Intelligent Methods in Engineering Sciences 5(1): 10-19, 2026

IMIENS

INTELLIGENT METHODS
IN ENGINEERING SCIENCES

International
Open Access

April, 2026

https://www.imiens.org

e-ISSN 2979-9236

Research Article

https://doi.org/10.58190/imiens.2026.167

Cancer Detection in Breast Histopathological Images Using Extremely Randomized Trees

Mahendra Kanojia [a,*]

[a] *Department of Computer Science; Sheth L.U.J. and Sir M.V. College, India*

ARTICLE INFO

Article history:
Received 13 February 2026
Accepted 18 April 2026

Keywords:
Breast Cancer Detection,
Extremely Randomized Trees,
Histopathological Images,
Image Processing,
Machine Learning,
Recursive Feature Elimination

ABSTRACT

This research aimed to develop an effective machine learning-based system for the automated detection of breast cancer using histopathological images, overcoming the limitations of manual examination. The study utilized a diverse dataset of 13,347 histopathological images from three secondary sources and one primary source. The inclusion of multiple image sources was intended to enhance the model's versatility. Initially, images underwent pre-processing to reduce noise using a median filter and were converted to grayscale. Otsu's thresholding was then applied to enhance nucleus edges and reduce background noise. A recursive feature elimination algorithm was employed to reduce the initial 98 features to these 48 key ones, focusing on the area and shape of the nucleus, color-based features, and image texture. For classification, the Extremely Randomized Trees Classifier was used. The model was trained to classify images as benign or malignant. The results demonstrated high performance, with the model achieving an accuracy of 98.95%. Further evaluation revealed a sensitivity of 99.48%, indicating a low false negative rate. Specificity was 94.67%, correctly identifying benign cases. The model also achieved precision of 98.97% and recall of 99.48%, with a Kappa statistic of 97.62%, suggesting substantial agreement beyond chance. The ROC performance was 98.67%, indicating robust performance. This study highlights the potential of machine learning, specifically the Extremely Randomized Trees Classifier, for automated and accurate breast cancer detection from histopathological images. The high-performance metrics suggest the model can enhance diagnostic accuracy and assist pathologists in clinical decision-making.

1. INTRODUCTION

Breast cancer is one of the leading causes of death worldwide. Breast cancer was the most common cancer in women in 157 out of 185 countries, accounting for 670,000 deaths globally, and occurs in every country worldwide, according to a fact sheet released by the World Health Organization [1]. This highlights the critical importance of early detection and accurate diagnosis for improved patient survival. Histopathological examination, which involves assessing cellular morphology and tissue structure under a microscope, is a fundamental method for identifying various malignancies [2]. This process typically uses tissue samples stained with Hematoxylin and Eosin (H&E). However, it is time-consuming and highly dependent on the pathologist's skill, which can lead to discrepancies and errors [3].

In recent years, computational methods, especially those involving image processing and machine learning algorithms, have gained popularity for automating this diagnostic process [4]. These methods aim to enhance the precision, speed, and repeatability of breast cancer diagnoses. The application of histopathological images for breast cancer detection presents challenges due to inherent variability in factors such as resolution, staining techniques, and tissue composition [4], [5], [6]. These variations can affect the performance of detection algorithms and limit the generalizability of models trained on specific datasets.

The goal of this research is to develop an efficient machine learning-based system for diagnosing breast cancer using histopathology images. The study employs a diverse dataset, including both secondary and primary sources. Integrating images from multiple sources aims to enhance the robustness and adaptability of the detection model. To improve the accuracy and reliability of the model, the study applies image processing techniques such as Otsu's thresholding, grayscale conversion, and noise reduction using a median filter[7], alongside comprehensive feature extraction. The proposed methodology employs the Extremely Randomized Trees Classifier (Extra Trees Classifier) [8], [9], [10], an ensemble learning algorithm, to classify histopathology

* **Corresponding Author:** kgkmahendra@gmail.com

images as benign or malignant. The model is trained on a large and diverse dataset of 13,347 histopathology images. Its performance is evaluated using standard classification metrics.

The ultimate goal is to support pathologists in the automated detection of breast cancer from histopathological images, potentially improving patient outcomes. This study demonstrates a successful implementation of a model that replicates the physical examination of tissue slides, taking into account all relevant computational aspects of histology.

2. Literature Review

Breast cancer remains a significant global health concern, and researchers have increasingly focused on leveraging computational methods, particularly machine learning, for its early and accurate detection. Early diagnosis is crucial for improving survival rates and treatment outcomes. This literature review synthesizes recent research efforts in breast cancer detection using various machine learning techniques, drawing exclusively from the provided sources.

[11] explored the use of Radial Basis Function Neural Networks (RBFN) combined with image processing techniques for detecting malignancy in histopathological images. This study aimed to provide a complete and automated detection system, unlike other works that focused solely on either image processing or classification based on online data. [12] conducted a comparative analysis of breast cancer detection and diagnosis using data visualization and various machine learning applications, including Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, and Rotation Forest. Using the Wisconsin Breast Cancer dataset, the study found that the Logistic Regression model with all features achieved the highest classification accuracy of 98.1%, highlighting the potential of these techniques in improving diagnostic accuracy. [13] proposed a new ensemble learning method combining the Multi-Verse Optimizer (MVO) with the Gradient Boosting Decision Tree (GBDT) to classify breast cancer as malignant or benign. They utilized the Wisconsin Breast Cancer datasets, and their GBDT–MVO model demonstrated higher precision and lower variance compared to other models. The issue of imbalanced datasets in breast cancer diagnosis was addressed by [14]; proposed a cost-sensitive Extreme Gradient Boosting (XGBoost) model. Applied to four imbalanced breast cancer datasets and optimized using hyperparameter tuning, the results indicated that the proposed model improved classification accuracy. A review of breast cancer classification models was presented by [15] based on Convolutional Neural Networks (CNNs) and hybrid-CNN architectures, noting their deep learning capabilities and promising results for this task .

