Contribution ID: **101** Type: **Talk**

Lung Health Screening by Automatic Cough Analysis with Applications to Tuberculosis and COVID-19

Wednesday, 2 December 2020 15:45 (30 minutes)

We have applied machine learning algorithms, including logistic regression (LR), support vector machines (SVM), k-nearest neighbour (KNN) and neural networks (DNN) including convolutional (CNN), recursive (LSTM) and Resnet50 architectures to classify the coughing sounds of tuberculosis (TB) and COVID19 patients.

To do this for TB, we have complied a dataset of cough recordings obtained in a real-world setting from 16 patients confirmed to be suffering from TB and 33 patients that are suffering from a respiratory condition that has been confirmed to not be TB. Among all classifiers considered, we find that best performance is achieved using a LR. In combination with feature selection by sequential forward search (SFS), our best system achieves an area under the ROC curve (AUC) of 0.94 using 23 features selected from a set of 78 high-resolution mel-frequency cepstral coefficients (MFCCs). This system is able to exceed the 90\% sensitivity at 70\% specificity specification considered by the WHO as a minimal requirement for an effective community-based triage test.

For COVID-19, gathering or own data has proved to be very challenging and hence we have developed initial systems using the publicly-available COSWARA dataset (https://coswara.iisc.ac.in/about), which currently includes recordings of the coughs by 1135 healthy and 95 COVID-19 positive patients. As this dataset is highly imbalanced, synthetic minority over-sampling (SMOTE) is applied before training CNN, LSTM and Resnet50 neural architectures. Our best system, which is a Resnet50, has achieved an AUC of 0.96. We would like to apply this system on a locally-compiled dataset. Therefore we are engaged in a data-gathering project (https://coughtest.online) which has so far collected cough sounds from 8 COVID positive and 14 COVID negative participants.

We conclude that, for TB, automatic classification of cough audio sounds is promising as a viable means of low-cost easily-deployable front-line screening, and we are actively pursuing improvements to our system. For COVID-19, cough classification also appears to hold much promise, but more extensive testing on locally-collected data is necessary to obtain more clarity. All classifiers were trained and evaluated using nested cross-validation to make best use of the small datasets for parameter estimation, hyperparameter optimisation and final testing.
This is a computationally extremely expensive process but easily parallelised, and hence the CHPC provided an ideal and key resource for performing this work.

Student?

No

Supervisor name

Supervisor email

Primary author: Dr PAHAR, Madhurananda (Stellenbosch University)

Presenter: Dr PAHAR, Madhurananda (Stellenbosch University)

Session Classification: HPC Applications

Track Classification: Cognitive Computing and Machine Learning