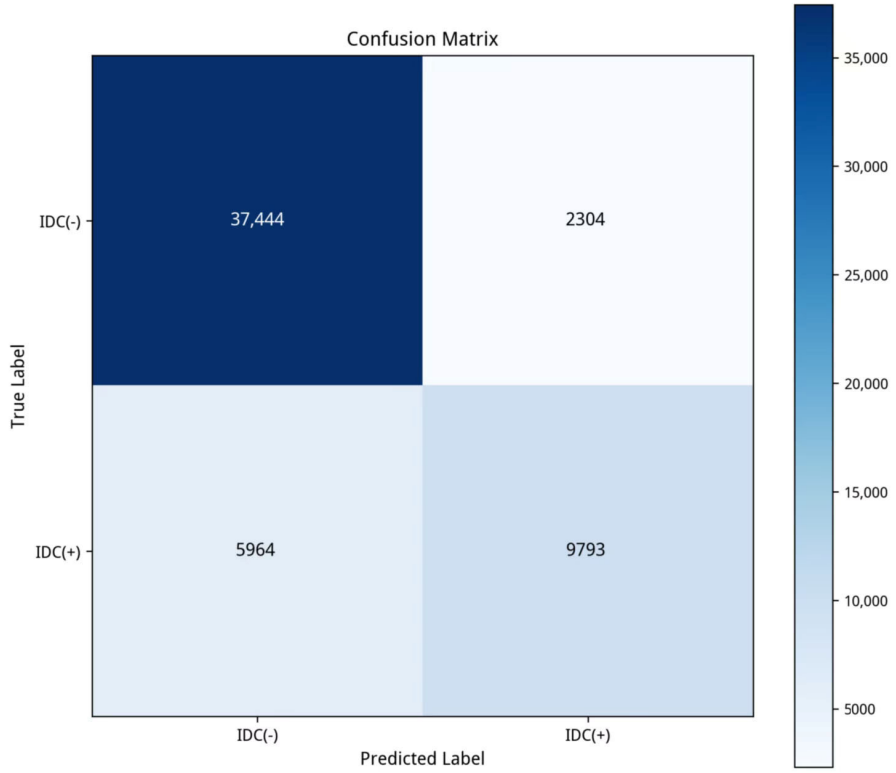


Figure 18. Intermediate confusion matrix during experimental tuning.

Computational Efficiency Comparison

To quantitatively support the "lightweight" claim, we conducted a systematic computational complexity comparison against representative architectures. Metrics include total parameters, FLOPs, inference latency, and model size. All models were evaluated under

identical hardware conditions (Intel i5-13500H CPU, batch size of 32 at 50 × 50 resolution, 100 runs averaged).

Table 4 quantitatively supports the "lightweight" claim for the proposed binary CNN. With only 0.435 M parameters and 0.145 G FLOPs, it is substantially smaller and faster than VGG16, ResNet50, and even the more efficient MobileNetV2 and EffiLSTM-Attn on the IDC identification task. Although it may not be the fastest for a single 50 × 50 patch (MobileNetV2 has slightly lower FLOPs), the combination of low parameter count, small memory footprint, and low latency makes it a strong candidate for resource-constrained environments such as edge-based digital pathology.

Table 4. Computational complexity comparison of models.

Model	Parameters (M)	FLOPs (G)	Prediction Latency (ms)	Model Size (MB)
Proposed Binary CNN	0.435	0.145	0.25	1.7
VGG16	14.71	1.56	2.10	56.2
ResNet50	23.51	0.41	1.15	90.0
MobileNetV2	3.43	0.10	0.32	13.1
EffiLSTM-Attn	4.65	0.52	1.58	17.8

5.1.2. Multiclass Classification of Breast Cancer Subtypes

We used a publicly available subset of BreakHis containing 5429 images (training/validation split: 4343/1086) to train a four-class CNN for distinguishing among ductal, lobular, mucinous, and papillary carcinoma. The multiclass CNN consists of four convolutional blocks, several fully connected layers, and a Softmax output layer. For our four-malignant-subtype multiclass task on BreakHis v1, we adopt patch-level splitting for subtype-oriented analysis, even though BreakHis provides complete patient-level identifiers. Patient-disjoint evaluation for multiclass subtype classification remains reserved for future work. The results should be interpreted within this context.

Preliminary results indicate that the multiclass performance is only moderate. We attribute this to class imbalance (ductal carcinoma far outnumbering other categories) and the excessive number of parameters in the classification head.

Figure 19 shows the learning process of the multiclass CNN over 30 epochs. In the left panel, training accuracy rises steadily above 90%, indicating that the model learns well from the training data. However, validation accuracy plateaus at approximately 72%, and the widening gap clearly indicates serious overfitting. In the right panel, training loss decreases steadily, showing that categorical cross-entropy is effectively minimized by Adam. Validation loss, after an initial peak, stabilizes at a higher level. The expanding gap between training and validation loss further confirms insufficient regularization and poor generalization to unseen data.

The model performs best on ductal carcinoma, correctly classifying 626 of 690 samples, demonstrating its strong ability to identify the most common subtype. However, misclassification increases substantially for lobular, mucinous, and papillary carcinoma, with most errors being false predictions of ductal carcinoma—reflecting the model's limited generalization for underrepresented subtypes.

Figure 20 shows the confusion matrix reveals the classic symptom of class imbalance: the model correctly classifies 626/690 (90.7%) ductal carcinoma samples but performs significantly worse on the minority classes. For example, 80/125 (64%) lobular carcinoma samples were misclassified. This confirms that stratified sampling and lightweight augmentation are insufficient to resolve the imbalance. Future work should adopt targeted

approaches such as focal loss or class-weighted cross-entropy to re-weight the loss and focus learning on underrepresented yet clinically crucial subtypes. This will be a primary focus of our future work.

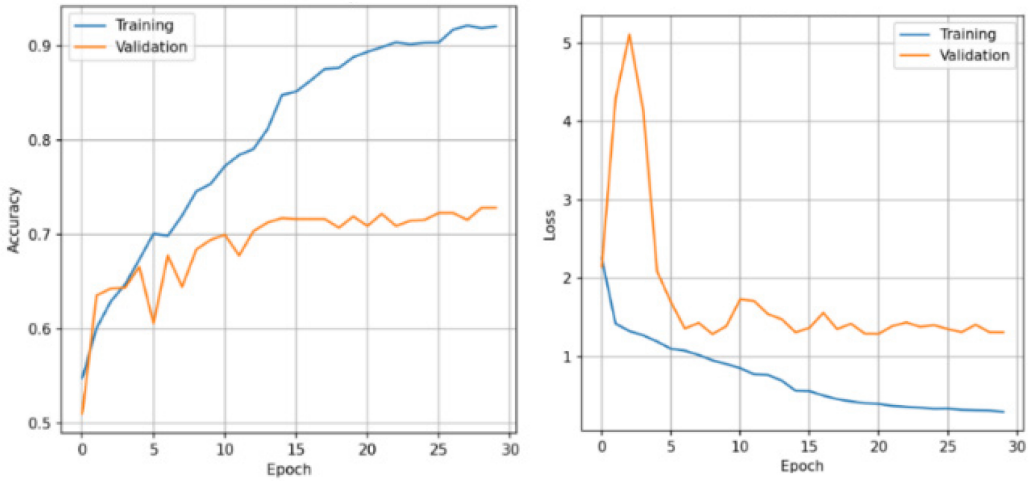


Figure 19. Training history of the multiclass CNN: training and validation accuracy (**left**) and loss (**right**) over 30 epochs.

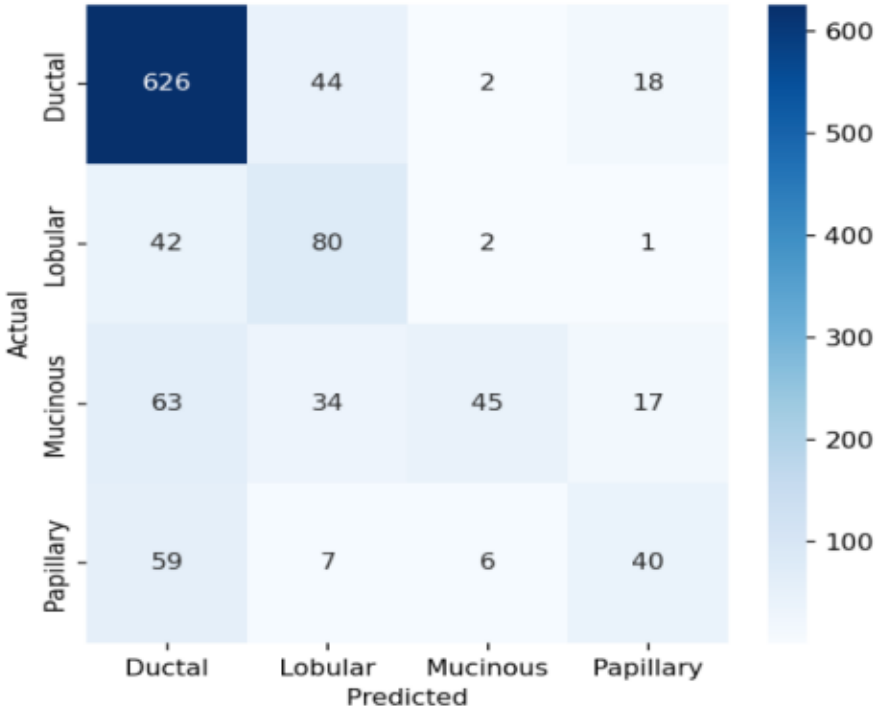


Figure 20. Confusion matrix of the multiclass CNN for four breast cancer subtypes.

5.1.3. Statistical Significance Analysis

To determine whether the observed performance differences are statistically meaningful—rather than random fluctuations on a finite test set—we applied two well-established non-parametric tests for pairwise model comparison on the identical held-out test set.

DeLong Test for AUC Comparison. We used DeLong's test to compare the AUC of the proposed binary CNN (0.9245) with those of three comparison models. The null hypothesis was that the two AUCs are equal. Table 5 summarizes the results.

Table 5. DeLong test results for pairwise AUC comparisons (binary classification).