In 2021, [3] reviewed image processing techniques used for breast cancer detection, emphasizing the complexity of histopathology images and the need for automated systems to aid pathologists. Texture features from histopathological images with the k-Nearest Neighbors (KNN) algorithm, employing dimensionality reduction techniques like Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) for breast cancer detection was reported by [16]. A Computer-based expert system was designed with a custom-designed Keras and U-Net Hybrid CNN (KUH-CNN) model to localize nuclei in histopathological images from the BreaKHis and Kaggle datasets, aiming to assist histopathologists and enable further machine learning analysis(M. G. Kanojia et al., 2021). BCD-WERT introduced by [18], a novel approach using the Extremely Randomized Tree and the Whale Optimization Algorithms (WOA) for efficient feature selection and classification. Evaluated on a comprehensive dataset, BCD-WERT achieved the highest accuracy rate of 99.30%, outperforming other machine learning algorithms like SVM and Random Forest. Another study in 2021, explored the Extremely Randomized Clustering Forests (ERCF) technique for breast cancer prediction, comparing its accuracy with k-NN, and concluded that ERCF performed significantly better [19].

In early 2022,[20] examined the potential of extremely randomized tree classifiers for breast cancer classification into malignant or benign tumors. The proposed model, with hyperparameter optimization and feature importance identification, achieved an accuracy 97.3% on the cross-validated model, calming its superiority over other state-of-the-art models. Another work [21] aimed to build a binary breast cancer classifier using blood test and anthropometric data from 116 subjects, comparing Decision Tree, Random Forest, KNN, Artificial Neural Networks, Support Vector Machines, and Logistic Regression. The random forest classifier achieved the best performance with 83.3% accuracy, 100% sensitivity, and 64% specificity. A work [22] on Mammographic classification of breast cancer microcalcifications using XGBoost on 51 extracted calcification features from mammograms, reported accuracy of 90.24%. [10] proposed a feature ensemble learning method based on Neural networks and Extra Trees the algorithm was tested on Breast Cancer Wisconsin dataset, reporting the accuracy of 99.74%. The imbalanced dataset issue was addressed by [23] with an engineered up-sampling method (ENUS), further reported that XGBoost Tree with ENUS achieved the best performance with an accuracy of 97.47%, and identified Cell_Shape and Nuclei as important predictive attributes. A fuzzy neural network expert system with improved Gini index random forest feature selection for early diagnosis in Saudi Arabia was

designed by [24]. The model attained 99.33% accuracy using the top five features from the Wisconsin dataset.

From the year 2023 we are reporting four prominent work wit in Breast cancer detection. An adaptive voting ensemble classifier was designed by [25]. They combined Extra Trees Classifier, LightGBM, Ridge Classifier, and Linear Discriminant Analysis, achieving an accuracy of 97.6% aging using the Wisconsin Breast Cancer dataset. Whereas [26] proposed a diffuse optical breast scanning (DOB-Scan) an ensemble of nine regression models (Polynomial Regression, SVM, Random Forest, KNN, Decision Tree, MLP, XGBoost, CatBoost, and Extra Trees) and predicted optical properties of breast-mimicking phantoms, achieving 93% accuracy with Extra Trees and 100% classification accuracy using Bagging with KNN. Another Random Forest model proposed [27] achieved an accuracy of 98%. The advanced study by [6] focused on choledochal cancer, using a U-Net architecture on hyperspectral tissue images, achieving an accuracy of 68.09% on the training data and 61.95% on the testing data. The work proposed the possibilities of working in hyperspectral tissue images.

The years 2024 and 2025 witnessed further advancements in the field of breast cancer detection, In 2024 a study aimed to develop a highly accurate machine learning model for the early prediction of breast cancer to reduce mortality rates. The methodology involved a comparative performance analysis of various machine learning classifiers. The key result was that the Extra Trees Classifier demonstrated superior performance, achieving the highest accuracy of 97.23% [28]. Authors [29] developed a hybrid deep learning model combining multipath transfer learning for feature extraction and an ensemble of classifiers for final classification, achieving improved accuracy in breast cancer detection using histopathological images. The study by [30] highlighted the effectiveness of tuned Random Forest with Gradient Boosting ensemble algorithm in enhancing both diagnosis and prognosis of breast cancer, with reported accuracy of 97.9%. Extreme Learning Machine (ELM) algorithm for breast cancer diagnosis was proposed by [31] and addressed the lack of statistical evaluation in prior reviewed. Their methodology demonstrated that the ELM algorithm achieved a high average accuracy of 94.52% on the WDBC dataset. Latest review by [5] to assessed different machine learning algorithms for breast tumor classification on models such as XGBoost, SVM, and Random Forest. Their analysis, spanning multiple datasets, showed that these models can achieve high diagnostic accuracy, with some reaching 99.2%. Notably,[32] developed a feature selection-based Extreme Gradient Boosting (XGB) classifier optimized by metaheuristic algorithms, Whale Optimization, Bald Eagle Search, and Sea Lion Optimization to enhance breast cancer prediction. Their approach reduced the Breast Cancer Coimbra dataset from nine to four features, achieving an F-score of 97.43% with the Sea Lion Optimization XGB model. SHAP analysis identified Glucose, Age, Resistin, and Adiponectin as key predictors.

Overall, an extensive literature exists on the use of ensemble learning techniques—such as Random Forest, XGBoost, and LightGBM—for breast cancer detection, the majority of these studies focus on clinical records (e.g., Wisconsin Diagnostic Breast Cancer dataset), mammograms, or ultrasound imagery. There remains a significant research gap regarding the specific application of the Extra Trees algorithm for analyzing histopathological images, despite its potential for high-dimensional feature handling. While deep learning architectures like Convolutional Neural Networks (CNNs) offer high performance, they often require significant computational resources and high-end GPUs, which may not be available in all clinical settings. This study bridges the research gap by utilizing a combination of handcrafted features and the Extremely Randomized Trees algorithm, providing a computationally efficient alternative that maintains high diagnostic accuracy without the need for specialized hardware.

3. RESEARCH METHODOLOGY

This section outlines the methods used to develop our machine learning model for detecting breast cancer in histopathological images. We begin with the Dataset Description, detailing the quantity and quality of the image data used. This is followed by an explanation of our Image Processing and Feature Extraction techniques, which prepared the images and extracted visual information. Finally, we describe the Extremely Randomized Trees Classifier, outlining its role as the core of our predictive model.

