Lung Cancer Classification with XGBoost, SVM, Random Forest, and k-Nearest Neighbor

Dr. Mitat Uysal
Software Engineering Department, Dogus University, Istanbul, Türkiye

Dr. Aynur Uysal
Software Engineering Department, Dogus University, Istanbul, Türkiye

Abstract

Early identification of lung cancer risk from tabular clinical measurements can support decision-making by prioritizing patients for imaging and further diagnostic workflows. This paper presents a comparative study of four widely used classifiers—k-Nearest Neighbor (kNN), Support Vector Machines (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost)—for lung-cancer classification. We summarize the mathematical foundations of each method, discuss practical considerations such as feature scaling, class imbalance, and overfitting, and provide a reproducible Python pipeline implemented **without** `sklearn`, `tensorflow`, or `networkx`. For data, we focus on the classic UCI Lung Cancer dataset (32 instances, 56 features) as a compact benchmark for algorithmic comparison, while emphasizing the limitations of small-sample medical datasets and the need for external validation.

Keywords

Lung cancer, classification, kNN, SVM, Random Forest, XGBoost, gradient boosting, CART, ROC-AUC, confusion matrix, medical ML.