**International Journal of Informatics and Data Science**
ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 81−88
DOI: doi.org/10.64366/ijids.v3i2.613
https://journals.adaresearch.or.id/ijids/index

# Patient-Disjoint Evaluation for Trustworthy and Explainable Breast Cancer Histopathology Classification

**Ronal Watrianthos[*], Arif Rizki Marsa, Dian Eka Putra, Rozi Meri, Novi Efendi, Sofia Yosse**

Teknologi Informasi, Politeknik Negeri Padang, Padang, Indonesia
Email: [1,*]ronalwatrianthos@pnp.ac.id, [2]arifrizkimarsa@pnp.ac.id, [3]dianekaputra@pnp.ac.id.id, [4]rozimeri@pnp.ac.id, [5]noviefendi@pnp.ac.id, [6]syosse@pnp.ac.id
Correspondence Author Email: ronalwatrianthos@pnp.ac.id
Received: 20/05/2026, Revised: 20/06/2026, Accepted: 27/06/2026, Published: 30/06/2026

**Abstract–**Breast cancer diagnosis from histopathological images increasingly relies on deep learning, with recent studies reporting classification accuracies above 95% on the benchmark BreaKHis dataset. However, much of this literature uses image-level splits that let the same patient's images appear in both training and test sets a source of inflation that is not merely statistical but potentially clinically misleading if taken as evidence of real-world reliability and rarely pairs high accuracy with a rigorous account of model interpretability. This study addresses that gap with an explainable deep learning pipeline for binary (benign/malignant) BreaKHis classification under a patient-disjoint split. An EfficientNetB0 backbone, pretrained on ImageNet, was fine-tuned in two phases with class-weighted loss and Reinhard-based stain normalization. Evaluated under this leakage-safe protocol, the baseline model achieved a test-set balanced accuracy of 0.765 markedly lower than the above-95% figures reported elsewhere, yet a more trustworthy generalization estimate together with an F1-score of 0.777 for the malignant class, a ROC-AUC of 0.849, and a PR-AUC of 0.940, with the best checkpoint selected before overfitting deepened during fine-tuning. A regularization ablation against a more heavily regularized variant showed a narrower generalization gap but lower test performance, confirming the unregularized configuration as the more suitable final model. Grad-CAM was applied to the model's final convolutional layer to visualize the regions driving individual predictions, supporting qualitative inspection of correct and misclassified cases. These findings argue for combining leakage-aware evaluation with explainability as a joint, rather than separate, requirement for trustworthy histopathological classification models.

**Keywords**: BreaKHis Dataset; Data Leakage; EfficientNet; Explainable Artificial Intelligence; Grad-CAM

## 1. INTRODUCTION

Breast cancer remains one of the most commonly diagnosed cancers among women worldwide [1], [2], accounting for nearly one-third of all new cancer diagnoses in women in the United States alone [3]. Histopathological examination of biopsied tissue remains the diagnostic gold standard [4], [5], [6], [7], allowing pathologists to assess malignancy at the cellular level. Yet this process is manual, time-consuming, and subject to inter-observer variability, a panel of 115 pathologists reached full diagnostic consensus on only 75.3% of breast biopsy cases [8] a pattern corroborated by a national-scale UK NHS Breast Screening Programme quality-assurance study and, more recently, by an expert-panel study of borderline lesions [24] a reminder that expert visual judgment, however skilled, is not an infallible ground truth. This variability, combined with the sheer volume of tissue pathologists must review, has motivated a large and growing body of work on computer-aided diagnosis for breast histopathology.

Artificial intelligence, and deep learning in particular, offers a natural response to the limitations just described: unlike a single reviewing pathologist, a trained model applies the same decision criteria consistently across thousands of slides, offering a scalable second opinion that could, in principle, curb exactly the inter-observer variability noted above The public release of the BreaKHis dataset gave this line of research a common benchmark [9], and convolutional neural networks quickly became the dominant approach for classifying its benign and malignant tissue samples. Progress has been rapid on paper: a recent ensemble of EfficientNetV1 and EfficientNetV2 backbones reports a binary classification accuracy of 99.58% on BreaKHis [10], and figures in the high-90s are now common throughout the literature. Many of these studies also incorporate Grad-CAM to visualize which image regions drive a given prediction [11], [12], echoing a concern raised throughout the medical imaging literature: a classifier that cannot explain itself is difficult for clinicians to trust, whatever its benchmark accuracy.