3.1. Dataset Description

This research is based on a comprehensive dataset of histopathological images compiled from three secondary and one primary source. The study material consists of Hematoxylin and Eosin (H&E) stained breast histology slides. It is important to note that image characteristics, such as file type, image quality, and resolution, differ across the three secondary sources. The twofold objective behind using these diverse resources is to enhance the detection algorithm's versatility by training on varied image types and to improve the prediction model, leveraging the principle that diverse input data strengthens machine learning outcomes. The three secondary data sources are: Kaggle Dataset : 2018 Data Science Bowl[33]. This dataset includes images of segmented nuclei and masks. BreaKHis Dataset [34] comprises breast histopathological images. The dataset contains a total of 3833 images, with 1211 benign and 2622 malignant samples. Center for Bio-Image Informatics [35] is another

breast histopathological images dataset including total of 58 images, with 31 benign and 27 malignant samples.

The primary histopathological image data was collected from the Department of Histopathology, BSES Municipal General Hospital, Mumbai, India. This image data, being unpublished and unprocessed, was specifically captured and compiled for this research. To enhance primary data diversity, augmentation techniques including rotation, shearing, cropping, and flipping were implemented. To ensure a rigorous and unbiased evaluation, the data augmentation techniques were applied exclusively to the training subset after the initial 70:30 partition. This methodology ensured that the testing set consisted entirely of original, unseen histopathological images, thereby preventing the model from evaluating its performance on synthetic or modified versions of the training data. As detailed in Table 1, the primary dataset contributed 9,456 images (6,352 benign, 3,104 malignant). This resulted in a grand total of 13,347 images for the research, consisting of 7,594 benign and 5,753 malignant samples. While the dataset exhibits a distribution of 7,594 benign and 5,753 malignant samples, no external balancing techniques such as SMOTE were required. The model relied on the inherent robustness of the Extremely Randomized Trees algorithm, which mitigates class imbalance through its ensemble based averaging and randomized feature selection, ensuring stable performance across both categories without the risk of overfitting synthetic data

Table 1. Distribution of dataset

Dataset	Total	Benign	Malignant
BreaKHis	3833	1211	2622
Center for Bio-Image Informatics	58	31	27
Primary Dataset	9456	6352	3104
Grand Total	13347	7594	5753

The feature extraction process initially yielded 98 features for each image. To prevent data leakage and ensure the integrity of the evaluation, the entire dataset was shuffled once at the beginning of the process. Subsequently, the 70:30 split was strictly enforced such that no individual image or its augmented variants used during the training phase appeared in the testing set. To refine the feature space, a recursive feature elimination (RFE) methodology [36], [37] was employed, which led to the exclusion of 50 features with low informational value. The final dataset comprised 48 principal features organized as feature vectors in a CSV file.

3.2. Tissue Image Processing and Feature Extraction

The comprehensive methodology for processing tissue imagery involved a sequence of essential stages to condition histopathological images for machine learning analysis. The pipeline began with image pre-processing to mitigate spurious noise using a median filter, followed by a conversion to grayscale to decrease computational overhead. As a crucial post-processing measure, Otsu's thresholding was employed to accentuate cell nuclei boundaries and minimize background interference, resulting in a dataset where nuclei appeared as distinct, high-intensity regions.

The computational algorithm subsequently iterated through these pre-processed images to isolate nuclei and extract specific features corresponding to histological criteria. This included twenty-five features describing nucleus geometry, ten color-derived features, and thirteen textural features. These values were organized into feature vectors and recorded in a CSV file, constructing a final dataset matrix of 13,347 samples and 48 feature dimensions, as outlined in Algorithm 1.

Algorithm 1: Feature Extraction from Histopathological Images

Objective: Extract morphometric, colour, and textural features from processed histopathological images to construct a dataset.
Input: D_{in}: A collection of N pre-processed histopathological images.
Output: D_{out}: A CSV file containing feature vectors for all images.
Procedure:
Begin
Initialize FeatureDataset as a new CSV file.
Define CSV_Header containing 48 feature names (25 Morphometric, 10 Colour, 13 Texture).
Write CSV_Header to the first row of FeatureDataset.
For Each CurrentImage in D_{in} Do
Initialize FeatureVector_V as an empty one-dimensional array.
SegmentedRegions ← Apply Otsu's thresholding on CurrentImage to identify nuclei.
AreaShape_Features ←Extract 25 morphometric features from SegmentedRegions.
Color_Features ←Extract 10 colour-based features from CurrentImage using SegmentedRegions as masks.
Texture_Features ← Extract 13 texture-based features from CurrentImage.
Append AreaShape_Features, Color_Features, and Texture_Features to FeatureVector_V.
Write FeatureVector_V as a new row to FeatureDataset.
End For
Return FeatureDataset
End

This image processing pipeline was developed to simulate a pathologist's expert examination, creating a quantitative feature set from qualitative observations in order to capture the essential diagnostic information contained within the tissue slides.

3.3. Extremely Randomized Trees Classifier

The Extremely Randomized Trees Classifier (Extra-Trees) is a sophisticated ensemble learning method that aggregates multiple decision trees to enhance classification accuracy. As a member of the averaging family, its primary objective is to reduce model variance, thereby improving generalization and minimizing the risk

of overfitting. While it shares a foundation with the Random Forest algorithm, Extra-Trees introduces a distinct modification in tree construction that significantly enhances randomness. [38]. The fundamental difference from Random Forest lies in two key aspects of the tree-building process. Firstly, Extra-Trees avoids bootstrap sampling, instead constructing each decision tree using the entire original training dataset. Secondly, and most critically, it selects split points completely at random, unlike the deterministic optimal split calculation in Random Forest. This combination of using the full dataset and introducing high randomness creates uncorrelated trees, culminating in a more robust and stable predictive model [38], [39]. The fundamental mechanism by which the Extra Trees classifier operates is systematically presented in Algorithm 2.

