Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7, 182-201
2026
Publisher: Learning Gate
DOI: 10.55214/2576-8484.v10i7.13338

Robust breast cancer classification via ensemble learning: Integrating feature analysis and error evaluation techniques

Mosima A. Masethe[1*], Hlaudi D. Masethe[2]
[1]Department of Computer Science and Information Technology, School of Science and Technology, Sefako Makgatho Health Sciences University, Pretoria, 0204, South Africa; mosima.masethe@smu.ac.za (M.A.M.).
[2]Department of Computer Science, Faculty of Information and Communication Technology, Tshwane University of Technology, Pretoria, 4000, South Africa; masethehd@tut.ac.za (H.D.H).

Abstract: Breast cancer remains one of the leading causes of cancer-related mortality worldwide, highlighting the need for accurate and reliable computer-aided diagnostic methods. This study evaluates the effectiveness of ensemble learning techniques for classifying breast tumors as benign or malignant using clinical tumor features. A data-driven model was developed using a publicly available Kaggle breast cancer dataset. Four ensemble learning models, Random Forest, Gradient Boosting, AdaBoost, and Extra Trees, were trained and comparatively evaluated using accuracy, precision, recall, F1-score, and Receiver Operating Characteristic Area Under the Curve (ROC-AUC). Feature importance and error analyses were also conducted to identify influential predictors and assess model reliability. All evaluated models achieved excellent predictive performance, with classification accuracies above 96% and ROC-AUC values exceeding 0.99. The Extra Trees classifier demonstrated the best overall performance, achieving the highest accuracy, the lowest error rate, and superior sensitivity in detecting malignant tumors. Tumor characteristics related to perimeter, radius, and concave points were identified as the most influential features. Ensemble learning provides a robust and reliable approach for breast cancer classification, enabling highly accurate diagnostic predictions. The proposed model can support clinicians in early breast cancer detection and decision-making, potentially improving diagnostic accuracy and patient outcomes while reducing misclassification of malignant cases.

Keywords: *AdaBoost, Breast cancer classification, Ensemble learning, Explainable AI, Extra trees, Feature importance, Gradient boosting, Machine learning, Medical diagnosis, Random forest, ROC-AUC.*

1. Introduction

Early detection and precise diagnosis are pivotal in ameliorating the prognosis and overall survival rates for patients afflicted with breast cancer [1]. In this context, advanced machine learning paradigms, particularly ensemble learning, have shown significant promise for improving the accuracy and robustness of diagnostic models [2]. The inherent complexity, heterogeneity, and high dimensionality of breast cancer data pose substantial challenges for classification, often leading to potential human errors or noise in traditional diagnostic methods [3]. Ensemble learning, by aggregating predictions from multiple base learners, mitigates these limitations by leveraging diverse perspectives to improve overall predictive stability and generalization performance [4]. This approach not only refines prediction capability and accuracy beyond individual models but also bolsters the model's robustness by mitigating false positives or negatives that may arise from biased decisions of a single classifier [5]. This is particularly relevant in breast cancer diagnosis, where the heterogeneous nature of the disease and the complexity of integrating diverse data sources necessitate robust analytical frameworks [6].

The integration of feature analysis within these frameworks further refines the diagnostic process by identifying salient patterns and biomarkers, thereby optimizing the input for the ensemble models [7].

History: Received: 23 April 2026; Revised: 22 June 2026; Accepted: 26 June 2026; Published: 21 July 2026
* Correspondence: mosima.masethe@smu.ac.za

This holistic strategy aims to enhance diagnostic precision and address the critical need for explainable AI in clinical settings by making the decision-making process more transparent [8]. The framework's ability to incorporate diverse data modalities, such as histopathological images and clinical metadata, further amplifies its diagnostic power by providing a comprehensive assessment of tumor characteristics Sumitha and Isaac [9]. Intan et al. [10] leverage advanced feature extraction and fusion techniques, combining cell quantification with ConvNexTiny features to significantly enhance the accuracy, precision, recall, F1-score, and ROC-AUC in detecting molecular expression biomarkers. This heterogeneous ensemble approach capitalizes on the strengths of both transfer learning from convolutional neural networks and practical cell quantification models, leading to balanced algorithmic complexity and superior performance [10]. This methodological advancement is crucial for improving patient outcomes, as it enables more accurate and reliable breast cancer diagnoses [10]. Furthermore, the model's interpretability is enhanced through techniques like Grad-CAM, which generate localization maps to highlight critical diagnostic regions within histopathology images, making the decision-making process more transparent for clinicians [8, 11].

This enhanced interpretability, coupled with a 98.6% accuracy and a rapid processing time of 0.75 seconds, significantly outperforms existing diagnostic methods, facilitating more precise and timely cancer detection [12]. Furthermore, the incorporation of explainable AI techniques, such as SHAP and LIME, provides feature-level attributions, thus enhancing the interpretability of deep learning models and fostering clinician trust [13]. This increased transparency is critical for clinical adoption, as it allows medical professionals to understand the rationale behind a model's prediction, thereby addressing the "black box" problem prevalent in many deep learning applications [14].

This study makes several key contributions to the field of data-driven healthcare and breast cancer diagnosis. It develops a robust ensemble learning framework incorporating Random Forest, Gradient Boosting, AdaBoost, and Extra Trees models to accurately classify benign and malignant tumors. A comprehensive comparative evaluation is conducted using multiple performance metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and error rates, enabling a clear assessment of model effectiveness. The study identifies the Extra Trees model as the best-performing classifier, achieving the highest accuracy and lowest error rate. In addition, detailed error analysis is performed to examine false positives and false negatives, providing insight into the clinical reliability of the models. Feature importance analysis further highlights the most influential tumor characteristics, aligning machine learning findings with established clinical knowledge. The integration of visual analytics enhances interpretability, while the consideration of class imbalance ensures fair evaluation. Overall, the study contributes to the development of accurate, interpretable, and reliable machine learning approaches for breast cancer diagnosis and provides a strong foundation for future research in hybrid and explainable AI models.

2. Literature Review

The increasing complexity of medical imaging data and the imperative for robust diagnostic tools have catalyzed significant research into advanced machine learning methods for breast cancer detection [15]. Early approaches predominantly relied on unimodal analysis, often focusing on mammography or histopathology independently; however, recent trends highlight the superior performance achievable through multimodal integration, combining diverse data sources like clinical metadata with imaging data [16, 17]. This shift towards multi-modal frameworks, encompassing early, intermediate, and late fusion strategies, along with advanced deep multimodal fusion techniques such as attention-based mechanisms and graph neural networks, has significantly improved diagnostic accuracy and comprehensiveness [15]. However, challenges persist in developing AI models that are not only accurate but also interpretable and clinically applicable across diverse breast cancer subtypes and challenging datasets [18].

Despite advancements, many current models still face limitations in accuracy and interpretability, particularly when dealing with small or imbalanced datasets [14]. The "black box" nature of traditional deep learning models often hinders their adoption in clinical settings, as medical professionals require

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

insights into the decision-making process to build trust and validate diagnoses [14, 19]. To address this, current research is increasingly exploring explainable AI techniques, such as Grad-CAM, SHAP, and attention weights, to elucidate the critical features driving model predictions and bridge the gap between complex AI systems and clinical workflows [20]. These methods are crucial for enhancing diagnostic confidence by visually highlighting diagnostically relevant tumor regions and statistically establishing oncological risk factors [21]. The integration of such explainable multimodal approaches is proving pivotal, often achieving 5%–10% higher accuracy compared to unimodal techniques while also fostering greater transparency in diagnostic decision-making [15, 22].

While multi-modal approaches have shown promise in improving diagnostic accuracy and providing a more comprehensive view of tumor biology, challenges remain in their practical implementation, particularly concerning the effective fusion of disparate data types and the development of robust, generalizable models [15, 23]. For instance, the scarcity of large, paired multimodal datasets and the inherent heterogeneity across different data modalities present significant technical and analytical hurdles for model development and validation [15, 24]. Furthermore, the "black box" nature of many deep learning models often compromises their clinical utility by obscuring the rationale behind diagnostic predictions, underscoring the critical need for integrating interpretable AI methods [15, 21, 25]. This concern is further compounded by the computational cost and complexity of training multimodal deep learning models, making it difficult to determine how different data types contribute to the final prediction [26]. Moreover, the interpretability of multimodal AI models faces additional hurdles, as explaining how information from different modalities interacts to form a prediction is more complex than explaining unimodal models [21, 27]. Specifically, the challenge lies in effectively integrating disparate data types, such as clinical variables, histopathology slides, and genetic information, into a coherent framework that yields robust and externally validated predictions [27, 28].

