Machine Learning Model for Early Identification of Vocal Patterns Associated With Glottic Carcinoma Risk by Analysis of Voice Biomarkers and Life Habits

Rivera, Inara (School: The San Juan Math, Science and Technology Center)

The purpose of this investigation was to develop a machine learning model capable of predicting early vocal patterns associated with glottic carcinoma risk using acoustic biomarkers and lifestyle factors. The issue arises with the difficulty of early diagnosis, as early symptoms are often subtle and current diagnostic methods can be invasive, costly, and less accessible. The hypothesis proposed that machine learning models trained on biomarkers such as jitter and shimmer combined with lifestyle factors such as smoking and alcohol consumption, could predict early-stage glottic carcinoma risk with moderate to high performance. The study involved twenty participants divided into diagnosed patients and healthy participants. Two selected recording exercises were taken per participant, which totals in forty voice recordings collected in a controlled clinical environment. Acoustic parameters were extracted from the recordings through a signal-processing pipeline, and classification models such as Logistic Regression, SVM and Random Forest were assessed with accuracy, sensitivity, precision, F1 score, and ROC-AUC metrics. Results indicated that acoustic biomarkers accounted for most of the predictive capacity, with jitter identified as the strongest feature. Logistic Regression achieved the highest overall performance (accuracy ˜ 0.82, F1 ˜ 0.88, ROC-AUC ˜ 0.82). Random Forest demonstrated extremely high sensitivity but extremely low specificity, while SVM showed a weaker overall performance. Findings supported the hypothesis that voice-based machine learning can function as a promising non-invasive screening approach for early glottic carcinoma risk, though broader validation with larger and more diverse datasets is required.

Awards Won: