PROJECT PROPOSAL

TABLE OF CONTENTS

1. Introduction 1
2. Problem Statement 2
3. Background 3
4. Data Exploration 7
5. Data Pipeline 10
6. Proposed Solutions 11
7. Risks and benefits of proposed solutions 12
8. Schedule 12
9. Team bios 14
10. References 16

INTRODUCTION

COVID-19 testing is inadequate, especially in developing countries. Testing is scarce, requires trained nurses with costly equipment, and is expensive, limiting how many people can obtain their results. Also, many people in developing countries cannot risk taking tests because results are not anonymous, and a positive result may mean a loss of day-to-day work income and starvation for their families, which further allows COVID-19 to spread.

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.

To combat these problems, we propose 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.

We propose to build an end-to-end system that preprocesses cough (audio signals) data and gives a binary output of whether COVID-19 is present or not. This model would involve audio engineering and extracting relevant features which shall be hosted on Amazon infrastructure and can be accessed through a REST API.

Due to the dynamic and fluid nature of the company, particularly regarding the organizational structure and hierarchy, our project goals are unclear to us at this instant, and are subject to change based on company requirements with time. Consequently, our particular project statement is not known to us, and we have a choice of two projects, as given below:

Project 1: To detect the presence of background noise in cough samples.
Project 2: To create a model to detect the presence of COVID in cough samples as a baseline for comparison.

The project may change in January. For the purpose of this assignment, we considered the “master” dataset provided to us, that we hypothesize will be useful for both projects. As a result, our analysis is rather general, and is not particularly focused to a particular project.

PROBLEM STATEMENT

COVID-19 testing is inadequate, especially in developing countries. Testing is scarce, requires trained nurses with costly equipment, and is expensive, limiting how many people can obtain their results. Also, many people in developing countries cannot risk taking tests because results are not anonymous, and a positive result may mean a loss of day-to-day work income and starvation for their families, which further allows COVID-19 to spread.

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.

To combat these problems, 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. They are building an end-to-end system that preprocesses cough (audio signals) data and gives a binary output of whether COVID-19 is present or not. This model involves audio engineering and extracting relevant features which shall be hosted on Amazon infrastructure and can be accessed through a REST API.

We have two possible directions for our capstone project.

Project 1: To detect the presence of background noise in cough samples.
Project 2: To create a model to detect the presence of COVID in cough samples as a baseline for comparison.

Technology stack:

- AWS
- Python
- Tensorflow

Data:
Audio (cough) samples

Challenges:

- It is under the trial stages and we will be a part of testing the proof of concept.
- Data processing may be cumbersome as there may be a lot of noise.
- Explainability and interpretability of the models are a challenge.

BACKGROUND

Over 800,000 people have died as a result of COVID-19 in the US and 5 million worldwide as of this writing. A record number of hospital beds have been occupied due to COVID-19's high infectiousness, exceeding hospital capacity and placing a tremendous burden on global healthcare systems. Despite ongoing global immunization campaigns, distribution attempts in low- and middle-income nations have been limited. Additionally, the efficacy of the current vaccinations in stopping the spread of COVID-19 has been reduced due to the emergence of new viral variations such as Omicron.

Emerging Artificial Intelligence (AI) technologies have demonstrated the capacity to develop quick, inexpensive, and available solutions. There is growing evidence that using machine learning and deep learning techniques, one may predict COVID-19 by listening to infected individuals' coughs. This study advances the development of an audio-based COVID-19 diagnostic tool by demonstrating adequate performance on a

sizable and varied dataset. Future work should aim at gathering large numbers of PCR-validated samples and determining the model's generalizability to additional datasets, in addition to enhancing model performance.

The approach to data is different in this study as this is a novel approach where the data is a sound sample. For COVID-19 patients of all ages, in various environments, symptomatic and asymptomatic, and at various times relative to symptom onset, numerous research groups have been collecting sound recordings.

There is a significant impact on the industry: a prospective, entirely digital COVID-19 detection approach would enable a smartphone-based quick, fair COVID-19 test with no danger of infection, financial burden, and supply chain problems—all features that are beneficial for reducing COVID-19 spread. Given the significance, it is necessary to investigate and create digital COVID-19 exams that are based on machine learning.

Using Deep Learning with Large Aggregated Datasets for COVID-19 Classification from Cough

To ensure minimal bias during model training, several publicly accessible datasets were merged and used. Their research app (https://virufy.org/study) was used to collect the crowdsourced dataset for Virufy, which asks users to input their PCR test results along with demographic data like age, gender, and comorbidities. The preprocessing stage takes the standardized data to filter and prepare usable samples by detecting background noise and removing any inadequate data, such as samples with no coughs detected from it. Several steps were involved in the preprocessing stage, namely, volume detection, clipping detection, cough detection, cough segmentation, and background noise detection. Each of these steps involves various resampling and filtering of audio signals based on various feature extraction and model training. The model outputs a 0 for absence of COVID and 1 for the presence of COVID. To map the SVM outputs to probabilities within the [0, 1] range, the SVM outputs are fed into logistic regression. Platt Scaling is used to transform the SVM output to these probability values. There are some complicated math algorithms used, for e.g., the calculation of Mel-frequency Cepstral Coefficients (MFCC which is based on the Fast Fourier Transform). The paper does not mention why a particular order of the derivative of the MFCC is used thus making it hard to understand the math behind the algorithm.

The results of this study are understood by the Area Under the Curve of the ROC plots. A thorough investigation of a self-supervised learning model gave the AUC = 0.807 and a Convolutional Neural Network (CNN) model gave the AUC = 0.802. The authors mentioned that they believe the performance can be further increased with an increase in training data size.

The authors have stated several tasks for future work. With various up-sampling techniques, hyperparameter tweaking, output threshold adjustment, and over-fitting prevention strategies, both models can be improved in several ways. Additionally, the pre-processing and feature extraction procedures can be strengthened by, for example, decomposing the cough segments into separate samples and concatenating them to provide a final prediction on a single audio file. Additionally, the Virufy team's and other teams' choice of data sources can impact the effectiveness of their models. The experimentation with a fusion model that incorporates predictions from the SSL-converter, CNN, SVM, and other techniques to enhance prediction performance is also another potential next step.

There are several limitations to be noted like the previous studies have shown a higher AUC than the results in this paper. Also, many academic papers regarding COVID-19 classification with AI suffer from reproducibility issues. I will need to see how the performance of this model by Virufy would stand in the performance of reproducibility. Another concern I see is that the majority of the labeled dataset used for training is done by crowdsourced workers and these are not verified hence leading to a possibility of errors/bias in the model.

Hierarchical Multi-modal Transformer for Automatic Detection of COVID-19

One relevant paper is *Hierarchical Multi-modal Transformer for Automatic Detection of COVID-19 (*https://dl.acm.org/doi/pdf/10.1145/3556384.3556414*).* This paper was written by members of the Virufy team.

The Hierarchical Multi-modal Transformer proposed in this paper improves the accuracy of COVID-19 detection on cough recording datasets. This model can also integrate into mobile devices without sacrificing accuracy like many other lightweight models built for mobile devices.

Instead of directly concatenating the different features together and throwing them in a model, this paper suggests a multi-model approach that detects and classifies COVID-19 with an MLP network for the 1D clinical and audio features and a transformer for the 2D audio spectrograms. This paper also proposes a cross-attention module that fuses the intermediate representations from each branch, effectively capturing the relations between each of them and their importance in contributing to the final prediction. That way only the important signal from the different types of cough data are used in the final prediction.

However, the model's accuracy is much below the necessary accuracy required for a COVID-19 detection application. The proposed model has an 81% accuracy on detection. "Analytic performance of many SARS-CoV-2 diagnostic PCR tests

approaches 100% at 500-5000 copies/mL" (https://www.cap.org/member-resources/articles/how-good-are-covid-19-sars-cov-2-diagnostic-pcr-tests)

If this paper along with future improvements achieves an accuracy similar to a PCR test, detecting COVID-19 will become much cheaper and easier. In turn, this will help temper the pandemic as this will remove a barrier of entry for COVID-19 testing.