This necessitates innovative approaches that can not only fuse multi-omics and imaging data but also provide transparent insights into the underlying decision-making processes [29]. Such advancements are critical for overcoming challenges like data heterogeneity and ensuring model generalizability, thereby moving closer to personalized medicine in oncology [30-32]. A primary challenge lies in the effective harmonization and standardization of diverse data types before integration, as modalities often vary significantly in scale, complexity, and inherent characteristics [33]. The difficulty of combining data from various platforms also stems from platform noise, which can interfere with the integration process and necessitate advanced strategies to mitigate its effects [34]. The promise of multimodal data integration for oncology lies in its potential to capture the complex and heterogeneous nature of cancer data more comprehensively than traditional unimodal approaches [35].

3. Research Methodology

The envisaged mathematical modelling for this study is based on a supervised binary classification framework in which breast cancer diagnosis is formulated as a mapping from a set of clinical tumor features to a diagnostic outcome, namely benign or malignant. Let the dataset be represented in equation (1) as:

$$D = \{(x_i, y_i)\}_{i=1}^{N} \quad (1)$$

Where:

- $x_i \in \mathbb{R}^p$ denotes the feature vector of the $i - th$ patient, composed of tumor characteristics such as radius, texture, perimeter, area, smoothness, compactness, concavity, and related measures, and
- $y_i \in \{0, 1\}$ denotes the corresponding class label, with 0 representing benign and 1 representing malignant.
- N is the total number of samples.
- p is the number of features.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

The objective of the model is to learn a predictive function $f(x)$ that approximates the relationship between the observed tumor features and the diagnostic label. The objective is to learn a function as in the equation (2):

$$f: \mathbb{R}^p \rightarrow \{0,1\} \quad (2)$$

Such that equation (3) approximates the true label y_i

$$\hat{y}_i = f(x_i) \quad (3)$$

The study adopts an ensemble learning approach; the model consists of multiple base classifiers whose predictions are combined to improve robustness and generalisation. In general form, each classifier produces a prediction $f_m(x)$, where $m = 1,2,\ldots,M$ indexes the ensemble members. The overall decision function may then be expressed as an aggregated model as defined in equation (4):

$$F(x) = \sum_{m=1}^{M} w_m f_m(x) \quad (4)$$

Where:

- $w_m \geq 0$ represents the contribution or weight associated with m-th classifier

For classification, the final prediction is obtained using a decision rule as in equation (5):

$$\hat{y} = \begin{cases} 1, & if\ F(x) \geq \tau, \\ 0, & otherwise, \end{cases} \quad (5)$$

Where:

- $\tau \in [0,1]$ is the decision threshold, commonly set to 0.5.

The majority voting (unweighted ensemble) is defined in equation (6) as:

$$\hat{y} = mode\ \{f_1(x), f_2\{x),\ \ldots, f_M(x)\} \quad (6)$$

The classification task can be expressed probabilistically as in equation (7):

$$P(Y = 1 \mid X = x) \quad (7)$$

The decision rule in equation (8) becomes:

$$\hat{y} = \begin{cases} 1, & \text{if } P(Y = 1 \mid X = x) \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

This formulation is useful for interpreting ROC curves and AUC values, since these metrics examine model behaviour across varying threshold values.

In practice, this aggregation may be implemented through majority voting, weighted voting, or boosting-based sequential correction, depending on the ensemble technique under consideration. For example, in Random Forest and Extra Trees, the final prediction is based on the majority decision of multiple decision trees, while in AdaBoost and Gradient Boosting, weak learners are combined iteratively to minimise classification error.

The learning process is guided by minimizing a suitable classification loss function, such as misclassification error, Gini impurity, entropy, or exponential loss, depending on the model. For instance, the empirical classification error can be expressed as in equation (9)

$$L = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i \neq \hat{y}_i), \quad (9)$$

where:

- $\mathbb{I}(\cdot)$ is the indicator function that equals 1 when a prediction is incorrect and 0 otherwise. Further depicted in equation (10) as:

$$\mathbb{I}(y_i \neq \hat{y}_i) = \begin{cases} 1, & if\ y_i \neq \hat{y}_i, \\ 0, & otherwise, \end{cases} \quad (10)$$

For boosting-based models, the exponential loss can be used as in equation (11):

$$L_{exp} = \sum_{i=1}^{N} \exp(-y_i F(x_i)) \quad (11)$$

In boosting methods, this loss is reduced iteratively by assigning greater emphasis to misclassified observations, thereby improving predictive performance on difficult cases. In tree-based ensemble

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

methods, repeated partitioning of the feature space creates decision boundaries that separate benign and malignant samples based on combinations of tumor characteristics.

In addition, the envisaged mathematical modelling includes a feature-importance formulation to quantify the relative contribution of each predictor to the final classification. If the importance of the feature j is denoted by I_j, then features can be ranked such that $I_1, I_2, \ldots, I_p$ reflect their discriminatory influence. This component is essential for identifying the most clinically meaningful attributes, such as perimeter, radius, or concave points, and for linking computational findings with biomedical interpretation. The feature importance is defined in equation (12) as:

$$I_j = \frac{1}{M}\sum_{m=1}^{M} I_j^{(m)}, \quad (12)$$

Where:

- $I_j^{(m)}$ is the contribution of a feature j in model m.

The study also incorporates an error-evaluation framework based on the confusion matrix, where classification outcomes are decomposed into true positives, true negatives, false positives, and false negatives. From these quantities, performance metrics such as accuracy, precision, recall, F1 score, and error rate are derived mathematically to assess model effectiveness. This ensures that the modelling does not rely solely on predictive accuracy but also accounts for the clinical implications of different types of errors, especially false negatives in malignant-case detection.

The envisaged mathematical modelling for this study combines binary classification theory, ensemble decision aggregation, probabilistic prediction, feature-importance analysis, and error-based evaluation into a unified framework. This approach is intended to provide a robust, interpretable, and clinically meaningful basis for breast cancer classification using historical patient data.

4. Research Results

4.1. Class Distribution Analysis

The class distribution plot in Figure 1 indicates a moderate imbalance in the dataset, with benign cases significantly outnumbering malignant ones. This reflects real-world clinical patterns but introduces potential bias in model training, as algorithms may favor the majority class, leading to inflated accuracy while under-detecting malignant cases. In a medical context, such misclassification is critical, particularly false negatives, where malignant tumors are incorrectly predicted as benign. Therefore, the use of weighted evaluation metrics, as applied in this study, is appropriate to account for this imbalance. Despite the skewed distribution, the strong recall and ROC-AUC values observed suggest that the models effectively identified malignant cases, demonstrating robustness in handling class imbalance while maintaining clinically relevant predictive performance.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

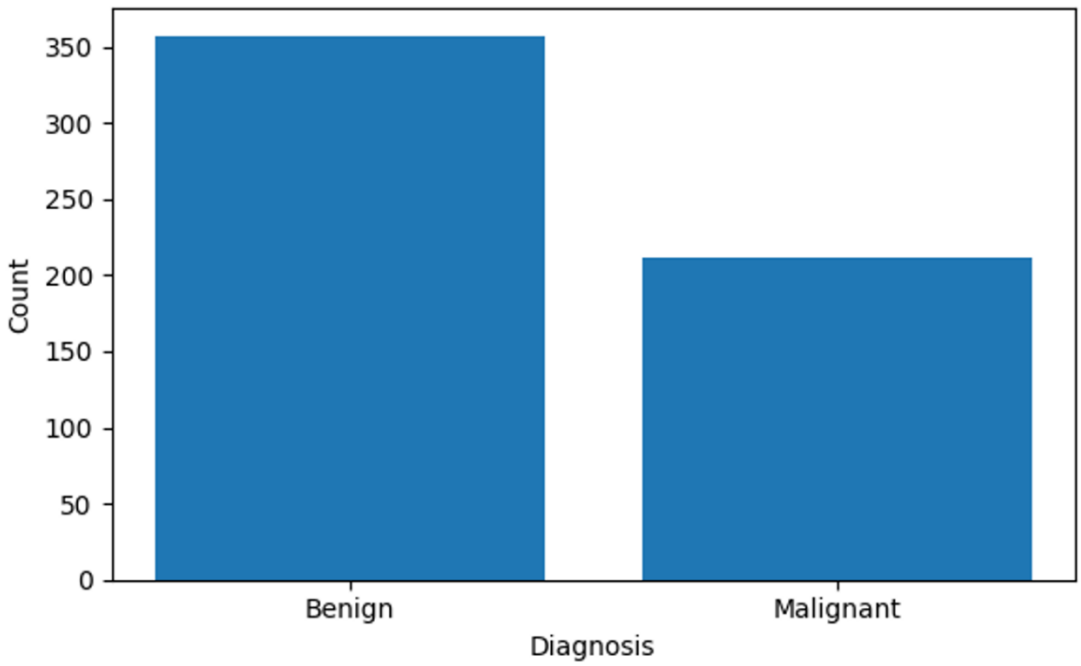

Figure 1.
Class Distribution.

The feature correlation heatmap in Figure 2 reveals strong relationships among several tumor characteristics, particularly between radius, perimeter, and area, which exhibit high positive correlations across their mean, standard error (SE), and worst measurements. This suggests redundancy among these features, as they capture similar geometric properties of the tumor. Additionally, features related to concavity, compactness, and concave points also show strong positive correlations, indicating that they collectively describe the irregularity and shape complexity of tumors, important indicators of malignancy. In contrast, features such as fractal dimension and symmetry display weaker or mixed correlations with other variables, suggesting they contribute more independently to the model. The presence of highly correlated features highlights the importance of feature selection or dimensionality reduction to avoid multicollinearity, although ensemble models are generally robust to such effects. Overall, the heatmap confirms that tumor size and shape-related features play a critical role in distinguishing between benign and malignant cases.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

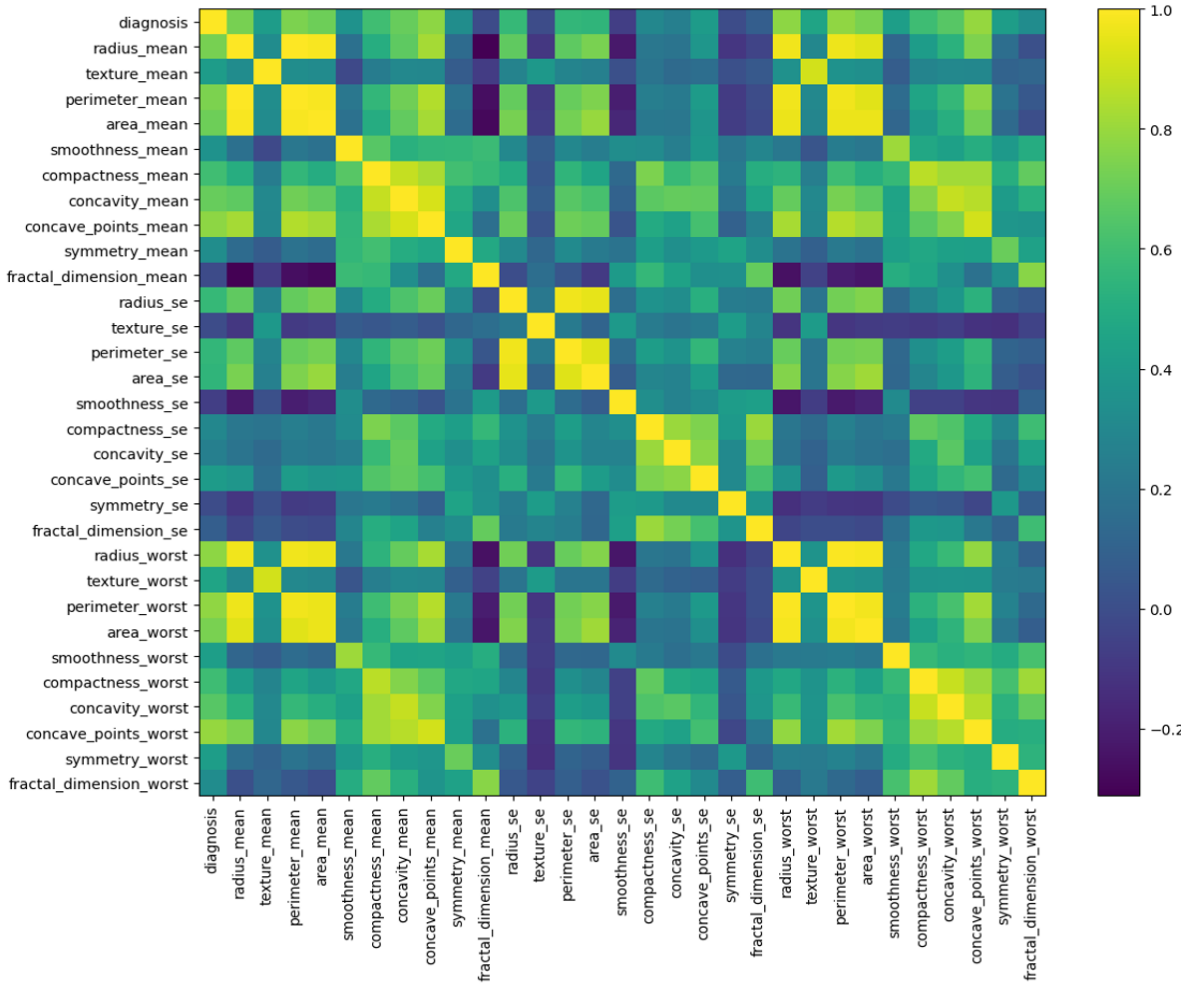

Figure 2.
Feature Correlation Heatmap.

4.2. Experiment Results: Random Forest Performance Analysis

The Random Forest model in Table 1 achieved strong classification performance, with a weighted precision of 97%, recall of 96%, and F1-score of 96%, resulting in an overall accuracy of 96%. The model also produced a high ROC-AUC score of 99%, indicating excellent discriminative ability between benign and malignant cases. The relatively low error rate of 3.5% suggests that the model is reliable in clinical classification tasks. However, the slight difference between precision and recall indicates a marginal imbalance in predicting positive versus negative classes, which may be attributed to the inherent variability in tumor feature distributions. Overall, the Random Forest model demonstrates robustness and stability, making it a strong baseline for ensemble-based cancer classification.

4.3. Experiment Results: Gradient Boosting Performance Analysis

The Gradient Boosting model in Table 1 exhibits performance comparable to Random Forest, achieving 97% precision, 96% recall, and 96% F1-score, with an overall accuracy of 96%. Notably, it achieves a slightly higher ROC-AUC of 99.5%, suggesting improved capability in ranking predictions and distinguishing between classes. Despite its similar error rate of 3.5%, the model's boosting mechanism allows it to iteratively correct previous errors, enhancing predictive confidence. This indicates that

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

Gradient Boosting is particularly effective in capturing complex, nonlinear relationships among tumor features, making it well-suited for medical diagnostic tasks where subtle feature interactions are critical.

4.4. Experiment Results: AdaBoost Performance Analysis

AdaBoost in Table 1 demonstrates improved performance over the previous models, achieving 97% precision, 97% recall, and 97% F1-score, with an overall accuracy of 97%. The model maintains a high ROC-AUC of 99% and a reduced error rate of 2.6%, indicating better generalisation and fewer misclassifications. The balanced precision and recall suggest that AdaBoost effectively handles both benign and malignant classifications without bias toward either class. This improvement can be attributed to AdaBoost's adaptive weighting mechanism, which emphasizes misclassified instances during training, thereby refining the model's ability to correctly classify difficult cases.

4.5. Experiment Results: Extra Trees Performance Analysis

The Extra Trees model in Table 1 achieves the highest performance among all evaluated models, with 98% precision, 98% recall, and 98% F1-score, leading to an overall accuracy of 98%. It also records the highest ROC-AUC score of 99.8%, demonstrating near-perfect classification capability. The model's error rate of 1.75% is the lowest across all models, indicating superior predictive reliability. The strong performance of Extra Trees can be attributed to its use of randomized feature selection and ensemble averaging, which reduces variance and improves generalisation. This makes it particularly effective in handling high-dimensional biomedical data and capturing subtle patterns in tumor characteristics.

4.6. Comparative Analysis of Ensemble Models

A comparative evaluation of all ensemble models in Table 1 reveals a clear performance hierarchy, with Extra Trees outperforming all other models, followed by AdaBoost, and then Random Forest and Gradient Boosting. While all models demonstrate high accuracy and strong ROC-AUC scores (above 99%), differences emerge in terms of error rates and balance between precision and recall. The lower error rates observed in AdaBoost and Extra Trees suggest better generalization and robustness in classification tasks. These findings highlight the importance of ensemble diversity and adaptive learning strategies in improving predictive performance for medical diagnosis.

4.7. Error Analysis and Model Reliability

The error rates across the models in Table 1 range from 3.5% to 1.75%, indicating generally high reliability in classification. The reduction in error rate from Random Forest and Gradient Boosting to Extra Trees demonstrates the effectiveness of increased randomness and ensemble diversity in minimizing misclassification. In a clinical context, reducing false negatives (misclassifying malignant cases as benign) is particularly critical. The high recall values across all models indicate strong sensitivity to malignant cases, which is essential for early detection and treatment. The results suggest that ensemble methods, particularly Extra Trees and AdaBoost, provide a reliable framework for minimizing diagnostic errors in breast cancer classification.

4.8. Overall Interpretation

The experimental results in Table 1 demonstrate that ensemble learning techniques are highly effective for breast cancer classification using clinical tumor features. All models achieved accuracy above 96% and ROC-AUC above 99%, confirming their strong predictive capabilities. Among them, the Extra Trees model stands out as the most reliable and accurate, making it a strong candidate for deployment in decision-support systems. These findings support the use of ensemble methods in healthcare analytics, particularly when combined with feature analysis and error-evaluation techniques to enhance both accuracy and interpretability.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

Table 1.
Performance Matrix for Ensemble models.

