DSCI 400 - COVID-19 DETECTION THROUGH COUGH SAMPLES

FINAL REPORT

German Gonzalez
Department of Electrical Engineering
Rice University
Houston, TX
gg28@rice.edu

Amar Moturu
Department of Electrical Engineering
Rice University
Houston, TX
aem14@rice.edu

Andrew Xavier
Department of Electrical Engineering
Rice University
Houston, TX
ahx1@rice.edu

James Beattie
Department of Mathematics
Rice University
Houston, TX
jbb7@rice.edu

Xuanyi (Jessica) Chen
Department of Cognitive Sciences
Rice University
Houston, TX
xc24@rice.edu

May 7, 2021

1 Introduction

1.1 Background

The Coronavirus Disease 2019 (COVID-19) is one of the greatest public health crises of the modern era. COVID-19 is known to cause flu-like symptoms such as fever, cough, and fatigue [1]. Its primary mode of transmission is through the respiratory tract by indirect means such as droplets and aerosols [2]. According to the most recently available statistics from the World Health Organization, COVID-19 has infected more than 150 million people and caused more than 3,230,000 deaths [3]. Throughout the pandemic, governments have often struggled to obtain adequate testing capacity, resulting in incomplete knowledge on the current status of the disease and hindering the effectiveness of containment measures [4]. Thus, developing a low-cost and easily accessible alternative to current testing methods is of significant public interest. Machine learning and artificial intelligence (AI) have been employed by researchers in a diverse range of fields to combat the current pandemic [5]. We believe that such areas may provide a method of diagnosing COVID-19 using coughing sounds.

1.2 Related Work

Various feature extraction methods have been used in the realm of audio classification. For instance, Oikarinen et al. developed a method of classifying sounds from the common marmoset using a spectrogram representation of the audio data [6]. Similarly, Emre et al. used audio recordings to classify ten different animal species and were able to obtain a classification accuracy of 75% [7]. In the realm of music processing, Mierswa et al. compared various machine learning schemes and found that a linear Support Vector Machine (SVM) had the lowest number of classification errors [8].

1.2.1 Prior Cough Detection Work

Research has also been conducted on cough-specific applications of audio signal processing. A cough consists of three distinct phases: the initial burst, the noisy airflow, and the glottal closure [9]. According to Amoh and Odame, there are two main approaches to cough detection and classification. First, the raw audio data can be transformed into a frequency-domain spectrogram, on which a feature selection procedure can be performed. The resulting feature vector can then be passed into a classifier, such as a Support Vector Machine (SVM). Second, the frequency-domain

spectrogram of the raw audio data can be passing directly into a Convolutional Neural Network (CNN), which is known to be well suited for such audio classification tasks [10].

Additionally, Pramono et al. describe a procedure by which audio clips can be cropped to a certain length by removing sections with excess silence. This is done by comparing the standard deviation of audio frames to the mean of the overall standard deviation and setting a threshold at which cropping will occur [11]. In addition, they listed several features of audio signals that can be used in processing, such as the zero-crossing rate (the frequency of sign-changes in the signal [12]), crest factor (ratio of the peak value to the root mean square), energy level, and Mel-Frequency Cepstral Coefficients (MFCC). Mel-Frequency Cepstral Coefficients are important features in the realm of speech recognition as they represent the amplitude spectrum concisely [13].

1.2.2 Cough Detection Efforts on COVID-19 Classification

Since the outbreak of COVID-19, machine learning methods have been used to detect COVID-19 cases [14] [15]. Specifically, various crowdsourced cough datasets were collected and made available for researchers interested in tackling such problems [16][17][18]. Previous results obtained by researchers attempting to detect COVID-19 using patient cough samples are encouraging [19]. Using a set of handcrafted features including duration, onset, and tempo, Brown et al. achieved an ROC-AUC of 80% with simplistic logistic regression and an SVM.

Better performance was achieved through the use of more sophisticated methods. Dunne et al. were able to construct a machine learning model which could distinguish between COVID-19 coughs and non-COVID-19 coughs with a classification accuracy of 97.5% [20]. For this purpose, the authors converted the audio data to a Mel-frequency Cepstral Coefficients (MFCC) spectrogram image, which was used as the input for a Convolutional Neural Network (CNN). Similarly promising results were achieved by Imran et al. using analogous methods [21]. However, a major shortcoming of these models is the relatively small sample sizes on which they were trained. Since more than a year has passed since the beginning of the pandemic, we now have access to much larger data sets containing more samples from individuals infected with COVID-19.

1.2.3 Transfer Learning and ImageNet

In addition to manually constructed models, transfer learning has also been proven valuable for building CNN models when the available dataset has limited samples due to time or availability constraints. Transfer learning refers to the practice of transferring previously learned knowledge from a related task to a new task in order to improve the learning of a new task [22]. Specifically for neural networks, transfer learning is often achieved by borrowing the structure and connection weights from a pre-trained neural network on a related task, while allowing the connection weights of the latter layers to be trained on the novel task while the connection weights of the previous layers remain set [23]. Transfer learning has been especially prevalent in image classification problems, especially with models trained on the ImageNet dataset [24], a hand-annotated visual object recognition dataset with more than 14 million entries. Previous models trained on ImageNet that are used for transfer learning included the famous AlexNet [25], VGG[26], ResNet [27], DenseNet [28], among others, and has been shown to have prominent transfer ability [29][30]. Transfer learning from these models pre-trained in non-medical domain has been proven successful in image classification problems in medical domain through computer-aided diagnosis (CADx), including the detection of thoraco-abdominal lymph nodes, interstitial lung disease, kidney problems, etc. [31] [32]. Due to the limited number of COVID-19 cough samples at the current stage, we employed transfer learning from ImageNet trained models to reduce and prevent overfitting.

1.3 Objectives

Our main objective is to create a model that can differentiate between the audio file of a cough from a person with COVID-19 compared to a cough from a person without COVID-19 with a high degree of sensitivity and specificity. The initial model receives cough samples as the sole input and produces a sigmoid output for the classification.

1.4 Data Science Pipeline

The following subsections outline our approach to achieving our project's objectives. We have structured the process in accordance with the data science pipeline in order to provide an organized format in which to work.

Datasets

The current project utilizes two publicly available datasets for COVID-19 cough detection: the Coswara Project [16] and the Coughvid Project [33]. As our primary dataset, the Coswara dataset is used to create a cough detection model.

In contrast, the Coughvid dataset is used to explore whether the cough detection model trained, validated, and tested on the Coswara dataset can be generalized to coughs collected through other platforms.

Data Wrangling

In order to clean our data, we remove subjects with erroneous audio files or ambiguous COVID-19 status. Subsequently, we pre-process our desired audio in order to feed it into our model. This pre-processing step consists of trimming the audio to 5 seconds and removing any silence seen in the audio.

For the Coswara dataset, we divide the data into train, validation, and test sets. We randomly selected 64% of the positive and negative samples as the training set, 16% as the validation set, and the remaining 20% as the test set. The relative distribution of positive and negative samples is preserved in each set.

Exploratory Data Analysis

To understand the basic demographic distribution of the datasets, we visualize the age distribution, gender distribution, and health status distribution among all participants for each dataset.

Additionally, visualizing and understanding the impact of the given health descriptions can help us remove the irrelevant descriptors for the individuals that included them. For example, it is important to distinguish the coughs of a healthy person, a person with respiratory diseases, and a person suffering from COVID-19. Therefore, we plot the distribution of each COVID-19 symptom against COVID-19 status to determine the symptoms that are more closely associated with COVID-19. We further analyze the influence of mask usage on COVID-19 status to understand whether mask usage can help predict COVID-19 infection.

Modeling

The purpose of our model is to take in an audio sample of an individual coughing and produce a sigmoid output which is able to predict whether or not that individual is infected with COVID-19.

To begin, we built a Convolutional Neural Network model that takes a Mel Frequency Cepstrum spectrogram and gives a prediction of the COVID-19 infection status. Consisting of several convolution layers alongside max pooling and batch normalization steps, we train and evaluate this model architecture and plan to make further changes if needed.

Additionally, we employ a model that consists of a Convolutional Neural Network alongside a Single-layer Perceptron (SLP). The Convolutional Neural Network takes in a Mel Frequency Cepstrum representation of the audio signal as before, while the Single-layer Perceptron takes in a vector of handcrafted features obtained using the Surfboard API [34]. The final layers of these two models are then concatenated together and passed through a dense layer to produce a final output.

Validation

Within the Coswara dataset, the validation set consists of 16% of the data. Parameters, including learning rate, epochs, and filters of the models are tuned based on the validation set performance and the best model selected.

2 Data & Data Exploration

2.1 Dataset Description

Two datasets were used to train and verify our COVID-19 detection model: the Project Coswara dataset [16] created by the Indian Institute of Science (IISc) Bangalore, and the Project Coughvid dataset [33] collected by École polytechnique fédérale de Lausanne (EPFL). Both datasets are crowdsourcing datasets that collect subjects' demographic information, cough samples, as well as other relevant data through an online website; therefore, they are comparable with each other. Both datasets were used due to their differences in available data for each subject and the total subject number.

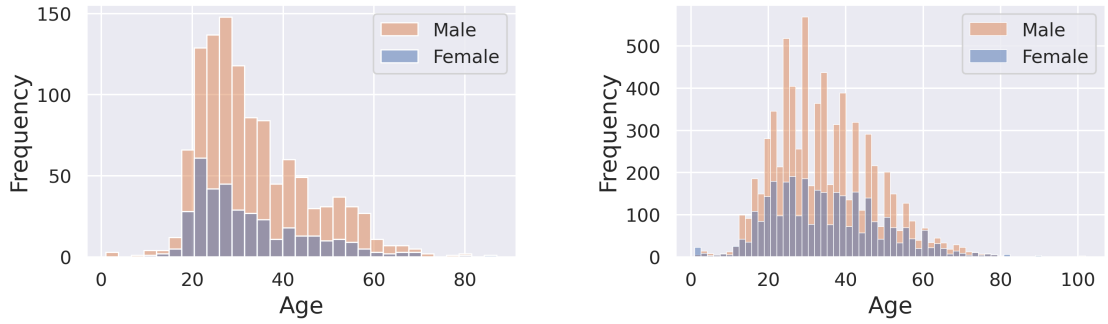
The Coswara dataset was created via the use of the Project Coswara website which had participants fill out a 5-7 minute survey of various health and demographic information as well as specified vocal sounds. The audio files, which are all in an uncompressed lossless format, consist of various breathing sounds, cough samples, vowels, and counting of numbers. Similarly, the Coughvid dataset was collected through the Project Coughvid website through which subjects recorded a cough sample and indicated their COVID-19 status, the presence or absence of other respiratory conditions and symptoms, age, and gender. The range of data collected by Coughvid is considerably limited in comparison to Project Coswara, but it contains a much larger dataset. Table 1 shows the information collected from the two datasets.

	Coswara	Coughvid
Audio samples	cough (heavy, shallow), breathing (deep, shallow), counting (fast, normal), vowels (a, e, o)	cough
Core demographic information	age, gender, residence area	age, gender
COVID-19 status	healthy, unidentified respiratory illness, COVID-19 exposed, COVID-19 positive (asymptomatic, mild, moderate), COVID-19 recovered	healthy, symptomatic, COVID-19 positive
COVID-19 test status	✓ (missing in many subjects)	N/A
Using mask	✓	N/A
Symptoms	cough, diarrhea, breathing difficulties, sore throat, fever, fatigue, muscle pain, loss of smell & taste	fever or muscle pain
Other health conditions	asthma, smoker, hypertension, cold, diabetes, Ischemic Heart Disease, lung disease, pneumonia	respiratory conditions
Other information	English proficiency, returning user	expert labels

Table 1: Comparison of Information Collected in Coswara and Coughvid Datasets

2.2 Metadata Description

The Coswara dataset (released as of February 6, 2021) includes 1503 participants. The subjects' ages ranged between 1 and 87 years old (mean 33.16). The Coughvid dataset (released as of September 24, 2020) includes 20072 samples, with an age range of 1 to 102 (mean 34.60). Both datasets consisted of predominantly male participants (1137, 75.6% in Coswara; 7226, 63.9% in Coughvid). It is important to note that many subjects in the Coughvid dataset lack critical COVID-19 labels or basic demographic information. Figure 1 shows the age distribution of the subjects in each of the datesets. The age distribution for both genders are right skewed in both datasets, but the age distribution in the Coughvid dataset is less skewed compared to that of the Coswara dataset.



(a) Coswara

(b) Coughvid

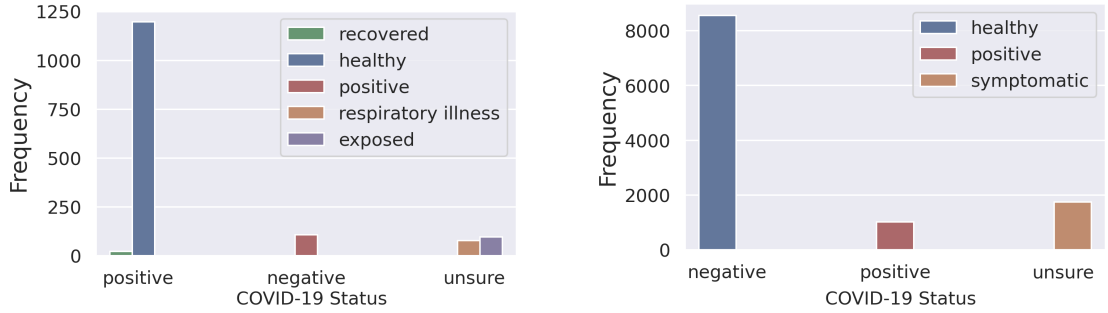
Figure 1: Distribution of Age By Gender

The datasets are also imbalanced in COVID-19 statuses (Figure 2). For the Coswara dataset, the majority of the subjects (1198; 79.7%) are healthy individuals, and the rest includes subjects who tested positive for COVID-19, those who were exposed but not yet diagnosed, those who recovered from COVID-19, and those with non-COVID-19 respiratory diseases. Although the original report of the dataset uses binary categorization of healthy vs. unhealthy [16], for this project, we group them based on the COVID-19 status. Both healthy and recovered subjects are grouped into COVID-19 negative, whereas subjects who had been exposed to COVID-19 but had yet to receive a definite diagnosis, as well as subjects with unidentified respiratory illness, are grouped into COVID-19 unsure category and discarded for the current study. The remaining 1329 subjects consist of 1221 COVID-19 negative patients and 108 COVID-19 positive ones. The asymmetrical distribution is also observed in the Coughvid dataset, with 8562 healthy COVID-19 negative samples and 1010 COVID-19 positive samples. The 1742 symptomatic subjects whose COVID-19 status were ambiguous are removed due to similar reasoning used in cleaning the Coswara datset.

As our primary data set, the Coswara dataset [16], is relatively imbalanced in regards to the higher proportion of negative cases to positive cases (1221 to 108 respectively) and has a small overall size (less than 1400 participants), we

address the class imbalance problem in our model by penalizing the false negatives more. We train, validate, and test our model on the Coswara dataset to obtain our main results.

As we are interested in seeing whether our model can be applied to other datasets, we include the Coughvid dataset [33] to further explore the performance of our model on a larger dataset collected through a different online platform once the model has been trained, validated, and tested on the Coswara dataset.



(a) Coswara (b) Coughvid

Figure 2: Distribution of COVID-19 Status

2.3 Data Wrangling

First, we removed all subjects with unsure COVID-19 labels, as described above, and those who have erroneous audio files. After extracting all the raw data from the set, we organize them in the following format, in which all audio files are renamed to include both the unique participant ID as well as the name of the type of sound they contain.

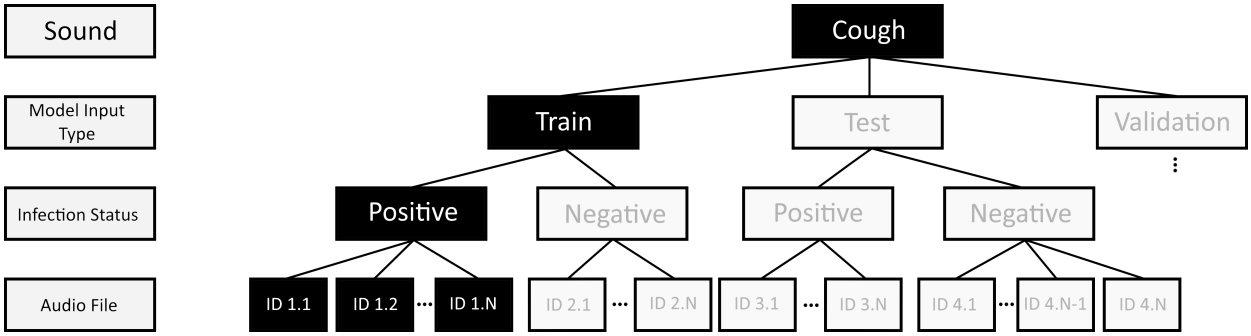


Figure 3: Reorganization of Data

As the Coswara dataset includes a compendium of nine sounds for each participant, we first reorganize all the audio files into each type of sound. Two of these nine audio files from the Coswara participants are cough samples, labeled as 'shallow cough' and 'heavy cough.' For the initial model focusing purely on cough samples, the two types of coughs are combined together to create the inputs. Subsequently, we split both datasets into positive and negative cases and divide them proportionally into training, validation, and test sets with 64%, 16% and 20% of the full cough sample set, respectively.

After dividing up our data, we employ pre-processing to standardize our audio. As most of our cough samples are two to three seconds long at most, we first crop out the beginning and ending silences by employing a reference decibel value to define the lowest threshold of sound. Anything below that is considered as silence. This is a common technique used in audio processing to remove extraneous noise and low-volume sounds [35]. After doing this, we standardize the length of all our audio files to five seconds via either cropping samples which are too long or by flipping and repeating samples which are too short.

2.4 Exploratory Data Analysis

First, we want to examine whether certain symptoms would be good predictors of a COVID-19 positive status. Figure 4 shows the distribution of different symptoms in COVID-19 negative vs. positive populations in the Coswara dataset. All symptoms, except for diarrhoea have a higher rate of appearing in COVID-19 positive subjects compared to the COVID negative population. Loss of smell and taste, fatigue, and breathing difficulties seem to be especially highly associated with a COVID-19 positive status.

