

AS/OP/05

Interpretable Breast Cancer Detection Using Shannon Entropy-Based Attribute Selection

Subawickrama HDAW¹, Udagedara UGIGK², Nishantha SAA³, Abeysundara SP⁴

¹*Department of Physical Sciences, Faculty of Applied Sciences, Rajarata University of Sri Lanka*

²*Lewis University, IL, United States*

³*Clarkson University, NY, United States*

⁴*Department of Statistics & Computer Science, Faculty of Science, University of Peradeniya*

Background: Breast cancer is a leading cause of death among women worldwide, with over 2.3 million new cases and 600,000 deaths reported annually. In Sri Lanka, nearly 3,000 new cases each year highlight the need for accurate and interpretable diagnostic methods.

Objective: This study aims to improve breast cancer detection by proposing a Shannon entropy-based attribute selection technique. Shannon entropy, a fundamental concept from information theory, quantifies the amount of uncertainty or information in a dataset. Since the Wisconsin Diagnostic Breast Cancer (WDBC) dataset (569 instances) consists of continuous features, attributes were discretized using an equal-width binning rule with 5 bins before calculating entropy and information gain. Information gain values were then used to rank the features, and the optimal subset was determined by selecting 13 attributes at the intersection point of the bar plot of information gain and the cumulative gain curve, ensuring a balanced trade-off between dimensionality and relevance. Using these selected features, three classification models Support Vector Machine (SVM), Logistic Regression, and Random Forest were evaluated.

Results: SVM achieved the highest accuracy (94.14%), followed by Logistic Regression (92.5%) and Random Forest (90.83%). SVM correctly identified 87 out of 90 benign and 26 out of 30 malignant cases. SVM outperformed the others due to its ability to exploit the clearer decision boundaries created by entropy-based selection. It also benefits from kernel functions that capture non-linear patterns, unlike Logistic Regression's linear assumptions. Random Forest may suffer from reduced tree diversity when fewer features are available, limiting its performance.

Conclusion: Overall, the proposed entropy-guided SVM model demonstrates a powerful, high-performing, and interpretable solution for enhancing diagnostic precision in real-world breast cancer detection tasks.

Keywords: Attribute selection, Breast cancer detection, Shannon entropy