Closer inspection of how these near-ceiling accuracies is obtained, however, reveals a methodological gap that an explainability layer alone does not resolve. BreaKHis contains many images per patient, and partitioning the dataset by randomly sampling individual images, rather than by patient, allows images from the same patient, and therefore the same staining batch and tissue morphology, to appear in both the training and test sets. This is not a hypothetical concern: the ensemble study reporting 99.58% accuracy states that its training, validation, and test sets were formed via random sampling, with no mention of grouping by patient [13], and image-level splitting of this kind has been shown, across digital pathology more broadly, to inflate reported performance by up to 41% relative to a patient-disjoint evaluation [14].

Explainability tools such as Grad-CAM are typically applied downstream of this split [11], [12], [15], which means a saliency map can look diagnostically plausible while the very number it is explaining remains inflated by leakage. In effect, much of the literature treats evaluation rigor and interpretability as separate concerns, when a model's explanations are only as trustworthy as the evaluation protocol used to validate it in the first place. To our knowledge, few studies report Grad-CAM-based explainability alongside a strictly patient-disjoint evaluation and an explicit accounting of the resulting generalization gap, this is the gap the present study addresses.

**International Journal of Informatics and Data Science**
ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 81−88
DOI: doi.org/10.64366/ijids.v3i2.613
https://journals.adaresearch.or.id/ijids/index

We build an EfficientNetB0 classification pipeline for BreaKHis under a patient-disjoint train/validation/test split [16], quantify the overfitting that emerges during fine-tuning rather than reporting only a final test score, run a regularization ablation to test whether that overfitting can be reduced without sacrificing generalization, and apply [11], [17] to qualitatively inspect the resulting model's predictions. The outcome is a lower headline accuracy than much of the ensemble and high-accuracy literature reports but one presented alongside the evaluation protocol needed to judge what that number actually means.

## 2. RESEARCH METHODOLOGY

This study follows the seven-stage pipeline shown in Fig. 1, proceeding from raw data acquisition through to model interpretability analysis. Each stage's output feeds directly into the next; in particular, the regularization ablation (Stage 5) and the explainability analysis (Stage 7) are shaped by diagnostics collected at the preceding stage rather than fixed in advance. This study uses the BreaKHis dataset [9], comprising 7,909 breast tumor histopathological images from 82 patients across four magnification factors (40×, 100×, 200×, and 400×), with 2,480 benign and 5,429 malignant samples. Each image is a 700×460-pixel, three-channel RGB PNG file at 8-bit depth per channel.

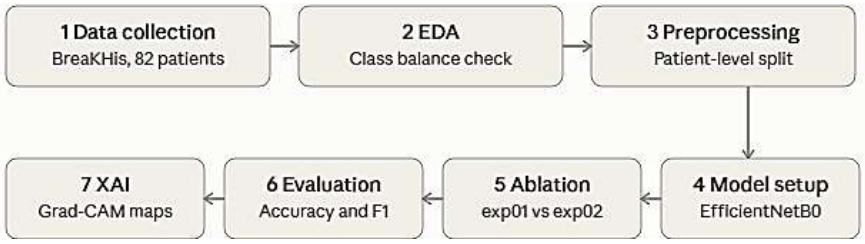


**Figure 1.** Research Stages

To prevent patient-level leakage, shown elsewhere to inflate reported performance by up to 41% in digital pathology when this precaution is omitted [14], the dataset was partitioned at the patient level using a class-stratified split (70%/15%/15% for training, validation, and test). Every image belonging to a given patient, regardless of magnification, was assigned to exactly one split. Hematoxylin-and-eosin (H&E) staining intensity varies across patients and acquisition batches. This variability is reduced using Reinhard color transfer, applied independently to each channel *c* of the LAB color space:

$$I'_c = \frac{I_c - \mu_{s,c}}{\sigma_{s,c}} \times \sigma_{t,c} + \mu_{t,c} \quad (1)$$

where *I_c* and *I'_c* denote the original and normalized channel intensities, *μ_{s,c}* and *σ_{s,c}* are the source image's channel mean and standard deviation, and *μ_{t,c}* and *σ_{t,c}* are target statistics computed as the mean over a random sample of training images, rather than a single reference image, for robustness to staining outliers.

### 2.1 Model Architecture and Training Configuration

We adopt EfficientNetB0 as the classification backbone [18], [19], [20], initialized with ImageNet-pretrained weights. EfficientNetB0 was selected for its compound scaling of network depth, width, and input resolution, which yields a strong accuracy-to-parameter-count ratio well suited to the moderate size of the BreaKHis training set. The original 1,000-class classification head is replaced with a fully connected layer projecting to two classes (benign, malignant).

