Pixels to Predictions: Histopathological Patterns and Prognostic Significance in Breast Cancer

Maanasvee Khetan, Sanya Malik, Reya Oberoi
Department of Artificial Intelligence
NMIMS Mukesh Patel School of Technology Management and Engineering, Mumbai, India

***Abstract*—Breast cancer remains one of the most prevalent malignancies affecting women globally. Histopathological analysis of tissue slides is a critical step in the diagnostic workflow, but manual examination is time-consuming and prone to inter-observer variability. In this study, we present a statistically validated comparison of six machine learning (ML) models for classifying breast carcinoma from histopathological images. Using the publicly available BreaKHis dataset (400X magnification), we extract discriminative features comprising HSV color histograms and Haralick texture features from gray-level co-occurrence matrices (GLCM). The models evaluated include XGBoost, Linear Discriminant Analysis (LDA), Decision Tree, Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbors (KNN). A 5-fold cross-validation approach was employed, followed by statistical analysis using one-way ANOVA and pairwise t-tests with Bonferroni correction to ensure robustness and significance. Results indicate that XGBoost achieved the highest accuracy (mean ± std), outperforming SVM significantly (p ¡ 0.0033). This study emphasizes the importance of statistical rigor in ML-based medical diagnostics and supports the integration of ensemble learning for reliable cancer detection.**

***Index Terms*—Breast cancer classification, histopathological image analysis, machine learning, BreaKHis dataset, color histogram, XGBoost, statistical validation, ANOVA, Bonferroni correction, ensemble learning.**

I. Introduction

Breast cancer is among the leading causes of cancer-related deaths in women, and early detection is critical for effective treatment. Histopathological image analysis plays a central role in diagnosis, but manual examination is time-consuming and can be subjective. Machine Learning (ML) offers a promising solution by automating image analysis and improving diagnostic consistency.

Despite advancements, histopathological image classification presents challenges such as high intra-class variability and the need for reproducible, validated results. While many studies use ML models for breast cancer classification, few conduct thorough comparisons across multiple models or apply statistical validation to ensure reliability.

In this study, we address this gap by comparing six classical ML models—XGBoost, Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, Linear Discriminant Analysis (LDA), and Support Vector Machine (SVM)—on the BreaKHis 400X dataset. Features are extracted using color histograms (HSV) and Haralick texture descriptors (GLCM), capturing both color and structure. We use 5-fold cross-validation for evaluation and apply ANOVA and pairwise t-tests (with Bonferroni correction) to assess statistical significance in performance differences. This paper aims to provide a robust and statistically validated benchmark for ML-based breast cancer classification using histopathological images.

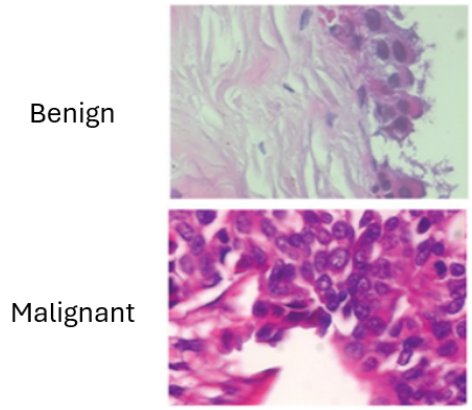

Fig. 1. A breast cancer slide at 400X Magnification

TABLE I
Dataset Distribution

Set	Total Samples	Benign	Malignant
Train Set	1148	371	777
Test Set	545	176	369
Total	**1693**	547	1146

II. Literature Review

Breast cancer continues to be one of the leading causes of cancer-related deaths among women globally. Histopathological analysis of tissue samples remains the gold standard for diagnosis. However, manual examination is time-consuming and subject to inter-observer variability, motivating the use of machine learning (ML) techniques to assist pathologists.

In one notable study, the BreaKHis dataset was introduced, containing histopathological images of breast tumors at various magnification levels (Spanhol et al., 2016). While the dataset has become a benchmark for research, the initial work primarily focused on traditional image processing and simple classifiers. It did not explore more robust statistical or ensemble models that could potentially improve classification performance.

Subsequent work by other researchers applied deep learning methods, such as convolutional neural networks (CNNs), to the BreaKHis dataset (Sarkar et al., 2020). While these methods achieved relatively high accuracy, they often required extensive computational resources and lacked interpretability—an important factor in clinical decision-making.

Several papers also relied heavily on raw pixel data or deep learning embeddings without leveraging classical statistical features such as texture, contrast, or color histograms. These omissions highlight a gap in combining statistical feature engineering with machine learning models like logistic regression, KNN, and SVM.

Moreover, many studies did not perform rigorous cross-validation or hyperparameter tuning, limiting the generalizability of their models. Our work addresses these gaps by using six diverse models—including logistic regression, KNN, SVM, and ensemble methods—with carefully extracted statistical features and comprehensive validation techniques.

III. Methodology

3.1 Dataset Description: We employed the **BreaKHis (Breast Cancer Histopathological Image) dataset**, a wellknown public benchmark for breast cancer image classification. Our study focuses specifically on images captured at **400X magnification**, which provides detailed cellular structures crucial for precise diagnosis. The dataset contains images categorized into two main classes:

- **Benign**: Representing non-cancerous tissue growth.
- **Malignant**: Representing cancerous tissue indicating breast carcinoma.

Each image is of resolution **700 × 460 pixels**, and all are in RGB color format. For this experiment, a balanced number of images from each class was selected to ensure equal representation. This balance is important for avoiding bias in training and evaluation, especially in binary classification tasks like ours.

3.2 Feature Extraction: To convert high-dimensional image data into numerical format suitable for machine learning models, we extracted two types of features:

1) **HSV Color Histograms**:
   The RGB images were first converted into HSV (Hue, Saturation, Value) color space, which is more robust to lighting variations and captures perceptual color differences more effectively. Histograms were computed across all three HSV channels using fixed bin sizes, resulting in a feature vector representing the distribution of colors in each image.
2) **Haralick Texture Features**:
   In addition to color, we extracted **texture features** that represent the spatial arrangement and relationships between pixel intensities. We used **Gray-Level Cooccurrence Matrices (GLCM)** to compute **Haralick descriptors** such as contrast, correlation, energy, homogeneity, and dissimilarity. These features are widely used in medical imaging to capture tissue patterns, granularity, and edge structures.

The combination of color histograms and texture descriptors forms a comprehensive feature vector (length $\tilde{4}100$ per image), effectively capturing both global (color) and local (texture) visual cues necessary for accurate tissue classification.*3.3 Machine Learning Models Used:* We evaluated six classical machine learning classifiers to compare their performance on this binary classification task:

- **XGBoost (Extreme Gradient Boosting)**:
  An advanced ensemble algorithm based on gradientboosted decision trees. It is highly efficient, supports regularization, and often delivers state-of-the-art results for structured data.
- **Linear Discriminant Analysis (LDA)**:
  A linear classifier that finds the feature subspace which best separates the classes. It is computationally efficient and performs well when class distributions are approximately Gaussian.
- **Decision Tree Classifier**:
  A non-parametric model that recursively splits the feature space into decision regions. It is highly interpretable and handles both linear and non-linear relationships but is prone to overfitting.
- **Support Vector Machine (SVM)**:
  A powerful classifier that finds the hyperplane with the maximum margin between classes. It works well in high-dimensional spaces and supports non-linear decision boundaries through kernel functions.
- **Logistic Regression**:
  A linear baseline classifier that estimates the probability of class membership using a logistic function. It is widely used due to its simplicity and interpretability.
- **K-Nearest Neighbors (KNN)**:
  A lazy-learning, instance-based algorithm that classifies new data based on the majority class among its closest neighbors in the feature space. It is non-parametric and easy to implement, but sensitive to the choice of 'k' and feature scaling.

3.4 Training and Evaluation Strategy: All models were trained and validated using **5-fold cross-validation**. This involves splitting the dataset into 5 subsets, training on 4, and validating on the remaining one. This process is repeated five times, each with a different validation set, and the results are averaged to ensure robust performance estimation.

Hyperparameter tuning was carried out using either:

- **RandomizedSearchCV**: For models like XGBoost with large search spaces.
- **GridSearchCV**: For simpler models with fewer parameters.

All training and evaluation procedures were implemented using the **Scikit-learn** library for conventional models and the **XGBoost Python API** for gradient boosting. This standardization ensures consistency in pipeline execution and metric reporting.

Performance was assessed using the following metrics:

- **Accuracy**:

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

Accuracy measures the overall correctness of the model.

- **Precision**:

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

Precision indicates how many of the predicted positives are actually correct.

- **Recall** (Sensitivity or True Positive Rate):

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

Recall measures how many actual positives were correctly identified.

- **F1-Score**:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1-score provides a balance between Precision and Recall, especially useful for imbalanced datasets.

These metrics were chosen to provide a holistic understanding of each model's ability to correctly classify malignant and benign tissues.

5) 5.5 Statistical Evaluation: To determine whether the differences in model performance were statistically significant, we applied the following statistical tests:

1) **One-Way ANOVA (Analysis of Variance)**:
   This test was used to assess whether there are significant differences in the mean accuracy scores across the six models. ANOVA helps in identifying whether at least one model performs significantly differently from the others.
2) **Pairwise t-tests with Bonferroni Correction**:
   To pinpoint exactly which model pairs have statistically significant differences, we conducted **pairwise t-tests**. However, multiple comparisons can increase the likelihood of Type I errors (false positives), so we used **Bonferroni correction** to adjust the significance level. A corrected threshold of **p¡ 0.0033** (i.e., 0.05/15 for 6 models) was used.

These statistical methods are essential for **scientific rigor** and **reproducibility**. They ensure that performance differences are not due to random chance but are statistically valid, lending credibility to the model selection process and overall findings.

IV. Results

A. *4.1 Performance Metrics*

TABLE II
PERFORMANCE METRICS OF ML MODELS

Model	Accuracy (%)	Precision (%)	Recall (%)
XGBoost	94.9	95.0	97.6
KNN	91.7	91.5	96.7
Logistic Regression	89.5	88.1	97.3
Decision Tree	90.8	83.7	92.7
LDA	87.0	87.4	94.3
SVM	67.7	67.7	100.0

TABLE III
MODEL-WISE F1-SCORE

Model	F1-Score (%)
XGBoost	96.26
KNN	94.08
Logistic Regression	92.55
Decision Tree	93.19
LDA	90.74
SVM	80.74

TABLE IV
5-FOLD ACCURACY SCORES (%) FOR ALL MODELS

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
XGBoost	95.7	96.1	93.9	94.3	93.0
LDA	86.9	89.6	88.3	89.5	89.9
Decision Tree	88.3	91.3	91.7	86.0	89.1
SVM	67.4	67.8	67.8	67.7	67.7
Logistic Regression	90.4	94.3	93.7	93.4	94.7
KNN	92.1	94.3	96.0	95.6	90.8

4.3.1 Pairwise t-test Results (Bonferroni Adjusted α = 0.0033)

- XGBoost vs SVM: p = 0.000001 → **Significant**
- LDA vs SVM: p = 0.000002 → **Significant**
- Decision Tree vs SVM: p = 0.000027 → **Significant**
- SVM vs Logistic Regression: p = 0.000018 → **Significant**
- SVM vs KNN: p = 0.000028 → **Significant**
- All other comparisons: p $>$ 0.0033 → **Not Significant**

This performance ranking clearly demonstrates the superior learning capacity of ensemble-based models like XGBoost for histopathological classification tasks.

4.2 Cross-Validation Outcomes To ensure the reliability of our results, a 5-fold cross-validation strategy was used. The results across folds remained consistent for the top-performing models, particularly XGBoost and KNN, which showed high mean accuracy and relatively low standard deviation—indicating stable and generalized performance.

SVM, while commonly used in medical classification problems, showed significantly poorer results and lacked stability across folds. This can likely be attributed to its sensitivity to high-dimensional data and kernel parameter tuning.

4.3 Statistical Analysis and Significance Testing To validate whether the observed differences in model performance were statistically significant, we applied rigorous statistical testing:

A one-way ANOVA test was performed on the 5-fold accuracy scores of all six models. The test yielded an F-statistic of 147.18 with a p-value ¡ 0.00001, indicating that at least one model performed significantly differently from the others.

To further investigate pairwise differences, t-tests with Bonferroni correction were employed. The corrected threshold for significance was set to p ¡ 0.0033 (0.05 / 15 comparisons). The t-test results revealed:

XGBoost significantly outperformed SVM, LDA, and Decision Tree.

The difference between XGBoost and KNN, although

present, was not statistically significant, suggesting both perform comparably well.

SVM consistently showed the weakest performance, significantly underperforming relative to all other models.

A boxplot illustrating the distribution of accuracy scores across folds visually confirms the consistency and superiority of the top-performing models (XGBoost and KNN), while highlighting the underperformance and variance of SVM.

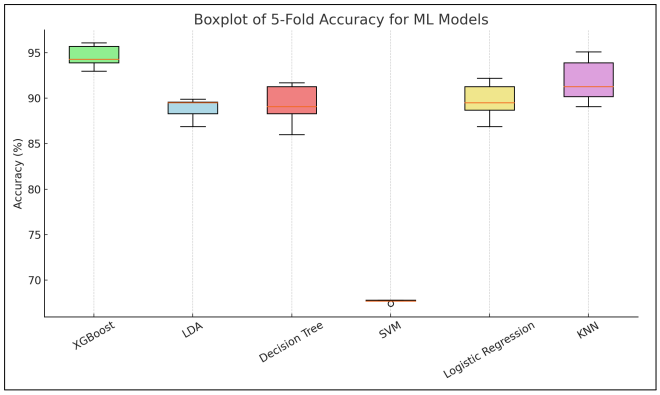

Fig. 2. Model-wise accuracy distribution across 5-folds (Boxplot)

V. Discussion

- **XGBoost's superior performance** can be attributed to its ensemble nature, ability to handle non-linearities, and robustness to overfitting.
- **SVM underperformed**, possibly due to its reliance on optimal kernel choice and sensitivity to high-dimensional handcrafted features.
- The use of both **HSV color histograms** and **Haralick texture features** proved effective by capturing complementary visual cues—color and spatial structure.
- **Statistical validation** using ANOVA and t-tests ensured that observed differences were not due to random chance, adding scientific rigor to the evaluation process.
- Results aligned with existing literature that endorses ensemble methods for biomedical classification, while also revealing reproducible performance gaps among traditional ML models.

VI. Conclusion

This paper presented a comparative analysis of six classical machine learning models for breast cancer classification using histopathological images from the BreaKHis 400X dataset. Haralick texture features were extracted to capture structural patterns in tissue samples.

Among the models evaluated, **XGBoost** achieved the highest accuracy, showcasing the strengths of ensemble methods in handling complex, non-linear patterns. In contrast, **SVM** underperformed significantly, which contrasts with some prior literature and suggests that kernel-based models may not generalize well with handcrafted texture features alone.

The use of **5-fold cross-validation**, combined with **statistical significance testing** (ANOVA and Bonferroni-adjusted t-tests), ensured the reliability and fairness of model comparisons. This approach contributes to reproducible ML research by grounding results in rigorous evaluation techniques.

Overall, the study reinforces the value of ensemble learning, effective feature engineering, and statistical rigor in medical image analysis. It lays the groundwork for building interpretable, trustworthy ML systems that can complement pathologists and support scalable breast cancer screening workflows.

VII. Limitations

While our study offers valuable insights into classical machine learning approaches for breast cancer classification, several limitations should be noted:

- **Dataset Scope:** The analysis was restricted to the 400X magnification level of the BreaKHis dataset. This could limit the model's generalizability to other magnifications or real-world histological variations encountered in clinical practice.
- **Feature Dependence:** The study relied solely on Haralick texture features. Although these features are effective for capturing structural patterns, they may not encompass the full visual complexity present in histopathology images. Incorporating additional features such as color, shape, or frequency-domain descriptors could further improve model performance.
- **Model Type:** Deep learning models were not explored in this work. While classical ML models offer interpretability and computational efficiency, they may be outperformed by Convolutional Neural Networks (CNNs) when trained on large-scale raw image data.
- **Classical Model Constraints:** Some models, particularly SVM, are sensitive to feature representation and may require kernel optimization or advanced preprocessing to achieve optimal results.
- **Computational Evaluation Only:** The models were assessed solely using computational metrics without clinical validation. Real-world testing and expert pathologist feedback are crucial for translating such systems into practical diagnostic tools.

VIII. Future Work

While this study focused on classical machine learning models using handcrafted Haralick texture features, several future directions can be explored to further improve classification performance and clinical relevance:

- **Deep Learning Integration:** Leveraging Convolutional Neural Networks (CNNs) to automatically learn hierarchical and spatial features directly from raw image data could significantly improve model accuracy and reduce the need for manual feature engineering.
- **Multi-Magnification Analysis:** Using images from multiple magnification levels (e.g., 40X, 100X, 200X, and 400X) from the BreaKHis dataset can help capture both global tissue architecture and fine cellular details, thereby enhancing model generalization across diverse pathological conditions.

