

Explainable AI for Breast Cancer Diagnosis Using Machine Learning

By Sravan Kumar Reddy Perugu

Abstract

Breast cancer remains one of the most prevalent and life-threatening diseases affecting women globally. Early and accurate diagnosis is essential to improve treatment outcomes and patient survival rates. In recent years, machine learning techniques have been increasingly applied to assist clinicians in diagnosing breast cancer due to their ability to analyze complex medical data efficiently. However, many high-performing models operate as black boxes, offering limited transparency into how predictions are made. This lack of interpretability hinders clinical trust and real-world adoption. This project proposes an Explainable Artificial Intelligence (XAI) framework for breast cancer diagnosis using Logistic Regression, Random Forest, and XGBoost models. To ensure transparency, SHAP is used for global feature importance analysis, while LIME provides patient-specific explanations. The results demonstrate that the proposed framework achieves high predictive accuracy while maintaining interpretability, making it suitable as a clinical decision-support system.

Dataset Description

The Wisconsin Breast Cancer Diagnostic Dataset is a widely used benchmark dataset in medical machine learning research. It consists of 569 patient records, each described by 30 numerical features derived from digitized images of fine-needle aspirates of breast masses. These features capture detailed characteristics of cell nuclei, including radius, texture, perimeter, area, concavity, concave points, smoothness, symmetry, and fractal dimension. Each instance is labeled as either malignant or benign, allowing the dataset to be used for binary classification. The dataset is well-balanced and free of missing values, making it suitable for reliable experimentation and evaluation.

Logistic Regression Results

Logistic Regression is used as a baseline model due to its simplicity, computational efficiency, and inherent interpretability. As a linear classifier, it establishes a clear relationship between input features and predicted outcomes, making it suitable for comparison with more complex models.

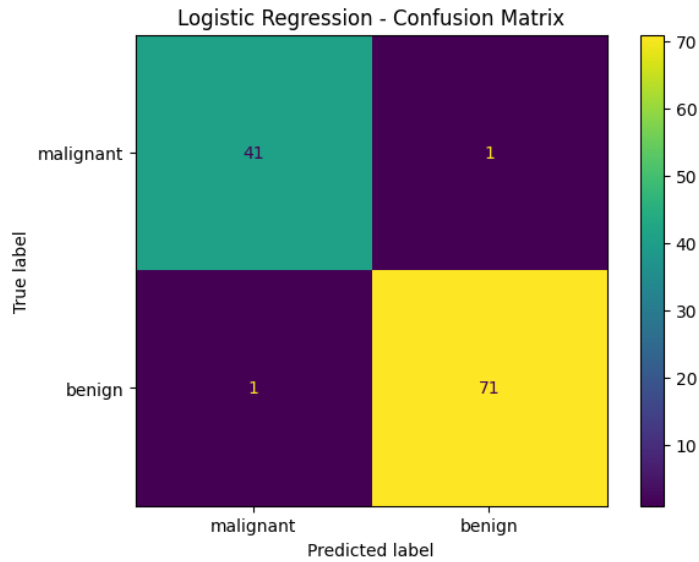


Figure 1: Logistic Regression Confusion Matrix

The confusion matrix indicates that the model correctly classifies the majority of both malignant and benign cases, with only a very small number of misclassifications. This demonstrates that Logistic Regression provides strong baseline performance despite its simplicity.

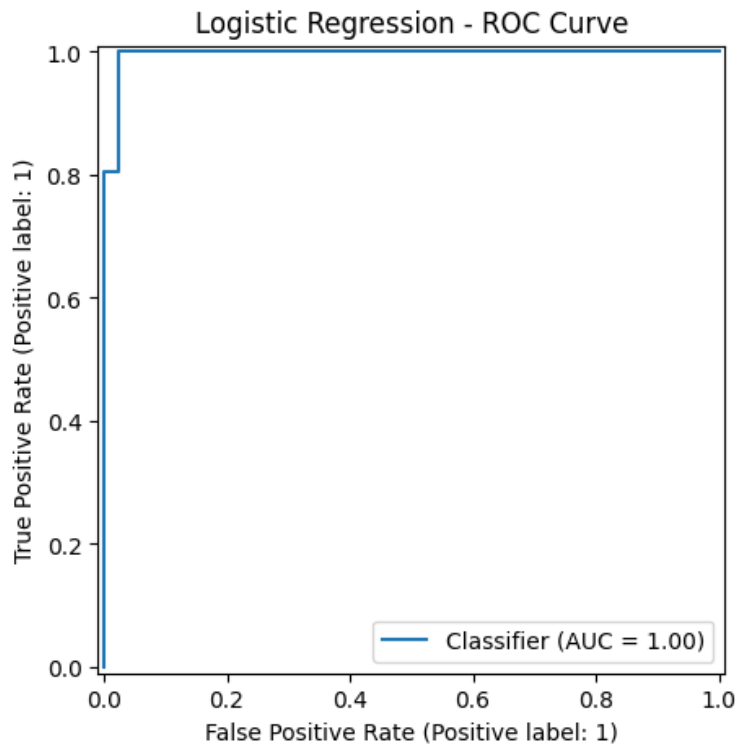


Figure 2: Logistic Regression ROC Curve

The ROC curve exhibits near-perfect class separation with an AUC close to 1.0, indicating that the model effectively distinguishes between malignant and benign tumors across different decision thresholds.

Random Forest Results

Random Forest is an ensemble learning algorithm that constructs multiple decision trees and aggregates their predictions to improve robustness and generalization. It is capable of capturing non-linear relationships and complex feature interactions that may not be modeled by linear classifiers.

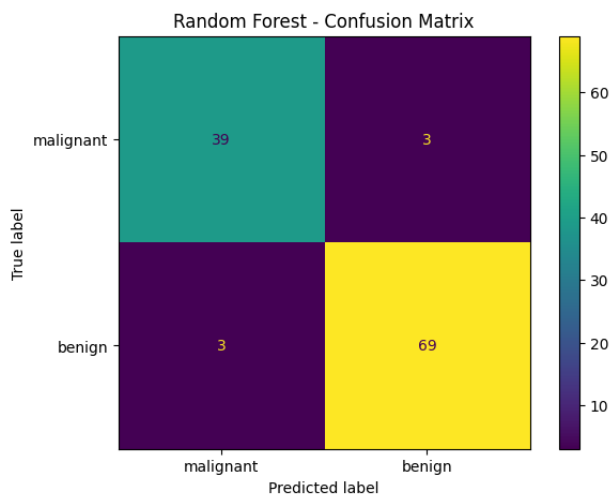


Figure 3: Random Forest Confusion Matrix

The confusion matrix shows high classification accuracy with slightly more misclassifications than Logistic Regression. This behavior is expected due to the model’s sensitivity to feature interactions.

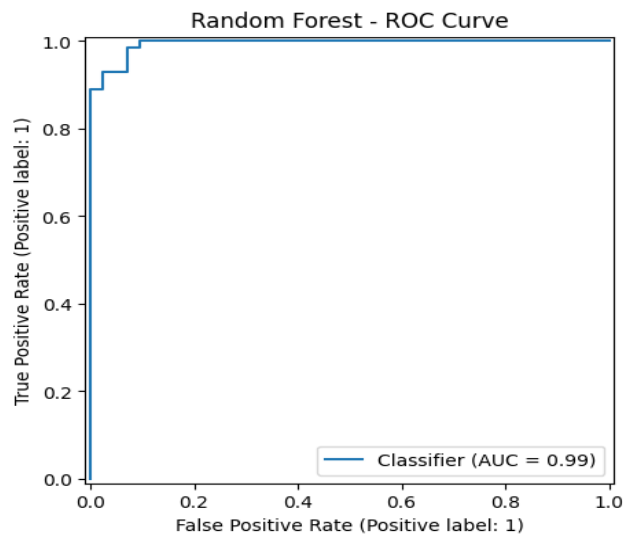


Figure 4: Random Forest ROC Curve

The ROC curve achieves an AUC of approximately 0.99, demonstrating strong discriminative power and confirming the effectiveness of the ensemble approach.

XGBoost Results

XGBoost is a powerful gradient boosting algorithm designed for high performance and efficiency. It sequentially builds decision trees to correct previous errors, enabling it to learn complex patterns within the data and often outperform traditional ensemble methods.

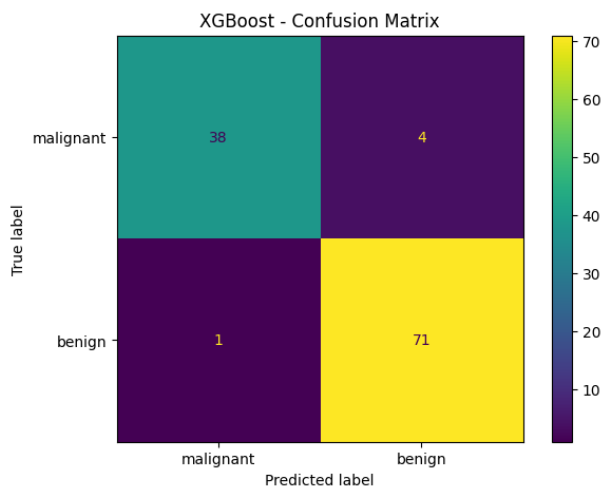


Figure 5: XGBoost Confusion Matrix

The confusion matrix demonstrates that XGBoost produces the most balanced classification results, minimizing both false positives and false negatives.