A comprehensive approach for classification of the cough type

In this paper, the authors explore the classification of cough types, such as dry or wet, as it is effective to detect the presence of sputum in the lungs (sputum is a mixture of saliva and mucus coughed up from the respiratory tract, typically as a result of infection or other disease and often examined microscopically to aid medical diagnosis). Detecting the presence of sputum is important to identify various diseases such as lung infection, pneumonia, and cancer. In this work, the authors propose a classification mechanism for cough sounds to detect the presence of sputum. This classification can be used to assist in the diagnosis of various respiratory diseases.

This work enables the objective classification of the presence of lung congestion based on the cough sound using a smartphone. Methods for cough type detection so far were subjective and at best very faulty. The existence of sputum in the lung is a strong indicator of numerous conditions from simple pulmonary infection to more serious diseases such as pneumonia and cancer. Therefore, they also establish the clinical significance of detecting sputum in the lungs, as it is useful for condition assessment and adverse pulmonary event prediction.

The findings of this work suggest that with thorough processes employed for annotation, data pre-processing, feature selection, and model fitting, healthcare can see significant changes with the application of machine learning. In this pulmonary domain of cough analysis, this work can further be improved and expanded upon for different kinds of diseases with similar base characteristics and symptoms, such as COVID-19. For instance, while classifying COVID-19 cough samples, a similar pre-processing mechanism can be employed to ensure the improved performance of classification algorithms.

DATA EXPLORATION

Link to detailed data exploration document:
https://docs.google.com/document/d/1civ8GaFIBq3np7ZVCensFMPmVHr06Wqmpq8bm9IqPc4/edit?usp=sharing

Data repository: https://github.com/iiscleap/Coswara-Data

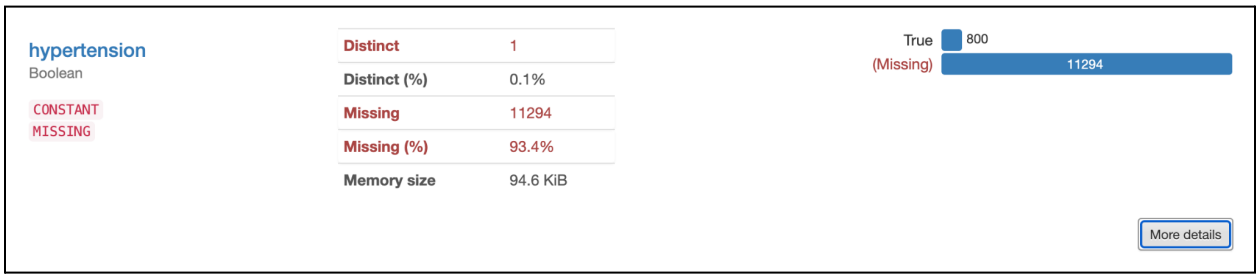
Project Coswara by the Indian Institute of Science (IISc) Bangalore is an attempt to build a diagnostic tool for COVID-19 detection using audio recordings such as breathing, cough, and speech sounds of an individual. The data was collected through crowdsourcing.
Our project(s) at Virufy will mainly utilize this data.

We assume the following:

- Positive test results taken over a month ago are treated as recovered infections.
- This data is crowdsourced and consequentially may contain biases from this collection procedure such as sampling bias, convenience sampling, domain specificity, and so on.
- We hypothesize that NaNs in our comorbidity features are equivalent to False, but cannot confirm the same before conferring with our stakeholders. Consequently, we have not used this assumption in our EDA.