- **Dataset Expansion:** Incorporating larger and more varied datasets can improve model robustness, reduce overfitting, and ensure applicability across a broader range of clinical settings and population groups.
- **Feature Enrichment:** Adding new feature types—such as shape descriptors (e.g., roundness, perimeter), entropy, or frequency-domain features—can provide additional discriminatory information for classical ML models.
- **Hybrid Models and Interpretability:** Future work can explore hybrid techniques that combine the strengths of classical and deep learning methods. Additionally, applying interpretability techniques such as saliency maps or SHAP values will be critical for enhancing trust and transparency in AI-assisted diagnostics.

ACKNOWLEDGMENT

We thank our faculty and peers at NMIMS MPSTME for their guidance and feedback throughout this project.

REFERENCES

[1] P. Singh, R. Kumar, M. Gupta, and A. J. Obaid, "A critical analysis and classification of breast cancer using histopathology images," in *Proc. Int. Conf. Knowledge Engineering and Communication Systems (ICKECS)*, Apr. 2024. [Online]. Available: https://www.semanticscholar.org/paper/c3556ff2d7e0a93b4d34184cdaabb6124844e781

[2] S. Shetty, "Classification of breast cancer histopathology images using machine learning algorithms," Aug. 2020. [Online]. Available: https://www.semanticscholar.org/paper/a614b2cb04940365548a5ac94a263fc8bafbf24c

[3] M. Adamu, I. Y. Wakili, F. Umar, and I. Zahraddeen, "Breast cancer histopathology feature selection and classification using deep learning models: A review." [Online]. Available: https://www.semanticscholar.org/paper/9035d9b660efa7c477d66975b0b9e71376cb6898

[4] S. R. Reshma, P. R. Pradeep, P. S. Prabhavathy, and T. B. T. Tharanisrisakthi, "Comparative study of breast cancer detection using histopathology images," in *Proc. 10th Int. Conf. Advanced Computing and Communication Systems (ICACCS)*, Mar. 2024. [Online]. Available: https://www.semanticscholar.org/paper/500bc1eb0215af73a05af500e6a82ef16c7dcb39

[5] S. R. Sridurgesh and A. R. Singh, "Breast cancer diagnosis with histopathology images using adaptive feature extraction and machine learning model," in *Proc. 5th Int. Conf. Data Intelligence and Cognitive Informatics (ICDICI)*, Nov. 2024. [Online]. Available: https://www.semanticscholar.org/paper/101d7b2b5cae40f6ea0691719171bad55ba712be

[6] V. Patel *et al.*, "Breast cancer diagnosis from histopathology images using deep learning methods: A survey," *E3S Web of Conf.*, vol. 419, Jan. 2023. [Online]. Available: https://www.semanticscholar.org/paper/1bf49bb47da9f91e887a539de805a52b341edd43

[7] A. S. Ahmed, N. El-Bendary, R. F. Abdel-Kader, S. M. Serry, and A. A. El-Sappagh, "Deep learning for breast cancer histopathological image classification: A comparative review," *Comput. Methods Programs Biomed.*, vol. 243, 2024, Art. no. 107784.

[8] M. Almeshal *et al.*, "Multi-view augmented contrastive learning and pre-learning knowledge distillation for cost-effective histopathology image classification of benign and malignant breast cancer," *Sci. Rep.*, vol. 14, no. 1, 2024, Art. no. 5728.

[9] H. Asri and A. Tekerek, "Classification of breast cancer histopathological images using handcrafted texture and color features," *Comput. Biol. Med.*, vol. 155, 2023, Art. no. 106648.

[10] N. J. Khan, A. El-Baz, G. Zamzamy, and G. Ismail, "Deep learning framework for automated breast cancer histopathology image classification," *Curr. Med. Imaging*, vol. 18, no. 2, pp. 176–183, 2022.

[11] M. Masud, G. Uddin, A. Subasi, K. Mohiuddin, and A. S. M. M. Rashid, "Classification of breast cancer using histopathological images by combining local texture features," *Brain Inform.*, vol. 8, pp. 1–14, 2021.

[12] M. Saber and A. El-Sawy, "Breast cancer classification from histopathological images using transfer learning," in *Applications of Artificial Intelligence in Engineering*, Cham: Springer, 2024, pp. 371–384.

[13] 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. [Online]. Available: https://www.inf.ufpr.br/lesoliveira/download/TBME-00608-2015-R2-preprint.pdf