ABSTRACK

Breast cancer is one of the most common cancers and the leading cause of death for women in the world. Many patients realize their condition when it is already at an advanced stage, so early detection is very important as an early warning and for strengthening diagnosis and treatment effectiveness. This study aims to apply machine learning with Random Forest algorithm for breast cancer prediction and to identify influential factors in the prediction. The data used in this study comes from medical records at Al-Ihsan Hospital which consists of two classes, namely Ca Mammae and Ca Mammae + Comorbidities. There is an imbalance of data between the two classes, namely 85.62% (941 data) is Ca Mammae and 14.38% (158 data) is Ca Mammae + Comorbidities. This data imbalance was addressed with random undersampling, random oversampling, and SMOTE to ensure a more optimal model. The Random Forest model building process is applied to three data splitting ratios, namely 70:30, 80:20, and 90:10 and grid search to determine the optimal model. Model evaluation was conducted using confusion matrix, showing that the optimal model with 90:10 data split and handling imbalance using SMOTE with parameters criterion=“entropy”, max_depth=20, min_samples_leaf=1, min_samples_split=2, n_estimators=200. The results show that Random Forest set with optimal parameters can provide good performance in breast cancer prediction, giving it the potential to be used as a tool in clinical diagnosis and medical decision making. This model achieved an accuracy of 87,27% with a precision of 64,71%, recall of 57,89%, f1-score of 61,02%, and the number of duplicate data of 145. Based on the results of the best model, it shows that the factors that influence the increase in breast cancer include Platelets, Hemoglobin, AST (SGOT), Ureum, and ALT (SGPT). This research is expected to be useful as an early warning to increase awareness of early detection of breast cancer.

*Keywords—***breast cancer, data balancing, machine learning, medical records, Random Forest algorithm**