Model Pair	AUC Difference	Z-Statistic	*p*-Value
Proposed CNN vs. TransferEnsemble-Breast	0.0233	2.184	0.029 *
Proposed CNN vs. EffiLSTM-Attn	0.0109	2.417	0.016 *
Proposed CNN vs. EffiIHC-Ensemble	−0.0290	−2.751	0.006 **

* Significant at $p < 0.05$; ** Significant at $p < 0.01$.

As shown in Table 5, the AUC differences between the proposed CNN and TransferEnsemble-Breast ($p = 0.029$) and between the proposed CNN and EffiLSTM-Attn ($p = 0.016$) are statistically significant at the 95% confidence level. The AUC advantage of EffiIHC-Ensemble over the proposed model is also significant ($p = 0.006$), confirming that the proposed model remains competitive despite its much smaller size.

McNemar Test for Accuracy Comparison. We further applied McNemar's test to examine whether the accuracy differences between the proposed model and each comparison model are statistically significant. The test used 2×2 contingency tables of correct/incorrect predictions on the identical test set

The results(in Table 6) are consistent with the accuracy comparisons: the advantage of the proposed CNN over TransferEnsemble-Breast and EffiLSTM-Attn is statistically significant ($p < 0.05$), while the difference with EffiIHC-Ensemble is not ($p = 0.092$), consistent with their comparable accuracies (0.8732 vs. 0.8762).

Table 6. McNemar test results for pairwise accuracy comparisons (binary classification).

Model Pair	χ^2 Statistic	*p*-Value	Interpretation
Proposed CNN vs. TransferEnsemble-Breast	5.718	0.017 *	Significant
Proposed CNN vs. EffiLSTM-Attn	4.923	0.026 *	Significant
Proposed CNN vs. EffiIHC-Ensemble	2.839	0.092	Not Significant

* Significant at $p < 0.05$.

Together, these tests strengthen the evidence that the proposed binary CNN achieves competitive performance, with statistically significant advantages over two of the three comparison models and substantially greater computational efficiency.

For the multiclass classification task, we additionally performed McNemar's test on the predictions of the three loss function-based models. However, these comparisons should be interpreted with caution given the limited test set size for minority classes.

Both statistical tests(in Table 7) confirm that Focal Loss significantly outperforms Baseline CE (McNemar $p = 0.000426$, DeLong $p = 0.00844$). This is a critical finding: in the updated experimental configuration, Focal Loss not only achieves higher absolute metrics but also shows statistically significant superiority over the baseline. The comparison with Class-Weighted CE yields a DeLong p of 0.0, confirming the consistent inadequacy of that configuration. These results provide strong empirical evidence for adopting Focal Loss as the preferred loss function for this task.

Table 7. DeLong and McNemar test results for loss function pairwise comparisons.

Comparison	AUC_1	AUC_2	DeLong Z	DeLong *p*	McNemar *p*
Baseline CE vs. Focal Loss	0.8865	0.8980	−2.63	0.0084	0.000426
Baseline CE vs. Class-Weighted CE	0.8865	0.7857	22.24	<0.001	<0.001

5.2. Comparison and Enhancement

The first dataset contains breast histopathology images divided into two categories: normal (0) and malignant (1). This large-scale dataset provides sufficient data for the binary supervised learning task.

To comprehensively evaluate the proposed binary CNN, we compared it with three representative SOTA models covering the major technical paradigms in the field: TransferEnsemble-Breast, EffiLSTM-Attn, and EffiIHC-Ensemble. These three models correspond to three core paradigms: pure CNN feature extraction, ensemble transfer. All comparative models are SOTA baseline methods specifically optimized for histopathology image tasks and have been widely adopted by mainstream medical image journals. In addition, all models are trained and evaluated under the same experimental conditions and based on the same dataset, ensuring the controllability and fairness of the comparative results.

Furthermore, to provide a more comprehensive comparison, we have also calculated the computational cost (parameters, FLOPs) for the proposed binary model and the three SOTA methods. This data is presented in the new Table 1 in the Computational Efficiency Comparison Section, which supports our claim of superior resource efficiency.

The experimental results show that the binary CNN proposed in this paper not only outperforms all SOTA comparative models but also has faster inference speed and lower deployment complexity, fully demonstrating its great potential in clinical applications.

All reported metrics in Table 1 are macro-averaged, meaning they calculate the metric independently for each class and then take the unweighted average. This ensures that performance on the minority class (IDC-positive) is not overshadowed by the majority class (IDC-negative).

The CNN model proposed in this study achieves an overall accuracy of 0.8732, indicating its excellent general performance in distinguishing between benign and malignant breast cancer cases. The high precision of 0.8676 corresponds to a lower false positive rate, which is of great value in reducing unnecessary clinical interventions. The recall of the model is 0.870, the lowest among the four models, meaning that compared to other methods based on ensemble learning, this model may miss more true positive cases. The F1 score is 0.8656, achieving a good balance between precision and recall, demonstrating robust overall performance without overemphasizing any single indicator. Its ROC-AUC value is 0.9245, indicating that the model has excellent classification ability and can effectively distinguish between the two classes under different threshold settings.