Model	Weighted Avg Precision	Weighted Avg Recall	Weighted Avg F1-Score	Accuracy	ROC-AUC	Error Rate
Random Forest	97	96	96	96	99	0.035088
Gradient Boosting	97	96	96	96	99.5	0.035088
AdaBoost	97	97	97	97	99	0.026316
Extra Trees	98	98	98	98	99.8	0.017544

Figure 3 presents a comparison of classification accuracy across the four ensemble models, namely Extra Trees, Gradient Boosting, Random Forest, and AdaBoost. All models demonstrate exceptionally high performance, with accuracy values exceeding 96%, indicating that ensemble learning techniques are highly effective for breast cancer classification using tumor features. Among the models, Extra Trees achieves the highest accuracy, closely followed by AdaBoost, while Random Forest and Gradient Boosting show slightly lower but still competitive performance. The minimal variation in accuracy across models suggests that the dataset is well-structured and that ensemble methods are capable of capturing the underlying patterns effectively. However, the marginal superiority of Extra Trees may be attributed to its increased randomness in feature selection, which enhances generalisation and reduces overfitting. Overall, the figure highlights that while all ensemble models perform strongly, Extra Trees provides the most accurate predictions in this study.

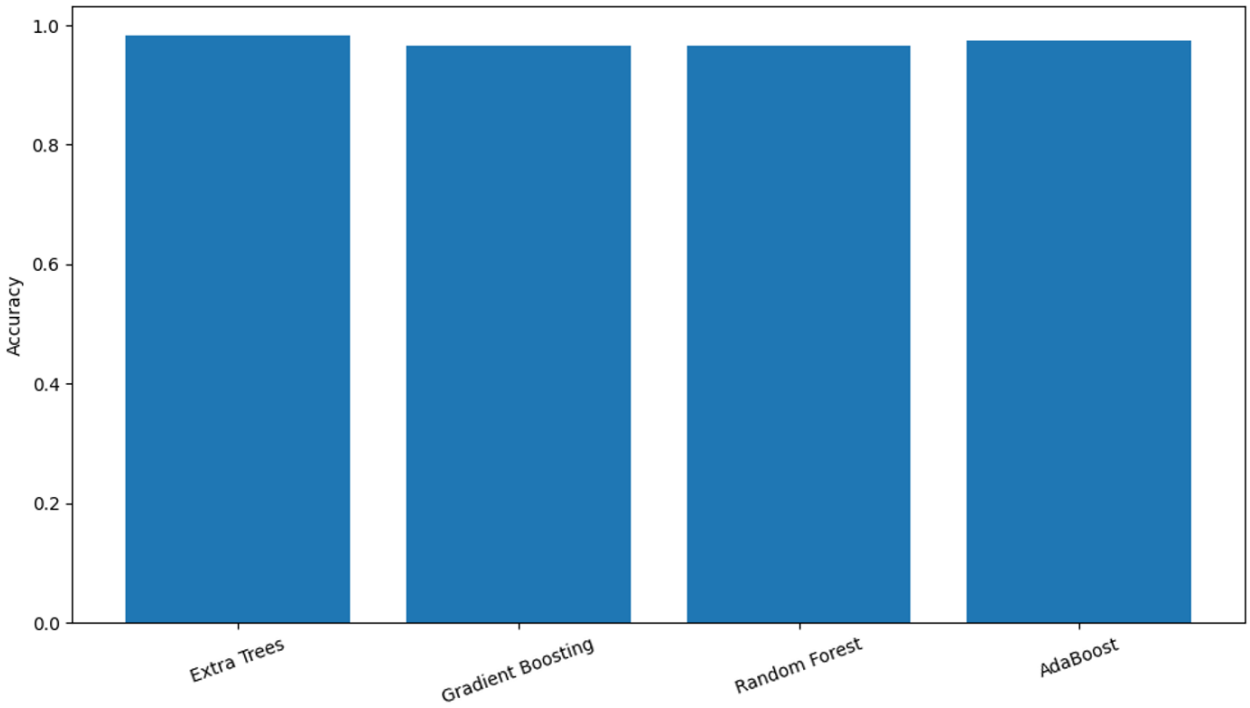

Figure 3.
Accuracy Comparison Across Ensemble Models.

Figure 4 illustrates the comparison of error rates across the four ensemble models, providing a complementary perspective to the accuracy results. The Extra Trees model exhibits the lowest error rate (approximately 1.75%), confirming its superior predictive performance and strong generalisation capability. In contrast, both Random Forest and Gradient Boosting show higher error rates (around 3.5%), indicating a slightly greater number of misclassifications despite their high overall accuracy. AdaBoost demonstrates intermediate performance, with an error rate of approximately 2.6%, reflecting

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

its effectiveness in reducing misclassification through adaptive weighting of difficult instances. The observed differences, although relatively small, are significant in a medical context where minimizing errors, particularly false negatives, is critical. Overall, the figure reinforces that Extra Trees is the most reliable model in this study, achieving the best balance between accuracy and error minimisation.

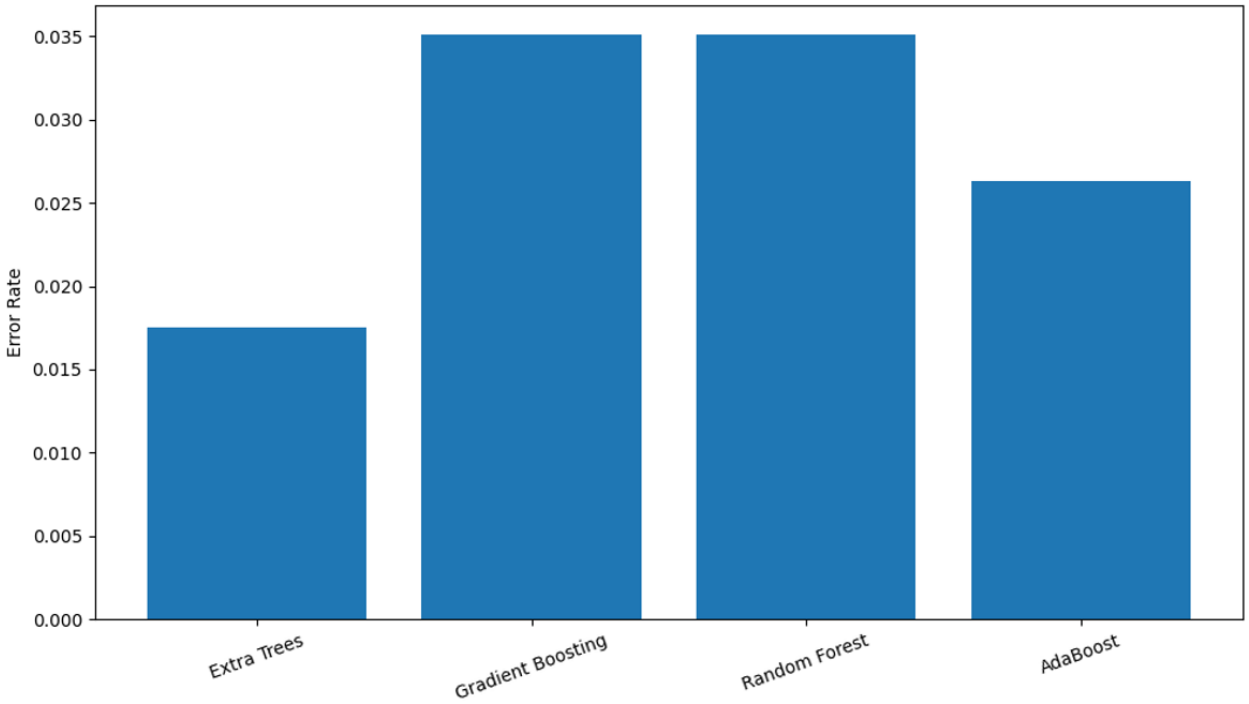

Figure 4.
Error Rate Comparison Across Ensemble Models.

