Two machine learning models for prostate cancer screening using urinary volatile organic compounds

Asante, Peter K.[1], Bustamante-Murguia, Pablo[1], Lee, Wen-Yee[3], Mariani, Maria C.[1,2,4], Orosz, Michael[2], and Pokojovy, Michael[1,2,4]

[1]Computational Science Program, The University of Texas at El Paso
[2]Data Science Program, The University of Texas at El Paso
[3]Department of Chemistry and Biochemistry, The University of Texas at El Paso
[4]Department of Mathematical Sciences, The University of Texas at El Paso

October 17, 2022

Abstract

Prostate cancer (PCa) is the most common cancer type in men and the second leading cause of male cancer-specific mortality in the United States. Early and accurate detection is crucial for successful prevention and treatment of PCa. Protein Specific Antigen (PSA) test remains the most widely used PCa screening method to decide if a patient needs to undergo a biopsy, an expensive and invasive procedure that may cause pain and other medical complications. However, the sensitivity of the PSA test is unacceptably low resulting in a 60-75% false positive rate in patients with elevated PSA levels. Computer aided diagnosis (CAD) systems have proved very useful in oncology with primary application to medical imaging data. Urinary volatile organic compounds (VOC) have recently been discovered as a powerful non-image biomarker for PCa giving rise to first machine learning models for non-invasive urinary VOC-based PCa screening. In this study, two machine learning models, namely, penalized logistic regression and nonparametric random forest, are trained, validated and tested on a dataset collected from a total of 183 patients (108 cancer patients and 75 controls). The performance of our both methods in terms of accuracy, specificity and sensitivity (0.90/0.76/0.88 and 0.96/0.79/0.94) is analyzed and compared to that of the PSA test (0.54/0.42/0.60) suggesting significant superiority of VOC-based PCa screening.