Identification of Predictors Helping the Diagnosis of Lung Cancer Using Machine Learning Models and Feature Selection Methods

Suejit Pechprasarn[1*], Nichapha Suechoey[2], Nutchareeya Pholtrakoolwong[2], Pattaraporn Tanedvorapinyo[2] and Yanisa Toboonliang[2]

[1]College of Biomedical Engineering, Rangsit University, Pathumthani, Thailand
[2]Satriwithaya School, Wat Bowon Niwet, Phra Nakhon, Bangkok, Thailand
The last two authors have contributed equally.
*Corresponding author; E-mail: suejit.p@rsu.ac.th

Abstract

As of 2023, approximately 238,000 individuals have been diagnosed with lung cancer, making it one of the most prevalent forms of cancer. Further, the number of new cases of lung cancer continues to rise steadily. This study employs clinical predictors with a dataset of 309 observances and 15 predictors to aid in the diagnosis of lung cancer, including (1) Swallowing Difficulty, (2) Peer Pressure, (3) Gender, (4) Allergy, (5) Yellow Fingers, (6) Anxiety, (7) Wheezing, (8) Alcohol Consuming, (9) Chronic Disease, (10) Chest Pain, (11) Coughing, (12) Fatigue, (13) Smoking, (14) Age, and (15) Shortness of Breath. This study aims to train, compare, and evaluate the performance of several types of machine learning models as well as identify the critical factors used to predict lung cancer from the 15 specified clinical features using unsupervised feature selection methods. By dividing the dataset into training and test datasets and preprocessing the dataset for unbiased training, it is possible to utilize this dataset to test several machine learning models and assess their accuracy in diagnosing lung cancer. Subsequently, machine learning models and feature selection can be utilized to establish the essential predictors for a malignant diagnosis. Next, the authors examine and contrast discriminant, logistic regression, naive Bayes, support vector machine (SVM), K-nearest neighbor (KNN), ensemble, neural network, and kernel. It was discovered that a Gaussian Naïve Bayes machine learning model, which only needs nine variables - Swallowing Difficulty, Peer Pressure, Gender, Allergy, Yellow Fingers, Anxiety, Wheezing, Alcohol Consumption, and Chronic Disease - could obtain the highest cross-validation classification accuracy of 82.81%. Thus, this study highlights the feasibility of model complexity reduction from 15 predictors to 9 predictors without losing classification capability.

Keywords**:** *Lung Cancer Classification, Machine Learning, Smart Diagnosis, Feature Selection, Artificial Intelligence*