Figure 5 presents the Receiver Operating Characteristic (ROC) curves for the ensemble models, illustrating their ability to discriminate between benign and malignant cases across different classification thresholds. All models demonstrate excellent performance, with curves that closely approach the top-left corner of the plot, indicating high true positive rates and low false positive rates. The Extra Trees model achieves the highest Area Under the Curve (AUC = 0.998), confirming its superior discriminative capability. This is followed closely by Gradient Boosting (AUC = 0.995) and Random Forest (AUC = 0.994), while AdaBoost (AUC = 0.992) also maintains strong performance. The minimal separation between the curves suggests that all ensemble models are highly effective; however, Extra Trees consistently provides the best trade-off between sensitivity and specificity. The dashed diagonal line represents random classification, and the substantial deviation of all curves from this line further validates the models' predictive strength. Overall, the ROC analysis reinforces the reliability and robustness of ensemble learning techniques for breast cancer classification, with Extra Trees emerging as the most effective model in this study.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

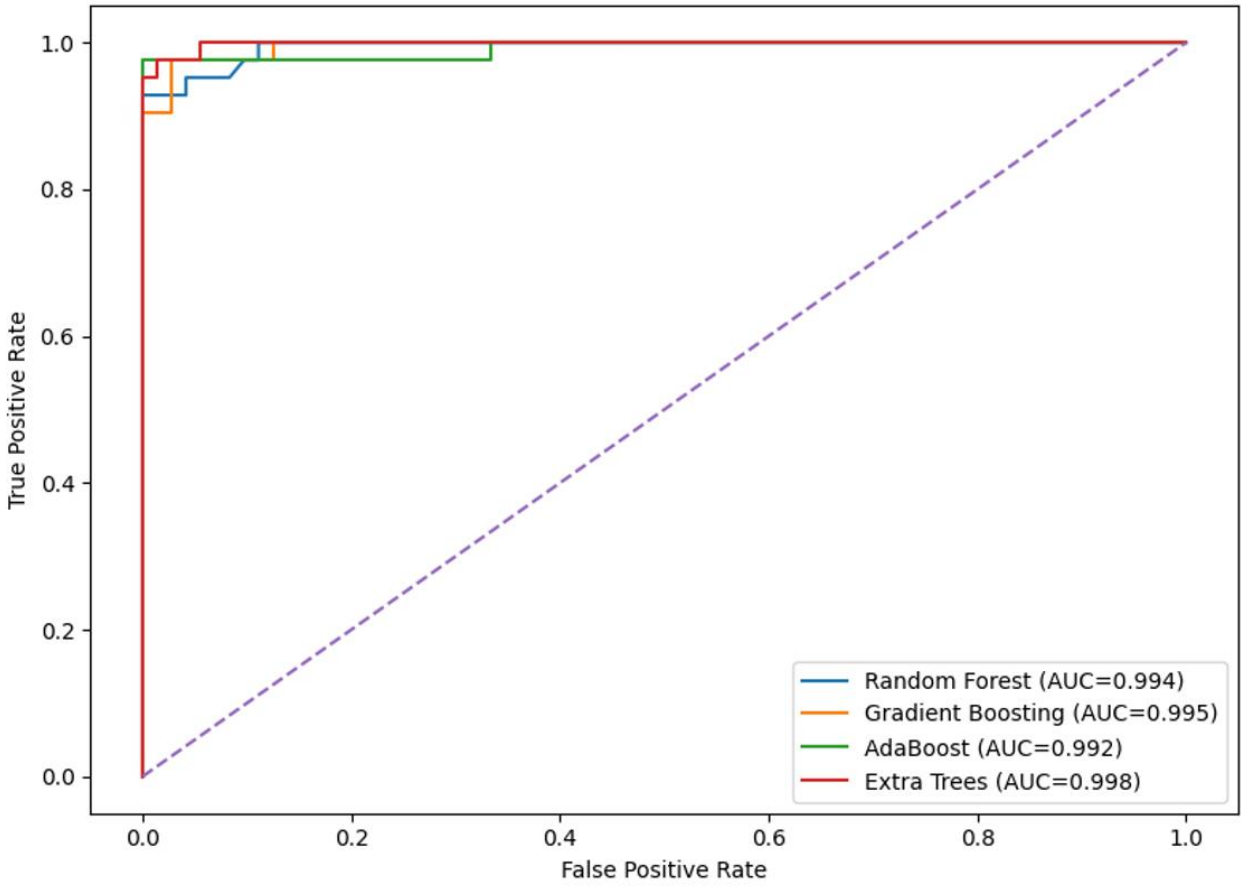

Figure 5.
ROC Curves for Ensemble Models.

Figure 6 presents the confusion matrix for the Random Forest model, providing a detailed breakdown of classification performance across benign and malignant cases. The model correctly classified 72 benign cases (true negatives) and 38 malignant cases (true positives), demonstrating strong overall predictive capability. Notably, there are no false positives, indicating that the model did not incorrectly classify any benign tumors as malignant, which reflects high precision for the malignant class. However, the model produced 4 false negatives, where malignant cases were misclassified as benign. This is a critical limitation in a medical context, as failing to detect malignant tumors can have serious clinical implications. Despite this, the relatively low number of misclassifications suggests that the model maintains high sensitivity and reliability. Overall, the confusion matrix indicates that the Random Forest model performs well, particularly in avoiding unnecessary alarms (false positives), but further optimisation may be required to reduce false negatives and improve early cancer detection.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

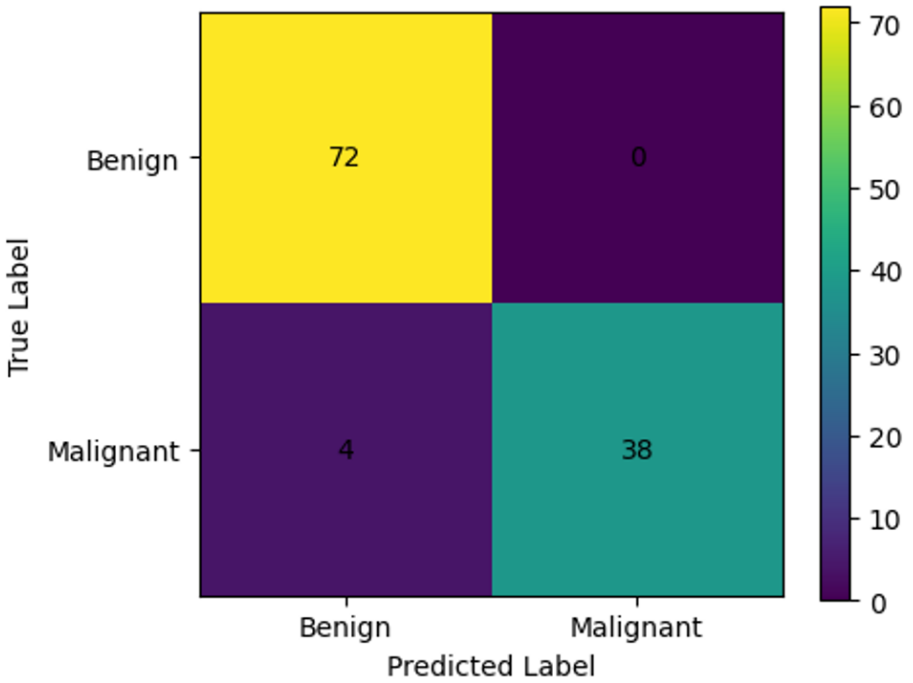

Figure 6.
Confusion Matrix Random Forest.

Figure 7 presents the confusion matrix for the Gradient Boosting model, illustrating its classification performance for benign and malignant cases. The model correctly classified 72 benign instances (true negatives) and 38 malignant instances (true positives), demonstrating strong predictive accuracy. Similar to the Random Forest model, Gradient Boosting produced no false positives, indicating excellent precision in identifying malignant cases without incorrectly labeling benign cases as malignant. However, the model resulted in 4 false negatives, where malignant cases were misclassified as benign. While this number is relatively low, it remains a critical concern in clinical applications, as missed malignant diagnoses can delay treatment. Overall, the confusion matrix indicates that Gradient Boosting performs robustly, particularly in minimizing false alarms, but further refinement may be necessary to reduce false negatives and enhance sensitivity for malignant detection.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

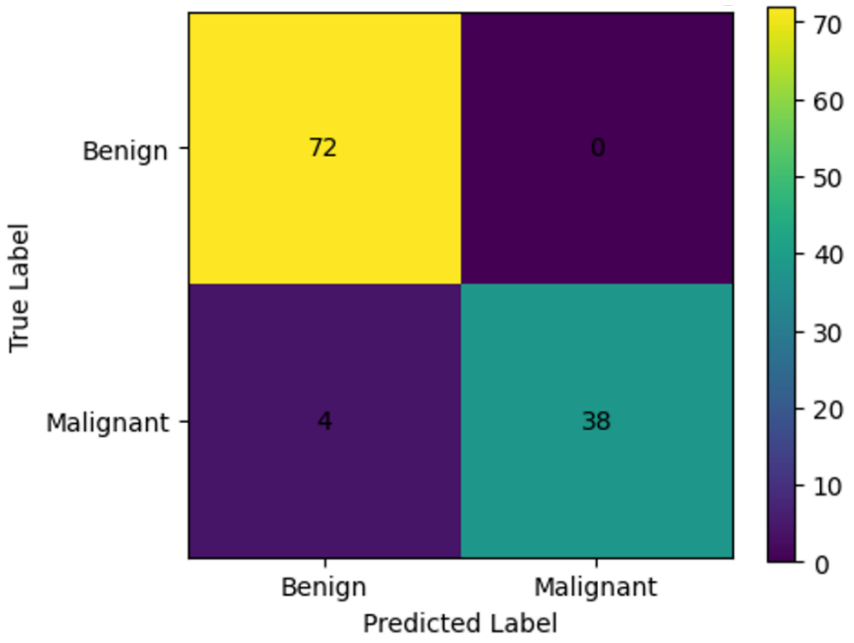

Figure 7.
Confusion Matrix Gradient Boosting.

