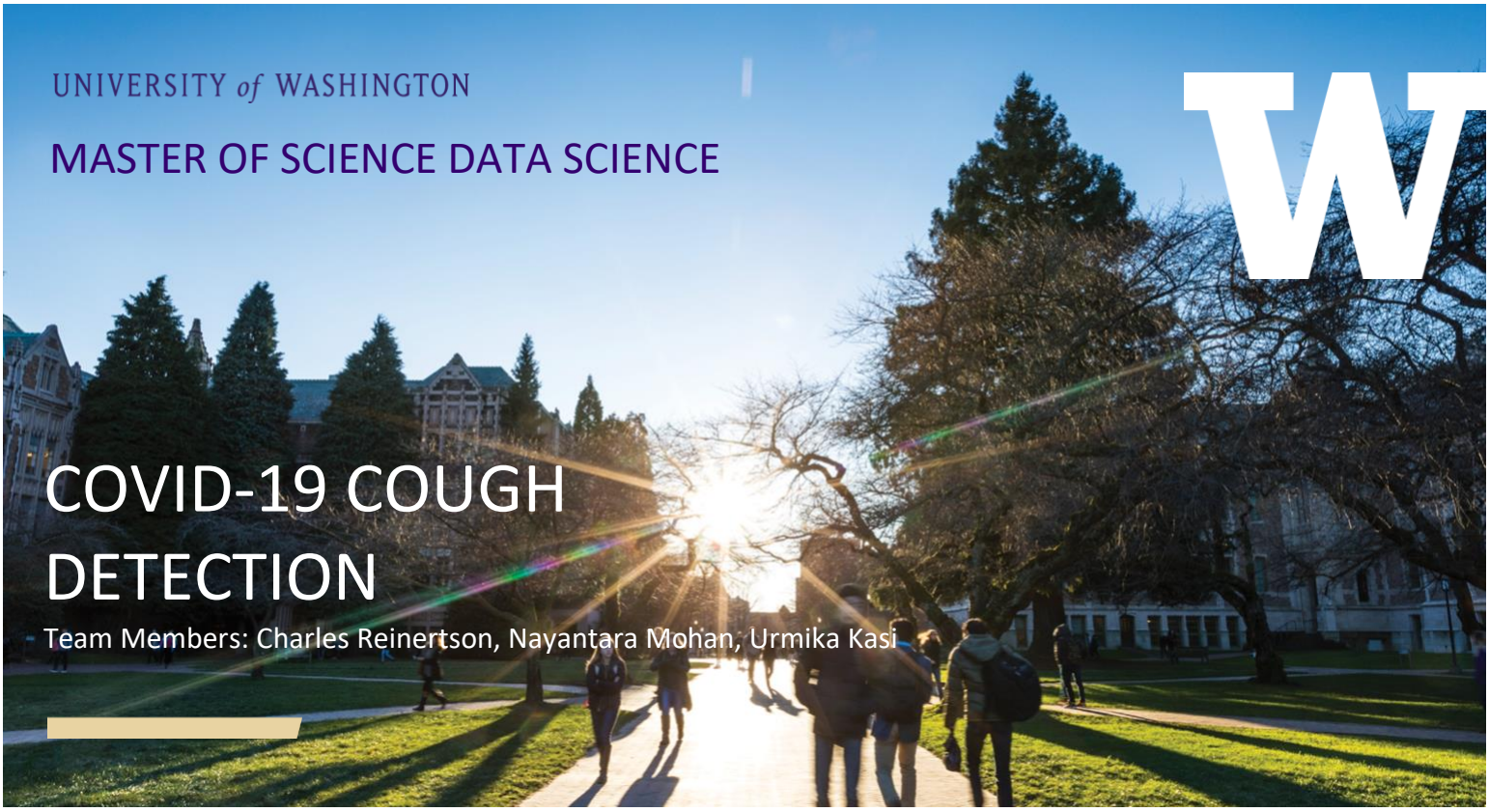

UNIVERSITY of WASHINGTON

MASTER OF SCIENCE DATA SCIENCE

COVID-19 COUGH DETECTION

Team Members: Charles Reinertson, Nayantara Mohan, Urmika Kasi

Virufy

PROBLEM STATEMENT

The problem of inadequate COVID-19 testing, particularly in developing countries, poses a significant challenge. The current testing methods are scarce, expensive, and require trained personnel and costly equipment, which limit the number of people who can obtain their results.

Numerous attempts have been made to solve this problem with partial success, including contact tracing apps which have not been widely adopted often due to privacy concerns. Pharmaceutical companies have also fast-tracked the development of vaccines, but they still will not be widely available in developing countries for some time.

As a solution to this problem, the creation of a cough detection classifier has been proposed, which would provide a binary output indicating the presence or absence of COVID-19.