Training proceeds in two phases. Phase 1 freezes the entire convolutional backbone and trains only the classifier head for 5 epochs (learning rate $1\times10^{-3}$), allowing the randomly initialized head to stabilize before any backbone weight is perturbed. Phase 2 unfreezes the final two convolutional blocks and fine-tunes them jointly with the head, using discriminative learning rates ($1\times10^{-5}$ for the backbone, $1\times10^{-4}$ for the head) for up to 20 epochs, with early stopping (patience = 5) monitored on validation balanced accuracy.

Training images are additionally augmented with random horizontal and vertical flips and random rotation (±20°) reasonable given histopathological tissue has no canonical orientation applied identically in exp01 and exp02. We did not extend to heavier geometric augmentation (e.g., random-resized cropping, elastic deformation), because Section [exp02 ablation] shows that this study's compute-constrained epoch budget is already sensitive to added regularization: exp02's combination of stronger regularization and a shortened schedule reduced test performance even as it narrowed the generalization gap. Adding further augmentation without a proportional increase in training epochs risks the same underfitting outcome; we treat this as future work once compute budget allows a matched-epoch comparison. Model outputs are converted to class probabilities via the softmax function:

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (2)$$

To counteract the benign/malignant class imbalance, each class c is assigned an inverse-frequency weight:

**International Journal of Informatics and Data Science**
ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 81−88
DOI: doi.org/10.64366/ijids.v3i2.613
https://journals.adaresearch.or.id/ijids/index

$$w_c = \frac{N}{K \times n_c} \quad (3)$$

where N is the total number of training images, K is the number of classes (K=2), and n_c is the number of training images belonging to class c. These weights enter the classification loss as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log(p_{i,y_i}) \quad (4)$$

where y_i is the true label of sample i and p_{i,y_i} is the predicted probability assigned to that label. The baseline configuration (exp01) uses the Adam optimizer without weight decay. A regularization ablation (exp02) instead uses AdamW [21], which decouples weight decay from the gradient-based update, together with an increased classifier dropout rate (0.2→0.45) and label smoothing (0.08).

## 3. RESULT AND DISCUSSION

The BreaKHis dataset used in this study comprises 7,909 breast tumor histopathological images collected from 82 patients across four magnification factors (40×, 100×, 200×, and 400×), consisting of 2,480 benign and 5,429 malignant samples [9]. Each image measures 700×460 pixels, three-channel RGB, 8-bit depth per channel, PNG format. Table 1 presents the full distribution across magnification factors. With a benign-to-malignant ratio of roughly 1:2.19 (Fig. 1), the dataset carries a non-trivial class imbalance. We address this through class-weighted loss during training.

**Table 1.** Distribution of the breakhis dataset by class and magnification factor

| Magnification | Benign | Malignant | Total |
|---|---|---|---|
| 40× | 625 | 1,370 | 1,995 |
| 100× | 644 | 1,437 | 2,081 |
| 200× | 623 | 1,390 | 2,013 |
| 400× | 588 | 1,232 | 1,820 |
| Total | 2,480 | 5,429 | 7,909 |

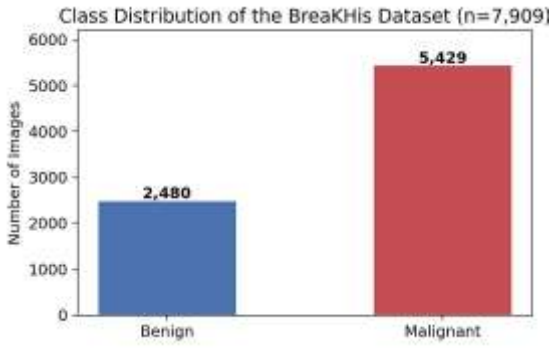


**Figure 2.** Class Distribution

Data were partitioned at the patient level rather than the image level, using a class-stratified split with a target ratio of 70%/15%/15% for training, validation, and test sets. This distinction matters: because each BreaKHis patient contributes many images, a naive image-level random split risks placing images from the same patient in both the training and test sets a leakage mode shown to inflate reported performance by up to 41% in digital pathology studies more broadly [14]. Stain normalization followed the Reinhard method in LAB color space to reduce inter-patient variability in hematoxylin-and-eosin (H&E) staining. Target statistics were computed as the mean over a sample of training images, rather than from a single reference image, for greater robustness to staining outliers.

### 3.1 Baseline Model Training Results (exp01)

An EfficientNetB0 backbone, pretrained on ImageNet, was fine-tuned in two phases: Phase 1 with the backbone frozen (5 epochs, classifier head only) and Phase 2 with the last two convolutional blocks unfrozen (up to 20 epochs, early stopping with patience 5). Table 2 reports test-set performance at the checkpoint with the highest validation balanced accuracy.