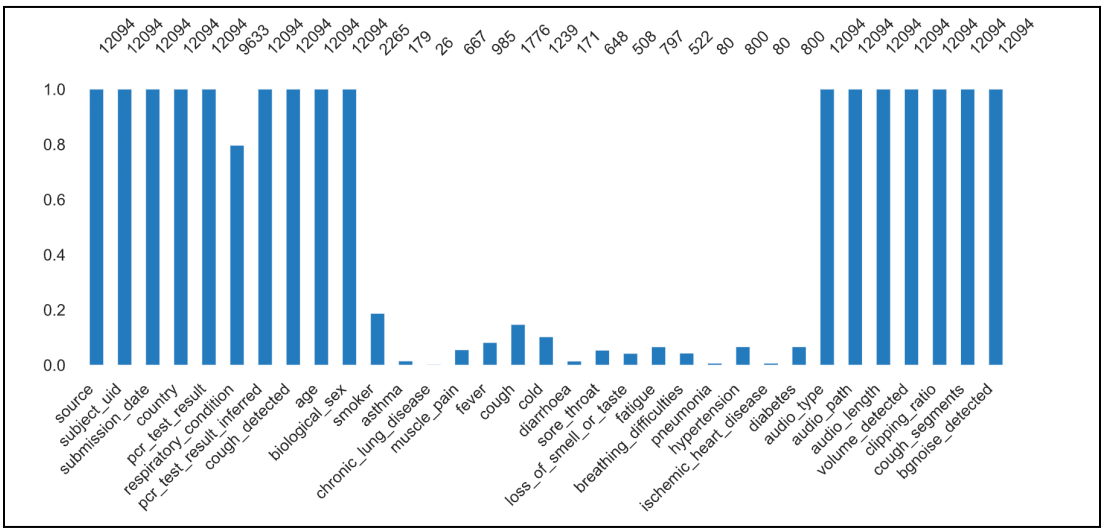
Preprocessing
For columns indicating comorbidities, we have a large number of NaN values - over 95%.
In this regard, this data cannot be used for analysis unless further conferred upon with the stakeholders - do NaNs translate to False, or is this data actually missing from the dataset?

The following picture demonstrates an example of one such comorbidity feature -



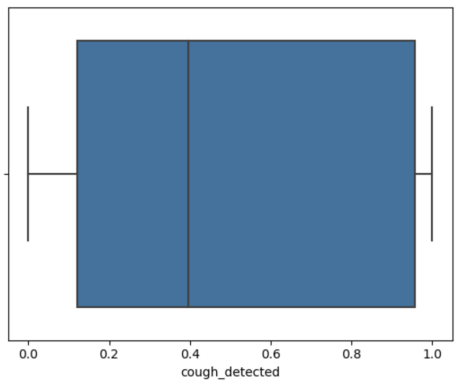
Missing Values

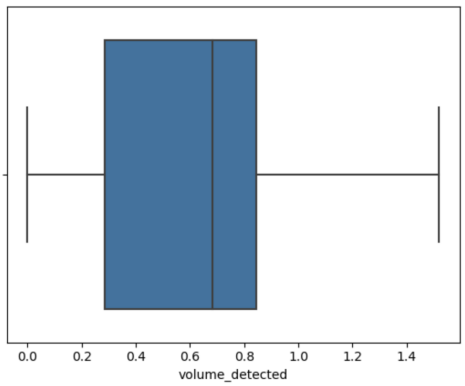
The graph below shows the percentage of missing values for each column attribute from our dataset.

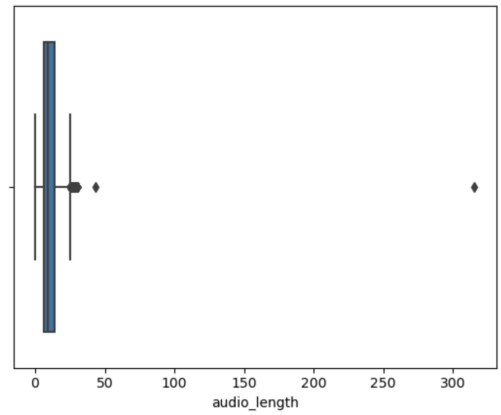


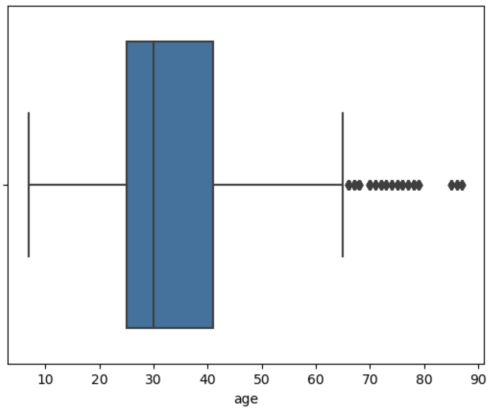
Outliers

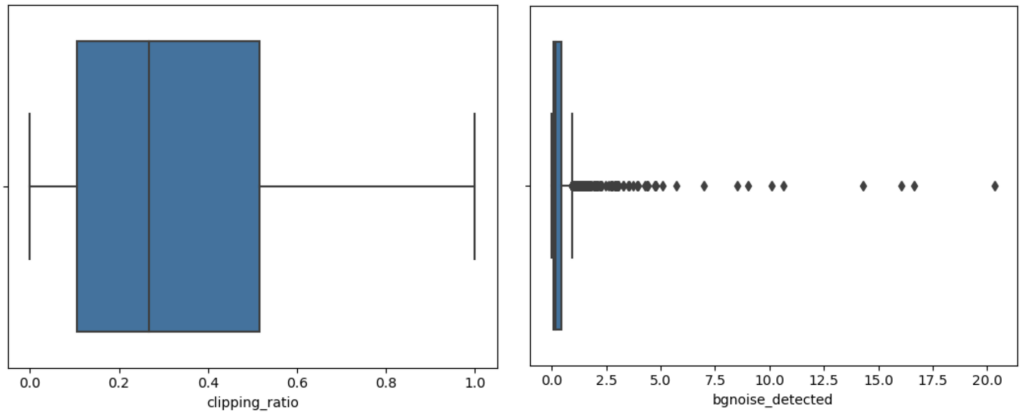
Outliers for the features of interest to us have been visualized as follows: -



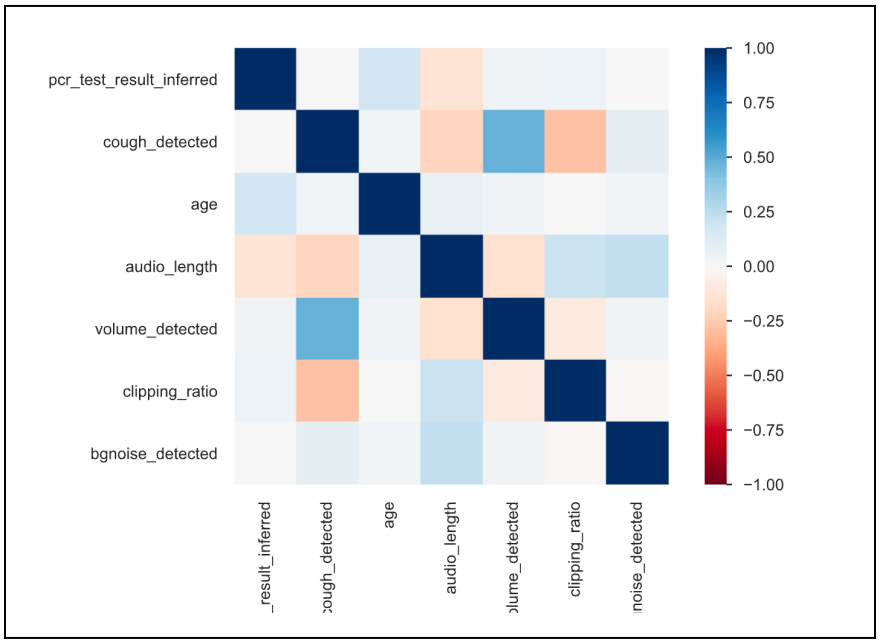








Correlated Features



Data Flaws

We have several datasets provided to us for exploration. There is one “master” dataset (the “Coswara” dataset), which includes demographic and COVID-specific symptomatic data about the cough samples collected for COVID detection. This is the dataset we will consider for exploration.

- While investigating these datasets, we noticed that some columns were split and combined differently across different datasets - there needed to be more consistency in this manner. There were also some naming inconsistencies - what was called “gender” in one dataset was called “biological_sex” in another.
- There are several columns that have very high percentages of missing values.
- No schema/dataset description provided.

- For one of the fields (bgnoise_detected), the values are real numbers with an unspecified range or magnitude. As a result, the meaning of these values aren't entirely clear.

DATA PIPELINE

We are going to load the data from S3. There is no specific mechanism to act as an interface to access and query the data yet. Although, the data warehouse is a work in progress by the team (AWS Glue) and it will be fully set up by January 2023. In the meantime, we will be pulling the data directly from S3. Data resides on AWS S3 buckets. Code sharing and version control happen on Github.