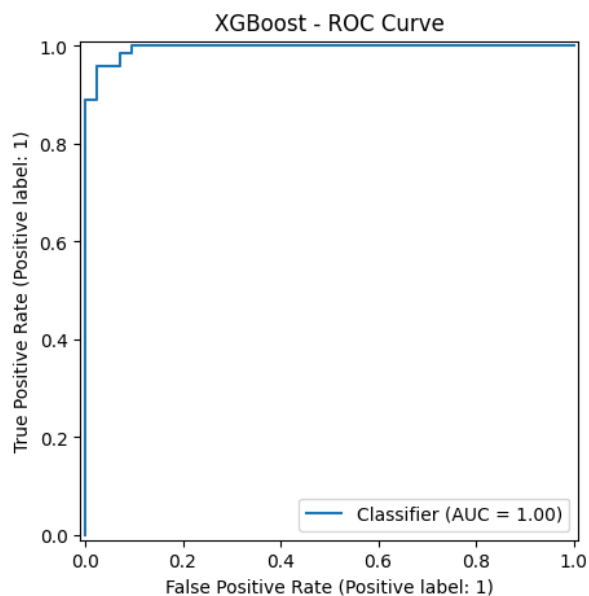


Figure 6: XGBoost ROC Curve

The ROC curve shows an AUC close to 1.0, indicating excellent classification performance and superior generalization capability compared to other models.

SHAP Explainability Analysis

SHAP (SHapley Additive Explanations) provides a global understanding of model behavior by assigning each feature a contribution value toward the final prediction. This approach ensures consistent and theoretically sound explanations based on cooperative game theory.

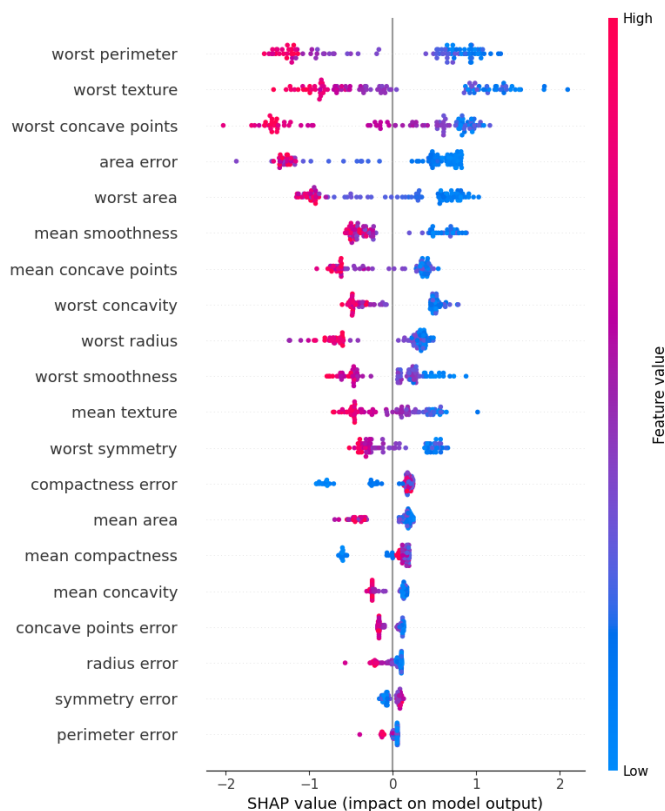


Figure 7: SHAP Summary Plot (Named Features)

The SHAP summary plot reveals that features such as worst perimeter, worst texture, worst concave points, and worst area have the strongest influence on malignancy predictions. High feature values consistently push predictions toward malignant outcomes, aligning with established clinical knowledge.

LIME Local Explanation

LIME (Local Interpretable Model-Agnostic Explanations) focuses on explaining individual predictions by approximating the complex model locally with a simpler interpretable model.

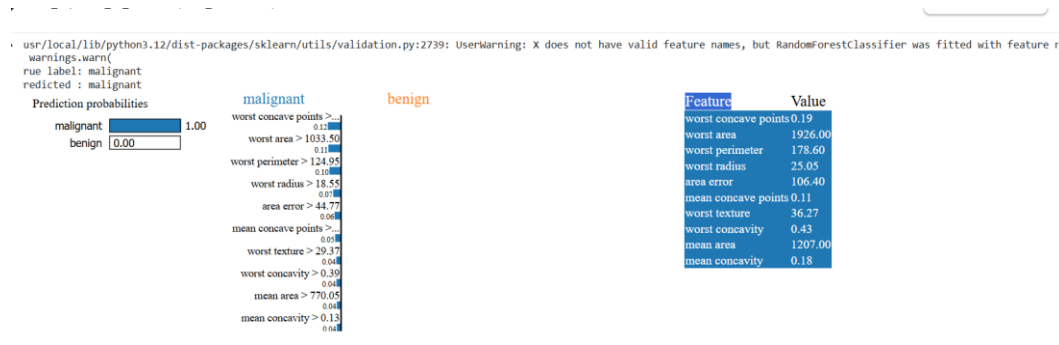


Figure 8: LIME Explanation for a Single Patient

The LIME visualization highlights the most influential features contributing to a single patient's prediction. This local explanation enhances clinician understanding and supports case-level decision making.

Conclusion

This project demonstrates that integrating explainable AI techniques with machine learning models enables accurate, transparent, and clinically meaningful breast cancer diagnosis. While XGBoost achieved the highest predictive performance, SHAP and LIME ensured interpretability at both global and local levels. The alignment of model explanations with known medical indicators strengthens clinical trust and highlights the potential of the proposed framework as a reliable decision-support tool. Future work may focus on integrating imaging data, addressing model bias, and deploying the system in real-world healthcare settings.