

OPEN ACCESS

Volume: 5

Issue: 4

Month: October

Year: 2026

ISSN: 2583-7117

Published: 15.10.2026

Citation:

Aastha Mishra, Prof. Brajesh Patel “A Deep Learning-Based Multi-Class Breast Cancer Histopathological Image Classification System Using CBAM” International Journal of Innovations in Science Engineering and Management, vol.5, no.4, 2026, pp. 144-156

DOI:

10.69968/ijsem.2026v5i4144-156



This work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

A Deep Learning-Based Multi-Class Breast Cancer Histopathological Image Classification System Using CBAM

Aastha Mishra¹, Prof. Brajesh Patel²

¹Research Scholar, Shri Ram Institute of Technology

²Professor, Shri Ram Institute of Technology

Abstract

Breast histopathological image classification is challenging because benign and malignant breast lesions can exhibit overlapping cellular morphology, texture, staining appearance, and tissue architecture. This paper presents a multi-class deep learning framework that integrates the BreakHis dataset, ImageNet-pretrained InceptionResNetV2, and the Convolutional Block Attention Module (CBAM) for eight-class histopathological image classification. The pipeline performs class-distribution analysis, mean-count-based oversampling and undersampling, CLAHE-based contrast enhancement in LAB space, resizing to 224×224 pixels, normalization, and horizontal and vertical training augmentation. InceptionResNetV2 provides the main convolutional representation, while CBAM sequentially applies channel and spatial attention to refine discriminative feature information before a dense classification head. The latest ten-epoch training run reaches 98.44% training accuracy. The reported test evaluation achieves 98% accuracy on 2,160 test images, with 270 images per class. Macro precision, recall, and F1-score are approximately 0.98. The results demonstrate strong performance for the integrated pipeline under the reported benchmark protocol. At the same time, the paper identifies methodological limitations concerning resampling before partitioning, image-wise rather than patient-wise splitting, the absence of component-level ablation, and the current hard-label ROC/AUC implementation. These limitations define the scope of the reported findings and motivate stricter patient-wise, external, and probability-based validation.

Keywords; Breast Cancer, Histopathology, BreakHis, InceptionResNetV2, Multi-Class Classification, Image Augmentation.

INTRODUCTION: METEOROLOGY AS LITERARY EXPERIENCE

Breast cancer diagnosis depends strongly on accurate tissue-level interpretation, particularly when the objective is to distinguish among multiple histopathological subtypes rather than only separate benign from malignant tissue. Histopathology provides a detailed view of cellular morphology, glandular architecture, stromal organization, nuclear characteristics, and other microscopic structures that contribute to subtype recognition. Reviews of breast cancer detection techniques have emphasized the continuing importance of reliable tissue and imaging-based diagnosis and the challenges involved in achieving consistent detection. [1] Digital pathology has made these tissue images available for computational analysis, creating an opportunity to develop image-based systems that can support pathologists as research tools or future decision-support systems, while requiring careful consideration of validation and clinical context. [2]

The computational problem is difficult because breast histopathological classes do not have clean visual boundaries. Different lesions can share similar color distributions, cell densities, tissue textures, and spatial arrangements. Systematic reviews of artificial intelligence in breast cancer detection describe the rapid expansion of automated approaches while also identifying dataset, methodological,

and validation challenges. [3] In a multi-class setting, the classifier must learn several decision boundaries at the same time, including boundaries between malignant subtypes and between visually similar benign lesions. Reviews of mammographic artificial intelligence further illustrate how task definition and data characteristics influence reported model performance. [4]

BreakHis is a widely used benchmark for breast histopathological image research and provides images across multiple tumor subtypes and magnification levels. Research on alternative breast cancer detection technologies, including electromagnetic approaches, demonstrates that automated diagnosis extends beyond a single imaging modality and that different sensing mechanisms capture different aspects of disease-related information. [5] The present study focuses on eight classes: adenosis, fibroadenoma, phyllodes tumor, tubular adenoma, ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma. Broader technological reviews of breast cancer detection similarly emphasize the diversity of diagnostic approaches and the importance of selecting methods appropriate to the clinical and imaging problem. [6]

Conventional convolutional neural networks can learn hierarchical image representations, but histopathological images contain substantial variation in contrast, staining intensity, morphology, and the spatial distribution of informative structures. Deep-learning research in breast cancer detection and classification supports the use of learned hierarchical representations for automated image analysis and provides a foundation for transfer-learning-based approaches. [7] For this reason, the proposed framework combines three complementary ideas: image-level preprocessing and balancing, transfer learning for high-capacity feature extraction, and attention-based feature refinement. These choices must still be considered alongside broader challenges in breast cancer detection and treatment, including accessibility, consistency, and translation beyond controlled research datasets. [8]

The central aim of this paper is not only to report a headline accuracy value, but to characterize model behavior through training and validation curves, class-wise precision, recall and F1-score, and a discussion of the principal methodological risks in the experimental protocol. Recent optimized deep-learning work in breast cancer detection illustrates the continuing development of specialized architectures for extracting discriminative image patterns. [9] The latest reported evaluation reaches 98% test accuracy on 2,160 images with equal support across the eight classes.

The paper therefore emphasizes both the strength of the observed benchmark result and the conditions under which that result should be interpreted. Research on dynamic model scaling also demonstrates the importance of adapting computational representation to the characteristics of the target imaging problem. [10]

The main contributions of the study are: an integrated eight-class BreakHis classification pipeline based on InceptionResNetV2 and CBAM; a preprocessing and balancing sequence combining class resampling, CLAHE and augmentation; a detailed class-wise evaluation that identifies ductal carcinoma as the most difficult category in the reported test results; and a methodological analysis that distinguishes the observed internal test performance from claims of general clinical generalization. Related work on multimodal, cross-patient breast cancer analysis further motivates careful consideration of how information from different sources and patients is separated during model development. [11] Recent reviews of breast cancer artificial intelligence also highlight trends toward more integrated systems, stronger validation, and broader future directions for AI-assisted diagnosis. [12]