We have 10,000 audio samples in our dataset. Some of the records have clinical and demographic data (country of origin) as well. Along with time of collection, a label determines whether or not there was a positive PCR result.

We are using the following tools and software:

- Python 3
- AWS
    - S3: to store data. Currently, data is directly accessed from S3 through Python.
    - Glue: to query data on S3. This is currently being set up, and will be completed by January 2023.
    - SageMaker: to perform ML operations on the data. This is currently being set up, and will be completed by January 2023.

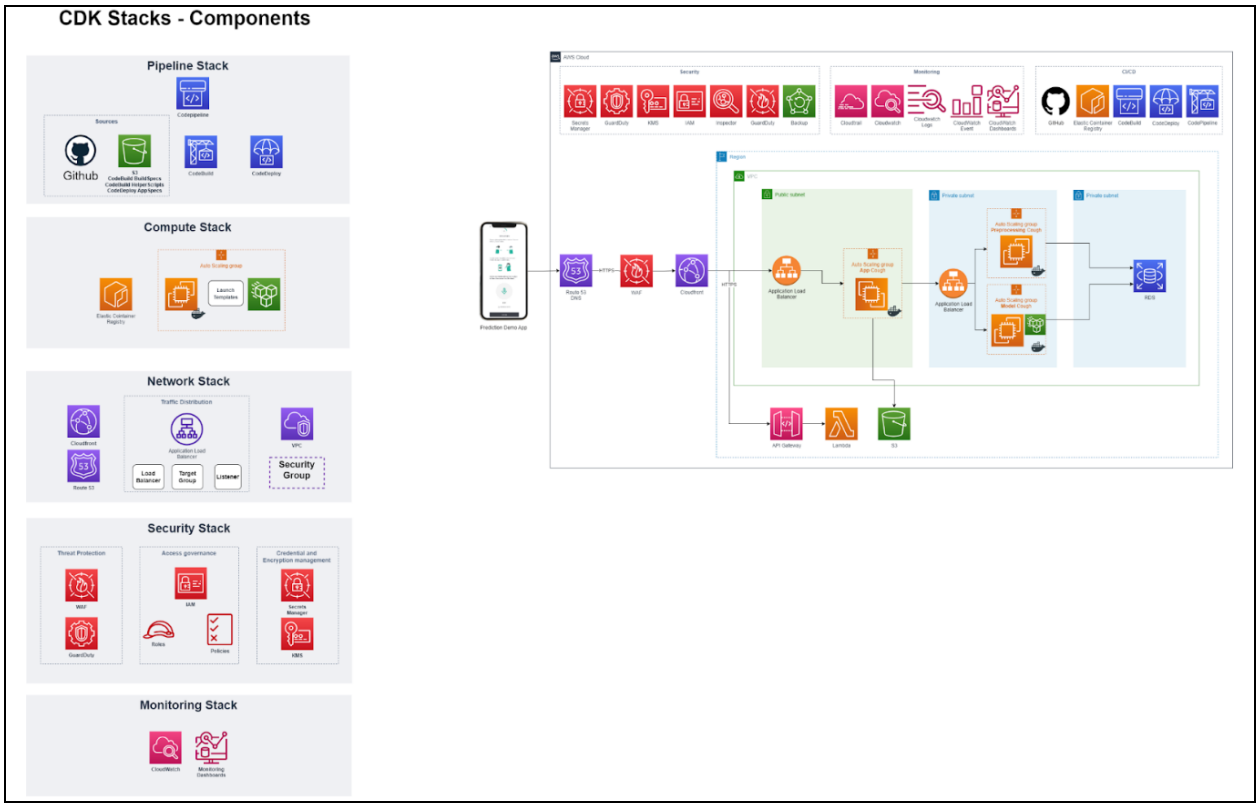


Figure of the CDK Stacks - Components

We have also been given temporary access to the data via .csv files for the purposes of data exploration. However, this is an interim, offline solution, and we hope to be able to access the data on S3 by the beginning of the next quarter.

