

Abstract

Lung cancer is the most common cause of death in someone who has cancer. This happens because of remembering the importance of lung function as a breathing apparatus and oxygen distribution throughout the body. Early identification of lung cancer is crucial to reduce its mortality rate. Accuracy is crucial since it indicates how accurately the model or system makes the right predictions. High levels of accuracy show that the model can produce trustworthy and accurate findings, essential for making effective decisions based on available data. In this research, ensemble learning approaches, namely bagging and boosting methods, were employed for classifying lung cancer. Hyperparameters, a class of parameters, are crucial to this model's effectiveness. In order to increase the lung cancer classification model's accuracy, a thorough investigation was conducted to identify the best hyperparameter combination. In this study, the dataset used is a medical dataset that contains a history of patients who have been diagnosed with lung cancer or not. The dataset is taken from Kaggle mysarahmadbhat and cancerdatahp from data world. To evaluate the model's accuracy, this study used the confusion matrix method which compares the model's prediction results with the ground truth. the study findings revealed that employing a dataset split ratio of 70:30 produced the best results, with the Random Forest, CatBoost, and XGBoost models achieving an impressive 98% accuracy, 0.98 precision, 0.98 recall, and 0.98 f1-score. but for AdaBoost, the best results were obtained on a dataset with a ratio of 80:20 with an accuracy of 96%, 0.97 precision, 0.96 recall, and 0.96 f1-score. Optimizing the selection of appropriate parameters and using a proportional division of training data can increase the effectiveness of the model in classifying lung cancer.