Lung Cancer Diagnosis Using Exhaled VOC Data and Machine Learning

Davis, Jonathan (School: Signature School)

Lung cancer is the most commonly diagnosed cancer worldwide and the leading cause of cancer-related deaths. It is difficult to diagnose accurately, especially at an early stage. No traditional non-invasive diagnosis method achieves both high performance and cost-effectiveness. Recent research has found that cancerous lung cells produce certain abnormal chemical byproducts, which enter the bloodstream and reach the lungs, leading to the presence of specific volatile organic compounds (VOCs) in exhaled breath. The cost of breath capture and analysis is low. Machine learning classifiers can process complex data and effectively rank the importance of features for the prediction. This study trained, fine-tuned, and evaluated models for each of eight classifiers using exhaled VOC data, compared the ROC-AUC scores of the best model for each classifier, identified the top-performing model (a fine-tuned Random Forest), retrained and retested it, and compared it to traditional non-invasive diagnostic methods using medical performance metrics. The results show that the retained best model achieves accuracy, precision, specificity, sensitivity, false positive rate, and false negative rate values that are similar to or better than those of traditional non-invasive methods. Additionally, the study identifies and ranks the top five most important exhaled VOCs related to lung cancer diagnosis using SHAP values. The findings of this study benefit both hospitals and lung cancer patients, particularly those in the early stages before symptoms appear. Furthermore, it contributes to the discovery of potential targets for new drug development.

Awards Won: