

AN ENTROPY-BASED ATTRIBUTE SELECTION TECHNIQUE FOR INTERPRETABLE BREAST CANCER DETECTION

H.D.A.W. Subawickrama^{1*}, U.G.I.G.K. Udagedara², S.A.A. Nishantha³ and S.P. Abeysundara⁴

¹*Department of Physical Science, Rajarata University of Sri Lanka, Mihintale, Sri Lanka.*

²*Department of Engineering, Computing and Mathematical Sciences, Lewis University, Romeoville, IL, United States.*

³*Department of Mathematics and Computer Sciences, University of St. Francis, Joliet IL, United States.*

⁴*Department of Statistics & Computer Science, University of Peradeniya, Peradeniya, Sri Lanka.*

*amandisubawickrama586@gmail.com

Breast cancer is one of the most common and life-threatening cancers among women worldwide, with approximately 2.3 million new cases and over 600,000 deaths reported annually, according to the World Health Organisation. In Sri Lanka alone, nearly 3,000 new cases are diagnosed each year, underscoring the urgent need for reliable and efficient diagnostic techniques. Recently, data-driven methods have shown great potential in improving diagnostic accuracy and reducing subjectivity. This study introduces a Shannon entropy-based attribute selection approach for breast cancer detection using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset of 569 instances. Unlike Principal Component Analysis (PCA), which transforms data into abstract components, the proposed method directly identifies informative attributes without altering the original data. Using the selected 13 attributes, a Support Vector Machine (SVM) classifier achieved 93.00% accuracy, with a precision of 0.95 (benign) and 0.86 (malignant), and a recall of 0.96 and 0.83, respectively. For comparison, PCA, with 13 principal components followed by SVM, yielded a slightly higher accuracy of 95.83%, but at the cost of interpretability, as PCA-derived features lack clinical relevance. When the number of components was chosen via scree plot ($k = 2$), accuracy further decreased to 91.67%. Unlike PCA, the proposed entropy-based approach retains clinically meaningful attributes, such as mean radius and texture, making the results more interpretable for medical professionals. Additionally, it demonstrated greater computational efficiency by avoiding the matrix decomposition required in PCA. These findings suggest that entropy-based feature selection provides a more interpretable, efficient, and clinically relevant alternative to transformation-based methods for breast cancer detection. Future work will involve testing the method on large-scale, real-time datasets to assess scalability and practical applicability in healthcare.

Keywords: Attribute selection, Breast cancer detection, Classification, Principal component analysis, Shannon entropy, Support vector machine