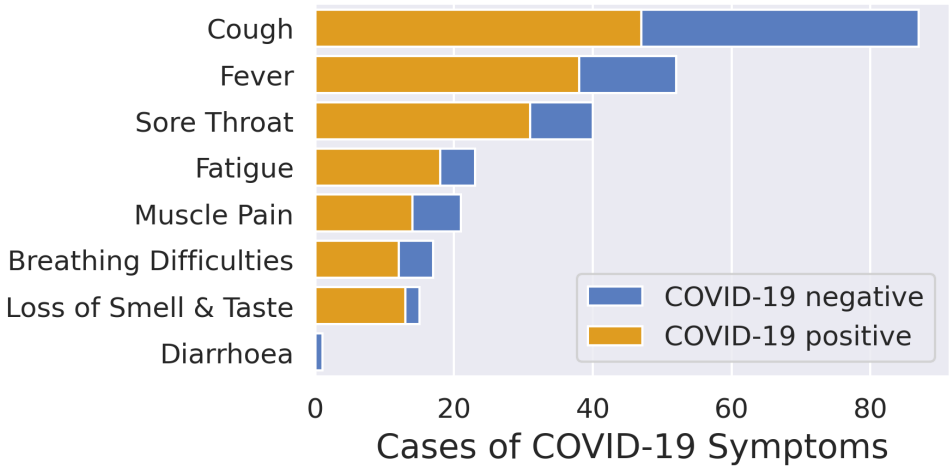


Figure 4: Distribution of COVID Related Symptoms

Additionally, we analyzed the effect of mask usage on COVID-19 status within the Coswara dataset. Figure 5 shows the distribution of mask usage and COVID-19 status. It is surprising that mask users has a higher likelihood for COVID-19 positive (42% positive) than people who do not use mask (23%). However, it is worth noting that only 410 (27%) of the subjects reported their mask usage status. Specifically, a large number of healthy subjects did not report their mask usage status. It is possible that people who are wearing masks and staying healthy did not feel the need of entering such information. Subjects who are not consistently wearing masks or are primarily staying at home might also have trouble answering this question. We further investigated the relationship between age and COVID-19 status by using logistic regression. As shown in Figure 6, for the population without regular mask usage, age has an insignificant negative trend ($\beta = -0.018, p = 0.112$) on the predicted COVID positive cases, with a 2% decrease in the predicted COVID-19 positive possibility for every 10 years older. This seemingly counter-intuitive pattern could be explained by the fact that younger people tend to be more socially active and therefore have more chance of being exposed to the virus. Note that the result is also influenced by the two outliers who are very young and tested COVID-19 positive. On the contrary, for the population that uses masks consistently, age has an insignificant positive trend ($\beta = 0.033, p = 0.084$) on the prediction of COVID-19 positive status. For each ten year increase in age, the probability of being detected as COVID-19 positive increases 7%. This pattern more closely resembles the common knowledge that the aging population is more susceptible to the virus. The results here suggest an interaction between mask usage and age on COVID-19 status, although the effect might not be significant.

We can visually analyze our data by converting the audio files to spectrograms (details are given in the following section.) As an example, the MFCC spectrogram of the cough of a 57-year-old man from India who had tested positive for COVID-19 is shown below in Figure 7a. For comparison, the MFCC spectrogram of the cough of a healthy 56-year-old female from India is shown below in Figure 7b. This builds on the methods of Dunne et al's analysis [20].

2.5 Data Processing

We obtain spectrograms by converting the raw audio data to Mel Scale via a series of transforms. The Mel Scale is based on how humans perceive audio and gives more useful information than simple frequency content [36]. MFCCs in particular are a time and frequency representation of audio. We first take the Fourier Transform of 10-30ms triangular overlapping windows of an audio clip, adjusting the outputs to Mel scale using Equation 1.

$$m = 2595 \log_{10}(1 + f/700) = 1127 \ln(1 + f/700) \tag{1}$$

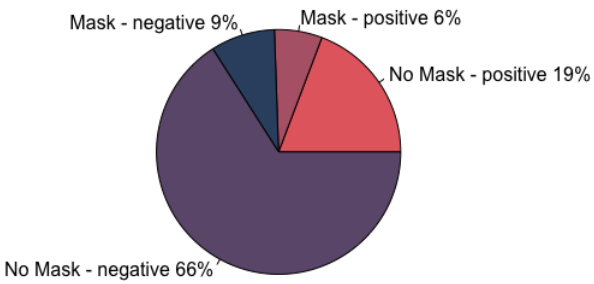


Figure 5: Pie Chart of Mask Usage and COVID-19 Status in Coswara

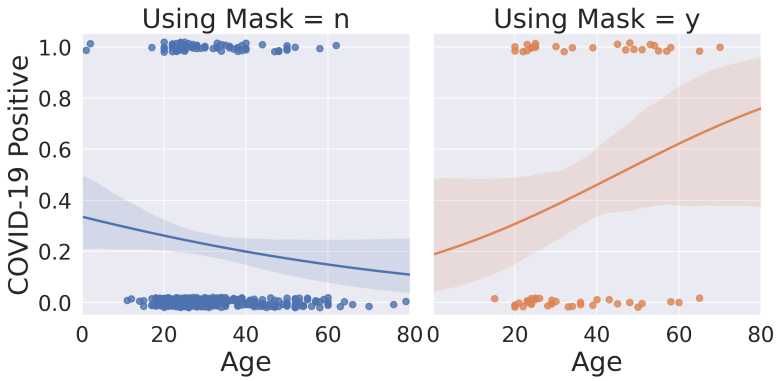
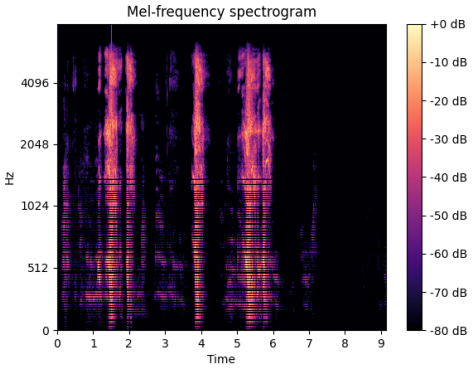
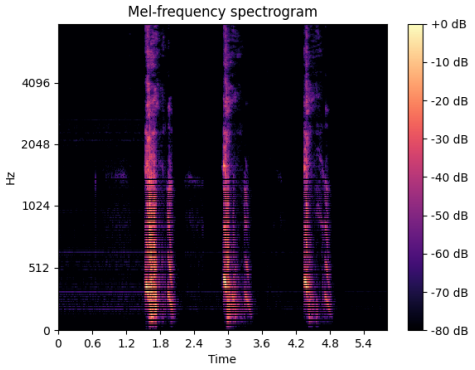


Figure 6: Logistic Regression between Age and Health By Mask Use



(a) Cough of a 57 year old man from India who had tested positive for Covid-19



(b) Cough of a healthy 56 year old female from India

Figure 7: Comparison of two coughing samples in our dataset

From there, a discrete cosine transform (a transform which represents a sequence of points in terms of a sum of cosine functions oscillating at various frequencies [37]) is taken and scaled. The MFCCs are the amplitudes of the resulting

spectrum. In each spectrogram, the x axis represents time, and the y axis represents frequency. The color or brightness on the graph represents the volume. As a result of the data pre-processing, all of the MFCC matrices are of the same dimension.

Additionally, we create a vector of handcrafted features that are extracted from the audio via the use of the *Surfboard* API [34]. The high level components extracted are as follows:

- 13 Components of MFCC: The first 13 components of the Mel-Frequency Cepstral Coefficients obtained from the short-term power spectrum , based on a linear cosine transform of the log power spectrum on a nonlinear Mel scale[17]
- Morlet Continuous Wavelet Transform: The frequency domain representation of the audio signal input after taking the Morlet Wavelet Transform [38]
  - The Morlet wavelet is a Gaussian-windowed complex exponential function defined by:

    $$\psi(t) = \exp(j\omega_0 t)\exp(-t^2/2),$$

    where ω_0 is the frequency. The continuous wavelet transform is obtained by taking the convolution of the original time series with a scaled and translated version of ψ. Unlike the Fourier transform, this procedure preserves information about the time location of features in the data [39].
- Log Melspectrogram: The log-scaled version of the power of the Mel Spectrogram, represented in decibels [40]
- Spectral Skewness: Measure for how much the shape of the spectrum below the center of gravity (the measure of how high the frequencies in a spectrum are on average) is different from the shape above the mean frequency [41]
- Spectral Kurtosis: The measure of how much the shape of the spectrum around the center of gravity is different from a Gaussian shape [41]
- Spectral Rolloff: The frequency below which a specified percentage of the total spectral energy lies [42]
- Root Mean Squared Energy: The root-mean-square of the magnitude of a short-time Fourier transform which provides the power of the signal [17]

The last four components as well as the MFCCs are extracted as time series. For these time series components, we also extract statistics that describe their distributions. These statistics include:

- Mean: The arithmetic average of the time series representation of the audio signal
- Standard Deviation: The measure of the amount of variation or dispersion of the time series values
- Skewness: Measure for how much the shape of the time series below the center of gravity (the measure of how high the amplitudes of a time series are on average) is different from the shape above the mean amplitude [41]
- Kurtosis: The measure of how much the shape of the time series around the center of gravity is different from a Gaussian shape [41]
- First Derivative Mean: The arithmetic average of the first derivative of the time series representation of the audio signal
- First Derivative Standard Deviation: The measure of the amount of variation or dispersion of the first derivative of the time series values
- First Derivative Skewness: Measure for how much the shape of the first derivative of the time series below the center of gravity (the measure of how high the amplitudes of the first derivative of the time series are on average) is different from the shape above the mean amplitude of the first derivative [41]
- First Derivative Kurtosis: The measure of how much the shape of the first derivative of the time series around the center of gravity is different from a Gaussian shape [41]

By compiling these features listed above, we end up with a handcrafted feature vector that captures the distributions of 4 different time series and a few other components not captured in the MFCC spectrograms. The handcrafted feature vectors are then reduced using Principal Component Analysis (PCA) [43] to have 256 dimensions due to the fact that they contain a high proportion of the variance.

3 Modeling

Using the MFCC spectrogram and the handcrafted feature vector as inputs to a Convolutional Neural Network and a Single-Layer Perceptron (SLP), respectively, we are able to implement two parallel models whose outputs are combined to create a prediction of COVID-19 status. This is in addition to the second model which consists solely of the Convolutional Neural Network. We compare both of these models to observe which works better.

3.1 Model Architecture

3.1.1 Convolutional Neural Network

The Convolutional Neural Network takes in the Mel-Frequency Cepstrum Spectrogram and passes it through three convolutional layers along with max pooling after the first. After each convolutional layer, batch normalization is applied, along with the ReLu activation function. After this, the data is flattened. This process functions as feature extraction via convolution while also training the model to learn what features to extract [44]. A diagram detailing the overall structure of the Convolutional Neural Network is shown below in Figure 8:

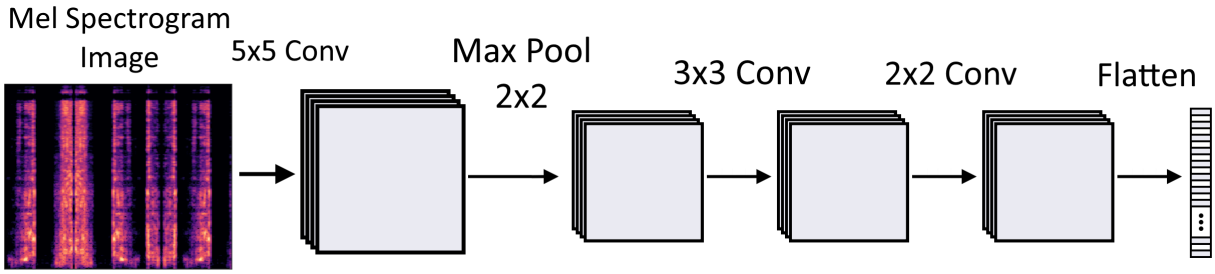


Figure 8: Convolutional Neural Network

While not explicitly shown in each case, every 2D Convolution filter is followed by a batch normalization layer, then a activation (ReLu) layer, and then a max-pooling layer.

3.1.2 Single-Layer Perceptron

Alongside the Convolutional Network, we implement a Single-Layer Perceptron which takes in a handcrafted feature vector as an input. This vector undergoes dimensionality-reduction via Principal Component Analysis (PCA) during the pre-processing stage, and is then passed into a dense layer. The layer of the Perceptron uses a dropout rate of 0.5 alongside L1 regularization. The flow of the PCA reduced feature vector (of dimension 256) to the dense layer (of size 64) to the output is shown below in Figure 9:

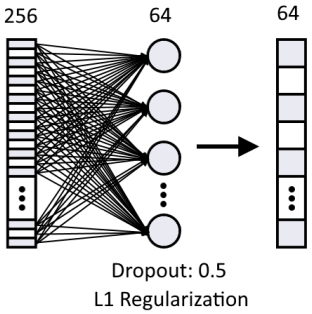


Figure 9: Handcrafted Features Through Perceptron

3.1.3 Combined Model

The nodes of the two models are concatenated and densely connected to a hidden layer. All the outputs of this layer are combined as inputs to the final node which uses a sigmoid activation function to produce a probability that represents how likely the initial input was from a COVID-19 positive person. This probability is then rounded to either '0' or '1'

using a threshold determined by the optimal point in the precision-recall curve or sensitivity-specificity curve. The final combination of the two models is shown below in Figure 10:

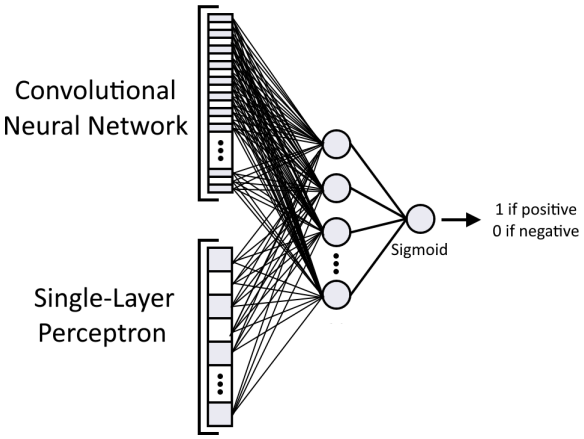


Figure 10: Combined Model

3.1.4 Transfer Learning from DenseNet

After obtaining our initial results with the CNN and CNN+SLP model, we looked into taking this transfer learning approach. From the applications API within Keras, we used the DenseNet model [28] with the preloaded weights trained on the ImageNet dataset. We modified the spectrogram inputs to 3-channel images by putting each channel to be the same as the original image. We trained the last 6 convolutional blocks out 15 that corresponded to the last 12 convolutional layers and added a dense layer with a sigmoidal activation function at the end which predicts COVID-19 status. During initial exploratory analysis, we also experimented transfer learning with ResNet [45] and VGG [26], both trained on ImageNet, but DenseNet showed the most promising and satisfactory result.

3.2 Training and Validation

As mentioned in the previous section, we divided the Coswara dataset into a training set, validation set, and test set containing 64%, 16%, and 20% of the full dataset, respectively. We train the models on the training set and report final results on the test set. We use the validation set during the training process in order to tune the model hyper-parameters and to detect possible over-fitting. Since the Coswara dataset suffers from a class imbalance problem with around 11 times more COVID-19 negative samples compared to COVID-19 positive samples, a class weight of approximately 11:1 is applied to put more importance on the underrepresented COVID-19 positive samples.

During the training process, we use binary cross-entropy as the loss function. More precisely, the loss function is given by the following formula:

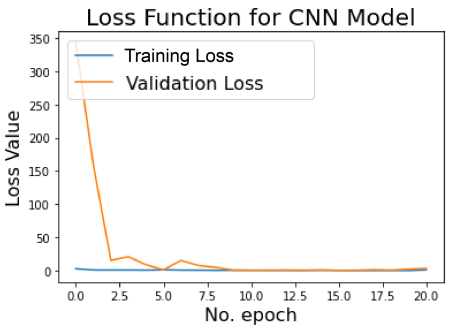
$$-\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1-y_i)\log(1-p_i), \tag{2}$$

where N is the total sample size, y_i is the true class label of the i^{th} observation (either 0 or 1), and p_i is the predicted probability of the i^{th} observation (a number between 0 and 1.) The binary cross-entropy loss function is commonly used for training deep neural networks on classification problems [46]. We use the Adam algorithm, which is a modified version of the classical stochastic gradient descent algorithm, to optimize the loss function [47]. Finally, early stoppage is used during the training process to prevent overfitting. The resulting loss curves for the different models are shown in Figures 11, 12, and 13 below.

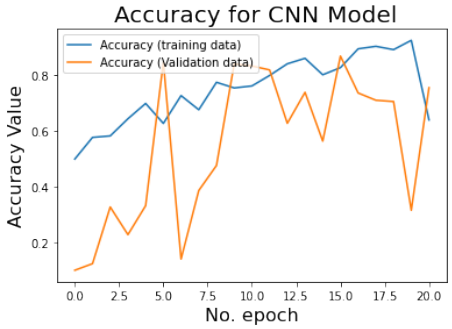
4 Results

4.1 Metrics

We report a number of metrics to evaluate the model performance on the test set. Due to the imbalanced nature of the dataset, we report the sensitivity-specificity curve and the precision-recall curve, which are common in clinical settings [48]. Sensitivity, also known as recall, refers to the true positive rate, which measures the proportion of positives that

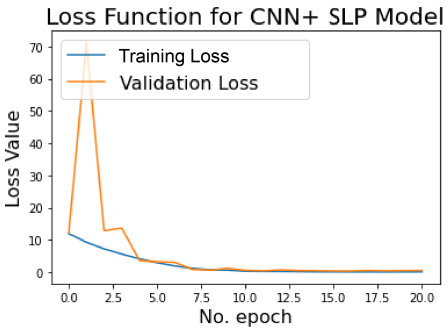


(a) Loss for both the test and validation sets while training

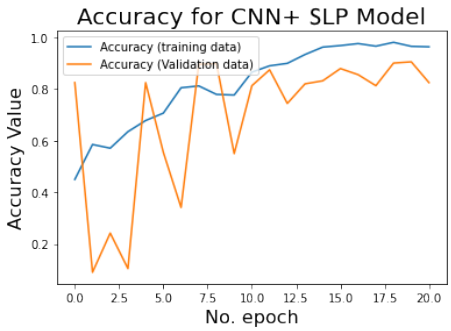


(b) Accuracy for both the test and validation sets while training

Figure 11: Training and Validation Results of the CNN model

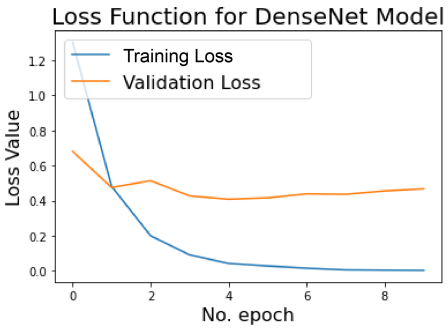


(a) Loss for both the test and validation sets while training

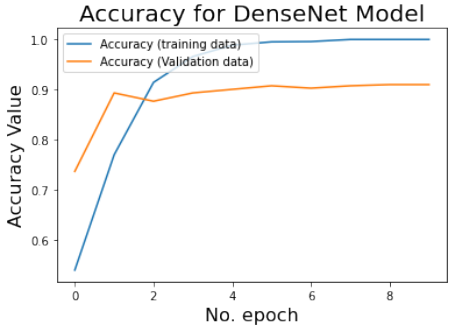


(b) Accuracy for both the test and validation sets while training

Figure 12: Training and Validation Results of the CNN + SLP model



(a) Loss for both the test and validation sets while training



(b) Accuracy for both the test and validation sets while training

Figure 13: Training and Validation Results of the DenseNet model

are correctly classified. On the other hand, specificity refers to the true negative rate, which measures the proportion of negatives that are correctly classified. Finally, precision, also known as the positive predictive value, is the ratio of true positives to the sum of true positives and false positives.

