Dry vs Wet Cough Automatic Classification using the COUGHVID Dataset: CS-433 Class Project

CONTACT PERSONS

- Ms. Lara Orlandic (lara.orlandic@epfl.ch)
- Dr. Tomas Teijeiro (tomas.teijeiro@epfl.ch)

DESCRIPTION

Cough sounds analysis is commonly performed by trained physicians to diagnose a variety of respiratory conditions (i.e. asthma, pneumonia, COPD, and potentially COVID-19). An important physiological phenomenon that can be detected from cough sounds is whether mucus is produced ("wet" cough) or not ("dry" cough), thus providing valuable medical information. The goal of this project is to automatically classify dry vs wet coughs based on cough sounds recorded through laptops and smartphones, which can enable quick and easy cough diagnoses without requiring trained medical professionals. The Embedded Systems Lab has published the COUGHVID dataset, a collection of crowdsourced cough sounds (https://zenodo.org/record/4048312#.X6y4vWhKg2w), 2,200 of which have been annotated by doctors to determine the type of cough. The lab will provide a set of ~70 features computed on each cough recording, which the students will use to perform automated cough type classification. A private test set has been excluded from publishing, and the students will contact the lab once they have developed a final model that is ready to be tested.

STUDENT TASKS

1. Perform wet vs dry cough classification using state-of-the-art ML classification algorithms (Logistic Regression, Support Vector Machines, Linear Discriminant Analysis, k Nearest Neighbors, Gaussian Naive Bayes, Decision Tree, Random Forest, and eXtreme Gradient Boosting) using the computed features and subject metadata.
    1. Compare the success of different classifiers, very carefully performing a fair model comparison (i.e. relying on validation scores using leave-n-subjects-out cross-validation).
    2. Make sure that data from a single subject does not end up in both the training and validation groups
    3. Deal with missing metadata appropriately
    4. Compare classification success on segmented and non-segmented cough recordings (all pre-processing will be done by the lab)
    5. Tune the model hyperparameters appropriately

    6. Report meaningful scores accounting for potential label imbalance (i.e. report AUC instead of classification accuracy)

2. Perform exploratory data analysis and feature engineering (ex. examine the effects of normalizing features, recursive feature elimination)
3. Assess the importance of different features to the classification result by analyzing the weights of the classifier or SHAP values.
4. (Bonus) Try using Deep Learning methods to perform cough classification directly on the raw audio signals