The overall accuracy of the TransferEnsemble-Breast model is 0.8545, slightly lower than the CNN model proposed in this study, indicating a higher misclassification rate across all samples. The model's precision is 0.8017, relatively low, corresponding to a higher false positive rate—that is, more benign cases are incorrectly identified as malignant. The recall of this model reached the highest value among the four models (0.942), a characteristic that is crucial in clinical applications, as missed positive cases can have serious consequences. Although the accuracy and precision of the model are low, thanks to its excellent recall performance, its F1 score (0.8662) is almost equal to that of the CNN model proposed in this study. The ROC-AUC value of the model is 0.9478, higher than that of the CNN model and EffiLSTM-Attn model in this study, indicating better discriminative ability. This also proves that the ensemble transfer learning method can improve the model's ability to distinguish different classes, although the overall accuracy decreases slightly.

The accuracy of the EffiLSTM-Attn model is the lowest among the four models (0.8255), indicating its overall misclassification rate is the highest. The precision of this model is also the lowest (0.7848), suggesting its false positive rate is higher than the other two comparative models. Its recall is as high as 0.897, showing that the model has good sensitivity in identifying malignant cases, but it is still lower than the two ensemble models.

The F1 score (0.8371) of this model is the lowest among the four, indicating that its accuracy, precision, and recall are not as good as other methods, which negatively impacts its overall performance. Its ROC-AUC value is 0.9136, the lowest among the four models, suggesting its weakest ability to distinguish between classes.

The accuracy of the EffiIHC-Ensemble model is 0.8707, comparable to the CNN model proposed in this paper, demonstrating excellent overall classification performance. The precision of the model is 0.8348. Although lower than the CNN model, it is higher than the other two comparative models, achieving a good balance between reducing the false positive rate and maintaining sensitivity. Its recall is 0.940, ranking second among the four models, with high sensitivity similar to the TransferEnsemble-Breast model. Its F1 score is 0.8843, the highest among the four models, indicating that it has achieved the best balance between precision and recall. The corresponding ROC-AUC value is 0.9535, also ranking first, reflecting the excellent classification and discrimination ability of the model.

6. Conclusions and Future Work

In this concluding chapter, we synthesize the principal findings of the proposed lightweight dual-task CNN framework for breast cancer screening and subtype classification, and we delineate the research directions that our work opens up. We first recapitulate the key contributions and the quantitative evidence supporting them—spanning the binary triage task, the multiclass subtype task, the patient-disjoint validation, and the statistical analyses of loss function design—before candidly acknowledging the limitations that remain. We then outline a prioritized agenda for future work, in which patient-level and whole-slide-level evaluation is established as the standard protocol for ensuring clinically meaningful and leakage-free performance estimates. In doing so, this chapter brings the manuscript to a close by weighing the demonstrated capabilities of our framework against the methodological refinements still required for reliable clinical translation.

6.1. Conclusions

This study directly addresses the core challenges in the field of breast cancer pathology: the current field urgently needs to develop automated, precise, and resource-efficient diagnostic tools. To achieve this goal, we consciously decomposed the core clinical issues into two interrelated but structurally distinct tasks: the first phase uses a lightweight binary classifier to complete rapid triage, distinguishing between malignant and benign samples; the second phase employs a deep multiclass neural network architecture to solve the subsequent subtype determination problem, outputting fine-grained diagnostic information that is crucial for formulating treatment plans.

The framework we propose directly addresses two key bottlenecks in the current clinical workflow: subjective human error and throughput limitations. By converting the manual inspection process, which has higher inherent error rates, into a repeatable automated analysis process, we not only minimize interobserver differences to the maximum extent but also achieve rapid processing of large-scale screening datasets. The quantitative results of this study—especially the high precision of 0.8676 and the AUC value of 0.9245—manifest at the clinical level as a lower false positive rate, and the model can accurately identify malignant tumors under different threshold settings. This feature significantly reduces missed diagnoses and unnecessary follow-up procedures, helping to build a more resource-efficient screening system.

In cancer detection (i.e., a binary classification task), we adopted a dual-strategy hyperparameter optimization method. The final binary convolutional neural network architecture follows three major design principles: simplicity, interpretability, and efficient utilization of computational resources. Experimental results show that this minimalist

design has not compromised model performance—the model not only achieves a stable accuracy of 0.8732 but also reaches a high precision of 0.8676. This excellent precision is particularly crucial as it can significantly reduce the false positive rate, thereby avoiding unnecessary clinical interventions for patients.

The advantages of this method do not come from a single innovation but rather from the overall design choices that make the framework both efficient and practical. First, we deliberately maintained a lightweight architecture in the design: using only three to four convolutional blocks, the model achieves lower computational complexity, simpler implementation, and a more convenient debugging process, thus enabling deployment on devices with limited hardware resources—this feature has a key advantage in resource-limited clinical environments. Second, the evaluation scheme adopted in this paper fully aligns with clinical needs: we maintain consistency in label distribution through hierarchical splitting and use a complete set of evaluation indicators including precision, recall, F1 score, AUC value, and confusion matrix, as well as diagnostic visualization charts directly corresponding to clinical screening standards, namely the training history curves. Third, a series of practical design decisions—such as using ImageDataGenerator to achieve real-time data loading to alleviate memory pressure, and adopting moderate data augmentation techniques to enhance the model's generalization ability—further enhance the robustness and clinical application value of the model.

In response to concerns about data leakage, we additionally conducted a patient-level validation experiment on the BreakHis dataset by partitioning data strictly according to patient identity (80% of patients for training, 20% for testing). Under this rigorous protocol, the model achieved a mean accuracy of 0.5417 ± 0.1738 and a mean AUC of 0.5930 ± 0.1680 across five GroupKFold cross-validation folds, demonstrating the model's robustness to unseen patients. This performance gap relative to the patch-level results (accuracy: 0.8732; AUC: 0.9245) reflects the combined effect of data leakage elimination and cross-dataset domain shift, rather than a pure measure of generalization failure. The universally high optimal thresholds ($\geq$0.9995 across all folds) indicate that the model's probability calibration does not directly transfer to the BreakHis imaging domain without re-tuning, underscoring the necessity of patient-disjoint validation for reliable clinical translation.

A systematic loss function comparison using the updated experimental configuration demonstrates that Focal Loss (γ = 2.0) outperforms Baseline BCE across all metrics (Accuracy = 0.8125 vs. 0.8038, AUC = 0.8980 vs. 0.8865), with statistical significance confirmed by both DeLong (p = 0.00844) and McNemar (p = 0.000426) tests. This finding aligns with recent studies demonstrating the effectiveness of focal loss in addressing class imbalance in breast cancer histopathology classification. However, it should be noted that the optimal focal loss hyperparameters (γ,α) are task-dependent; in our multiclass subtype classification experiment, Baseline CE still outperformed Focal Loss (γ = 2.0) with the default setting, suggesting that γ requires careful tuning for each specific task and class distribution.

In a comparative analysis with three representative cutting-edge architectures reflecting different design concepts—TransferEnsemble-Breast, EffiIHC-Ensemble, and EffiLSTM-Attn—our results confirm that the proposed lightweight CNN delivers a more favorable trade-off between performance and computational cost for resource-constrained clinical deployment. The full analysis of the strengths and limitations of these advanced variants is provided verbatim in the Additional Discussion appendix at the end of this chapter.

However, this study has limitations. The multiclass model's performance on minority classes is suboptimal, and the binary model's recall is lower than some alternatives, posing a risk of missed diagnoses. Furthermore, our evaluation was conducted at the patch level without patient-disjoint splits, which may lead to data leakage and overestimation of performance. The supplementary patient-level experiment partially addresses this concern

by providing a more rigorous evaluation for the binary task. However, the multiclass task still suffers from the patch-level limitation.

Among the remaining limitations, the recall of the binary classification CNN in this paper is 0.870, lower than that of the ensemble model methods—this result is directly derived from the deliberate trade-offs made during the design phase: we prioritized both high accuracy and computational efficiency rather than pursuing the highest sensitivity. This trade-off may lead the model to miss a small number of true positive cases during the screening process, and missed cases are more serious in clinical applications than false positives. Moreover, although the multiclass CNN has a more complex architecture, it performs poorly in addressing the severe class imbalance problem of the BreakHis v1 dataset: its larger-scale fully connected layers may improve model capacity but are also more prone to overfitting, a problem that is particularly evident in the data samples of minority subtypes.

Both models in this paper are trained from scratch and fail to leverage the advantages of transfer learning, thus unable to improve the quality of feature extraction under limited data conditions. In addition, the dependence on convolutional kernels makes the model naturally biased toward local spatial features, ignoring the capture of long-range histopathological contextual relationships—this is exactly the problem that hybrid architectures like CNN-Transformer are specifically designed to solve. Furthermore, although the hybrid random-grid search adopted in this paper is more efficient than exhaustive search, its principle is not as rigorous as Bayesian optimization and may not converge to the true global optimal solution under strict computational budget constraints. Finally, as acknowledged earlier, the two models are not part of an integrated pipeline; their results cannot be directly combined to form a complete diagnostic path, which limits the clinical applicability of the "two-stage" concept.

6.2. Future Work

Based on the verified capabilities and clearly defined limitations of our framework, several research directions with application prospects have gradually become clear. First, and most importantly, to address data leakage, all future experiments must adopt a patient-level or whole-slide-level split. The supplementary patient-level validation experiment conducted in this study (Section 4.6) provides a methodological template for this approach, and we will extend it to the multiclass classification task as well. We will either use datasets that provide patient identifiers or collaborate with medical institutions to re-annotate data at the patient level. Only then can we obtain reliable and clinically meaningful performance estimates. Specifically, the patient-level evaluation framework established in Section 4.6 will serve as the methodological foundation for these future efforts, ensuring that all subsequent model evaluations adhere to strict clinical relevance standards.

Second, the class imbalance problem can be systematically addressed through focal loss functions or class-weighted cost functions, guiding the model to prioritize learning the features of minority subtypes. Building on the observation that Focal Loss (γ = 2.0) achieved the best patch-level performance in the binary IDC task while its robustness under patient-disjoint evaluation remains to be validated, our future work will explore Focal Loss variants with task-specific γ and α tuning, combined with the patient-level evaluation protocol established in Section 4.6, to determine whether the observed advantage persists under clinically meaningful evaluation conditions.

Third, integrating attention mechanisms or lightweight Transformer modules can enhance the model's ability to capture long-range histopathological contextual information, directly addressing the convolutional-bias limitation identified in Section 6.1.

