Breast Cancer Prediction with Transparent Decision Trees

Farheen Khan
Guru Nanak Khalsa College of Arts, Science & Commerce, Mumbai, Maharashtra, India

Bhagyashree Bonde
Associate Professor, Guru Nanak Khalsa College of Arts, Science & Commerce, Mumbai, Maharashtra, India

Abstract:

Breast cancer is a leading cause of mortality worldwide, and early, accurate diagnosis is critical for improving patient outcomes. This research addresses the dual challenge of predictive performance and interpretability by applying machine learning techniques to the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The proposed approach leverages the Multi surface Method-Tree (MSM-T), which integrates linear programming discrimination within a decision-tree framework. Unlike conventional black-box models, MSM-T produces transparent, mathematically grounded decision rules that clinicians can interpret and validate, ensuring trust and accountability in diagnostic workflows.

The study implements an end-to-end pipeline comprising data preprocessing, feature engineering, and feature selection using filter, wrapper, and metaheuristic methods such as Modified Bat Algorithm and Particle Swarm Optimization. Model training employs nested cross-validation for unbiased performance estimation, with evaluation metrics including accuracy, precision, recall, F1-score, ROC-AUC, and calibration measures. Explainability is further enhanced through SHAP-based interpretation, providing global and local feature attributions to support clinical decision-making.

Preliminary findings indicate that MSM-T achieves robust classification accuracy while maintaining interpretability, outperforming traditional models in transparency without sacrificing diagnostic reliability. Key predictors such as worst radius and mean concave points demonstrate strong clinical relevance, aligning with established pathological indicators. The research contributes a reproducible methodological framework for integrating interpretable machine learning into healthcare, addressing ethical considerations such as fairness, privacy, and bias mitigation. Future work will focus on scaling the approach to larger, heterogeneous datasets and exploring hybrid models that combine MSM-T with ensemble techniques. By balancing technical rigor with clinical applicability, this study underscores the potential of transparent decision trees as a cornerstone for trustworthy AI in medical diagnostics.