To assess feature importance, the algorithm calculates the total reduction in a metric like the Gini index that a feature contributes across all trees. This aggregated value, known as Gini importance, is used to rank features for effective selection. Extra-Trees algorithms are widely valued in supervised learning for their high accuracy, computational efficiency, and ability to model complex, non-linear relationships. They perform particularly well in high-dimensional settings, remaining robust even in the presence of noisy or less relevant features [38], [39]. Ensemble methods, such as Extra-Trees, are designed to combine predictions from multiple base models into a single, improved composite model. These techniques generally fall into two main categories: Averaging and Boosting. Averaging methods, which include Bagging, Random Forests, and Extra-Trees, work by constructing multiple independent base models and then averaging their outputs. This approach is particularly effective at reducing variance and mitigating the risk of overfitting.

Algorithm 2: The Extra-Trees Classifier

Objective: To build an ensemble of extremely randomized decision trees.
Input: A training dataset D containing N samples and M features.
Parameters:
T: The total number of decision trees to build in the forest.
k: The size of the random feature subset to consider at each node (k <= M).
Output: An ensemble model capable of making predictions on new data.
Procedure:
BEGIN
INITIALIZE 'Forest' as an empty collection of trees
FOR i from 1 to T DO:
  IF 'StoppingCriteria' THEN
    Convert the node to a leaf and halt splitting.
  ELSE
    FeatureSubs ← RANDOMLY_SELECT 'k' features from the M available features.
    INITIALIZE Best Split as null
    FOR EACH feature in Feature Subset DO:
     RandomThreshold ← GENERATE_RANDOM_VAL.
     CurrentSplitQuality ← EVALUATE_SPLIT (feature, RandomThreshold)
       using Gini Index.
     IF CurrentSplitQuality is better than BestSplit's quality
      BestSplit ← (feature, RandomThreshold)
     END IF
    END FOR
    LeftChildData, RightChildData ← PARTITION_DATA using BestSplit.
  END IF
  Forest ← Tree_i
END FOR
TO PREDICT for a new sample `x`: Prediction ← MAJORITY_VOTE
RETURN `Prediction`
END

This approach ensures that Extra Trees generate diverse decision trees, leading to a strong, stable, and accurate predictive model.

4. PROPOSED MODEL

This research presents a novel Machine Learning framework for automated breast cancer classification using histopathological images. As illustrated in Figure 1, the pipeline integrates image curation, normalization, feature engineering, and classification via the Extremely Randomized Trees (Extra-Trees) Classifier. The foundation of the model is a heterogeneous dataset of 13,347 histopathological images compiled from primary and secondary sources. Each image undergoes pre-processing for standardization and enhancement before entering the feature engineering phase. This process extracts diagnostic information across three distinct categories: Nucleus Area and Shape (25 features), Color-based properties (10 features), and Image Texture (13 features). From this initial set of 98 candidate features, a Recursive Feature Elimination (RFE) algorithm is applied to reduce dimensionality. This retains the 48 most discriminative features, mitigating the risk of overfitting and ensuring computational efficiency.

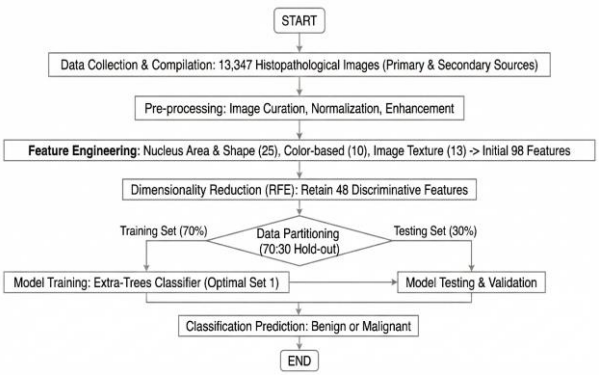

Figure 1. Proposed Model Workflow

The predictive core is the Extra-Trees Classifier, selected for its ability to lower variance through the use of

the full dataset and randomized split points. To determine the optimal configuration, a hyperparameter tuning experiment systematically evaluated four parameter sets, refer Table 2. As observed in Table 2, the classification accuracy consistently converged to 98.95% across all simulated hyperparameter sets. This indicates that the model reached a performance plateau where the selected 48 principal features provided sufficient discriminative power, making the accuracy resilient to further increases in forest depth or tree count. However, subtle variations were observed in the computational overhead, with Learning Time increasing from 1.95 to 2.20 seconds as the ensemble grew more complex; consequently, Set 1 was selected for the final implementation

Table 2. Extra Tree algorithm simulation results

Parameter	Set 1	Set 2	Set 3	Set 4
Number of trees in forest	10	30	60	70
Random state	1	4	6	7
Max. Features	sqrt	sqrt	auto	auto
Max. Depth	5	8	10	15
Criterion	Gini	Entropy	Entropy	Gini
Learning Time	1.95	2.08	2.19	2.20
Accuracy	98.95	98.95	98.95	98.95

The model was trained on 70% of the data to map feature vectors to binary labels (benign or malignant), while the remaining 30% was reserved as an unseen testing set to rigorously assess generalization. Algorithm 3 outlines this end-to-end operational flow, from data preparation to the final classification prediction.

Algorithm 3: Proposed Framework for Histopathological Image Classification

Input:
D_raw: Initial dataset of N raw histopathological images.
Y: Ground truth labels corresponding to D_raw, where yi € [Benign, Malignant].
T: Hyperparameter defining the number of trees in the Extra-Trees ensemble.
N_f: Number of features to select via RFE.
Output:
M_ET: Trained Extremely Randomized Trees classifier model.
P_metrics: Performance validation metrics on unseen data.
Procedure:
1: Data Preparation and Preprocessing
1. Initialize processed dataset D={}
2. for each image I_i in D_raw do :
Apply median filter to I_i for impulse noise suppression while preserving edges.
Convert I_i from RGB domain to grayscale domain.
Apply Otsu's thresholding to enhance nucleus boundaries and minimize background.
Add processed image I'_i to D'.
7. end for
2: Feature Extraction and Selection
8. Initialize feature matrix F_raw ={}
9. for each processed image I'_i in D' do :
Extract vector v_i consisting of 98 candidate features:
a) Nucleus shape and area features
b) Color-based statistical features
c) Texture descriptors
Append v_i to F_raw.
12. end for
13. Apply Recursive Feature Elimination (RFE) on F_raw using associated labels Y.
14. Select top N_f discriminative features to construct final feature matrix F.
3: Dataset Partitioning
15. Shuffle feature matrix F and labels Y randomly.
16. Split F into training set F_train(70%) and testing set F_test (30%) using hold-out methodology.
4: Model Training (Extremely Randomized Trees)
17. Initialize Extra-Trees ensemble model M_ET with T decision trees.
18. for each tree t in {1,..., T} do:
Load the F_train into the root node of tree t.
repeat (Recursive Node Splitting)
Select a random subset of k candidate features available at the current node.
For each selected feature, generate split points completely at random.
Evaluate split quality based on Gini Index.
Select the random split that maximizes Gini reduction and partition the node until termination criteria are met.
19. end for
5: Validation
20. Generate predictions Y_test using trained model M_ET on F_test.
21. Calculate performance metrics P_metrics by comparing Y_test with ground truth test labels.
22. return M_ET, P_metrics

5. RESULTS AND DISCUSSION

This section presents a detailed exposition of the experimental outcomes derived from the proposed machine learning-based model for automated breast cancer detection using histopathological images. It includes a comprehensive discussion of the results, explaining their relationship to real-world clinical utility and the primary goals of the research; a summary of these classification evaluation metrics and their clinical relevance is presented in Table 3. A core objective was to create an affordable, maintainable solution for early breast cancer detection. By utilizing open-source tools, the model avoids prohibitive licensing costs, ensuring accessibility for diverse healthcare facilities. Python was selected as the foundation for its versatility, using libraries like OpenCV for image processing and scikit-learn for machine learning to provide a robust computational backbone

To thoroughly assess the efficacy and reliability of the proposed model, its performance was subjected to comprehensive evaluation using a suite of well-established classification metrics. These metrics are fundamentally

derived from the Confusion Matrix, which quantifies the model's predictions into four critical categories relevant to clinical diagnosis. True Positives (TP) represent cases where malignant histopathological images were correctly identified by the model, signifying a correct cancer diagnosis. Conversely, False Positives (FP), also known as Type I errors, occur when benign images are incorrectly predicted as malignant, potentially leading to unnecessary follow-up procedures and patient anxiety. True Negatives (TN) denote instances where benign images were correctly identified as healthy tissue. Finally, False Negatives (FN), or Type II errors, represent the most critical error in this context, occurring when truly malignant images are incorrectly predicted as benign, signifying a missed cancer diagnosis with potentially severe prognostic implications.

Table 3. Classification Metrics and Clinical Relevance

Metric	Clinical Relevance	Metric	Clinical Relevance
True Positive	Correct identification of cancer.	Sensitivity (Recall)	Crucial for not missing cancer (minimizing Type II errors).
False Positive	Unnecessary biopsies, patient anxiety.	Specificity	Crucial for not misdiagnosing healthy tissue as cancerous (minimizing Type I errors).
True Negative	Correct identification of healthy tissue.	Precision	Minimizing unnecessary interventions from false alarms.
False Negative	Missed cancer diagnosis, delayed treatment, severe consequences.	F1-Score	Balanced measure of accuracy, especially when considering false positives and negatives.
Accuracy	Overall correctness, but can be misleading with imbalanced data.	ROC AUC	Overall diagnostic ability, independent of classification threshold.
Kappa Statistic	Robust measure of agreement, especially for imbalanced datasets.		

The subsequent sections with Figure 2 elaborates on the individual performance metrics, connecting their numerical values to the model's efficacy and clinical implications [40].

Accuracy, a fundamental measure quantifying correct classifications, reached 98.95% in the proposed model, indicating that nearly 99 out of 100 images were identified correctly. While high accuracy suggests a reliable system, interpreting it requires caution due to potential class imbalances common in medical datasets. As summarized in Figure 2, the Kappa Statistic was employed to account for chance agreement. The model yielded an exceptional Kappa value of 97.62%, indicating almost perfect agreement between predictions and true labels beyond random chance, ensuring the classifications are genuinely informed.

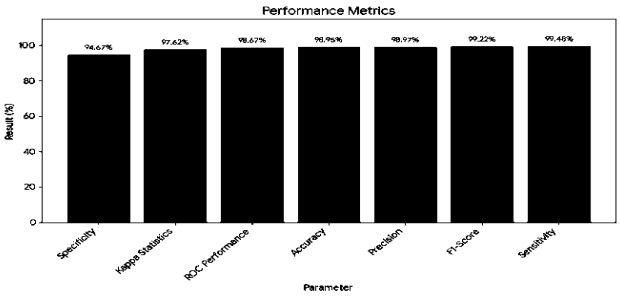

Figure 2. Performance Metrics of the Proposed Method

In clinical diagnostics, Sensitivity (Recall) is critical for minimizing False Negatives that lead to missed diagnoses. The model achieved an outstanding Sensitivity of 99.48%, ensuring nearly all malignant cases were correctly identified. Complementing this, Specificity measures the ability to identify benign samples and prevent False Positives. The model demonstrated a strong Specificity of 94.67%, exhibiting excellent precision in distinguishing healthy tissue and reducing unnecessary follow-up procedures. Precision, quantifying the proportion of true positive predictions, stood at 98.97%, ensuring that a malignant flag carries a high likelihood of accuracy. The F1-Score, the harmonic mean of precision and recall, reached 99.22%, reflecting a robust balance between detecting positive cases and minimizing false alarms. Despite the high performance, a critical error analysis of the 0.52% false negative rate reveals that missed detections primarily occurred in images with significant nuclear overlapping or variations in staining intensity from the primary dataset. Furthermore, a primary limitation of the current framework is its focus on binary classification (benign vs. malignant). While this is a vital first step, future iterations will aim to transition toward multi-class classification to identify specific histopathological subtypes, such as ductal or lobular carcinoma, further enhancing its clinical utility. To further validate discrimination efficiency, the Receiver Operating Characteristic (ROC) curve was analyzed. The model achieved an impressive Area Under the Curve (AUC) of 98.67%, visually represented in Figure 3, confirming its superior ability to distinguish between benign and malignant cases.

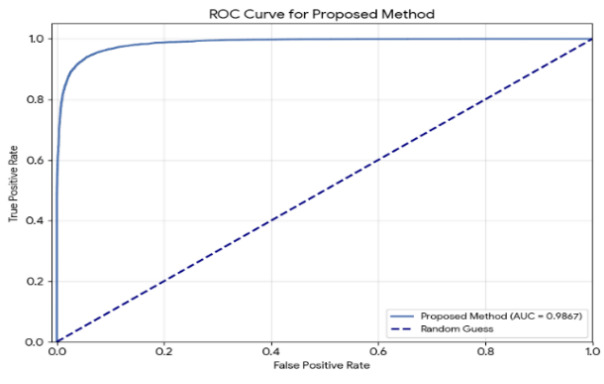

Figure 3. ROC Curve of the Proposed Method

6. CONCLUSION

The primary objective of this research was to address the complexity of automated breast cancer detection from histopathological images, a domain often overshadowed by mammogram-based analysis. By strategically leveraging Python and its open-source ecosystem, this study successfully developed an affordable and maintainable software solution, eliminating proprietary cost barriers for diverse clinical settings. A review of contemporary literature identified a gap in utilizing the Extremely Randomized Trees (Extra Trees) Classifier for this specific application. This research bridges that gap, demonstrating the algorithm's unique ability to reduce variance and handle the complex feature interactions inherent in heterogeneous tissue data.

The framework's success is substantiated by its exceptional quantitative performance. Combining robust pre-processing with an optimized feature set, the model achieved an overall accuracy of 98.95%. Crucially, it yielded a sensitivity of 99.48%, minimizing the risk of false negatives and ensuring the reliable detection of malignancies. This was balanced by a strong specificity of 94.67%, effectively identifying benign tissue to prevent unnecessary patient anxiety. The model's robustness was further confirmed by a Kappa Statistic of 97.62% and a Receiver Operating Characteristic (ROC) performance of 98.67%. To contextualize these findings, Table 4 presents a comparative analysis with recent studies, highlighting the competitive and superior performance of our proposed approach.

Table 4. Comparative Analysis with Recent Breast Cancer Detection Studies

Reference	Proposed Method	Dataset Size	Accuracy (%)	Sensitivity / F1 (%)
[41]	Ensemble of Fine-Tuned VGG16 & VGG19	7,909	95.29	97.73
[42]	DarkNet19 with Attention Branch Network (ABN-DCN)	>7,000	98.70	High Sensitivity Reported
[43]	Vision Transformer (ViT) Ensemble	7,909	97.50	98.80
[44]	DenseNet121-based Deep Learning Model	7,909	98.50	F1: 98.60 (Malignant)
Proposed Study	Image Processing + RFE + Extra Trees	13,347	98.95	99.48

As evidenced in Table 4, the proposed model utilizes a significantly larger and more diverse dataset while achieving higher sensitivity than comparable studies, underscoring its clinical reliability. Furthermore, the proposed Extra Trees framework offers a significantly lighter footprint, requiring minimal RAM and standard CPU processing compared to the high-end GPU requirements of the VGG and DenseNet-based models cited in this comparison. The combination of comprehensive feature engineering and a powerful ensemble classifier has proven to be a highly effective strategy. This research successfully implements a complete pipeline that translates raw histopathological images into accurate predictions by computationally replicating key diagnostic attributes; such as nuclear shape, colour, and texture, effectively automating expert pathological assessment.

The system's robust performance indicates strong potential as a clinical assistive tool, providing a rapid, objective "second opinion" to reduce inter-observer variability and pathologist workload. The inclusion of primary data further reinforces confidence in the model's capability to handle real-world samples. Ultimately, this study makes a significant contribution to computational pathology by presenting a robust and accessible framework for breast cancer detection. The results validate the methodology, supporting the system's capacity to enhance diagnostic speed and accuracy, leading to improved patient outcomes.

7. RECOMMENDATIONS

Based on the insights gained from this experimentation, several recommendations are proposed to guide future research in breast cancer detection. First, future studies should prioritize advanced, adaptive image processing techniques; such as non-local means denoising and color normalization, to minimize inter-slide variability and enhance feature consistency. Concurrently, attention must be directed toward optimizing feature selection by exploring novel handcrafted metrics or employing sophisticated algorithms like genetic algorithms to isolate the most critical diagnostic cues. To transcend the limitations of manual feature engineering, we strongly advocate for the integration of deep learning architectures; utilizing transfer learning can automatically learn robust, hierarchical features directly from raw pixels. Finally, to better replicate routine clinical practice, future models should be trained on multi-magnification images, enabling the simultaneous assessment of both overall tissue architecture and fine cellular morphology.

Ethical Statement

The author of this research adheres to all established ethical guidelines for academic publishing. This includes maintaining the integrity of data reporting, ensuring the originality of the research presented, and providing accurate citations for all referenced works. The study specifically states that it does not involve experiments conducted on human or animal subjects

CRediT Authorship Contribution Statement

Mahendra Kanojia: Conceptualization, Methodology, Formal Analysis, Data Curation, Writing Original Draft, and Visualization.

Declaration of Competing Interest

The sole author declares that there are no financial or personal relationships with other people or organizations that could inappropriately influence or bias the work. There are no known conflicts of interest regarding the publication of this manuscript.

Funding Source

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The study was developed using open-source tools to ensure a cost-effective and maintainable solution.

Acknowledgements

The author expresses gratitude to the Department of Histopathology, BSES Municipal General Hospital, Swami Andheri West, Mumbai, Maharashtra, India, for their essential assistance in the collection of primary histopathological image data used in this study.

Data Availability Statement

The research utilizes a combination of primary and secondary data sources. Secondary data from the 2018 Data Science Bowl (Kaggle), BreaKHis, and the Center for Bio-Image Informatics are available through their respective public repositories. The primary histopathological image data collected for this research is available from the corresponding author upon reasonable request.