RELATED WORK

Breast cancer detection research has evolved from conventional imaging and statistical analysis toward increasingly sophisticated machine learning and deep learning systems. Recent histopathological research has examined ensemble deep-learning strategies for breast cancer image detection and emphasizes the value of combining complementary learned representations. [13] Systematic reviews of deep learning in breast cancer further summarize the growing evidence across segmentation, detection, and classification tasks while emphasizing dataset characteristics and validation methodology. [14]

Within medical image analysis, transfer learning has become a practical strategy for datasets where the amount of annotated target-domain data is smaller than the parameter capacity of a modern convolutional architecture. Pretrained networks provide a strong initialization and can be adapted to domain-specific visual patterns. Research on artificial intelligence for breast screening, including mammographic density assessment, demonstrates how modern AI systems are being evaluated in realistic screening-oriented settings. [15] Broader decade-scale reviews likewise show that breast cancer AI is expanding across detection, classification, prognosis, and explainability, with increasing attention to methodological robustness. [16]

Histopathological image classification introduces a particularly fine-grained challenge. The model must discriminate between lesions that may have overlapping visual structures while remaining insensitive to image-level variations that are unrelated to the target label. Recent subgroup-performance research in breast imaging demonstrates why aggregate metrics should be supplemented with analysis of performance across relevant subgroups and operating conditions. [17] In parallel, breast cancer detection research increasingly explores sensing technologies such as graphene-assisted surface plasmon resonance biosensors, illustrating the wider movement toward diverse computational and physical diagnostic modalities. [18]

Attention mechanisms have become important in image classification because not every learned feature channel or spatial region is equally useful for every class. CBAM applies channel attention followed by spatial attention. Channel attention reweights feature channels according to their inferred importance, while spatial attention emphasizes locations within the feature map. Recent work on machine-learning-assisted terahertz biosensor optimization illustrates a related objective: identifying informative representations while controlling computational complexity in diagnostic systems. [19] Other deep-learning research has explored specialized architectures such as hybrid capsule networks for breast cancer computer-aided diagnosis, further demonstrating the diversity of representation-learning strategies used in medical imaging. [20]

Image preprocessing is another recurring consideration in computational pathology. Staining variation and local contrast differences may change the appearance of cellular structures even when the underlying morphology is similar. Recent research on deep-learning-based optimization of two-dimensional material biosensors shows how learned models are being combined with engineered sensing pipelines to improve breast cancer detection performance. [21] The present framework applies CLAHE in LAB space before resizing and normalization, providing a consistent preprocessing stage for all images. [22] Related artificial-intelligence-enhanced terahertz metasurface biosensor research further illustrates the broader trend toward integrating advanced sensing with machine learning for biomarker-oriented detection. [23]

Class imbalance is also widely recognized as a potential source of bias in medical image classification. A classifier trained on an uneven class distribution can become better adapted to high-frequency categories while producing

weaker performance on less frequent classes. Research on breast cancer screening has examined the trade-off between sensitivity and recall decisions in clinical screening workflows, reinforcing the importance of evaluating more than a single aggregate metric. [24] The present framework uses a path-level mean-count balancing procedure that oversamples classes below the mean count and undersamples classes above the mean. Reviews of machine-learning-enhanced optical spectroscopy provide another example of how breast cancer diagnosis is increasingly evaluated through multiple signals and complementary computational measures rather than accuracy alone. [25]

A closely related study in the source thesis is Sharmin et al., which combines ResNet50V2 feature extraction with ensemble machine-learning classifiers for binary invasive ductal carcinoma detection. [26] The reported task differs from the present eight-class BreakHis problem, so the reported accuracies are contextual rather than directly comparable. The comparison nevertheless illustrates how transfer learning and downstream classifiers have been used in breast histopathology and provides a useful reference point for the present study. [27]

The reviewed literature collectively indicates that breast cancer AI is moving toward increasingly specialized architectures, multimodal and complementary sensing strategies, subgroup-aware evaluation, and stronger methodological validation. [28] These themes motivate the evaluation strategy adopted in the present paper, where class-wise results are considered alongside overall accuracy and where limitations in the current partitioning protocol are stated explicitly. [29]

METHODOLOGY

Overall System Architecture

The proposed framework follows a sequential image-classification pipeline. The dataset is first organized into class and magnification directories. Image paths are balanced through mean-count-based oversampling and undersampling. Images are then processed with LAB-space CLAHE, resized to 224×224 pixels, normalized to the [0,1] range, and augmented during training through random horizontal and vertical flips. The data are divided into training, validation and test subsets using the reported 70%, 15% and 15% partitioning configuration. The image representation is learned by an ImageNet-pretrained InceptionResNetV2 backbone, refined with CBAM, and converted into class probabilities by a dense classifier with 512, 256 and 128 units followed by an eight-unit softmax layer.

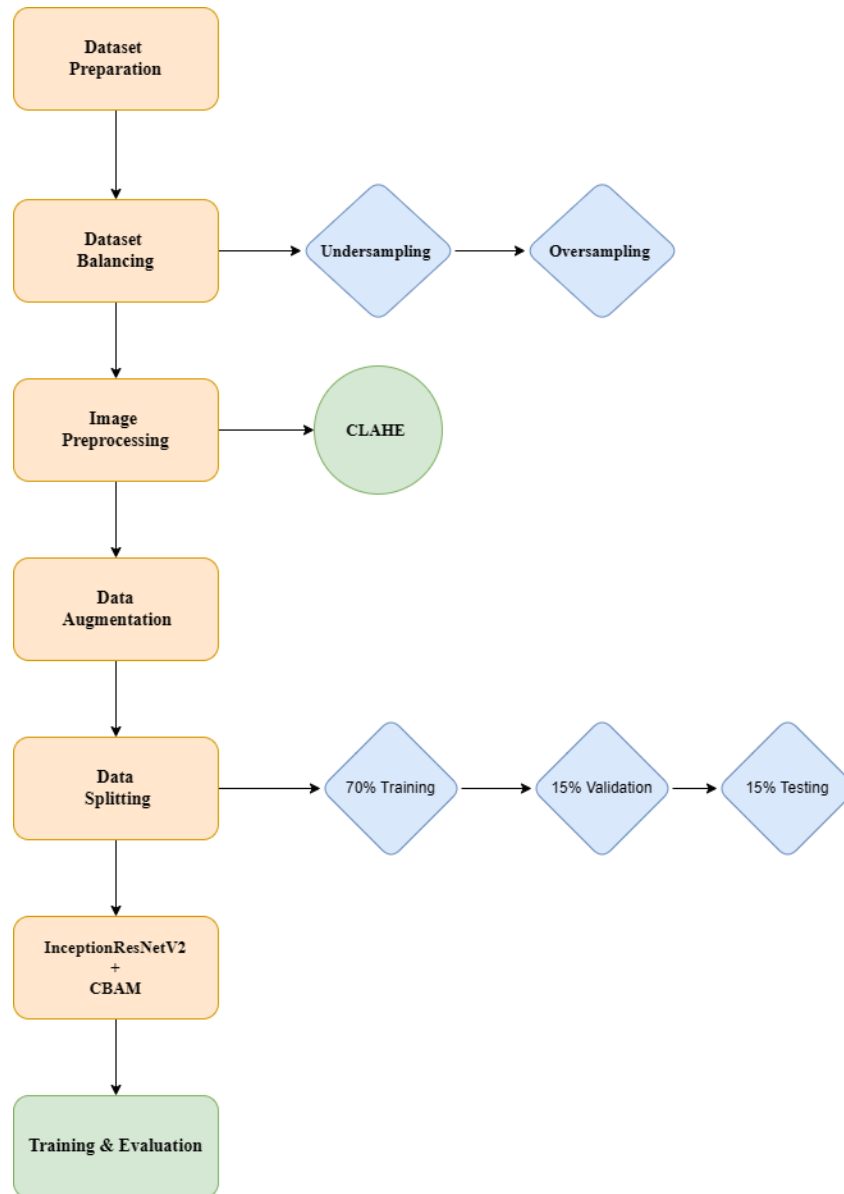


Figure. 1. Proposed InceptionResNetV2-CBAM framework and evaluation flow.

Dataset and Class Composition

The study uses the multiclass portion of the BreakHis breast histopathological image benchmark as organized in the supplied implementation. The target problem contains eight classes with four benign and four malignant categories. The image paths are collected from the dataset directory hierarchy and mapped to integer class labels. Because the implementation retains multiple magnification levels, the model is exposed to tissue structures represented at different spatial scales. The class names used for evaluation are adenosis, lobular carcinoma, phyllodes tumor, mucinous carcinoma, ductal carcinoma, papillary carcinoma, tubular adenoma, and fibroadenoma.

Table 1. Eight-class label configuration used in the classification task.

Group	Class	Code label
Benign	Adenosis	adenosis
Benign	Fibroadenoma	fibroadenoma
Benign	Phyllodes tumor	phyllodes tumor / dataset name
Benign	Tubular adenoma	tubular adenoma
Malignant	Ductal carcinoma	ductal_carcinoma
Malignant	Lobular carcinoma	lobular_carcinoma
Malignant	Mucinous carcinoma	mucinous_carcinoma
Malignant	Papillary carcinoma	papillary_carcinoma

Class Balancing and Data Partitioning

The class-balancing stage operates on image paths rather than synthesizing new image pixels. The mean class count is calculated, minority classes are oversampled with replacement up toward the mean, and majority classes are undersampled without replacement toward the same target. The purpose is to reduce differences in representation before model training. Because the current implementation performs this path-level resampling before splitting, the experimental protocol carries a potential leakage risk: duplicated paths generated by oversampling can theoretically occur in more than one later partition. This limitation is retained in the analysis rather than treated as insignificant.

The reported final evaluation contains 2,160 test images with 270 images per class. Under the stated 70%/15%/15% partitioning of the balanced image-path pool, the corresponding training and validation partitions contain 10,080 and 2,160 images, respectively. These counts are useful for reproducing the latest ten-epoch training history, which reports 315 steps per epoch with a batch size consistent with the supplied implementation.

Image Preprocessing and Augmentation

Histopathological images can vary in local contrast, illumination and staining appearance. The implementation applies Contrast Limited Adaptive Histogram Equalization (CLAHE) in the LAB color space. The L channel is enhanced while the chromatic components are retained, after which the image is converted back to the working representation. The preprocessing stage uses a clip limit of 2.0 and an 8×8 tile grid according to the methodology described in the thesis. This operation is intended to make local tissue structures more visible without relying on global histogram equalization.

Following contrast enhancement, images are resized to 224×224 pixels and normalized to [0,1]. Training-time augmentation uses random horizontal and vertical flips. These transformations were selected because the tissue patches do not have a fixed semantic orientation, so flipping can increase appearance diversity without changing the underlying histopathological class. Importantly, augmentation is conceptually distinct from balancing: balancing changes sampling frequency, whereas augmentation changes image appearance.

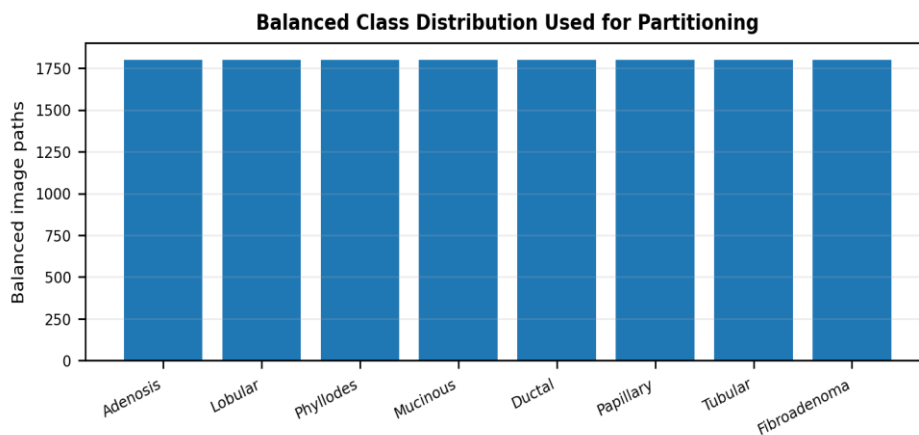


Figure. 2. Balanced class distribution used for the reported partitioning configuration.

InceptionResNetV2 Feature Extraction

InceptionResNetV2 is used as the primary feature extractor with ImageNet-pretrained weights. The original classification layer is replaced by the task-specific eight-class head. Inception-style modules allow processing at multiple receptive-field scales, while residual connections support stable optimization in a deep network. Transfer learning provides a starting representation learned from large-scale natural-image data and allows the model parameters to adapt to the target histopathological domain during training.

The choice of InceptionResNetV2 is consistent with the study objective of learning complex visual representations from microscopic images. Shallow and intermediate layers can respond to edges, color transitions and textures, whereas deeper layers can capture more complex combinations of local structures. The subsequent CBAM stage is therefore applied to a rich convolutional feature representation rather than to raw pixels.

Convolutional Block Attention Module

CBAM performs two sequential forms of attention: channel attention and spatial attention. For channel

attention, global average pooling and global maximum pooling summarize the incoming feature maps. These complementary descriptors are passed through shared dense transformations and combined to form channel-wise attention weights. The weights are applied multiplicatively to the feature map, increasing or decreasing the contribution of individual channels.

Spatial attention then aggregates information across channels using average and maximum projections. The two spatial descriptors are concatenated and passed through a 7×7 convolution followed by a sigmoid function. The resulting attention map reweights the spatial representation. In a histopathological context, this mechanism is intended to emphasize image regions and feature channels that contain more discriminative tissue information, such as local cellular organization or structural patterns relevant to subtype recognition.

Classification Head and Optimization

The refined representation is passed to a dense classification head consisting of fully connected layers with 512, 256 and 128 units. Batch normalization and dropout are used to support training stability and regularization. The final layer contains eight neurons with softmax activation, producing a probability distribution across the target classes. The training objective uses categorical cross-entropy with label smoothing of 0.05. The optimizer is Adam with a learning rate of 1×10^{-5} . The implementation also uses early stopping and learning-rate reduction based on validation behavior.

The current implementation assigns a class weight of 1.0 to ductal carcinoma, which is equivalent to the baseline weight used for the other classes. Therefore, the present experiment should not be described as applying an additional loss penalty specifically to ductal carcinoma. The principal imbalance intervention is the path-level resampling operation, while the loss function and label smoothing provide regularized optimization.

Preprocessing Configuration and Image Standardization

Table 3. Image preprocessing and augmentation configuration.

Component	Configuration	Purpose
Color processing	LAB-space conversion	Separate luminance from chromatic information
CLAHE	Clip limit 2.0; 8×8 grid	Improve local contrast while limiting amplification

Image size	224×224 pixels	Match backbone input requirement
Normalization	Pixel values scaled to [0,1]	Stabilize numerical optimization
Training augmentation	Random horizontal and vertical flips	Increase appearance diversity

The preprocessing sequence is designed to make the input distribution more stable before deep feature extraction. CLAHE is particularly relevant for histopathology because local nuclei, glandular boundaries and stromal patterns may be difficult to distinguish when contrast varies across an image. Applying the operation consistently allows the network to see a more standardized representation. The preprocessing stage is deliberately separated from augmentation so that the effects of deterministic enhancement and stochastic training transformations remain conceptually distinct.

Normalization to [0,1] reduces the numerical scale of input values and places training examples on a common intensity range. Resizing to 224×224 is used consistently with the selected feature extractor. Although resizing improves computational consistency, it can also reduce very fine cellular detail. This is an important trade-off for histopathology: the model gains a standardized input size but may lose microscopic information when structures are represented at a lower effective resolution. Multi-scale or patch-specific future work could investigate this issue more directly.

Model Architecture and Principal Hyperparameters

Table 3. Model architecture and principal components.

Stage	Configuration	Role
Backbone	InceptionResNetV2; ImageNet-pretrained	Hierarchical feature extraction
Attention	CBAM; channel then spatial attention	Feature refinement
Dense layer 1	512 units	High-dimensional task representation
Dense layer 2	256 units	Progressive compression
Dense layer 3	128 units	Compact discriminative representation
Output	8-unit softmax	Eight-class probability distribution

The architecture is intentionally modular. InceptionResNetV2 provides the base representation,

CBAM refines that representation, and the dense head translates the refined features into class probabilities. This organization makes the system suitable for future ablation because individual components can be removed or replaced while retaining the rest of the pipeline. Such ablations were not part of the reported experiment, so the present results should be described as performance of the integrated architecture rather than evidence that each module is independently necessary.

The dense head contains progressively smaller hidden layers of 512, 256 and 128 units. This gradual reduction can be interpreted as a transition from high-dimensional backbone features toward a compact task-specific representation. Batch normalization contributes to training stability, while dropout provides regularization. The final softmax layer ensures that the eight outputs form a normalized categorical distribution.

Optimization and Training Controls

Table 3. Optimization and training configuration for the reported run.

Training component	Setting
Optimizer	Adam
Learning rate	1×10^{-5}
Label smoothing	0.05
Epochs in latest reported run	10
Training steps per epoch	315
Batch size implied by run	32
Early stopping	Validation-loss monitoring
Learning-rate scheduling	Validation-loss based reduction

The latest run shows that the optimization process continues to improve training accuracy throughout the ten epochs. However, validation accuracy peaks early and then fluctuates within a narrow interval. This behavior illustrates why training accuracy alone should not be used to select a final model checkpoint.

In future experiments, the checkpoint corresponding to the best validation criterion should be retained and evaluated on the test set only once after the experimental design and hyperparameters are fixed.

Label smoothing is included to reduce overconfident target distributions during optimization. Early stopping and learning-rate reduction are intended to avoid unnecessary training once validation loss stops improving. The combination of these controls provides regularization at the optimization stage, although it cannot correct a data-partition leakage pathway introduced before model training.

RESULTS AND DISCUSSION

Experimental Configuration and Evaluation Criteria

The final reported experiment uses the eight-class BreakHis classification setting described above. The latest training execution runs for ten epochs with a learning rate of 1×10^{-5} . Each epoch contains 315 training steps, corresponding to 10,080 training images at the supplied batch configuration. Validation is performed on 2,160 images and the final test evaluation contains 2,160 images, with 270 images per class. The main result reported here is the 98% test accuracy together with the class-wise metrics supplied by the classification report.

Table 4. Overall test-set performance of the proposed framework.

Metric	Value
Test accuracy	98%
Test support	2,160 images
Images per class	270
Macro precision	0.98
Macro recall	0.98
Macro F1-score	0.98
Weighted precision	0.98
Weighted recall	0.98
Weighted F1-score	0.98

Overall Classification Performance

The proposed framework achieves an overall test accuracy of 98% on 2,160 images. The same evaluation reports macro precision, recall and F1-score of approximately 0.98. Because each class contributes 270 test images, the equal-support test configuration reduces the possibility that a single highly represented category dominates the aggregate score. The result therefore provides evidence of strong internal performance across the eight-class benchmark.

The agreement between the overall accuracy and the macro-averaged F1-score is particularly informative. Accuracy summarizes the fraction of all images classified correctly, whereas macro F1 gives each class equal weight. When these values are close, the aggregate result is less likely to be explained solely by excellent performance on a subset of dominant categories. Nevertheless, the class-wise results show that performance is not perfectly uniform: ductal carcinoma is clearly weaker than adenosis and phyllodes tumor, and lobular carcinoma also has lower precision.

Training and Validation Performance

The latest ten-epoch learning history begins with 96.83% training accuracy and 97.18% validation accuracy. Training

accuracy increases throughout the run and reaches 98.44% at epoch 10. Validation accuracy fluctuates in a narrower band, reaching its highest value of 97.50% in epoch 2 and ending at 97.36%. The training-loss curve decreases from 0.4397 to 0.3953, indicating continuing optimization of the training objective, while validation loss reaches its lowest observed value of 0.4205 in epoch 3 and later increases to 0.4578 by epoch 10.

The epoch-10 training accuracy exceeds validation accuracy by 1.08 percentage points. This gap is modest in absolute terms but larger than the separation observed in the earlier training history. The pattern suggests some degree of late-training overfitting, because additional optimization continues to improve the training set while the validation curve remains comparatively stable. The result does not

imply that the model has failed to generalize; rather, it suggests that the validation curve should be considered when selecting the final checkpoint in future runs.

Table 5. Training history for the latest ten-epoch execution.

Epoch	Train Acc.	Val. Acc.	Train Loss	Val. Loss
1	96.83%	97.18%	0.4397	0.4530
2	96.92%	97.50%	0.4379	0.4289
3	97.16%	97.27%	0.4273	0.4205
4	97.26%	96.90%	0.4242	0.4312
5	97.74%	97.31%	0.4179	0.4549
6	98.04%	97.27%	0.4122	0.4210
7	97.98%	97.31%	0.4080	0.4257
8	98.07%	97.22%	0.4041	0.4488
9	98.25%	97.45%	0.3980	0.4310
10	98.44%	97.36%	0.3953	0.4578

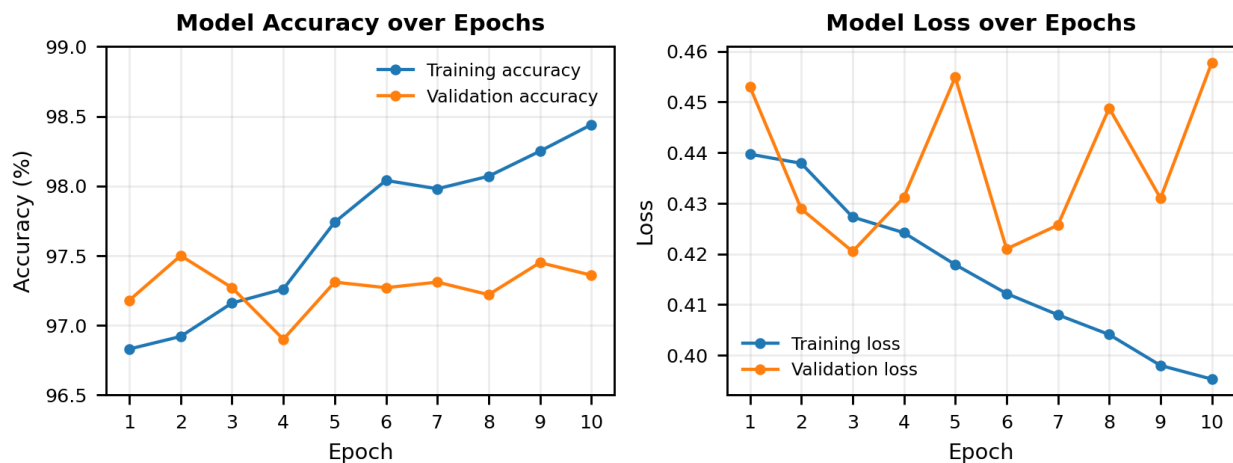


Figure 3. Training and validation accuracy and loss over ten epochs.

The training precision trajectory increases from 97.37% at epoch 1 to 98.90% at epoch 10, while training recall rises from 95.89% to 97.80%. Validation precision remains in the narrow range of 97.39%–97.71%, and validation recall remains between 96.76% and 97.27%. This pattern is consistent with a model that is learning increasingly selective representations on the training set while maintaining relatively stable validation behavior.

Class-Level Classification Performance

Table 6. Final class-wise classification performance on the reported 2,160-image test set.