**Table 2.** Baseline Model (exp01) Test-Set Performance

| Metric | Value |
|---|---|
| Balanced Accuracy | 0.7651 |
| F1-score (Malignant) | 0.7773 |
| ROC-AUC | 0.8489 |

 

**International Journal of Informatics and Data Science**
ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 81−88
DOI: doi.org/10.64366/ijids.v3i2.613
https://journals.adaresearch.or.id/ijids/index

| Metric | Value |
|---|---|
| PR-AUC | 0.9397 |
| Loss | 0.6134 |

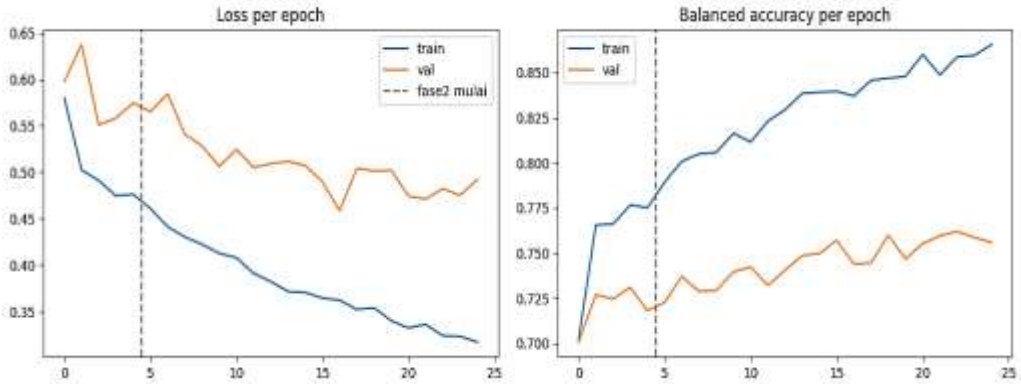


**Figure 3.** Training Validation Loss and Balanced Accuracy Per Epoch

The training history shows that the best checkpoint occurred at epoch 18 of Phase 2 (validation balanced accuracy = 0.762). Beyond this point, the model showed clear signs of overfitting: training balanced accuracy kept climbing to 0.866 by epoch 20, while validation balanced accuracy plateaued in the 0.72–0.76 range. Validation loss likewise plateaued around 0.47–0.51 from roughly epoch 11 onward, even as training loss continued falling to 0.317 a generalization gap of about 0.10–0.11 by the end of training. The checkpoint-on-best mechanism nonetheless captured the model's peak generalization before overfitting deepened further: test-set performance (balanced accuracy = 0.765) closely tracks the validation score at that checkpoint (0.762), which suggests the test result is not artificially inflated by residual leakage.

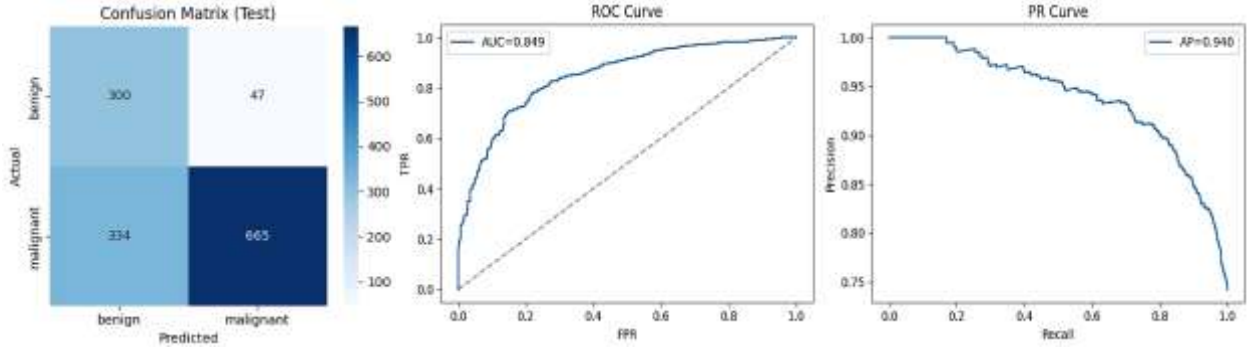


**Figure 4.** Confusion Matrix and ROC Precision-Recall Curves

A balanced accuracy of 0.765 deserves careful framing. Many BreaKHis studies reporting 95–98% accuracy rely on image-level splitting, which risks leaking patient identity across training and test sets — a bias shown to inflate performance scores by up to 41% in digital pathology more generally [14]. Because this study enforces a strict patient-level split (Section IV-B), 0.765 more plausibly reflects the model's true generalization ability, even though it appears numerically modest next to comparison studies that do not control for this leakage.

### 3.2 Regularization Ablation: exp01 vs. exp02

To test whether the generalization gap observed in exp01 could be narrowed, we retrained the model with additional regularization (exp02): AdamW with weight decay, classifier dropout raised from 0.2 to 0.45, label smoothing of 0.08, a shortened Phase 2 (20→15 epochs, patience 5→3), and an optional light ColorJitter augmentation. Data splits, stain-normalization targets, and class weighting were kept identical to exp01 to keep the comparison apples-to-apples.

**Table 3.** Test-set performance comparison: exp01 vs. Exp02

| Metric | exp01 (baseline) | exp02 (regularized) | Δ |
|---|---|---|---|
| Balanced Accuracy | 0.7651 | 0.7482 | −0.0169 |
| F1-score (Malignant) | 0.7773 | 0.7692 | −0.0081 |
| ROC-AUC | 0.8489 | 0.8398 | −0.0091 |
| PR-AUC | 0.9397 | 0.9345 | −0.0052 |

**International Journal of Informatics and Data Science**
ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 81−88
DOI: doi.org/10.64366/ijids.v3i2.613
https://journals.adaresearch.or.id/ijids/index

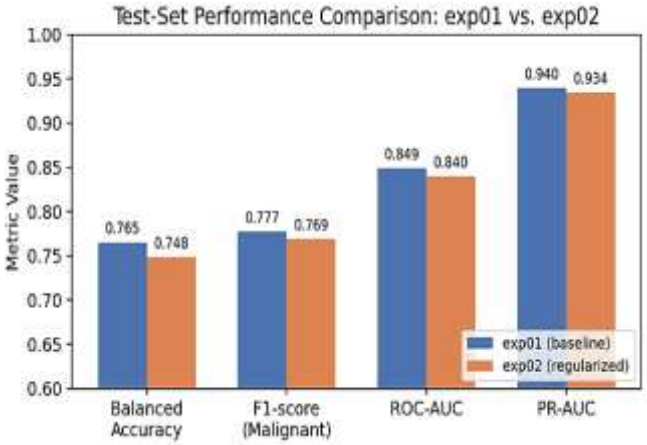


**Figure 5.** Confusion Matrix and ROC Precision-Recall Curves

Every metric declined under exp02 (Table 3), a consistent direction across all four metrics, indicating a systematic effect rather than run-to-run noise. The exp02 training history (Fig. 6) shows Phase 2 stopping early at epoch 8, triggered by early stopping with patience 3, with the best checkpoint at Phase 2 epoch 5 (validation balanced accuracy = 0.761). The generalization gap at exp02's best checkpoint narrowed to roughly 0.055, compared with ≈0.10–0.11 for exp01, yet test-set performance dropped to 0.748. This pattern suggests that the added regularization did curb overfitting, but the combination of tighter early stopping and simultaneous regularization cut training short before the model could reach a capacity comparable to exp01 — a case of relative underfitting rather than a genuine improvement. Based on this ablation, the unregularized exp01 recipe was retained as the final model for this study.

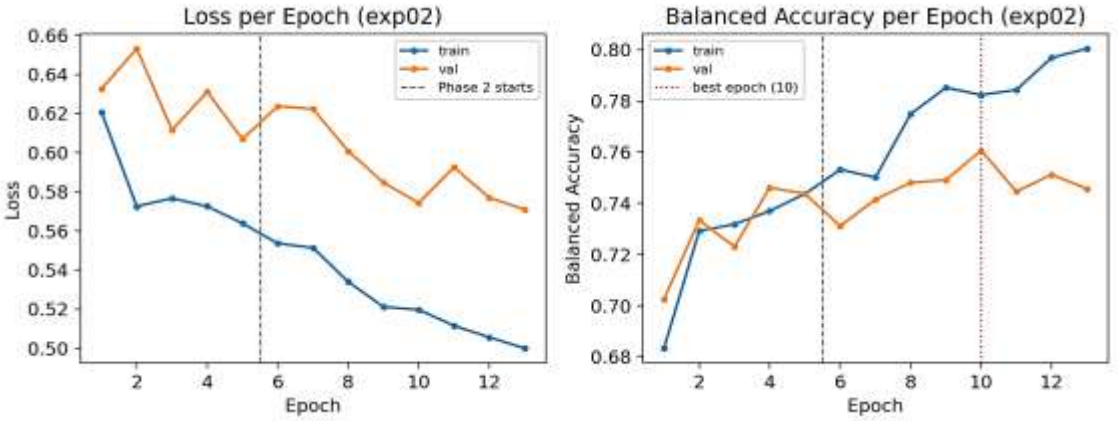


**Figure 6.** Exp02 Training History

### 3.3 Interpretability Analysis (Grad-CAM)

Previous findings have indicated that pathologists often disagree in a significant number of cases [8]; therefore, the results generated by a classification system will be most useful if clinicians can also review the evidence on which it is based. Therefore, we supplemented the quantitative evaluation with Grad-CAM [11], which was applied to the last convolutional layer of EfficientNetB0 (model.features[-1], 1,280 channels) in accordance with Equations (7)–(8). Grad-CAM was chosen over perturbation-based alternatives such as LIME or SHAP primarily due to its computational efficiency. This method reuses gradients that have already been computed during the forward/backward process, rather than requiring repeated model evaluations on perturbed inputs, and has proven effective particularly for CNN-based visual classification tasks such as the one used here.

We examine three case groups drawn from the test set, each serving a distinct diagnostic purpose rather than being an arbitrary sample: (i) correctly classified benign cases and (ii) correctly classified malignant cases, to check whether correct predictions are driven by plausible tissue evidence rather than an incidental shortcut; and (iii) misclassified cases, since inspecting where the model looked when it was wrong is arguably more informative for error auditing than inspecting only its successes. Models that suffer from overfitting are known to rely more on spurious and non-causal correlations in the training data than on features relevant to the task. The “checkpoint-on-best” selection described above is not only an accuracy-related decision but also a readability-related decision, where the training balanced accuracy reaches 0.866 while the validation balanced accuracy plateaus at ~0.75.

International Journal of Informatics and Data Science
ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 81−88
DOI: doi.org/10.64366/ijids.v3i2.613
https://journals.adaresearch.or.id/ijids/index

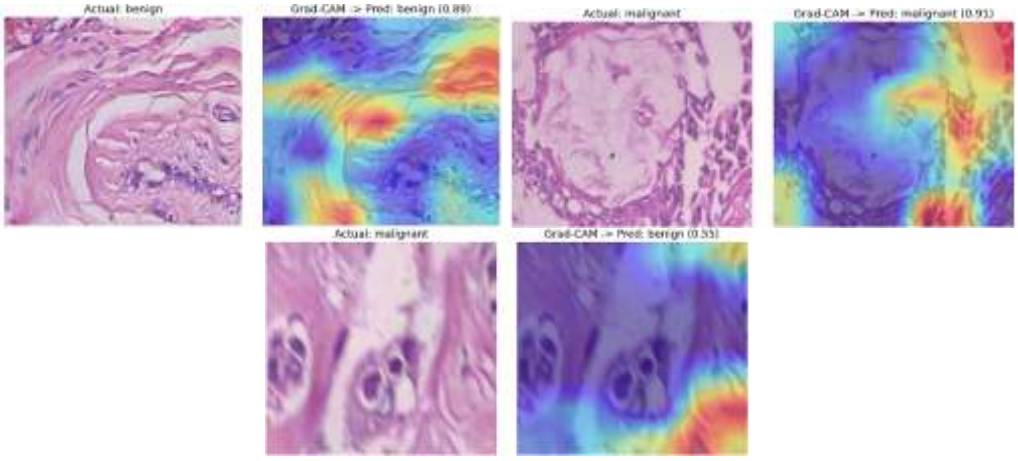


**Figure 7.** Grad-CAM Visualizations

Concretely, a pathologically plausible result would show Fig. 7's highlighted regions concentrated on pleomorphic cell nuclei, irregular nuclear contours, hyperchromasia, or elevated nuclear-to-cytoplasmic ratio in the malignant cases \u2014 the features a pathologist would cite for that diagnosis u2014 rather than on non-tissue background or staining artifacts. This analysis has clear technical limitations. The saliency map is derived from the feature map of the final convolutional layer, whose resolution has been significantly reduced compared to the original 700×460 input. As a result, Grad-CAM highlights broad areas rather than specific nuclei or cell boundaries [11]. Therefore, this visualization is interpreted qualitatively at the network region level, rather than as precise pixel-level segmentation. This analysis remains qualitative because there has been no quantitative validation of the saliency's reliability, such as the use of subtraction/addition metrics to assess changes in predicted probabilities when the highlighted regions are altered. This represents a potential area for future research, along with cross-validation at the patient level (k=3).

### 3.4 Discussion

The baseline balanced accuracy of 0.765 is well below the 95%+ figures commonly found in the BreaKHis literature; however, this discrepancy is a consequence of the evaluation protocol, not because the underlying model is weaker. The precise, non-overlapping patient separation eliminates the leakage pathways that have been shown elsewhere to improve these figures; therefore, the two figures do not measure the same thing and should not be compared literally. The narrower generalization gap in exp02, coupled with lower test performance, indicates that overfitting in exp01 is indeed present, but can only be partially mitigated through regularization alone. This makes exp01 a more reliable baseline, not because it has less overfitting, but because its "checkpoint-on-best" procedure captured its best generalization point before the overfitting became more severe.

Grad-CAM analysis extends this same evaluation logic to the realm of interpretability. Saliency maps computed on models that are safe from leakage are inherently more reliable objects of study than those computed on inflated models, regardless of whether their content is ultimately deemed medically plausible. The current limitation of this analysis lies in its qualitative scope, not in the method itself. Validation of removal/addition, as well as a true comparison between saliency maps on overfit models and those on the final model, remain unresolved issues, and neither will be meaningful without a pre-established evaluation framework. Overall, these results show that evaluation accuracy, the interpretation of ablation-based training dynamics, and interpretability are not standalone criteria, but rather an interdependent chain, with direct implications for whether the resulting model can be trusted as more than just a research artifact

## 4. CONCLUSION

This study developed and evaluated a deep learning pipeline for the binary classification of breast cancer histopathological images from the BreaKHis dataset, built around a transfer-learned EfficientNetB0 backbone, a strictly patient-level data split, Reinhard-based stain normalization, and Grad-CAM-based explainability. The baseline model achieved a test-set balanced accuracy of 0.765, an F1-score of 0.777 for the malignant class, and a ROC-AUC of 0.849. A subsequent regularization ablation confirmed that this configuration, rather than a more heavily regularized alternative, offered the better trade-off between generalization and predictive capacity. While these figures sit below the 95–98% accuracy commonly reported in the BreaKHis literature, the gap is best read as a methodological correction rather than a shortcoming. Grad-CAM visualizations further show that the model's predictions can be qualitatively inspected rather than treated as a black box, laying the groundwork for the quantitative interpretability analysis, patient-level cross-validation, and multi-backbone comparison identified as directions for future work. These results also carry a direct implication for how this class of model should be

 

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 81−88
DOI: doi.org/10.64366/ijids.v3i2.613
https://journals.adaresearch.or.id/ijids/index

positioned relative to clinical practice. A test-set balanced accuracy of 0.765 does not support using this model as an autonomous diagnostic or decision-support tool: it falls well short of the sensitivity and specificity levels expected of regulatory-cleared breast-imaging AI systems, and it has been evaluated on a single, retrospective, single-institution dataset without the external, multi-site validation such regulatory pathways require before any deployment claim.

# 5. DECLARATIONS

### 5.1 Author Contributions

Conceptualization: R.W;
Methodology: R.W, A.R.M;
Formal Analysis: D.E.P;
Investigation: R.M;
Writing Original Draft Preparation: N.E;
Visualization: S.Y;
All authors have read and agreed to the published version of the manuscript.

### 5.2 Data Availability

The data presented in this study are available on request from the corresponding author.

### 5.3 Funding

This research was funded by the DIPA funding from the Politeknik Negeri Padang through the 2026 Fundamental Research Grant Program. The authors express their sincere gratitude for the support provided.

### 5.4 Institutional Review Board Statement

Not applicable.

### 5.5 Informed Consent Statement

Not applicable.

### 5.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### 5.7 Generative AI and AI-assisted Technologies in The Writing Process

The authors used Generative AI to improve the writing clarity of this paper. They reviewed and edited the AI-assisted content and took full responsibility for the final publication. Thankyou to those who have supported the implementation of this research.

# REFERENCES

[1] R. Watrianthos, R. Rayendra, E. Asri, Y. Yuhefizar, and H. Humaira, "Threshold-Optimized and Calibrated Logistic Regression for Breast Cancer Classification," *Salud, Ciencia y Tecnología*, vol. 5, p. 2241, Oct. 2025, doi: 10.56294/saludcyt20252241.

[2] Agus Perdana Windarto, Anjar Wanto, S Solikhun, and Ronal Watrianthos, "A Comprehensive Bibliometric Analysis of Deep Learning Techniques for Breast Cancer Segmentation: Trends and Topic Exploration (2019-2023)," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 5, pp. 1155–1164, Oct. 2023, doi: 10.29207/resti.v7i5.5274.

[3] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024. Cancer Facts & Figures 2024.," *CA Cancer J. Clin.*, 2024.

[4] B. Jiang, L. Bao, S. He, X. Chen, Z. Jin, and Y. Ye, "Deep learning applications in breast cancer histopathological imaging: diagnosis, treatment, and prognosis," *Breast Cancer Research*, vol. 26, no. 1, p. 137, Sep. 2024, doi: 10.1186/s13058-024-01895-6.

