ABSTRACT

Coughing is one of the primary symptoms of respiratory illnesses, including COVID-19. With the advancement of machine learning, cough sound classification has emerged as a promising method for non-invasive health screening. This study proposes a lightweight and interpretable approach to classify cough sounds into COVID-19 and non-COVID-19 categories by transforming Mel-Frequency Cepstral Coefficients (MFCCs) into statistical features. The dataset used comprises over 20,000 cough recordings, with approximately 4,000 expert-labeled samples, each represented by 11 statistical descriptors such as mean, standard deviation, skewness, and kurtosis. Three classical classifiers—Decision Tree, Random Forest, and XGBoost—were evaluated using multiple data balancing techniques (undersampling, oversampling, SMOTE, SMOTE-ENN, SASMOTE) and outlier removal (Z-score and IQR). The best experimental scenario, using SMOTE-ENN and statistical features (Q25, Q75, IQR), yielded macro-averaged F1-scores in the range of 0.45 to 0.57, depending on the test data distribution. While balanced test scenarios allowed for more equitable classification across both classes, performance declined significantly under imbalanced conditions—highlighting the real-world challenge of data skew. These results emphasize the need for robust preprocessing and fair evaluation protocols. The proposed approach demonstrates that interpretable, low-complexity models can provide meaningful diagnostic value, especially in low-resource environments where computational simplicity and transparency are essential.

Keywords: COVID-19, cough classification, MFCC, statistical features, machine learning, class imbalance, XGBoost.