PROPOSED SOLUTIONS

Project 1: To detect the presence of background noise in cough samples.

- For binary noise background noise classification we will start simple and make more complex models if necessary. Logistic regression will be our baseline model to judge all future “advanced” models. Those “advanced” models will be variations of neural networks.
- To do so, we will employ Python, Tensorflow, and host our solution on AWS. Some libraries we would employ are scikit-learn, scipy, pandas, and other machine learning libraries.

Project 2: To create a model to detect the presence of COVID in cough samples as a baseline for comparison.

- We would create a baseline model that outputs a binary “yes” or “no” that indicates the presence of COVID in the given audio sample.
- To do so, we would use a neural network of some sort. The exact structure will be determined if and when we begin working on the same.
- This project would involve feature selection, network specification, and hyperparameter tuning to produce an acceptable and functioning model.
- To do so we envision using Python and Tensorflow, and will host our solution on AWS.

RISKS AND BENEFITS OF PROPOSED SOLUTIONS

The use of AI for predicting if someone has COVID-19 could potentially have several benefits. For example, it could help healthcare providers more quickly and accurately diagnose patients, potentially leading to earlier treatment and better outcomes. It could also help healthcare systems prioritize the allocation of resources, such as testing and personal protective equipment, to those who are most in need.

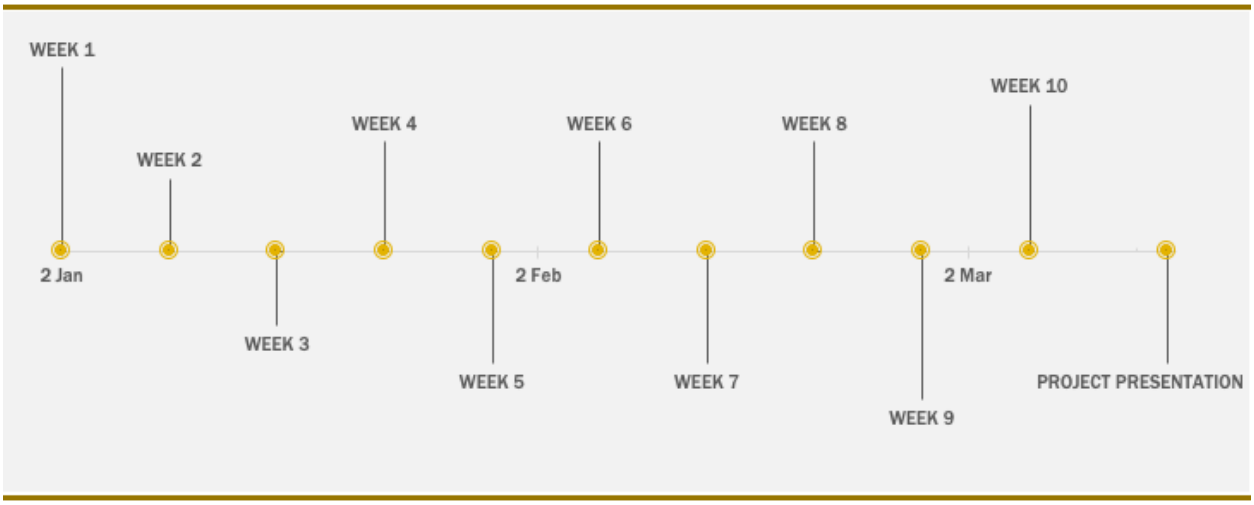
However, there are also potential risks and limitations to using AI for COVID-19 prediction. One major concern is the accuracy of the predictions. AI algorithms are only as good as the data they are trained on, and the accuracy of their predictions can be affected by a variety of factors, such as the quality and diversity of the training data. There is also the potential for bias in the data or the algorithm, which could lead to unfair or inaccurate predictions.

Another potential risk is the ethical concerns around using AI for medical diagnosis. Some people may be uncomfortable with the idea of relying on a machine to make important medical decisions, and there are valid concerns about the potential for errors or misdiagnoses.

Overall, the use of AI for predicting COVID-19 infections could have potential benefits, but it is important to carefully consider the risks and limitations and ensure that any AI systems are developed and used responsibly.

