

Stain Normalization for Cross-Dataset Generalization in Breast Cancer Histopathology

Sajal Kumar Kar*, Achyut Pandey**

*Department of Computer Science & Engineering, APS University, Satna, Madhya Pradesh, India

** Department of Physics, T.R.S. University, Rewa, Madhya Pradesh, India

ABSTRACT

Models trained on one breast-histopathology dataset are rarely tested on a genuinely different one, so how much of their reported accuracy survives a change of acquisition source is often unknown. We trained an EfficientNet-B0 classifier on BreakHis under a strictly patient-level split (validation F1 = 0.917, AUC = 0.929) and evaluated it, unmodified, on the independent BACH dataset, first on raw images and then after Macenko stain normalization against a single fixed BreakHis reference image. Raw BACH performance collapsed to near-chance ranking (AUC = 0.597) with benign-class recall of only 0.18, indicating the model had effectively learned to default to the majority class outside its training distribution. Normalization raised AUC to 0.647, but a paired DeLong test on the same 400 images shows this difference is not statistically significant ($p = 0.070$), and accuracy and F1 were essentially unchanged. We report this as a genuinely mixed, preliminary result rather than a confirmed benefit: a single reference image may not represent the method's best case, and a multi-reference sensitivity analysis is already underway to determine whether the pattern seen here is stable or an artefact of the one reference image used. We see value in publishing this intermediate, honestly-reported finding, including the parts that did not turn out as expected, over withholding it until a more favourable number appears.

Keywords:— breast cancer histopathology, stain normalization, Macenko method, domain shift, cross-dataset generalization, BreakHis, BACH, deep learning

I. INTRODUCTION

Breast cancer is diagnosed by a pathologist examining stained tissue under a microscope, and deep learning has been applied to this task with considerable success on public benchmarks [20], most heavily on BreakHis, which provides 7,909 images from 82 patients. What is reported far less often is how a model trained this way behaves once it leaves the dataset it was trained on, which matters if the work is meant to generalize beyond the lab that produced the training images.[35]

Part of the reason cross-dataset performance drops is not biological at all: two histopathology labs rarely stain tissue the same way, since reagent batches, incubation times, and scanner colour response all differ, so genuinely identical tissue can look different side by side. This paper asks a narrow, answerable question about that gap: how much of the drop from BreakHis to the independent BACH dataset can be recovered by a simple, training-free colour correction, and how much is left over? We use Macenko stain normalization, applied only at evaluation time to the target-domain images, with the source-trained model held completely fixed, so that any difference in outcome can be attributed specifically to colour rather than to a change in the model.

To be clear about what is and is not new here: Macenko normalization itself is an established method [3], not a contribution of this paper. What we contribute is a controlled BreakHis-to-BACH evaluation of test-time Macenko normalization on the same frozen classifier and the same 400 target images, reported with confusion matrices, class-wise metrics, bootstrap confidence intervals, and a paired

significance test rather than headline accuracy alone — and reported honestly even where, as here, the result is more modest and less conclusive than a first informal look suggested.

II. RELATED WORK

A. Stain Normalization in Digital Pathology

Colour normalization for H&E images predates deep learning. Reinhard et al. [7] matched colour-channel statistics to a reference but is not stain-aware; Ruifrok and Johnston [8] introduced colour deconvolution using known stain absorption spectra. Macenko et al. [3], the method used here, instead estimates stain vectors per image via SVD of the optical-density values, making it adaptive rather than dependent on fixed absorption coefficients. Vahadane et al. [5] separate stains with sparse non-negative matrix factorization while preserving tissue structure; Khan et al. [6] proposed an image-specific, classifier-driven variant; and Gavrilovic et al. [15] approached the same decomposition as a blind source-separation problem. More recent work has moved toward learned normalization: Kang et al. [12] trained a lightweight network (StainNet) to mimic reference-based normalization cheaply, and Hetz et al. [11] used a cycle-consistent GAN to normalize across multiple stain domains at once. Tosta et al. [10] combined dictionary learning with stain normalization specifically for breast histology. Tellez et al. [9], in one of the more thorough comparisons in this space, found that normalization and stain-specific augmentation are not interchangeable, and that normalization is aimed specifically

at bridging a known, fixed distributional gap — exactly the situation addressed here.

B. Cross-Dataset Generalization in Breast Histopathology

Direct BreakHis-to-BACH or comparable cross-site evaluations remain less common than within-dataset benchmarking. Sani et al. [17] trained a contrastive-learning classifier on BreakHis and evaluated transfer to BACH using the same binary label mapping adopted here, finding that generalization depended on how the representation was learned. Stanescu and Stoica-Spahiu [29] compared threshold recalibration, head-only fine-tuning, and full fine-tuning for cross-site breast-histopathology transfer under strict patient-level separation, and found even simple recalibration recovered a meaningful share of lost performance; a related study [30] examined the same kind of transfer through the lens of false-negative behaviour specifically, arguing that aggregate accuracy can mask a clinically important asymmetry between missed malignant and missed benign cases — a concern this paper's own benign-recall finding echoes directly. Mahbod et al. [16] showed that even the timing of stain normalization (train-time versus test-time) affects how well a segmentation model generalizes, and the MIDOG challenge [13] made scanner-level domain generalization the primary evaluation axis for mitosis detection. What distinguishes the present study is less the presence of a cross-dataset evaluation, which these works already include, than its deliberately single-variable design (Section III-E) and its willingness to report a statistically inconclusive result rather than only a favourable one.

C. Patient-Level Evaluation and Data Leakage

Benhammou et al. [21] documented that a substantial share of published BreakHis results rely on an image-level rather than patient-level split, letting multiple images of one patient's tissue appear on both sides of a split and inflating reported accuracy. Stanescu and Stoica-Spahiu [29], [30] make the same point in their own cross-site work. The baseline here is trained and validated under an equivalent patient-level protocol (Section III-B), which matters because the within-dataset number the cross-dataset comparison is measured against would otherwise be suspect.

III. MATERIALS AND METHODS

A. Datasets

BreakHis [1] provides 7,909 H&E-stained breast tumour images from 82 patients at four magnifications, labelled benign or malignant; it is the source domain used for training. BACH [2], from the ICIAR 2018 Grand Challenge, contributes 400 images across four classes (normal, benign, in situ, invasive) collected under acquisition conditions entirely independent of BreakHis. Following [17], normal and benign BACH images are mapped to benign and in situ/invasive to malignant. BACH is used exclusively as a target-domain test set: no BACH image is ever used for training or checkpoint selection.

B. Patient-Level Splitting

All BreakHis splitting is done at the patient level to avoid the leakage problem in Section II-C. The 82 patients are partitioned into training, validation, and test groups at an approximate 70:15:15 ratio, stratified by benign/malignant patient ratio, and only afterward are a patient's images assigned to their group; exact per-split patient and image counts are being finalized for a revised version of this paper rather than stated here without having re-verified them.

C. Baseline Classification Model

The classifier is an EfficientNet-B0 backbone [4], pretrained on ImageNet [23], trained on the full pool of BreakHis training images across all four magnifications (each image an independent sample, magnification not used as a model input[36]; magnification-specific versus magnification-independent training is compared in [18], [19]). We chose EfficientNet-B0 over a residual [14] or Vision Transformer [22] backbone for its parameter efficiency under a 6 GB GPU budget. Training runs for 20 epochs, batch size 32, AdamW [24] at learning rate 1×10^{-4} and weight decay 1×10^{-4} , a ReduceLROnPlateau schedule halving the rate after 3 stagnant epochs, class-weighted cross-entropy for BreakHis's ~69:31 imbalance, and mixed-precision training [25] on a single 6 GB RTX 3050 GPU. No early stopping is used; the epoch with the best validation F1 is kept. Images are resized to 224×224 with ImageNet normalization; flipping, rotation ($\leq 20^\circ$), and mild colour jitter apply during training only. Code is in PyTorch [26] with scikit-learn [27] for metrics. No fixed random seed is currently set, which we note as a reproducibility limitation.

D. Macenko Stain Normalization

Macenko normalization [3] works in optical-density (OD) space, since stain concentrations combine additively under the Beer–Lambert law:

$$OD = -\log_{10}((I + 1) / 256) \quad (1)$$

Pixels with OD below $\beta = 0.15$ in any channel are treated as background. For the remaining tissue pixels, the covariance of the OD values is eigen-decomposed; the top two eigenvectors define the plane containing the hematoxylin and eosin directions. Each pixel's OD is projected onto this plane and its angle computed; the α -th and $(100-\alpha)$ -th percentiles ($\alpha = 1$), rather than the extremes, give two robust stain vectors. Per-pixel concentrations are recovered by least squares against the stain matrix. To normalize a source image against a reference, both images' stain matrices and 99th-percentile concentrations are estimated this way; the source's concentrations are rescaled to match the reference's 99th percentile and recombined with the reference's stain matrix, then converted back to RGB via the inverse of Equation (1). A single, fixed BreakHis image is used as the reference for the result reported in Section IV; the method is implemented as a self-contained NumPy routine rather than an existing library, with $\alpha = 1$ and $\beta = 0.15$ fixed throughout. Five of the 400 BACH images could not be normalized (too few tissue pixels above β) and fall back to their raw form in the “normalized”

condition; both conditions are otherwise built from the identical 400 images.