References

[1] World Health Organization, "Breast cancer," World Health Organization, Fact sheet, Aug. 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] C. G. Yedjou, S. S. Tchounwou, R. A. Aló, R. Elhag, B. Mochona, and L. Latinwo, "Application of Machine Learning Algorithms in Breast Cancer Diagnosis and Classification," *Int. J. Sci. Acad. Res.*, vol. 2, no. 1, pp. 3081–3086, Jan. 2021.

[3] M. G. Kanojia, Mohd. A. Mohd. H. Ansari, N. Gandhi, and S. K. Yadav, "Image Processing Techniques for Breast Cancer Detection: A Review," in *Intelligent Systems Design and Applications*, vol. 1181, A. Abraham, P. Siarry, K. Ma, and A. Kaklauskas, Eds., in Advances in Intelligent Systems and Computing, vol. 1181. , Cham: Springer International Publishing, 2021, pp. 649–660. doi: 10.1007/978-3-030-49342-4_63.

[4] R. Krithiga and P. Geetha, "Breast Cancer Detection, Segmentation and Classification on Histopathology Images Analysis: A Systematic Review," *Arch. Comput. Methods Eng.*, vol. 28, no. 4, pp. 2607–2619, Jun. 2021, doi: 10.1007/s11831-020-09470-w.

[5] A. S. Boddu and A. Jan, "A systematic review of machine learning algorithms for breast cancer detection," *Tissue Cell*, vol. 95, p. 102929, Aug. 2025, doi: 10.1016/j.tice.2025.102929.

[6] S. Nabajja, M. Kanojia, and T. Yadav, "Choledochal Cancer Region Detection in Hyperspectral Tissue Images Using U-Net," in *Intelligent Systems Design and Applications*, vol. 1046, A. Abraham, A. Bajaj, T. Hanne, and P. Siarry, Eds., in Lecture Notes in Networks and Systems, vol. 1046. , Cham: Springer Nature Switzerland, 2024, pp. 316–325. doi: 10.1007/978-3-031-64813-7_33.

[7] A. Kuşcu and H. Erol, "Diagnosis of Breast Cancer by K-Mean Clustering and Otsu Thresholding Segmentation Methods," *Osman. Korkut Ata Üniversitesi Fen Bilim. Enstitüsü Derg.*, vol. 5, no. 1, pp. 258–281, Mar. 2022, doi: 10.47495/okufbed.994481.

[8] G. Alfian *et al.*, "Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method," *Computers*, vol. 11, no. 9, p. 136, Sep. 2022, doi: 10.3390/computers11090136.

[9] Sajiv. G and G. Ramkumar, "A Robust Breast Cancer Classification Model using Extra-Trees Classifier for Histopathological Image," in *2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, Chennai, India: IEEE, May 2023, pp. 1–7. doi: 10.1109/ACCAI58221.2023.10199852.

[10] D. Sharma, R. Kumar, and A. Jain, "Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning," *Meas. Sens.*, vol. 24, p. 100560, Dec. 2022, doi: 10.1016/j.measen.2022.100560.

[11] M. G. Kanojia and S. Abraham, "Breast cancer detection using RBF neural network," in *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, Greater Noida, India: IEEE, Dec. 2016, pp. 363–368. doi: 10.1109/IC3I.2016.7917990.

[12] M. F. Ak, "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications," *Healthcare*, vol. 8, no. 2, p. 111, Apr. 2020, doi: 10.3390/healthcare8020111.

[13] H. Tabrizchi, M. Tabrizchi, and H. Tabrizchi, "Breast cancer diagnosis using a multi-verse optimizer-based gradient boosting decision tree," *SN Appl. Sci.*, vol. 2, no. 4, p. 752, Apr. 2020, doi: 10.1007/s42452-020-2575-9.

[14] M. Phankokkruad, "Cost-Sensitive Extreme Gradient Boosting for Imbalanced Classification of Breast Cancer Diagnosis," in *2020 10th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, Penang, Malaysia: IEEE, Aug. 2020, pp. 46–51. doi: 10.1109/ICCSCE50387.2020.9204948.

[15] G. N. Gurav and M. G. Kanojia, "A review on classification of breast cancer histopathological images using convolutional neural networks," *Spec. Issue Int. J. Comput. Sci. Appl.*, vol. 13, no. 1, 2020.

[16] T. D. Murugan and M. G. Kanojia, "Breast Cancer Detection Using Texture Features and KNN Algorithm," in *Hybrid Intelligent Systems*, vol. 1375, A. Abraham, T. Hanne, O. Castillo, N. Gandhi, T. Nogueira Rios, and T.-P. Hong, Eds., in Advances in Intelligent Systems and Computing, vol. 1375. , Cham: Springer International Publishing, 2021, pp. 793–802. doi: 10.1007/978-3-030-73050-5_77.

[17] M. G. Kanojia, Mohd. A. Mohd. H. Ansari, N. Gandhi, and S. K. Yadav, "Computer Aided System for Nuclei Localization in Histopathological Images Using CNN," in *Proceedings of the 11th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2019)*, vol. 1182, A. Abraham, M. A. Jabbar, S. Tiwari, and I. M. S. Jesus, Eds., in Advances in Intelligent Systems and Computing, vol. 1182. , Cham: Springer International Publishing, 2021, pp. 226–234. doi: 10.1007/978-3-030-49345-5_24.

[18] S. Abbas *et al.*, "BCD-WERT: a novel approach for breast cancer detection using whale optimization based efficient features and extremely randomized tree algorithm," *PeerJ Comput. Sci.*, vol. 7, p. e390, Mar. 2021, doi: 10.7717/peerj-cs.390.

[19] A. Gupta *et al.*, "Prediction of Breast Cancer Using Extremely Randomized Clustering Forests (ERCF) Technique: Prediction of Breast Cancer," *Int. J. Distrib. Syst. Technol.*, vol. 12, no. 4, pp. 1–15, Dec. 2021, doi: 10.4018/IJDST.287859.

[20] T. Elizabeth Mathew, "An optimized extremely randomized tree model for breast cancer classification," *J. Theor. Appl. Inf.*

Technol., vol. 100, no. 16, pp. 5234–5246, Aug. 2022.