SCHEDULE

NOTE: The schedule is at a high level as Virufy shall confirm the exact project details by the end of week 1.



Week 1:

- **Finalize project.**
- Meet with the team to discuss the project and its goals.
- Identify the tasks that need to be completed and assign them to team members.
- Create a project plan that outlines the timeline and milestones for the project.

Week 2:

- Begin gathering data for the project.
- Start preprocessing the data, including cleaning and formatting it.

Week 3:

- Begin building the model, including choosing the appropriate algorithms and techniques.
- Continue preprocessing the data and training the model.

Week 4:

- Continue training the model, including fine-tuning its parameters.
- Start evaluating the model's performance on the training data.

Week 5:

- Continue evaluating the model's performance and making improvements as needed.
- Begin testing the model on new, unseen data.

Week 6:

- Continue testing the model and evaluating its performance.
- Identify any problems or limitations of the model and work on addressing them.

Week 7:

- Continue improving the model's performance and addressing any issues.
- Begin preparing the final project report and presentation.

Week 8:

- Continue working on the final report and presentation.
- Finalize the project plan and timeline.

Week 9:

- Complete the final report and presentation.
- Prepare for the final presentation and review the project with the team.

Week 10:

- Deliver the final presentation to the project stakeholders.
- Celebrate the successful completion of the project!

TEAM BIOS

Charles Reinertson: I received my Bachelor's degree in Data Science (Computer Science) from the University of Michigan. I am currently a graduate student at the University of Washington working towards a Master's degree in Data Science. As a young professional, I enjoy applying Machine Learning and AI models at scale to solve complex real-world problems that bring value to a business. I am competent in analytical skills and software engineering with a detail-oriented mindset and quick learning capabilities. I will be a Machine Learning Engineer working at HBO Max upon my graduation.	

Nayantara Mohan: I am a graduate student at the University of Washington and am pursuing my master's in Data Science. I have a keen interest in problem-solving with the means of human-centered design and data. I have previously worked as a business intelligence analyst at an investment bank and as a Founding Member of an early-stage med-tech startup. My technical expertise includes data analytics, machine learning, R, Python, SQL, and statistical analysis.	
Urmika Kasi: I am a graduate student in Data Science at the University of Washington. I love to look for everyday problems - small or large and frame possible technical solutions with the skills available at my disposal. My fluency in R, Python, SQL and a host of other languages allow me to do so with ease. Up and coming developments in the fields of Analytics, Deep Learning and Machine Learning have also piqued my interest through numerous courses, internships and jobs.	

REFERENCES

[1] Haritaoglu, E. D., Rasmussen, N., Tan, D. C. H., J., J. R., Xiao, J., Chaudhari, G., Rajput, A., Govindan, P., Canham, C., Chen, W., Yamaura, M., Gomezjurado, L., Broukhim, A., Khanzada, A., & Pilanci, M. (2022, March 30). Using deep learning with

large aggregated datasets for covid-19 classification from cough. arXiv.org. Retrieved November 22, 2022, from https://arxiv.org/abs/2201.01669

[2] Nemati, E., Rahman, M. M., Nathan, V., Vatanparvar, K., & Kuang, J.
(2020, July). A comprehensive approach for classification of the cough type. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 208-212). IEEE.

[3] Mouawad, P., Dubnov, T., & Dubnov, S. (2021). Robust detection of COVID-19 in cough sounds. SN Computer Science, 2(1), 1-13.

[4] Chaudhari, G., Jiang, X., Fakhry, A., Han, A., Xiao, J., Shen, S., & Khanzada, A. (2020). Virufy: Global applicability of crowdsourced and clinical datasets for AI detection of COVID-19 from cough. arXiv preprint arXiv:2011.13320.

[5] Fakhry, A., Jiang, X., Xiao, J., Chaudhari, G., Han, A., & Khanzada, A. (2021). Virufy: A multi-branch deep learning network for automated detection of COVID-19. arXiv preprint arXiv:2103.01806.

[6] Feng, K., He, F., Steinmann, J., & Demirkiran, I. (2021, March). Deep-learning based approach to identify covid-19. In SoutheastCon 2021 (pp. 1-4). IEEE.