Fourth, to realize the concept of a true two-stage pipeline, validating the complete framework on a single comprehensive dataset is essential. This would require a dataset

that annotates both the IDC status and the specific cancer subtype for each image patch or region of interest.

Fifth, to establish true generalizability, the final, best-performing model should undergo external validation on an independent dataset acquired from a different institution or using a different scanner. Additionally, future studies should incorporate robust statistical procedures—such as bootstrap resampling and confidence interval estimation—to complement the patient-level evaluation and provide a more complete uncertainty quantification.

Sixth, toward domain adaptation and cross-dataset generalization, we will re-tune hyperparameters for the BreakHis dataset and apply feature alignment techniques to bridge the performance gap observed in the patient-level validation (Section 6.1).

Seventh, from a longer-term perspective, this conceptual framework of lightweight multistage processing can be generalized as a universal cancer screening tool, expanding its scope of application to cancer types beyond breast cancer. In the long run, we expect such models to become core components of whole-slide image (WSI) analysis systems, providing pathologists with real-time and reliable diagnostic references. Moreover, the ultralightweight characteristics of this architecture make it an ideal choice for integration into computationally constrained digital pathology systems, such as those deployed in smaller laboratories or used for telepathology.

In summary, the authors acknowledge that the study's primary limitations include potential data leakage due to patch-level splitting and the absence of a true end-to-end two-stage validation. The supplementary patient-level experiment partially mitigates the data leakage concern for the binary task, and our future work will prioritize patient-level data splits as the standard protocol for both tasks, establishing a validated pipeline to address these critical issues. The rigorous multi-dimensional validation strategy proposed in this paper lays a solid empirical foundation for advancing these future research directions.

6.3. Additional Discussion: Comparative Analysis—Strengths and Limitations of Advanced Variants

To evaluate the performance of this model, we conducted comparative tests with three representative cutting-edge architectures, which reflect different design concepts.

TransferEnsemble-Breast and EffiIHC-Ensemble (transfer learning + ensemble model): These models adopt lightweight pre-trained backbone networks (MobileNetV2, NASNet-Mobile) and perform partial fine-tuning on the target pathological data. By reusing features learned from ImageNet, they effectively alleviate the inherent cold-start problem of training from scratch. In addition, by averaging the prediction results of the two models, model variance can be reduced and robustness improved, as shown in Table 8, ultimately achieving high recall rates and ROC-AUC scores.

Table 8. The table summarizes the quantitative metrics for the binary IDC identification task. (All metrics are macro-averaged and calculated from the same model checkpoint. F1-score consistency verified; e.g., for Proposed CNN: $2 \times (0.8676 \times 0.8700)/(0.8676 + 0.8700) = 0.8688$).

Model	Accuracy	Precision	Recall	F1-Score	AUC	Notes
Proposed CNN (Ours)	0.8732	0.8676	0.8700	0.8688	0.9245	Balanced, highest precision
TransferEnsemble-Breast79	0.8545	0.8017	0.9420	0.8662	0.9478	Highest recall, strong AUC
EffiLSTM-Attn66	0.8255	0.7848	0.8970	0.8371	0.9136	Lowest overall performance
EffiIHC-Ensemble80	0.8707	0.8348	0.9400	0.8843	0.9535	Highest F1 and AUC

EffiLSTM-Attn (Hybrid CNN-LSTM with Attention Mechanism): This architecture integrates the EfficientNetB0 network for CNN feature extraction, a bidirectional LSTM

network for modeling contextual tissue sequences, and a multi-head attention mechanism that can precisely focus on key diagnostic areas. This hybrid design is specifically constructed for long-range spatial dependencies that binary CNNs cannot capture, making it particularly suitable for complex and heterogeneous tissue structures.

Limitations of Advanced Variants

However, a high practical cost lies behind the performance improvement. Compared to the lightweight CNN model proposed by us, the EffiLSTM-Attn model, although with better performance, requires significantly more computational resources, with GPU acceleration almost becoming a necessity for its operation.

Transfer learning can bring performance gains, but it also poses domain adaptation risks: features learned from ImageNet natural scene data are not optimally tuned for histopathological textures, and this feature mismatch problem may limit the model's ability to finely differentiate cancer subtypes. Ensemble models further exacerbate these problems: they not only multiply training time and storage requirements but also reduce the feasibility of the model in actual clinical scenarios. Moreover, while the attention layer and LSTM layer have improved the model's feature representation ability, they have also reduced model interpretability. Without specialized feature visualization tools (such as Grad-CAM or attention heatmaps), it is difficult to correlate the model's prediction results with specific histopathological features—a correlation that is a core requirement for obtaining clinical trust and regulatory approval.

Author Contributions: Conceptualization, S.Z. and P.Z.; methodology, S.Z.; software, S.Z.; validation, S.Z., P.Z. and G.Z.; formal analysis, S.Z. and P.Z.; investigation, S.Z.; resources, S.Z.; data curation, S.Z.; writing—original draft preparation, S.Z.; writing—review and editing, S.Z., P.Z. and G.Z.; visualization, S.Z.; supervision, P.Z. and G.Z.; funding acquisition, G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Yunnan Province Agricultural Basic Research Joint Special Project "Estimation of aboveground biomass of Simao pinus forest based on multisource remote sensing data" (Grant No. 202401BD070001-117).

