

Integrating EHR, Environmental, and Genomic Data for Precision Lung-Cancer Risk Classification in Southern California



This project integrates electronic health records (EHRs), environmental exposures, and genomic data to predict lung cancer diagnosis in Southern California with the goal of better understanding features influencing lung cancer risk. We assembled a retrospective cohort of 7,151 UC San Diego Health patients (50.4% female; 65% ≥ 65 y) and linked 40+ clinical variables to environmental pollutant data from CalEnviroScreen and socioeconomic indices via geocoded addresses. A model-zoo of fourteen learners including logistic regression, random forest, XGBoost, and eleven PyTorch neural-network variants were evaluated with five-fold cross-validation (CV). XGBoost yielded the top mean Area Under the Receiver Operating Characteristic Curve (ROC-AUC) (0.877 ± 0.009) and was then optimized with Bayesian hyper-parameter search in Optuna, boosting CV ROC-AUC to 0.879 and test ROC-AUC to 0.869 while raising held-out accuracy, precision, and F1 to 0.802, 0.744, and 0.694. Shapley (SHAP) analysis reveals that cardiovascular comorbidity, current smoking status, body weight, hypertension, insurance coverage, latitude, ethnicity, age, height, and detailed tobacco burden (pack-years and cigarettes per day) drive predictions, whereas environmental variables such as neighbourhood unemployment, pesticides, and ozone exert smaller but measurable effects. We found a statistically significant difference in mean age-at-diagnosis for those with certain genomic mutations. These findings demonstrate that an optimized gradient-boosting model integrating EHR and environmental data can meaningfully stratify risk, expose pollution-linked inequities, and inform precision public-health interventions.

Gianni Pucillo, Sandip Patel, and Uduak George

This research is supported by the National Cancer Institute of the National Institutes of Health grant (U54CA285117 at SDSU & U54CA285115 at UCSD) and the Computational Science Research Center (CSRC) at San Diego State University