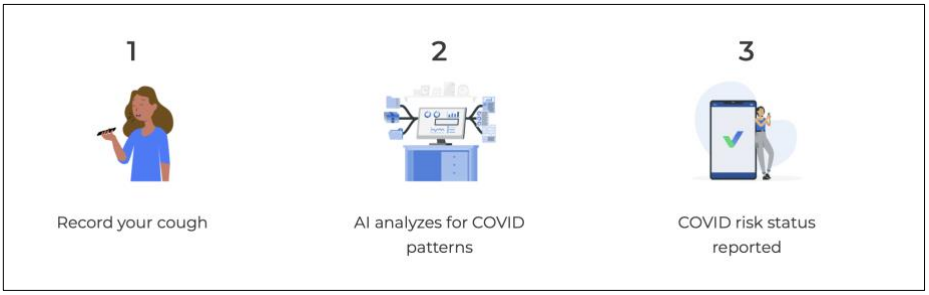

Figure 1: A representation of how the Virufy app can be used to detect COVID-19

MOTIVATION

> Machine learning and deep learning approaches can detect COVID-19 by analyzing cough sounds.

> Audio-based COVID-19 diagnostic tools could be developed to enable a completely digital and risk-free detection solution.

> Given the potential benefits, it is crucial to research and develop machine learning-based digital COVID-19 tests.

About Virufy

Virufy aims to build a free smartphone app to detect COVID-19 from cough recordings through machine learning analysis of audio signals, which would allow for mass-scale testing and could effectively stop the spread of the virus.

DATA

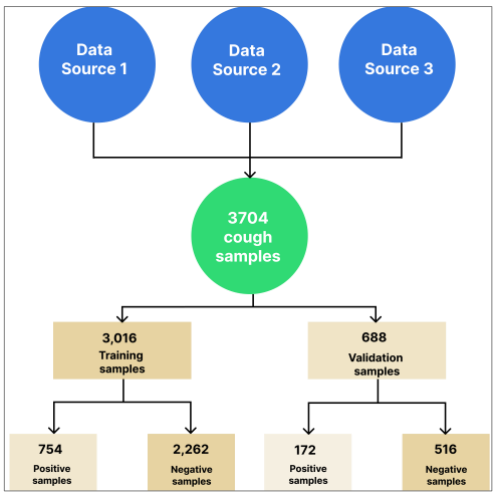

Figure 2: Illustration of the data used for the model classification.

> Three different data sources have been combined and collected through crowdsourcing. This work is licensed under a Creative Commons Attribution 4.0 International License.

> The features extracted for this project are the cough recordings and the corresponding PCR test result (positive or negative).

> The dataset is heavily skewed to negative samples.

PROPOSED SOLUTION

> We fine-tuned a HuBERT model for sequence classification on our audio samples.
> To prepare for training, we pre-processed the samples by encoding them into NumPy arrays, addressing class imbalances, and extracting features using Wav2Vec2.
> We split the data into training and validation sets.
> We chose HuBERT for its self-supervised speech representation learning, which involves offline clustering and a BERT-like prediction loss. The model was trained and evaluated on metrics like Accuracy, Precision, Recall, and AUC.

RESULTS

> Our HuBERT model accurately predicts COVID-19 from cough samples with a high true positive rate and low false positive rate.
> However, the model fails to predict positive COVID-19 cases about ⅕ times, which may cause users to overlook preventive measures.
> Validation dataset accuracy improves over iteration, but the model appears to overfit after Epoch 14.
> Lack of data and an imbalanced dataset are major issues in training the model. We only had 754 positive COVID-19 samples, which is not enough for proper tuning.
> We expect HuBERT to generalize better and false positive rate and AUC to increase as new positive samples are collected.
> Although our AUC score falls below the best-performing model's score, we cannot rule out the usefulness of HuBERT until more positive samples are collected.

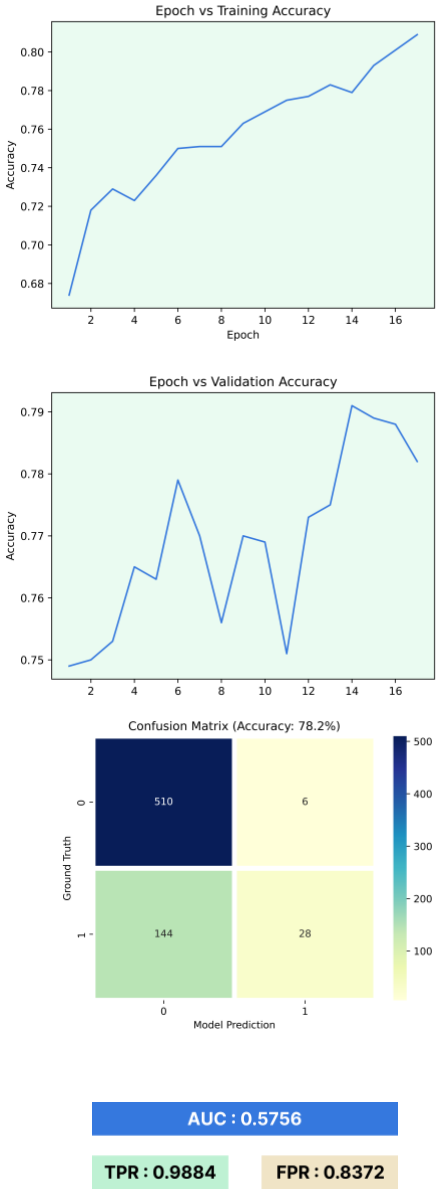

Figure 3: Confusion matrix and the performance of the model

FUTURE WORK

> HuBERT model's poor generalization to unseen data is due to an imbalanced dataset.

> Insufficient positive COVID-19 samples prevent proper tuning of the model.

> The main goal of future work is collecting more positive.