CSCI-SHU 360 Machine Learning: Final Project Proposal

Team members:

- Ivana Li yl8060
- Lihan Feng lf2383

Short description of the project: *[What is the topic? What is the main goal of the project?]*

The topic is respiratory disease prediction by inspecting the respiratory soundtracks of individuals. In this case, early diagnosis can be done automatically instead of using human efforts to achieve cost-efficient alerting diagnosis purposes.

The main goal of the project is to predict whether someone has respiratory diseases and what kind of disease someone may have by using machine learning techniques, based on the audio data about their respiratory cycles. The project also aligns with one of the WHO guidelines that aims at investing in the use of decision-support tools on digital devices to strengthen primary healthcare.

Data set: *[What kind of data do you need? How do you plan to get it? Do you already have some sources?]*

We use an audio dataset mainly containing 920 *.wav sound files recorded from 126 patients' respiratory sounds with 920 corresponding annotation files. In each file, there are multiple cycles of the respiratory sounds, in total 6898 cycles (5.5 hours). The dataset also contains the diagnosis results of the patients for training and testing and other demographic information.

We already have this potential open dataset found in Kaggle. The dataset is also based on the paper called: A Respiratory Sound Database for the Development of Automated Classification.

Citation: Rocha, B. M., Filos, D., Mendes, L., Vogiatzis, I., Perantoni, E., Kaimakamis, E., Natsiavas, P., Oliveira, A., Jácome, C., Marques, A., Paiva, R. P., Chouvarda, I., Carvalho, P., & Maglaveras, N. (2017). A respiratory sound database for the development of Automated

Classification. *Precision Medicine Powered by PHealth and Connected Health*, 33–37. https://doi.org/10.1007/978-981-10-7419-6_6

Project plan: *[What will be the main steps to develop your project? What kind of techniques do you plan to use? This does not need to be your final plan, but the earlier you have a plan, the earlier we can provide feedback and support.]*

1. Clean and preprocess data.
2. Learn about packages such as librosa to process audio data and convert it to image form such as spectrogram, etc.
3. Read some related papers and use specific CNN models to transform the image data into feature vectors.
4. Based on the feature vectors, train the data by different machine learning classifiers, such as decision trees, logistic regression, and neural networks, to detect whether someone has respiratory diseases and what kind of disease someone may have.
5. Compare the results from each model, and identify which model works the best for this problem.