Class	Precision	Recall	F1-score	Support
Adenosis	1.00	1.00	1.00	270

Lobular carcinoma	0.92	0.99	0.95	270
Phyllodes tumor	1.00	1.00	1.00	270
Mucinous carcinoma	0.97	1.00	0.98	270
Ductal carcinoma	0.99	0.87	0.93	270
Papillary carcinoma	0.98	1.00	0.99	270
Tubular adenoma	0.99	1.00	0.99	270
Fibroadenoma	0.99	0.97	0.98	270
Macro average	0.98	0.98	0.98	2160
Weighted average	0.98	0.98	0.98	2160

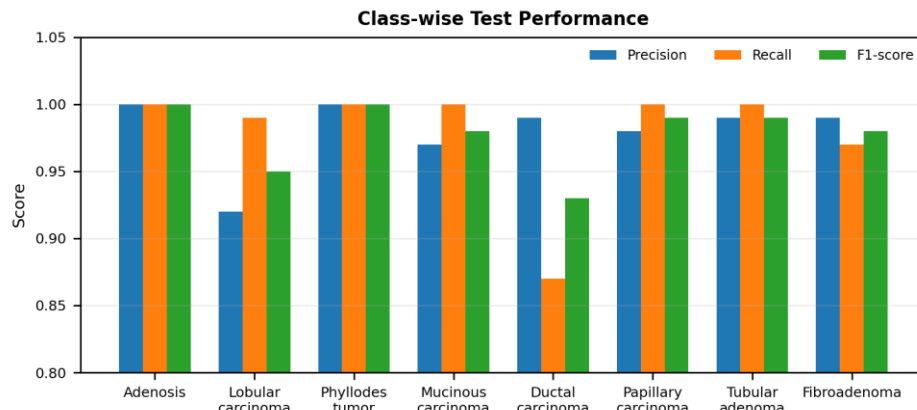


Figure. 4. Class-wise precision, recall and F1-score.

Adenosis and phyllodes tumor achieve perfect precision, recall, and F1-score of 1.00, indicating that all 270 test images in each category were correctly classified in the reported evaluation. Tubular adenoma and papillary carcinoma also demonstrate strong performance, with F1-scores of 0.99, while fibroadenoma and mucinous carcinoma achieve 0.98. Lobular carcinoma records an F1-score of 0.95.

Ductal carcinoma represents the main remaining challenge, with precision of 0.99, recall of 0.87, and F1-score of 0.93. Its high precision indicates that most images predicted as ductal carcinoma are correct, whereas the lower recall shows that some actual ductal carcinoma images are misclassified as other categories. Lobular carcinoma shows the opposite pattern, with precision of 0.92 and recall of

0.99, indicating that most true cases are detected but some non-lobular images are incorrectly assigned to this class.

Mucinous carcinoma achieves precision of 0.97, recall of 1.00, and F1-score of 0.98, while papillary carcinoma and tubular adenoma achieve recalls of 1.00 with F1-scores of 0.99. Fibroadenoma records precision of 0.99, recall of 0.97, and F1-score of 0.98.

Overall, the results demonstrate strong and consistent classification across most categories. The primary areas requiring further investigation are ductal carcinoma recall and lobular carcinoma precision. Since no external dataset was evaluated, these results should be considered benchmark performance under the current test distribution rather than evidence of universal generalization.

Evaluation Metrics and Statistical Interpretation

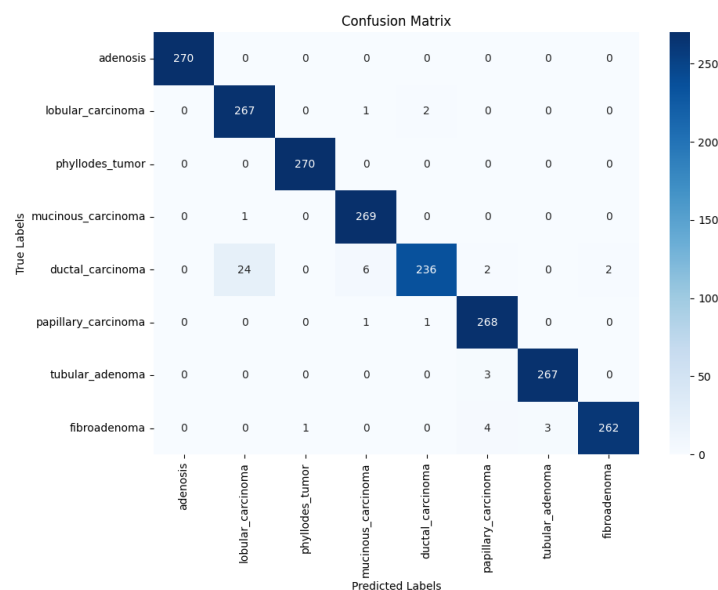


Figure 4 Confusion Matrix

Accuracy is computed as the number of correct predictions divided by the total number of predictions. For each class, precision measures the fraction of predictions assigned to that class that are correct, while recall measures the fraction of true samples of that class that are recovered. The F1-score combines the two measures using the harmonic mean. In the present eight-class setting, these metrics are reported separately because the direction of an error matters: ductal carcinoma, for example, has very high precision but lower recall, while lobular carcinoma has high recall but lower precision.

Macro averaging assigns equal importance to each of the eight classes, while weighted averaging reflects support. Since the final test support is 270 for every class, the two averages are expected to be close. This property is useful for the present evaluation because it allows the overall result to be viewed without a large support imbalance. Nevertheless, equal support in a balanced benchmark should not be confused with equal clinical importance; a future cost-sensitive analysis could explicitly model the consequences of different error types.

ROC/AUC Considerations

The source thesis includes one-vs-rest ROC/AUC evaluation, but the current implementation derives the curves from hard argmax predictions rather than the full softmax probability outputs. This distinction matters because a conventional ROC curve evaluates a continuous score over varying thresholds, whereas hard labels contain only the final class decision. A hard-label curve can therefore summarize the behavior of the implemented decision output, but it does not provide the same information as a probability-based ROC analysis.

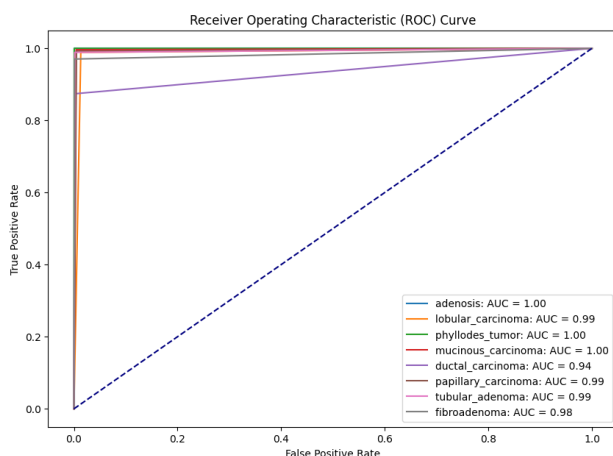


Figure 5 Receiver Operating Characteristic (ROC) Curve

For the final manuscript, accuracy and the class-wise precision, recall and F1-score are therefore treated as the primary quantitative test results. A future revision should preserve the full softmax matrix before applying argmax and calculate one-vs-rest ROC/AUC directly from each class probability vector. That change would allow threshold-independent discrimination analysis and would also make it possible to explore whether a different thresholding strategy could improve recall for the ductal carcinoma class without an uncontrolled increase in false positives.

Comparison with Existing Studies

The source thesis identifies Sharmin et al. as a relevant study using ResNet50V2 for deep feature extraction followed by ensemble machine-learning classifiers on an invasive ductal carcinoma dataset. Their study reports 95% accuracy for a binary detection problem. The present study reports 98% accuracy for an eight-class BreakHis classification problem. The two figures cannot be interpreted as an apples-to-apples ranking because the datasets, class definitions, task difficulty, preprocessing operations and evaluation protocols are different.

Table 7. Contextual comparison with a relevant breast histopathology study identified in the source thesis.

Study / System	Dataset / Task	Model / Strategy	Reported Result
Proposed framework	BreakHis; eight-class histopathology	InceptionResNetV2 + CBAM + CLAHE + balancing	98% accuracy
Sharmin et al. [27]	IDC histopathology; binary detection	ResNet50V2 + ensemble machine learning	95% accuracy

The contribution of the proposed framework is therefore better described in methodological terms than as a simple numerical advantage. The system combines a high-capacity transfer-learning backbone, explicit channel and spatial attention, image contrast enhancement, class balancing and multi-metric evaluation in one eight-class pipeline. The observed 98% test accuracy provides evidence of strong performance for that combined system on the reported benchmark. It does not establish that any one component is independently responsible for the result, nor does it establish superiority over studies using different datasets or validation designs.

The broader literature reviewed in the thesis includes work on optimized CNN systems, ensemble approaches, multimodal fusion, explainability, subgroup evaluation and alternative sensing modalities. The present contribution sits specifically within benchmarked histopathological image classification and emphasizes a focused combination of

transfer learning and attention. The literature also reinforces the need for patient-independent validation and external datasets before clinical claims are made.

Integrated Interpretation of the Quantitative Findings

The quantitative evidence can be grouped into three complementary layers. The first layer is convergence behavior, represented by the training and validation trajectories. Training accuracy rises from 96.83% to 98.44%, while validation accuracy remains near 97% throughout the run.

This establishes that the optimization procedure successfully learns a high-performing representation without a large collapse in validation performance. The second layer is aggregate test performance, where the complete system reaches 98% accuracy on 2,160 images. The third layer is class-level behavior, where the majority of classes achieve F1-scores of 0.98 or higher, but ductal carcinoma and lobular carcinoma remain comparatively weaker.

These three layers provide a more informative result narrative than the headline accuracy alone. A model can produce a high training accuracy without generalizing, or a high test accuracy while performing poorly on an important minority class. In the present balanced test set, the close agreement between accuracy and macro F1-score suggests that the aggregate result is reasonably representative of the eight-class distribution. At the same time, the lower ductal carcinoma recall demonstrates that the model still has a specific sensitivity gap that remains hidden if only overall accuracy is reported.

The result also illustrates the importance of separating observed facts from component-level explanations. The experiment establishes that the integrated InceptionResNetV2-CBAM pipeline, together with the reported preprocessing and balancing configuration, produces the stated test metrics.

It does not establish how much accuracy is contributed individually by CBAM, CLAHE, balancing or label smoothing. This distinction is central to reproducible research because a complex system can achieve strong performance while several modules may be redundant. Controlled ablation is therefore the natural next experiment rather than assuming that every component is equally responsible for the observed outcome.

Table 8. Integrated interpretation of the main quantitative and methodological findings.

Evidence layer	Observed finding	Interpretation
Training dynamics	98.44% train; 97.36% validation at epoch 10	Strong learning with a modest late-training gap
Test accuracy	98% on 2,160 images	High internal benchmark performance
Macro F1	0.98	Performance remains broadly balanced across classes
Strongest classes	Adenosis and phyllodes tumor: F1 1.00	Clear separation under the reported test protocol
Primary weakness	Ductal carcinoma: recall 0.87, F1 0.93	Most important class-specific error direction
Methodological constraint	Pre-split resampling and image-wise partitioning	Generalization may be optimistic; patient-wise validation needed

CONCLUSION AND FUTURE WORK

This study presented an eight-class breast histopathological image classification framework combining CLAHE-based preprocessing, mean-count class balancing, ImageNet-pretrained InceptionResNetV2 and CBAM attention with a regularized dense classifier. The latest ten-epoch training run reaches 98.44% training accuracy and 97.36% validation accuracy at epoch 10, with a maximum validation accuracy of 97.50% at epoch 2. The reported test evaluation reaches 98% accuracy on 2,160 images, with equal support of 270 images for each class and macro precision, recall and F1-score of approximately 0.98.

Class-wise evaluation provides a more detailed picture of the result. Adenosis and phyllodes tumor achieve F1-scores of 1.00, while tubular adenoma and papillary carcinoma reach 0.99. Mucinous carcinoma and fibroadenoma reach 0.98, lobular carcinoma reaches 0.95, and ductal carcinoma is the lowest at 0.93, with recall of 0.87. The class-wise results therefore show strong overall consistency while identifying a concrete area for additional analysis.

The study also demonstrates why a high benchmark score should be interpreted together with the experimental protocol. The current pipeline performs balancing before partitioning and uses image-wise rather than patient-wise splitting, which can increase the risk of optimistic

evaluation. In addition, the current ROC/AUC routine is based on hard labels, and no ablation study isolates the contribution of CBAM or CLAHE. These limitations do not erase the reported 98% test result, but they restrict the conclusions that can responsibly be drawn from it.

Future work should therefore focus on patient-wise or specimen-wise data partitioning, training-only resampling, independent external validation, probability-based ROC/AUC, systematic component ablation, calibration analysis and uncertainty estimation. Misclassification analysis should pay particular attention to ductal and lobular carcinoma, including inspection by magnification level and tissue morphology. With these improvements, the framework could be evaluated more rigorously as a reproducible research prototype for multi-class breast histopathology classification.