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were generated in this study. Two publicly available third-party datasets were analyzed in this work. The invasive ductal carcinoma (IDC) breast cancer histology dataset is available on Kaggle: https://www.kaggle.com/code/paultimothymooney/predict-idc-in-breast-cancer-histology-images/input (accessed on 4 September 2025). The Breast Cancer Histopathological Database (BreakHis) can be accessed at https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/ (accessed on 4 September 2025). Restrictions apply to the reuse of the BreakHis dataset; researchers must comply with its original license and obtain proper permission before use.

Acknowledgments: The author would like to sincerely thank Mark Vogelsberger for providing inspirational ideas and valuable guidance throughout the research and manuscript preparation process. Special gratitude is extended to classmates Megan and Jason for their valuable assistance and insightful discussions regarding model research and experimental exploration, which greatly facilitated the progress of this study. Additionally, heartfelt thanks are given to my family for their continuous support, care, and encouragement during the entire research period, which provided solid spiritual backing for the completion of this work. We hereby disclose the use of artificial intelligence tools during the preparation of this manuscript. First, the Doubao AI large model was utilized to assist in debugging and optimizing code and troubleshooting code logical errors. All experimental parameters and core code logic were independently designed by the authors, and all AI-generated code snippets were fully manually reviewed and revised. Second, AI literature summarization tools were adopted to sort out the core contributions of relevant references, while the literature collation,

comparative analysis, and academic discussion were independently completed by the authors. Third, AI tools were used for minor linguistic translation assistance for partial manuscript content. All AI-generated content serves only as auxiliary materials for manuscript writing. The experimental design, data analysis, graphical production, and core academic conclusions of this study are independently completed, reviewed, and verified by the authors, who take full responsibility for all the content of this published work.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
AUC	Area Under the Curve
IARC	International Agency for Research on Cancer
DCIS	Ductal Carcinoma In Situ
IDC	Invasive Ductal Carcinoma
ILC	Invasive Lobular Carcinoma
AlexNet	Alex Network
VGG	Visual Geometry Group
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
BCDCNN	Breast Cancer Diagnostic Convolutional Neural Network
MobileNet	Mobile Network
BC-GAHOCNN	Breast-Cancer-Guided Attention High-Order Convolutional Neural Network
Thermo-CAD	Thermography-Computer-Aided Diagnosis
Grad-CAM	Gradient-weighted Class Activation Mapping
SOTA	State-of-the-Art
TransferEnsemble-Breast	Transfer-learning Ensemble for Breast cancer diagnosis
EffiLSTM-Attn	EfficientNet-LSTM-Attention
EffiIHC-Ensemble	EfficientNet-Immunohistochemistry Ensemble
ResNet	Residual Network
BC-MBINet	Breast-Cancer Multi-Branch Information Network
ES-CNN	Enhanced Squeeze Convolutional Neural Network
EfficientNet	Efficient Network
GAN	Generative Adversarial Network
ResViT-GANNet	Residual-Vision-Transformer-Generative Adversarial Network
TAMA	Transformer-based Attention Multi-scale Aggregation
MFF-ClassificationNet	Multi-Feature Fusion Classification Network
CBAM-SE	Convolutional Block Attention Module-Squeeze-and-Excitation
BACH	Breast Cancer Histopathological Database
CMAF-Net	Cross-Modal Attention Fusion Network
DenseNet	Densely connected Convolutional Network
Inception-Net	Inception Network
CTNet	Context-Transformer Network

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Kook, Y.; Lee, Y.J.; Chu, C.; Jang, J.S.; Baek, S.H.; Bae, S.J.; Cha, Y.J.; Gong, G.; Jeong, J.; Lee, S.B.; et al. Differentiating HER2-low and HER2-zero tumors with 21-gene multigene assay in 2,295 HR+HER2− breast cancer: A retrospective analysis. *Breast Cancer Res.* **2024**, *26*, 154. [CrossRef] [PubMed]