E. Cross-Dataset Evaluation Protocol

The fixed BreakHis-trained baseline is evaluated on the same 400 BACH images under two conditions — raw, and Macenko-normalized — with no retraining or fine-tuning between them and identical ground-truth labels and evaluation code. Only the colour representation of the input changes, which is what lets any outcome difference be attributed specifically to normalization.

F. Evaluation Metrics and Statistical Analysis

We report accuracy, F1 for the malignant class at the default 0.5 threshold, AUC, and per-class precision/recall/F1, since both datasets are class-imbalanced and accuracy alone is unreliable [28]. For each condition we compute a 2,000-resample bootstrap 95% confidence interval for accuracy, F1, and AUC, and we compare the two conditions' AUC values with a paired DeLong test [31], appropriate here because both conditions are evaluated on the same 400 images.

IV. RESULTS

A. Within-Dataset Baseline

On the BreakHis validation split, the baseline reaches F1 = 0.917 and AUC = 0.929 at its best checkpoint, confirming the training pipeline behaves like a reasonably competent classifier before any cross-dataset question is introduced.

B. Raw BACH

On raw BACH images, the model scores accuracy = 0.578, F1 = 0.697, AUC = 0.597 (95% CI [0.540, 0.653]), close to chance-level ranking. The confusion matrix (Table I) shows why the F1 figure is misleading on its own: of 200 benign images only 37 are correctly identified (recall 0.18); of 200 malignant images 194 are correctly identified (recall 0.97). The model has, in effect, learned to default to malignant outside its training distribution, which a single F1 number does not reveal.

TABLE I
CONFUSION MATRICES, RAW VS. MACENKO-NORMALIZED BACH (N = 400)

Condition	True class	Pred. benign	Pred. malignant
Raw	Benign (n=200)	37	163
Raw	Malignant (n=200)	6	194
Normalized	Benign (n=200)	40	160
Normalized	Malignant (n=200)	10	190

A. Macenko-Normalized BACH

After normalization against the single fixed reference, the identical model scores accuracy = 0.575, F1 = 0.691, AUC = 0.647 (95% CI [0.590, 0.699]). Benign recall rises slightly, from 0.18 to 0.20 (Table I) — a small improvement, and still poor in absolute terms. Accuracy and F1 are essentially unchanged from the raw condition (Table II), and the 95% confidence intervals for accuracy and F1 overlap almost

completely between the two conditions. The AUC gain of 0.050 is the only metric that moves by more than the width of its own confidence interval, but a paired DeLong test comparing the two AUC curves on the same 400 images gives $z = -1.81$, $p = 0.070$ — not significant at the conventional 0.05 threshold, though close to it.

TABLE II
CROSS-DATASET RESULTS WITH 95% BOOTSTRAP CONFIDENCE INTERVALS (2,000 RESAMPLES)

Metric	Raw BACH (95% CI)	Normalized BACH (95% CI)
Accuracy	0.578 [0.528, 0.625]	0.575 [0.528, 0.625]
F1 (malignant)	0.697 [0.652, 0.738]	0.691 [0.647, 0.733]
AUC	0.597 [0.540, 0.653]	0.647 [0.590, 0.699]
Benign recall	0.18	0.20
Malignant recall	0.97	0.95

TABLE III
CLASS-WISE PRECISION, RECALL, AND F1

Condition	Class	Precision	Recall	F1
Raw	Benign	0.86	0.18	0.30
Raw	Malignant	0.54	0.97	0.70
Normalized	Benign	0.80	0.20	0.32
Normalized	Malignant	0.54	0.95	0.69

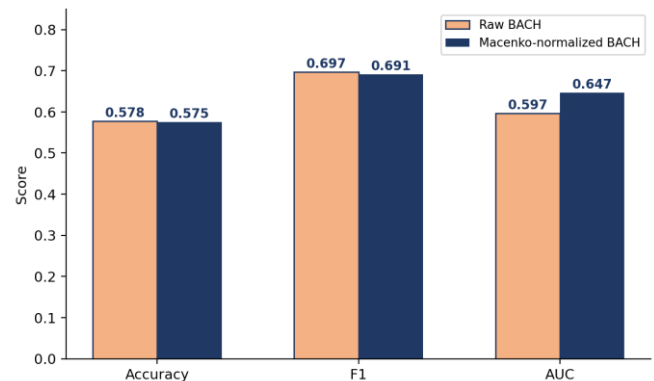


Fig. 1 Accuracy, F1, and AUC on raw versus Macenko-normalized BACH

B. Summary

Across every metric except AUC, raw and normalized BACH are statistically indistinguishable in this single-reference test; the one metric that shifts is not significant at $p < 0.05$. This is a materially weaker result than the point estimates alone (0.597 vs. 0.647) would suggest, and is why Section V treats it as preliminary rather than conclusive.