REFERENCES

- [1] Jaglan, Poonam, Rajeshwar Dass, and Manoj Duhan. "Breast cancer detection techniques: issues and challenges." *Journal of The Institution of Engineers (India): Series B* 100.4 (2019): 379–386.
- [2] Díaz, Oliver, Alejandro Rodríguez-Ruíz, and Ioannis Sechopoulos. "Artificial Intelligence for breast cancer detection: Technology, challenges, and prospects." *European Journal of Radiology* 175 (2024): 111457.
- [3] Nassif, Ali Bou, et al. "Breast cancer detection using artificial intelligence techniques: A systematic literature review." *Artificial Intelligence in Medicine* 127 (2022): 102276.
- [4] Gardezi, Syed Jamal Safdar, et al. "Breast cancer detection and diagnosis using mammographic data: Systematic review." *Journal of Medical Internet Research* 21.7 (2019): e14464.
- [5] Hassan, Ahmed M., and Magda El-Shenawee. "Review of electromagnetic techniques for breast cancer detection." *IEEE Reviews in Biomedical Engineering* 4 (2011): 103–118.
- [6] Nover, Adam B., et al. "Modern breast cancer detection: a technological review." *International Journal of Biomedical Imaging* 2009.1 (2009): 902326.
- [7] Hamed, Ghada, et al. "Deep learning in breast cancer detection and classification." *The International Conference on Artificial Intelligence and Computer Vision*. Cham: Springer International Publishing, 2020.
- [8] Barrios, Carlos H. "Global challenges in breast cancer detection and treatment." *The Breast* 62 (2022): S3–S6.
- [9] [Khan, Mudassir, et al. "Enhancing breast cancer detection through optimized thermal image analysis using PRMS-Net deep learning approach." *Journal of Imaging Informatics in Medicine* 39.1 (2026): 864–883.
- [10] Akbari, Younes, et al. "Dynamic model scaling based on segmented tumor size for breast cancer detection." *Biomedical Signal Processing and Control* 113 (2026): 109118.
- [11] Akbari, Younes, et al. "A novel virtual patient approach for cross-patient multimodal fusion in enhanced breast cancer detection." *Computerized Medical Imaging and Graphics* 127 (2026): 102687.
- [12] Mondal, Jyotirmoy, et al. "Artificial intelligence (AI) for breast cancer detection: Trends, challenges, and future directions." *Nano theragnostics in breast cancer: Advances, challenges, and future prospects*. Singapore: Springer Nature Singapore, 2026. 707–750.
- [13] Ahmad, Faizan, et al. "Ensemble Deep Learning-Based High-Precision Framework for Breast Cancer Detection from Histopathological Images." *Diagnostics* 16.5 (2026): 653.
- [14] Gao, Shuo, et al. "Application of deep learning technology in breast cancer: a systematic review of segmentation, detection, and classification approaches." *BioMedical Engineering OnLine* 25.1 (2026): 19.
- [15] Larsen, Marthe, et al. "Mammographic density assessment by an artificial intelligence model for breast cancer detection in BreastScreen Norway." *European Radiology* (2026): 1–12.
- [16] Owotogbe, Joshua, et al. "Machine learning and deep learning for breast cancer: A decade systematic review of detection, classification, prognosis, and explainability." *Informatics in Medicine Unlocked* (2026): 101756.
- [17] Brown-Mulry, Beatrice, et al. "Subgroup performance of a commercial digital breast tomosynthesis model for breast cancer detection." *Nature Communications* 17.1 (2026): 4249.
- [18] Toha, Tasnim Ferdous, et al. "Graphene assisted multilayer SPR biosensor with superior specificity for breast cancer cell detection." *Results in Engineering* (2026): 109739.
- [19] Wertani, Hamza, et al. "Advanced optimization of THz biosensor for breast cancer detection through

- machine learning.” IEEE Transactions on Plasma Science 54.4 (2026): 1750–1761.
- [20] Hoseini, Farnaz, and Masume Kheyri. “A Hybrid Capsule Network–Based Computer-Aided Diagnosis Framework for Breast Cancer Detection in Mammographic Images.” *Intelligence-Based Medicine* (2026): 100381.
- [21] Hlali, Aymen, et al. “Deep learning-based performance optimization of 2D material biosensors for breast cancer detection.” *Microsystem Technologies* 32.5 (2026): 53.
- [22] Bhardwaj, Kumar Kinjalk, and Vishal Pathak. “Nanostructured Materials and Their Physical Applications.” *International Journal of Innovations in Science, Engineering And Management* (2025): 413-420.
- [23] Appadurai, Jothi Prabha, et al. “Artificial intelligence-enhanced terahertz metasurface biosensor for breast cancer biomarker detection.” *Plasmonics* 21.1 (2026): 605–618.
- [24] Van Der Waal, Danielle, et al. “Finding the optimal recall rate in breast cancer screening: results from the ROCS study.” *European Radiology* 36.7 (2026): 5595–5603.
- [25] Tiwari, Sanjay Kumar, and Shashwat Tiwari. “A Philosophical Evaluation of the Role of Yoga in the Treatment of Cancer.” *International Journal of Innovations in Science Engineering And Management* (2026): 610-619.
- [26] Nakul, Mihir, et al. “Machine learning enhanced optical spectroscopy for breast cancer diagnosis: A review.” *Lasers in Medical Science* 41.1 (2026): 86.
- [27] Sharmin, Selina, et al. “A hybrid dependable deep feature extraction and ensemble-based machine learning approach for breast cancer detection.” *IEEE Access* 11 (2023): 87694–87708.
- [28] Suryawanshi, Pooja, and Pankaj Richhariya. “Maximizing Accuracy in Image Classification using Transfer Learning and Random Forest.” *International Journal of Innovations in Science, Engineering And Management* (2023): 39-48.
- [29] Sarasija, M. “Applications of Nanotechnology in Medicinal Chemistry Impact.” *AG Volumes* (2025): 57-63.