Abstract

Lung Cancer is the leading cause of cancer-related death worldwide and has a significant socioeconomic impact on patients, families, and society as a whole. In the diagnosis of cancer, the classification of different types of tumors is very important. Accurate prediction of different types of tumors allows for better treatment and minimizes patient toxicity. To analyze cancer classification problems using gene expression data, for feature selection and predictive models. This study aims to predict NSCLC by applying the ensemble method to microarray data. The author uses three ensemble methods to predict NSCLC, namely Random Forest, Adaptive Boosting (AB), and Extreme Gradient Boosting (XG). Feature selection is carried out using the variance threshold and chi-square parameter and then continued by building a prediction model using an ensemble. The result of the best model validation based on those diagnosed with cancer are the AB model with 10 features, XG with 10 features, and XG with 20 features that produce the same accuracy, recall, and f1-score values, namely 0.93, 1.00, and 0.93 respectively.

Keywords: chi-square, ensemble, microarray, variance threshold