V. DISCUSSION

A. Interpretation

The result here does not confirm the colour-domain-shift hypothesis in the way we initially expected. AUC moved in the expected direction and by a magnitude (5 points) that is not trivial, but it does not clear a conventional significance threshold on this 400-image test set, and accuracy and F1 barely moved at all. We think three explanations are plausible and not mutually exclusive: the true effect of normalization in this setting may simply be small; a single reference image

may be a poor representative of the method's achievable benefit, given that Macenko normalization is known to be sensitive to reference choice; or 400 images may not give this comparison enough statistical power to detect a real but modest effect. We cannot distinguish between these with the data reported here, which is precisely why we are running the multi-reference sensitivity analysis described in Section V-C rather than treating this single run as final.

B. Comparison with Prior Work

Tellez et al. [9] found normalization and augmentation are not interchangeable and that normalization's value depends on how different the two distributions already are; our near-chance raw-BACH AUC suggests a large distributional gap, yet the recovery we observe is smaller and less certain than that framing might predict. Sani et al. [17] and Stanescu and Stoica-Spahiu [29] both report that the specific adaptation strategy, not just the presence of some correction, materially changes cross-dataset outcomes — consistent with a normalization effect that is real but modest and reference-dependent rather than a complete fix.

C. Limitations and Ongoing Work

The result in Section IV-C rests on one fixed reference image; we are currently repeating the normalization and evaluation against 4–5 additional BreakHis reference images to test whether the AUC gain and its non-significance are stable across reference choice or specific to the one used here, and will report the resulting mean $\pm$ SD in a revision. The 400-image BACH set limits statistical power for any comparison of this kind. Five of the 400 images could not be normalized and fall back to their raw form (Section III-D). Exact per-split BreakHis patient and image counts, and exact BACH class counts, are still being finalized rather than stated provisionally. The benign/malignant mapping applied to BACH's four native classes is a simplification consistent with prior work but coarser than a dedicated four-class model. No fixed random seed is set for training (Section III-C). Finally, this study isolates colour as one variable and does not attempt to isolate tissue-framing or resolution differences between the two datasets, which remain plausible contributors to the residual gap.

VI. CONCLUSION

A BreakHis-trained classifier that performs well within its own dataset ($F1 = 0.917$, $AUC = 0.929$) collapses to near-chance ranking on raw BACH images ($AUC = 0.597$) and defaults heavily toward the majority class, missing 82% of benign cases. Macenko normalization against a single fixed reference image raises AUC to 0.647, but this difference is not statistically significant (paired DeLong test, $p = 0.070$), and accuracy and F1 are essentially unchanged. We report this as a genuine, preliminary finding rather than a confirmed remedy, and are extending the evaluation to multiple reference images to determine whether a stable effect exists before drawing a firmer conclusion.

ACKNOWLEDGMENT

The authors thank the creators of the BreakHis and BACH (ICIA 2018 Grand Challenge) datasets for making them publicly available, and A.P.S. University, Rewa, for their institutional support.

REFERENCES

- [1] 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, 2016.
- [2] G. Aresta et al., "BACH: Grand challenge on breast cancer histology images," *Medical Image Analysis*, vol. 56, pp. 122–139, 2019.
- [3] M. Macenko et al., "A method for normalizing histology slides for quantitative analysis," in *Proc. IEEE ISBI*, 2009, pp. 1107–1110.
- [4] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019.
- [5] A. Vahadane et al., "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE Trans. Med. Imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.
- [6] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1729–1738, 2014.
- [7] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 34–41, 2001.
- [8] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution," *Anal. Quant. Cytol. Histol.*, vol. 23, no. 4, pp. 291–299, 2001.
- [9] D. Tellez et al., "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Medical Image Analysis*, vol. 58, p. 101544, 2019.
- [10] T. A. A. Tosta et al., "A stain color normalization with robust dictionary learning for breast cancer histological images processing," *Biomed. Signal Process. Control*, vol. 85, p. 104978, 2023.
- [11] M. J. Hetz, T.-C. Bucher, and T. J. Brinker, "Multi-domain stain normalization for digital pathology: a cycle-consistent adversarial network for whole slide images," *Medical Image Analysis*, vol. 94, p. 103149, 2024.
- [12] H. Kang et al., "StainNet: a fast and robust stain normalization network," *Frontiers in Medicine*, vol. 8, p. 746307, 2021.
- [13] M. Aubreville et al., "Mitosis domain generalization in histopathology images — the MIDOG challenge," *Medical Image Analysis*, vol. 84, p. 102699, 2023.

- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.
- [15] M. Gavrilovic et al., "Blind color decomposition of histological images," *IEEE Trans. Med. Imaging*, vol. 32, no. 6, pp. 983–994, 2013.
- [16] A. Mahbod, G. Dorffner, I. Ellinger, R. Woitek, and S. Hatamikia, "Improving generalization capability of deep learning-based nuclei instance segmentation by non-deterministic train time and deterministic test time stain normalization," *Comput. Struct. Biotechnol. J.*, vol. 23, 2024.
- [17] M. M. Sani, A. Royat, and M. Soleymani Baghshah, "Classification of breast cancer histopathology images using a modified supervised contrastive learning method," *Med. Biol. Eng. Comput.*, 2024.
- [18] S. Taheri and Z. Golrizkhatami, "Magnification-specific and magnification-independent classification of breast cancer histopathological image using deep learning approaches," *Signal Image Video Process.*, vol. 17, no. 2, pp. 583–591, 2023.
- [19] S. Taheri, Z. Golrizkhatami, A. A. Basabrain, and M. S. Hazzazi, "A comprehensive study on classification of breast cancer histopathological images," *IEEE Access*, 2024.
- [20] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Medical Image Analysis*, vol. 67, p. 101813, 2021.
- [21] Y. Benhammou, B. Achchab, F. Herrera, and S. Tabik, "BreakHis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights," *Neurocomputing*, vol. 375, pp. 9–24, 2020.
- [22] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv:2010.11929*, 2020.
- [23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE CVPR*, 2009, pp. 248–255.
- [24] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.
- [25] P. Micikevicius et al., "Mixed precision training," in *Proc. ICLR*, 2018.
- [26] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, vol. 32, 2019, pp. 8026–8037.
- [27] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [28] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
- [29] L. Stanescu and C. Stoica Spahiu, "Clinically oriented evaluation of transfer learning strategies for cross-site breast cancer histopathology classification," *Applied Sciences*, vol. 15, no. 23, p. 12819, 2025.
- [30] L. Stanescu and C. Stoica Spahiu, "Reducing false negatives in AI-based breast histopathology: A clinically oriented evaluation of deep learning models under domain shift," *Diagnostics*, 2026.
- [31] Gautam, C. S., & Pandey, P. (2022). A review on genetic algorithm models for Hadoop MapReduce in big data. *International Journal of Recent Scientific Research*, 13(3E), 771–775. <https://doi.org/10.24327/ijrsr.2022.1303.0166>
- [32] Gautam, C. S., Soni, L. N., & Pandey, P. (2022). Clustering of big data using genetic algorithm in Hadoop MapReduce. *European Chemical Bulletin*, 12, 963–973.
- [33] Gautam, C. S., & Pandey, P. (2019). A review of big data environment, tools and challenges. *Journal of Emerging Technologies and Innovative Research*, 6, 569–575.
- [34] Chaudhari, S., Gautam, C. S., & Wao, A. A. (2024). Enhancing heart disease prediction accuracy: A comparative study of machine learning models with ensemble method. *JARIIE*, 10, 4827–4833.
- [35] Kar, S. K., Pandey, A., & Gautam, C. S. (2025). A review of machine learning techniques for breast cancer prediction. *International Journal of Current Engineering and Technology*, 15(3).
- [36] Shrivastava, P., Gautam, C. S., & Kar, S. K. (2024). Assessing the performance of Cataract Net and other deep learning systems for automated cataract detection. *ShodhKosh: Journal of Visual and Performing Arts*, 5(5).
- [37] Shrivastava, R., & Gautam, C. S. (2026). An optimized hybrid classification approach for early detection of heart disease. *International Journal of Computer Science Trends and Technology (IJCTST)*, 14(1), 25–31.
- [38] Shrivastava, R., Gautam, C. S., & Kar, S. K. (2024). Promoting a website with the help of SEO using PPC (pay per click). *ShodhKosh: Journal of Visual and Performing Arts*, 5(5), 133–140. <https://doi.org/10.29121/shodhkosh.v5.i5.2024.361>
- [39] Shrivastava, R., & Gautam, C. S. (2026). Enhanced heart disease prediction using ensemble of machine learning models. *International Journal of Computer Applications*, 187(104), 40–46.