3. Lakhani, S.R. Diagnostic challenges in breast pathology. *Pathology* **2013**, *45*, S7–S8. [CrossRef]
4. Singh, K.; Sudershan, A.; Kumar, S.; Shastri, S.; Mansotra, V. BC-MBINet: A novel architecture for accurate classification of breast cancer with microscopic biopsy images using deep convolutional neural networks. *Int. J. Pattern Recognit. Artif. Intell.* **2025**, *39*, 2557015. [CrossRef]
5. Atitallah, A.B.; Kamoun, J.; Alanazi, M.D.; Alanazi, T.M.; Albekairi, M.; Kaaniche, K. An advanced medical diagnosis of breast cancer histopathology using convolutional neural networks. *Comput. Mater. Contin.* **2025**, *83*, 5761–5779. [CrossRef]
6. Jaincy, D.E.M.; Pattabiraman, V. BCDCNN: Breast cancer deep convolutional neural network for breast cancer detection using MRI images. *Sci. Rep.* **2025**, *15*, 29014. [CrossRef] [PubMed]
7. Hassani, F.Z.E.; Daoudy, A.E.; Joudar, N.E.; Haddouch, K. Improving early breast cancer detection: A hybrid approach with convolutional neural networks and genetic algorithm-based optimization. *Clust. Comput.* **2025**, *28*, 590. [CrossRef]
8. Attallah, O. Harnessing infrared thermography and multi-convolutional neural networks for early breast cancer detection. *Sci. Rep.* **2025**, *15*, 27464. [CrossRef] [PubMed]
9. Asif, S.; Zhu, L.Y.; Yan, D.; Xu, L.; Huang, Z.; Xu, H.; Yan, R.; Cai, L.; Zheng, C.; Lin, J.; et al. BREAST-RANKNet: A fuzzy rank-based ensemble of CNNs with residual learning for enhanced breast cancer detection from ultrasound and mammogram images. *J. Big Data* **2025**, *12*, 194. [CrossRef]
10. Alshehri, H.M. BreNet: Attention-enhanced multi-scale CNN framework for breast cancer classification in histopathological images. *IEEE Access* **2025**, *13*, 143377–143391. [CrossRef]
11. Yusuf, M.; Kana, A.F.D.; Bagiwa, M.A.; Abdullahi, M. Multiclass of breast cancer histopathological image using enhanced shallow convolutional neural network. *J. Eng. Appl. Sci.* **2025**, *72*, 24. [CrossRef]
12. Shah, P.B.; Ganesan, S.; Ling, S.; Pokharel, S.; Somasiri, N. Histopathology image augmentation using generative adversarial models for breast cancer diagnostics. In *Proceedings of the 2025 International Aegean Conference on Electrical Machines and Power Electronics (ACEMP) & 2025 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM), Timisoara, Romania, 14–17 May 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–7. [CrossRef]
13. Zhou, Y.; Jin, F.; Suo, G.; Yang, J. ResViT-GANNet: A deep learning framework for classifying breast cancer histopathology images using multimodal attention and GAN-based augmentation. *BMC Med. Imaging* **2025**, *25*, 401. [CrossRef] [PubMed]
14. Wang, X.; Wang, G.; Li, L.; Zou, H.; Cui, J. MFF-ClassificationNet: CNN-Transformer hybrid with multi-feature fusion for breast cancer histopathology classification. *Biosensors* **2025**, *15*, 718. [CrossRef] [PubMed]
15. Ativi, W.X.; Chen, W.; Kwao, L.; Ayivi, W.; Sam, F.; Alqahtani, A.; Gu, Y.H.; Al-Antari, M.A. CMAF-Net: Cross-modal attention fusion with information-theoretic regularization for imbalanced breast cancer histopathology. *Sci. Rep.* **2026**, *16*, 9607. [CrossRef] [PubMed]
16. Nakach, F.Z.; Zerouaoui, H.; Idri, A. Deep hybrid AdaBoost ensembles for histopathological breast cancer classification. In *World Conference on Information Systems and Technologies (WorldCIST 2022)*; Rocha, A., Adeli, H., Dzemyda, G., Moreira, F., Eds.; Lecture Notes in Networks and Systems; Springer: Cham, Switzerland, 2022; Volume 468, pp. 486–496. [CrossRef]
17. Azmoodeh-Kalati, M.; Shabani, H.; Maghareh, M.S.; Barzegar, Z.; Lashgari, R. Leveraging an ensemble of EfficientNetV1 and EfficientNetV2 models for classification and interpretation of breast cancer histopathology images. *Sci. Rep.* **2025**, *15*, 21541. [CrossRef] [PubMed]
18. Zhang, S.; Zhang, Y.; Zhang, D.; Wang, Q.; Yang, W. DMFF: A Deep Learning-Based Multi-Omics Fusion Framework for Survival Prediction and Subtype Classification in Breast Cancer. *Sci. Rep.* **2026**, *16*, 14203. [CrossRef]
19. Kandan, S.U.; Abul, O. A Robust Machine Learning Approach for Breast Cancer Subtype Classification Using Relative Gene Expression Order Representations. *Comput. Methods Programs Biomed.* **2026**, *285*, 109510. [CrossRef] [PubMed]
20. Lodh, E.; Chowdhury, T.; Shaw, A.; Bedajna, D.; Bera, S.; Majumder, S.; De, M. A Stacked Ensemble Meta-Classifier Framework for PAM50 Breast Cancer Subtype Prediction from Gene Expression Signatures. *Netw. Model. Anal. Health Inform. Bioinform.* **2026**, *15*, 78. [CrossRef]
21. Ab Mumin, N.; Liew, C.-H.; Ong, S.-Q.; Wong, J.H.D.; Ramli Hamid, M.T.; Rahmat, K.; Ng, K.H. MRI-Based Texture Analysis for Breast Cancer Subtype Classification in a Multi-Ethnic Population. *Magn. Reson. Mater. Phys. Biol. Med.* **2026**, *39*, 121–135. [CrossRef] [PubMed]
22. Ali, M.M.; Shabbir, A.; Akbar, A.; Angelopoulou, A.; Aftab, M.; Zia, Q.; Wang, C.; Zhenfei, W. CTNet: Multimodal channel attention transformer network for breast cancer image classification. *Biomed. Signal Process. Control* **2026**, *113*, 109108. [CrossRef]
23. Zhao, T.; Ren, Y.; Lu, H.; Kong, Y. Decision level scheme for fusing multi-omics and histology slide images using deep neural network for tumor prognosis prediction. *Sci. Rep.* **2025**, *15*, 25479. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.