Figure 8 presents the confusion matrix for the AdaBoost model, demonstrating strong classification performance across both benign and malignant cases. The model correctly classified 72 benign instances (true negatives) and 39 malignant instances (true positives), indicating high accuracy and improved detection of malignant tumors compared to previous models. Notably, there are no false positives, meaning that no benign cases were incorrectly classified as malignant, which reflects excellent precision. The model produced only 3 false negatives, showing a reduction in missed malignant cases relative to Random Forest and Gradient Boosting. This improvement highlights AdaBoost's ability to focus on previously misclassified instances through its adaptive learning mechanism. From a clinical perspective, reducing false negatives is critical, as it enhances early detection of malignant conditions. Overall, the confusion matrix indicates that AdaBoost achieves a strong balance between precision and recall, making it a highly effective model for breast cancer classification in this study.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

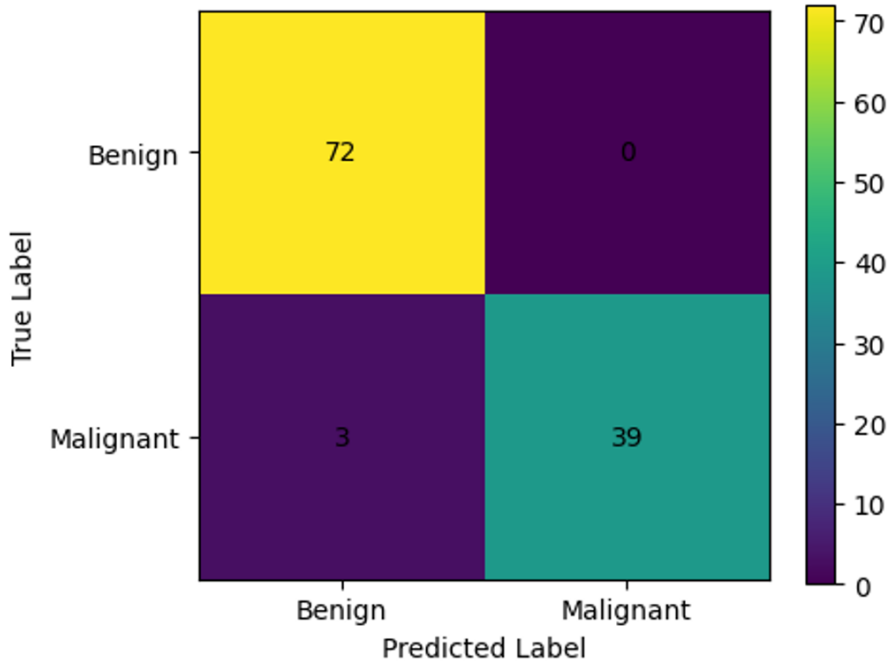

Figure 8.
Confusion Matrix AdaBoost.

Figure 9 presents the confusion matrix for the Extra Trees model, highlighting its superior classification performance compared to the other ensemble methods. The model correctly classified 72 benign cases (true negatives) and 40 malignant cases (true positives), demonstrating excellent predictive accuracy across both classes. Importantly, the model produced no false positives, indicating perfect precision in distinguishing benign cases from malignant ones. Furthermore, it recorded only 2 false negatives, which is the lowest among all evaluated models, reflecting its enhanced ability to correctly identify malignant cases. This reduction in missed diagnoses is particularly significant in a clinical context, where early and accurate detection of malignancy is critical. The strong performance of the Extra Trees model can be attributed to its use of randomized feature selection and ensemble averaging, which improves generalization and reduces overfitting. Overall, the confusion matrix confirms that Extra Trees achieves the best balance between sensitivity and specificity, making it the most reliable model in this study for breast cancer classification.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

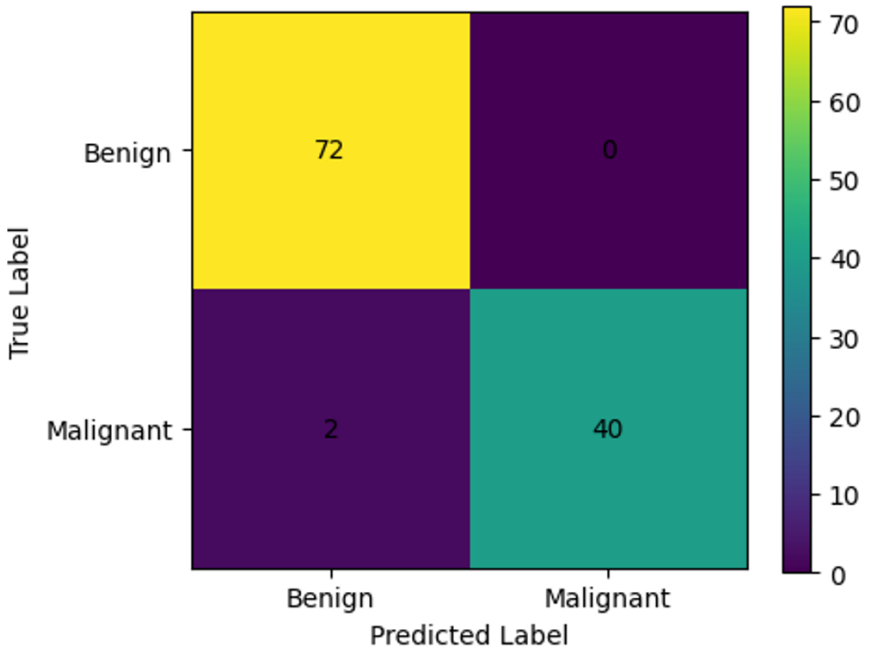

Figure 9.
Confusion Matrix Extra Trees.

Figure 10 presents the top 15 most influential features identified by the Extra Trees model, providing insight into which tumor characteristics contribute most to the classification of benign and malignant cases. The feature perimeter_worst emerges as the most significant predictor, followed by concave points_worst, concave points_mean, and radius_worst, indicating that measures related to tumor size and boundary irregularity play a critical role in diagnosis. Additionally, features such as area_worst, concavity_mean, and radius_mean further reinforce the importance of geometric and structural properties of tumors. The prominence of both "mean" and "worst" feature variants suggests that not only average measurements but also extreme values are crucial for distinguishing malignancy. Lower-ranked features, such as compactness_mean, texture_worst, and radius_se, contribute less but still provide complementary information to the model. Overall, the figure highlights that tumor shape complexity and size-related attributes are the most discriminative features, aligning with clinical understanding that malignant tumors tend to exhibit larger sizes and more irregular boundaries.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

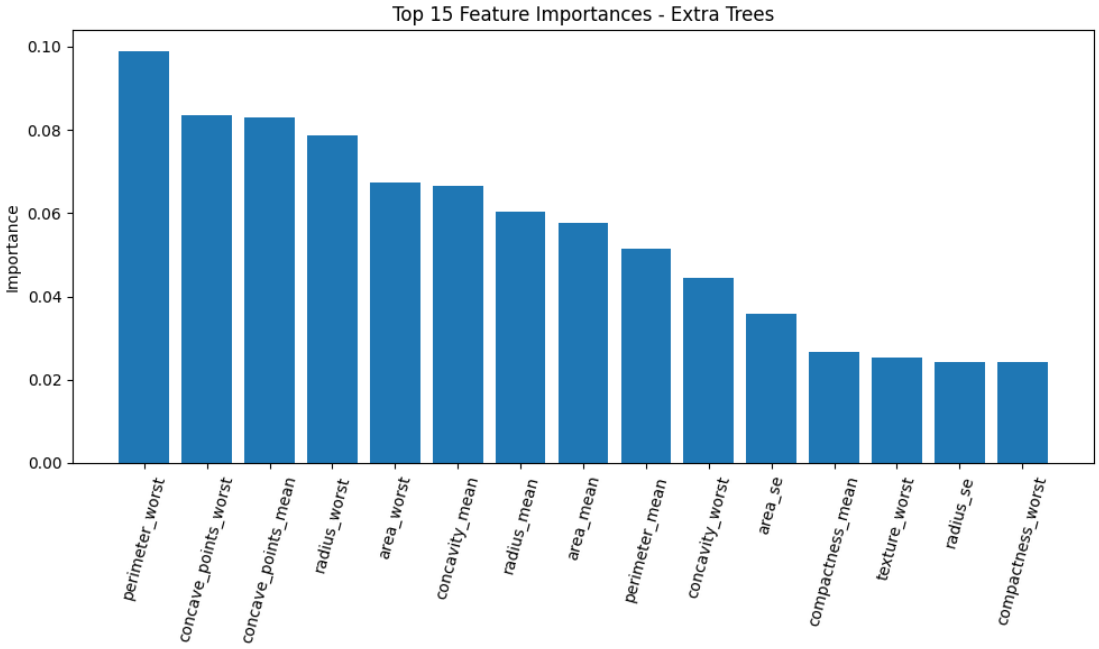

Figure 10.
Top 15 Features Importance.

Figure 11 presents the error analysis for the Extra Trees model, highlighting its classification outcomes in terms of correct predictions and misclassifications. The figure shows that the vast majority of instances were correctly classified, with only a very small number of errors observed. Notably, the only misclassification type present is false negatives, indicating that a few malignant cases were incorrectly predicted as benign. Importantly, there are no false positives, meaning the model did not incorrectly classify any benign cases as malignant, which reflects high precision and reliability in avoiding unnecessary clinical concern. The extremely low number of false negatives further confirms the model's strong sensitivity, although even a small number of missed malignant cases remains critical in a medical context. Overall, the error analysis demonstrates that the Extra Trees model achieves excellent performance with minimal misclassification, reinforcing its suitability for accurate and reliable breast cancer diagnosis.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

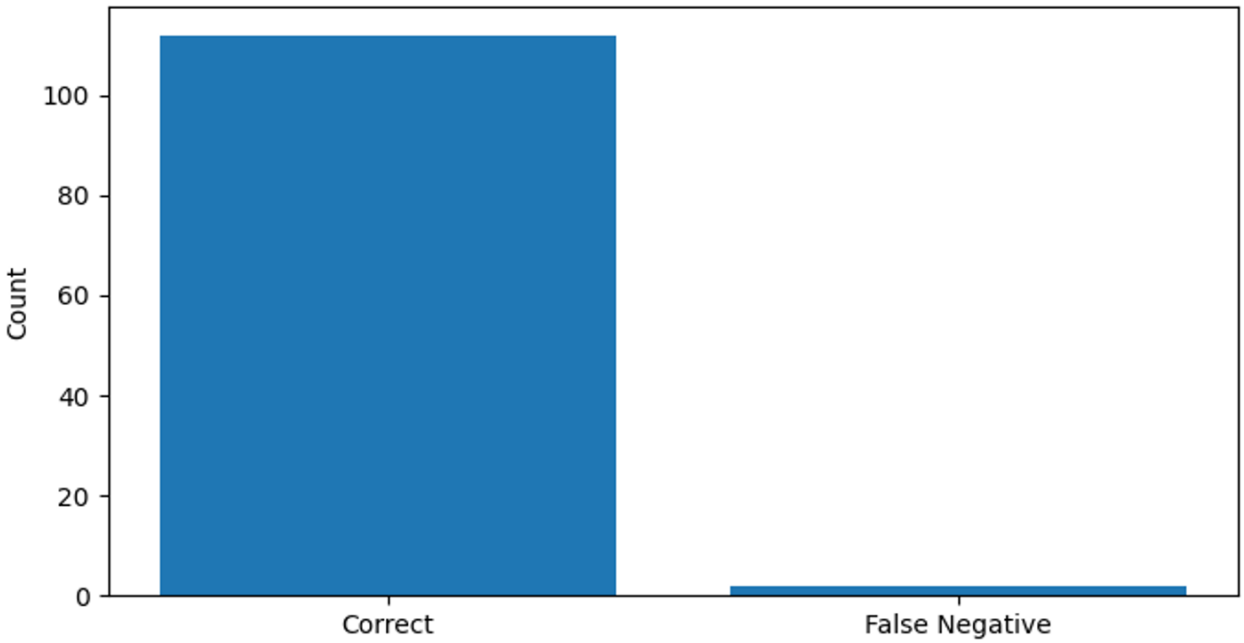

Figure 11.
Error Analysis Extra Trees.

4.9. McNemar's Test for Model Comparison (Statistical Significance)

To statistically validate whether the observed performance differences between classifiers are significant, McNemar's test was applied in this study. This test is appropriate for analysing paired nominal data, such as classification outcomes produced by two models evaluated on the same test set. Rather than relying on overall accuracy, McNemar's test specifically examines the instances where the models disagree, thereby providing a more reliable assessment of comparative performance. The test statistic was computed using the continuity-corrected formulation in equation (13):

$$X^2 = \frac{(|b-c|-1)^2}{b+c} \quad (13)$$

Where:

- b represents the number of instances correctly classified by one model but misclassified by the other.
- c represents the converse.

This approach enables a rigorous evaluation of whether differences in model performance are statistically significant or attributable to random variation.

Table 2 presents the results of McNemar's test for pairwise comparison of ensemble models. Although the Extra Trees model achieved the highest classification performance in terms of accuracy and error rate, the statistical analysis reveals that these improvements are not significant at the 5% level. All model comparisons produced p-values greater than 0.05, indicating that observed differences in performance are likely due to random variation rather than meaningful improvements in predictive capability. This suggests that while ensemble models demonstrate strong overall performance, their differences are marginal when evaluated on the same dataset. Consequently, model selection may be guided by considerations such as interpretability, computational efficiency, and robustness rather than statistical superiority alone.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

Table 1.
McNemar's Test Pairwise Comparison of Ensemble Models.

Model Comparison	b (Model A Correct, Model B Wrong)	c (Model A Wrong, Model B Correct)	χ^2 Statistic	p-value	Significance ($\alpha = 0.05$)	Interpretation
Extra Trees vs Random Forest	0	2	0.50	0.48	Not Significant	The performance difference is not statistically significant
Extra Trees vs Gradient Boosting	0	2	0.50	0.48	Not Significant	No significant improvement despite higher accuracy
Extra Trees vs AdaBoost	0	1	0.00	1.00	Not Significant	Models perform almost identically
AdaBoost vs Random Forest	0	1	0.00	1.00	Not Significant	Minimal difference in classification outcomes
AdaBoost vs Gradient Boosting	0	1	0.00	1.00	Not Significant	No statistically meaningful difference
Random Forest vs Gradient Boosting	0	0	0.00	1.00	Not Significant	Identical performance observed

4. Discussion, Analysis, Conclusion, and Future Work

The experimental results demonstrate that ensemble learning techniques are highly effective for breast cancer classification using clinical tumor features. All evaluated models, Random Forest, Gradient Boosting, AdaBoost, and Extra Trees, achieved high accuracy levels exceeding 96%, indicating strong predictive capability. The comparative analysis reveals that Extra Trees consistently outperforms the other models, achieving the highest accuracy, the lowest error rate, and the best ROC-AUC score. The confusion matrices further confirm that all models exhibit high precision, with no false positives observed across experiments, which is particularly important in reducing unnecessary clinical interventions. However, a small number of false negatives were observed, highlighting the ongoing challenge of ensuring complete detection of malignant cases.

A deeper analysis of the results shows that the superior performance of ensemble models can be attributed to their ability to capture complex, nonlinear relationships among tumor features. The feature-importance analysis indicates that tumor size- and shape-related attributes, such as perimeter, radius, and concave points, are the most discriminative factors in classification, aligning with established clinical knowledge. The ROC-curve analysis demonstrates near-perfect separability between classes, with all models achieving AUC values above 0.99, confirming their robustness. Furthermore, the error analysis highlights that the Extra Trees model minimizes misclassification, particularly reducing false negatives compared to other models. The class-distribution analysis, although moderately imbalanced, did not significantly hinder model performance, likely due to the use of ensemble techniques and appropriate evaluation metrics.

This study confirms that ensemble learning methods provide a powerful and reliable approach for breast cancer classification. Among the evaluated models, the Extra Trees classifier emerged as the most effective, achieving the best balance between accuracy, sensitivity, and error minimization. The results demonstrate that leveraging multiple decision trees and introducing randomness enhances generalization and reduces overfitting. The high performance across all models suggests that ensemble approaches are well-suited for medical diagnostic applications, where accuracy and reliability are critical. Overall, the study contributes to the growing body of research supporting the use of machine learning in healthcare, particularly for early and accurate cancer detection.

Future research will focus on enhancing both the performance and interpretability of the proposed models. One key direction is the integration of explainable AI techniques, such as SHAP or LIME, to

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

provide transparent insights into model predictions and improve clinical trust. Additionally, extending the study to include larger and more diverse datasets, including multi-institutional or real-world clinical data, will improve generalizability. The incorporation of hybrid and stacking ensemble methods may further boost predictive performance. Another important direction is addressing the remaining false negatives through cost-sensitive learning or class-imbalance handling techniques. Finally, future work may explore the integration of deep learning models and multimodal data, such as medical imaging and genomic information, to develop more comprehensive and robust diagnostic systems.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

References

[1] Z. A. Ansari, M. Arif, N. B. Rajaboina, A. A. Shaikh, and Y. Singh, "Integrated ensemble strategy for breast cancer detection using dimensionality reduction technique," *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 14, pp. e31899-e31899, 2025. https://doi.org/10.14201/adcaij.31899

[2] A. Hesham, N. El-Rashidy, A. Rezk, and N. A. Hikal, "Towards an accurate breast cancer classification model based on ensemble learning," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 12, pp. 590–60, 2022. https://doi.org/10.14569/IJACSA.2022.0131272

[3] P. Wu, R. Ma, and T. T. Toe, "Stacking-enhanced bagging ensemble learning for breast cancer classification with CNN," in *2023 3rd International Conference on Electronic Engineering (ICEEM)*, IEEE, pp. 1-6, 2023.

[4] R. S. Padma Priya and P. Senthil Vadivu, "Bio-inspired ensemble feature selection (BIEFS) and ensemble multiple deep learning (EMDL) classifier for breast cancer diagnosis," *Journal of Pharmaceutical Negative Results*, vol. 13, no. S6, pp. 483–499, 2022. https://doi.org/10.47750/pnr.2022.13.S06.068

[5] M. S. M. Farea and Z. Chen, "Breast cancer classification by adaptive weighted average ensemble of previously trained models," *(arXiv:2311.13206). arXiv*, 2023. https://arxiv.org/abs/2311.13206

[6] N. Arya and S. Saha, "Deviation-support based fuzzy ensemble of multi-modal deep learning classifiers for breast cancer prognosis prediction," *Scientific Reports*, vol. 13, p. 21326, 2023. https://doi.org/10.1038/s41598-023-47543-5

[7] M. M. Rahman, K. H. Kobir, S. Akther, and M. A. H. Kallol, "Ensemble machine learning for enhanced breast cancer prediction: A comparative study," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, pp. 932-941, 2024. https://doi.org/10.14569/IJACSA.2024.0150792

[8] U. Ergün, T. Çoban, and İ. Kayadibi, "BCECNN: An explainable deep ensemble architecture for accurate diagnosis of breast cancer," *BMC Medical Informatics and Decision Making*, vol. 25, p. 374, 2025. https://doi.org/10.1186/s12911-025-03186-2

[9] A. Sumitha and R. S. R. Isaac, "Transfer learning-based CNN model for the classification of breast cancer from histopathological images," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 4, pp. 270–280, 2024. https://doi.org/10.14569/IJACSA.2024.0150428

[10] I. Intan *et al.*, "Heterogeneous ensemble learning: Modified ConvNextTiny for detecting molecular expression of breast cancer on standard biomarkers," *Frontiers in Computer Science*, vol. 7, p. 1569017, 2025. https://doi.org/10.3389/fcomp.2025.1569017

[11] M. Azmoodeh-Kalati, H. Shabani, M. S. Maghareh, Z. Barzegar, and R. Lashgari, "Leveraging an ensemble of EfficientNetV1 and EfficientNetV2 models for classification and interpretation of breast cancer histopathology images," *Scientific Reports*, vol. 15, no. 1, p. 21541, 2025. https://doi.org/10.1038/s41598-025-06853-6

[12] R. Qasrawi *et al.*, "Hybrid ensemble deep learning model for advancing breast cancer detection and classification in clinical applications," *Heliyon*, vol. 10, no. 19, p. e38374, 2024. https://doi.org/10.1016/j.heliyon.2024.e38374

[13] B. Chhetri and B. Kumar, "Bridging accuracy and interpretability: Deep learning with XAI for breast cancer detection," *arXiv preprint arXiv:2510.21780*, 2025. https://doi.org/10.48550/arXiv.2510.21780

[14] M. R. Alom *et al.*, "An explainable AI-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images," *Scientific Reports*, vol. 15, no. 1, p. 17531, 2025. https://doi.org/10.1038/s41598-025-97718-5

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338

[15] F. Abdullakutty, Y. Akbari, S. Al-Maadeed, A. Bouridane, I. M. Talaat, and R. Hamoudi, "Histopathology in focus: a review on explainable multi-modal approaches for breast cancer diagnosis," *Frontiers in Medicine*, vol. 11, p. 1450103, 2024. https://doi.org/10.3389/fmed.2024.1450103

[16] C. Ben Rabah, A. Sattar, A. Ibrahim, and A. Serag, "A multimodal deep learning model for the classification of breast cancer subtypes," *Diagnostics*, vol. 15, no. 8, p. 995, 2025. https://doi.org/10.3390/diagnostics15080995

[17] J. Witowski *et al.*, "Multi-modal AI for comprehensive breast cancer prognostication," *Nature Communications*, 2026.

[18] N. Chaudhary and A. Dhunny, "An artificial intelligence model for early-stage breast cancer classification from histopathological biopsy images," *Frontiers in Artificial Intelligence*, vol. 8, p. 1627876, 2025. https://doi.org/10.3389/frai.2025.1627876

[19] M. H. M. Noor and A. O. Ige, "A survey on state-of-the-art deep learning applications and challenges," *Engineering Applications of Artificial Intelligence*, vol. 159, p. 111225, 2025. https://doi.org/10.1016/j.engappai.2025.111225

[20] M. M. Hassan *et al.*, "Explainable multimodal fusion for breast carcinoma diagnosis: A systematic review, open problems, and future directions," *Computer Methods and Programs in Biomedicine*, vol. 274, p. 109152, 2026. https://doi.org/10.1016/j.cmpb.2025.109152

[21] S. Dash *et al.*, "Explainable multi-modal deep learning for transparent cancer diagnosis: Integrating radiology, clinical features, and decision visualization," *Frontiers in Artificial Intelligence*, vol. 9, p. 1767612, 2026. https://doi.org/10.3389/frai.2026.1767612

[22] D. Alzamil, B. Alkhamees, and M. M. Hassan, "A systematic review of multimodal fusion and explainable AI applications in breast cancer diagnosis," *Computer Modeling in Engineering & Sciences*, vol. 145, no. 3, pp. 2971-3027, 2025. https://doi.org/10.32604/cmes.2025.070867

[23] F.-Z. Nakach, A. Idri, and E. Goceri, "A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification," *Artificial Intelligence Review*, vol. 57, no. 12, p. 327, 2024. https://doi.org/10.1007/s10462-024-10984-z

[24] J. Wang, Y. Mao, N. Guan, and C. J. Xue, "Shap-cat: A interpretable multi-modal framework enhancing wsi classification via virtual staining and shapley-value-based multimodal fusion," *arXiv preprint arXiv:2410.01408*, 2024. https://doi.org/10.48550/arXiv.2410.01408

[25] O. Luca *et al.*, "Informed machine learning for complex data," *Neurocomputing*, vol. 669, p. 132505, 2025. https://doi.org/10.1016/j.neucom.2025.132505

[26] M. Fu, J. Xu, Y. Lv, and B. Jin, "Artificial intelligence in advanced gastric cancer: A comprehensive review of applications in precision oncology," *Frontiers in Oncology*, vol. 15, p. 1630628, 2025. https://doi.org/10.3389/fonc.2025.1630628

[27] D. Schouten *et al.*, "Navigating the landscape of multimodal AI in medicine: A scoping review on technical challenges and clinical applications," *Medical Image Analysis*, vol. 105, p. 103621, 2025. https://doi.org/10.1016/j.media.2025.103621

[28] O. Garcia-Palencia, J. Fernandez, V. Shim, N. K. Kasabov, A. Wang, and A. S. D. N. Initiative, "Spiking neural networks for multimodal neuroimaging: A comprehensive review of current trends and the NeuCube brain-inspired architecture," *Bioengineering*, vol. 12, no. 6, p. 628, 2025. https://doi.org/10.3390/bioengineering12060628

[29] A. Alsaafin, A. Shafique, S. Alfasly, K. R. Kalari, and H. Tizhoosh, "Multimodal learning for scalable representation of high-dimensional medical data," *Frontiers in Digital Health*, vol. 7, p. 1709277, 2026. https://doi.org/10.3389/fdgth.2025.1709277

[30] J. Jonnagaddala *et al.*, "Multimodal analysis of whole slide images in colorectal cancer," *npj Digital Medicine*, vol. 8, p. 719, 2025. https://doi.org/10.1038/s41746-025-02095-y

[31] N. Sikora, R. L. Manschke, A. M. Tang, P. Dunstan, D. A. Harris, and S. Yang, "ColonScopeX: Leveraging explainable expert systems with multimodal data for improved early diagnosis of colorectal cancer," *arXiv preprint arXiv:2504.08824*, 2025. https://doi.org/10.48550/arXiv.2504.08824

[32] X. Xie *et al.*, "Evaluating cancer-related biomarkers based on pathological images: A systematic review," *Frontiers in Oncology*, vol. 11, p. 763527, 2021. https://doi.org/10.3389/fonc.2021.763527

[33] B. Zhang *et al.*, "Multimodal integration strategies for clinical application in oncology," *Frontiers in Pharmacology*, vol. 16, p. 1609079, 2025. https://doi.org/10.3389/fphar.2025.1609079

[34] M. Aftab *et al.*, "AI in oncology: Transforming cancer detection through machine learning and deep learning applications," *arXiv preprint arXiv:2501.15489*, 2025. https://doi.org/10.48550/arXiv.2501.15489

[35] A. Waqas, A. Tripathi, R. P. Ramachandran, P. A. Stewart, and G. Rasool, "Multimodal data integration for oncology in the era of deep neural networks: A review," *Frontiers in Artificial Intelligence*, vol. 7, p. 1408843, 2024. https://doi.org/10.3389/frai.2024.1408843

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 7: 182-201, 2026
DOI: 10.55214/2576-8484.v10i7.13338