[5] E. M. Senan, F. W. Alsaade, M. I. A. Al-Mashhadani, T. H. H. Aldhyani, and M. H. Al-Adhaileh, "Classification of histopathological images for early detection of breast cancer using deep learning," *Journal of Applied Science and Engineering (Taiwan)*, vol. 24, no. 3, 2021, doi: 10.6180/jase.202106_24(3).0007.

[6] L. Aldakhil, H. Alhasson, and S. Alharbi, "Attention-Based Deep Learning Approach for Breast Cancer Histopathological Image Multi-Classification," *Diagnostics*, vol. 14, no. 13, p. 1402, Jul. 2024, doi: 10.3390/diagnostics14131402.

[7] T. A. Toma *et al.*, "Breast Cancer Detection Based on Simplified Deep Learning Technique With Histopathological Image Using BreaKHis Database," *Radio Sci.*, vol. 58, no. 11, Nov. 2023, doi: 10.1029/2023RS007761.

[8] J. G. Elmore *et al.*, "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *JAMA - Journal of the American Medical Association*, vol. 313, no. 11, 2015, doi: 10.1001/jama.2015.1405.

[9] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A Dataset for Breast Cancer Histopathological Image Classification," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1455–1462, Jul. 2016, doi: 10.1109/TBME.2015.2496264.

 

ISSN 3026-7315 (media online)
Volume 3, No 2, June 2026, Page: 81−88
DOI: doi.org/10.64366/ijids.v3i2.613
https://journals.adaresearch.or.id/ijids/index

[10] T. T. Brunyé *et al.*, "Machine learning classification of diagnostic accuracy in pathologists interpreting breast biopsies," *Journal of the American Medical Informatics Association*, vol. 31, no. 3, 2024, doi: 10.1093/jamia/ocad232.
[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.
[12] A. Alqutayfi *et al.*, "Explainable Disease Classification: Exploring Grad-CAM Analysis of CNNs and ViTs," *Journal of Advances in Information Technology*, vol. 16, no. 2, 2025, doi: 10.12720/jait.16.2.264-273.
[13] M. Azmoodeh-Kalati, H. Shabani, M. S. Maghareh, Z. Barzegar, and R. Lashgari, "Leveraging an ensemble of EfficientNetV1 and EfficientNetV2 models for classification and interpretation of breast cancer histopathology images," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-06853-6.
[14] N. Bussola, A. Marcolini, V. Maggio, G. Jurman, and C. Furlanello, "AI Slipping on Tiles: Data Leakage in Digital Pathology," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2021. doi: 10.1007/978-3-030-68763-2_13.
[15] S. Liu, G. M. S. Himel, and J. Wang, "Breast Cancer Classification With Enhanced Interpretability: DALAResNet50 and DT Grad-CAM," *IEEE Access*, vol. 12, pp. 196647–196659, 2024, doi: 10.1109/ACCESS.2024.3520608.
[16] H. Guo, J. Zhang, Y. Li, X. Pan, and C. Sun, "Advanced pathological subtype classification of thyroid cancer using efficientNetB0," *Diagnostic Pathology* , vol. 20, no. 1, 2025, doi: 10.1186/s13000-025-01621-6.
[17] H. T. Vo, N. N. Thien, K. C. Mui, and P. P. Tien, "Enhancing Confidence in Brain Tumor Classification Models With Grad-CAM and Grad-CAM++," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 12, no. 4, 2024, doi: 10.52549/ijeei.v12i4.5977.
[18] M. Harahap *et al.*, "Skin cancer classification using EfficientNet architecture," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 4, 2024, doi: 10.11591/eei.v13i4.7159.
[19] M. I. Rosadi, L. Hakim, and M. Faishol A., "Classification of Coffee Leaf Diseases using the Convolutional Neural Network (CNN) EfficientNet Model," *Conference Series*, vol. 4, no. 1, 2023, doi: 10.34306/conferenceseries.v4i1.627.
[20] I. N. Alam, I. H. Kartowisastro, and P. Wicaksono, "Transfer Learning Technique with EfficientNet for Facial Expression Recognition System," *Revue d'Intelligence Artificielle*, vol. 36, no. 4, 2022, doi: 10.18280/ria.360405.
[21] H. Ali, N. Shifa, R. Benlamri, A. A. Farooque, and R. Yaqub, "A fine tuned EfficientNet-B0 convolutional neural network for accurate and efficient classification of apple leaf diseases," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-04479-2.

 