[21] N. Binsaif, "Application of Machine Learning Models to the Detection of Breast Cancer," *Mob. Inf. Syst.*, vol. 2022, pp. 1–8, Mar. 2022, doi: 10.1155/2022/7340689.

[22] H. Liang, J. Li, H. Wu, L. Li, X. Zhou, and X. Jiang, "Mammographic Classification of Breast Cancer Microcalcifications through Extreme Gradient Boosting," *Electronics*, vol. 11, no. 15, p. 2435, Aug. 2022, doi: 10.3390/electronics11152435.

[23] T. Tran, U. Le, and Y. Shi, "An effective up-sampling approach for breast cancer prediction with imbalanced data: A machine learning model-based comparative analysis," *PLOS ONE*, vol. 17, no. 5, p. e0269135, May 2022, doi: 10.1371/journal.pone.0269135.

[24] E. A. Algehyne, M. L. Jibril, N. A. Algehainy, O. A. Alamri, and A. K. Alzahrani, "Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia," *Big Data Cogn. Comput.*, vol. 6, no. 1, p. 13, Jan. 2022, doi: 10.3390/bdcc6010013.

[25] A. Batool and Y.-C. Byun, "Toward Improving Breast Cancer Classification Using an Adaptive Voting Ensemble Learning Algorithm," *IEEE Access*, vol. 12, pp. 12869–12882, 2024, doi: 10.1109/ACCESS.2024.3356602.

[26] M. Momtahen, S. Momtahen, R. Remaseshan, and F. Golnaraghi, "Early Detection of Breast Cancer using Diffuse Optical Probe and Ensemble Learning Method," in *2023 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO)*, Winnipeg, MB, Canada: IEEE, Jun. 2023, pp. 139–142. doi: 10.1109/NEMO56117.2023.10202520.

[27] S. Naveed, "Prediction of Breast Cancer Through Random Forest," *Curr. Med. Imaging Rev.*, vol. 19, no. 10, p. e300922209414, Sep. 2023, doi: 10.2174/1573405618666220930150625.

[28] R. Sinha, M. Patel, S. Gupta, K. K. Sinha, and Prateeksha, "Performance Analysis of Breast Cancer Predictor using Machine Learning Techniques," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India: IEEE, Jun. 2024, pp. 1–5. doi: 10.1109/ICCCNT61001.2024.10724436.

[29] S. Sasidharan Nair and M. Subaji, "Automated Identification of Breast Cancer Type Using Novel Multipath Transfer Learning and Ensemble of Classifier," *IEEE Access*, vol. 12, pp. 87560–87578, 2024, doi: 10.1109/ACCESS.2024.3415482.

[30] B. N. Ravi Kumar, Naveen Chandra Gowda, A. B. J., V. H. N., B. Ben Sujitha, and D. Roja Ramani, "An Efficient Breast Cancer Detection Using Machine Learning Classification Models," *Int. J. Online Biomed. Eng. IJOE*, vol. 20, no. 13, pp. 24–40, Oct. 2024, doi: 10.3991/ijoe.v20i13.50289.

[31] I. Kadhim Ajlan, H. Murad, A. A. Salim, and A. Fadhil Bin Yousif, "Extreme Learning machine algorithm for breast Cancer diagnosis," *Multimed. Tools Appl.*, vol. 84, no. 15, pp. 14739–14758, Jun. 2024, doi: 10.1007/s11042-024-19515-y.

[32] P. Sarker, A. Ksibi, M. M. Jamjoom, K. Choi, A. A. Nahid, and M. A. Samad, "Breast cancer prediction with feature-selected XGB classifier, optimized by metaheuristic algorithms," *J. Big Data*, vol. 12, no. 1, p. 78, Apr. 2025, doi: 10.1186/s40537-025-01132-7.

[33] J. C. Caicedo *et al.*, "Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl," *Nat. Methods*, vol. 16, no. 12, pp. 1247–1253, Dec. 2019, doi: 10.1038/s41592-019-0612-7.

[34] 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.

[35] E. Drelie Gelasca, B. Obara, D. Fedorov, K. Kvilekval, and B. Manjunath, "A biosegmentation benchmark for evaluation of bioimage analysis methods," *BMC Bioinformatics*, vol. 10, no. 1, p. 368, Dec. 2009, doi: 10.1186/1471-2105-10-368.

[36] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, Jan. 2002, doi: 10.1023/A:1012487302797.

[37] S. R. Vupulluri and J. K. Munagala, "Histopathological Image Analysis Using Deep Learning Framework," in *RAiSE-2023*, MDPI, Dec. 2023, p. 132. doi: 10.3390/engproc2023059132.

[38] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[39] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[40] Jude Chukwura Obi, "A comparative study of several classification metrics and their performances on data," *World J. Adv. Eng. Technol. Sci.*, vol. 8, no. 1, pp. 308–314, Feb. 2023, doi: 10.30574/wjaets.2023.8.1.0054.

[41] Z. Hameed, S. Zahia, B. Garcia-Zapirain, J. Javier Aguirre, and A. María Vanegas, "Breast Cancer Histopathology Image Classification Using an Ensemble of Deep Learning Models," *Sensors*, vol. 20, no. 16, p. 4373, Aug. 2020, doi: 10.3390/s20164373.

[42] S. Krishna, S. S. Suganthi, A. Bhavsar, J. Yesodharan, and S. Krishnamoorthy, "An interpretable decision-support model for breast cancer diagnosis using histopathology images," *J. Pathol. Inform.*, vol. 14, p. 100319, 2023, doi: 10.1016/j.jpi.2023.100319.

[43] E. M. Othman, "Breast Cancer Multi-Class Classification Using ViT Model," *Int. J. Comput. Appl.*, vol. 186, no. 13, pp. 13–18, 2024.

[44] A. Rafiq, A. Jaffar, G. Latif, S. Masood, and S. E. Abdelhamid, "Enhanced Multi-Class Breast Cancer Classification from Whole-Slide Histopathology Images Using a Proposed Deep Learning Model," *Diagnostics*, vol. 15, no. 5, p. 582, Feb. 2025, doi: 10.3390/diagnostics15050582.