ABSTRACT

Breast cancer is one of the most common types of cancer affecting the mammary gland tissue and remains a leading cause of death among women worldwide. Determining the cancer stage is crucial for selecting the appropriate treatment strategy. This study aims to classify breast cancer stages using the Random Forest algorithm and to identify the most influential clinical features in the classification process. The dataset used is the *SEER Breast Cancer Data – Labeled*, published by the National Cancer Institute (NCI) and available on the Kaggle platform. It consists of 4,024 female patient records with complete clinical information, including cancer size, lymph node status, and cell differentiation grade. After preprocessing and model training, the Random Forest algorithm achieved a high classification performance with an accuracy of 77,01% on a test set of 805 entries. The model successfully classified all four cancer stages, showing excellent performance for stages 0 and 1, and satisfactory results for stage 2 and 3. This study demonstrates that the Random Forest algorithm is effective as a data-driven diagnostic support tool for automatic classification of breast cancer stages.

Keywords: Breast Cancer, Random Forest, Cancer Staging, Machine Learning, Feature Importance, Classification, SEER Dataset.