The sensitivity-specificity curve is obtained by plotting the sensitivity and specificity values at varying classification thresholds, where sensitivity is plotted on the y-axis and specificity on the x-axis. The precision-recall curve is obtained in a similar way. The Area Under the Curve (AUC) for each of these two curves summarizes the performance of the

model across different thresholds, where an AUC of 1 represents a perfectly skilled model. The AUC value is a useful metric for comparing the performance of different models, because it does not depend on any particular choice of classification threshold.

4.2 Performance

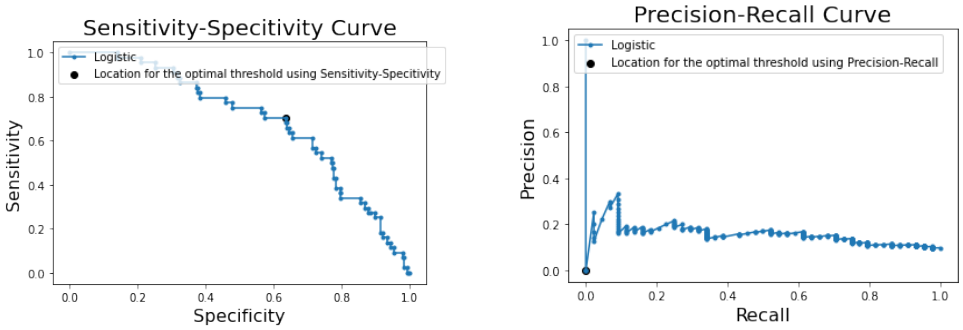
After training these models, we plot both the Precision-Recall curve and Sensitivity-Specificity curve to evaluate model performance and compute optimal thresholds for classification. To find the threshold that yields the best balance between precision and recall, we seek to maximize the F-measure, which is the harmonic mean of precision and recall:

$$\text{F-Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

Similarly, to find the threshold that yields the best balance between sensitivity and specificity, we seek to maximize the Geometric Mean or G-Mean, which is given by the following formula:

$$\text{G-Mean} = \sqrt{\text{Sensitivity} * \text{Specificity}}. \tag{4}$$

Figures 14-19 show the sensitivity-specificity curve, precision-recall curve, confusion matrix at the optimal threshold for sensitivity-specificity, and confusion matrix at the optimal threshold for precision-recall for each of the three models we employed. Across the three models, it is apparent that thresholds obtained through sensitivity-specificity provide better classification performance, with tolerable false positive and false negative rates. In contrast, thresholds obtained through precision-recall minimize false positives but result in a large number of false negatives and very few (if any) true positives. This is a particularly serious problem because the goal of our model is to identify COVID-19 infections and help contain the spread of the disease. Therefore, we use the threshold which maximizes the G-mean of sensitivity and specificity for classification.



(a) Sensitivity-Specificity of CNN model

(b) Precision-Recall of CNN model

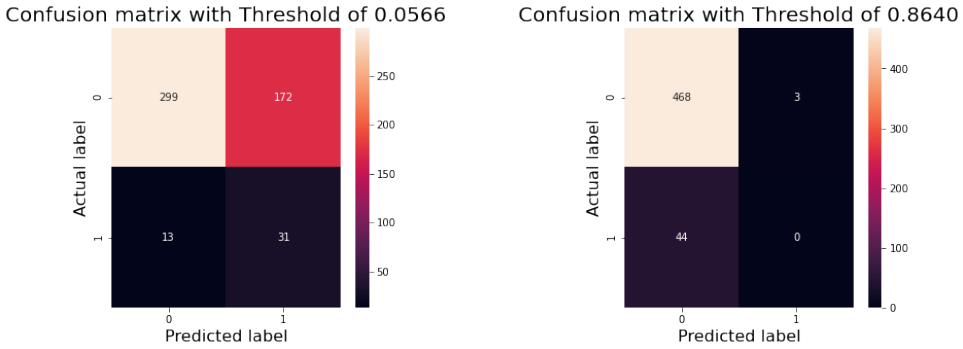
Figure 14: Sensitivity-Specificity and Precision-Recall curves of the CNN model

The performance of the three models with the selected threshold is displayed in Table 2. In particular, we report the sensitivity, specificity, precision-recall AUC, ROC-AUC, and accuracy. Although ROC-AUC and accuracy are less suitable for data with class imbalance, the two measures are reported here for ease of comparison with the past literature.

For the CNN model, at the optimal threshold, we obtain a sensitivity of 63% and a specificity of 70%. At the optimal threshold, we obtain a sensitivity of 71% and specificity of 55% for the CNN+SLP model. Note that we get a high sensitivity with low specificity. This is not ideal as we have a high number of false negatives which would be problematic in the medical domain and lead to delayed treatment and spread of the disease. Transfer learning from the DenseNet results in the best result overall. With the optimal threshold, the DenseNet model classified cough samples with a sensitivity of 72% and specificity of 68%. Although AUC for Precision-Recall curve is low across the three models, the DenseNet model still showed the best performance of 0.24.

4.3 Generalizeability

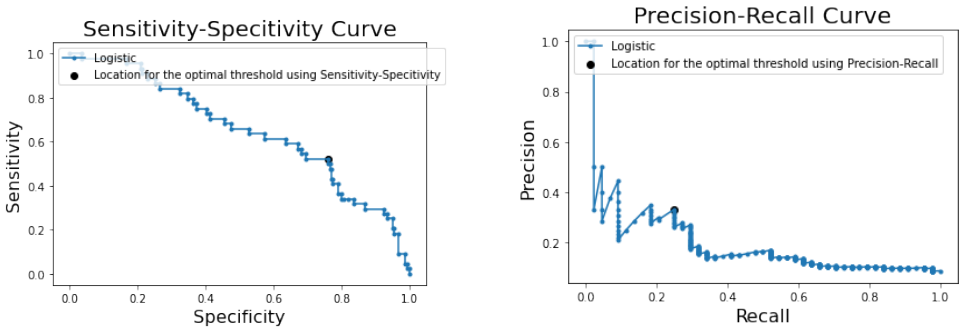
In order to test whether the model trained on the current Coswara dataset [16] can be generalized to cough samples collected through another similar online platform, we tested the performance on the Coughvid dataset [33] with DenseNet, the best performing model, trained on Coswara dataset. As shown in Figure 20 and Table 2, the model



(a) Confusion Matrix using threshold given by Sensitivity-Specificity curve

(b) Confusion Matrix using threshold given by Precision-Recall curve

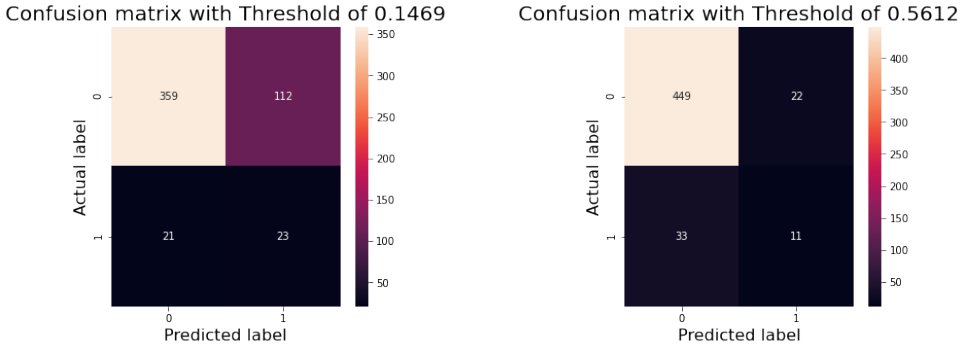
Figure 15: Confusion Matrices of the CNN model



(a) Sensitivity-Specificity of CNN + SLP model

(b) Precision-Recall of CNN + SLP model

Figure 16: Sensitivity-Specificity and Precision-Recall curves of the CNN + SLP model

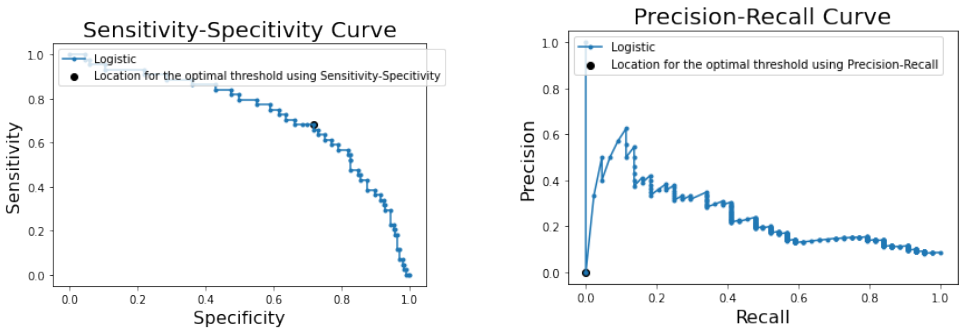


(a) Confusion Matrix using threshold given by Sensitivity-Specificity curve

(b) Confusion Matrix using threshold given by Precision Recall curve

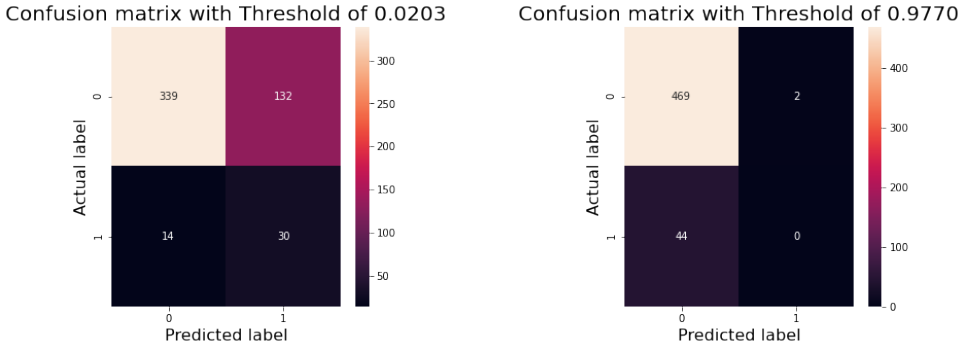
Figure 17: Confusion Matrices of the CNN + SLP model

obtained a sensitivity of 0.54 and specificity of 0.56, with a large number of both false positives and false negatives. The barely above chance performance suggests that the current model cannot be generalized to cough sample collected through a different platform. Additionally, the discrepancy in performance might also be attributed to difference in sample size, with around 400 samples in the test set of Coswara, and around 10,000 samples in Coughvid.



(a) Sensitivity-Specificity of DenseNet model (b) Precision-Recall of DenseNet model

Figure 18: Sensitivity-Specificity and Precision-Recall curves of the DenseNet model



(a) Confusion Matrix using threshold given by Sensitivity-Specificity curve

(b) Confusion Matrix using threshold given by Precision Recall curve

Figure 19: Confusion Matrices of the DenseNet model

Model	Sensitivity	Specificity	Precision-Recall AUC	ROC AUC	Accuracy
CNN	0.63	0.70	0.16	0.65	0.64
CNN+SLP	0.71	0.55	0.19	0.69	0.74
DenseNet	0.72	0.68	0.24	0.73	0.72
DenseNet - Coughvid	0.54	0.56	-	0.54	0.55

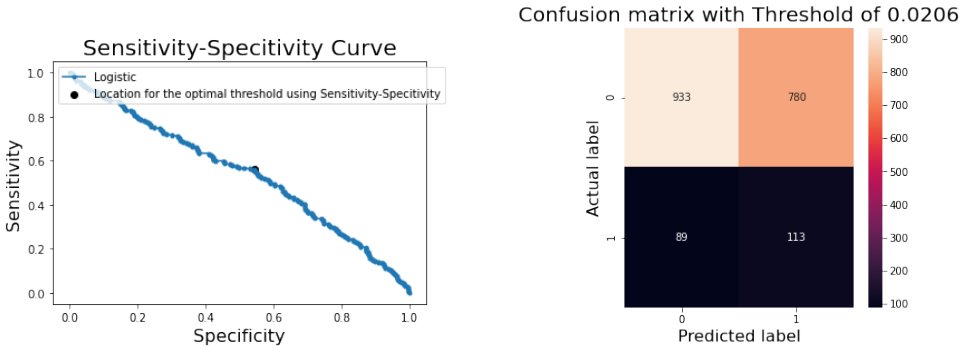
Table 2: Model Performance

5 Discussion

We prefer to design our model to tolerate a high number of false positives and to have a low number of false negatives. This is because it is better to diagnose someone who is healthy as sick rather than to classify people who are actually sick as healthy. If a healthy person is diagnosed as sick, then they potentially need to self-isolate and are subject to increased stressed from their diagnostic along with false self-perception of vulnerability [49]. Comparatively, a false negative potentially results in said case to be at risk of illness, death and spreading COVID-19 to others around them. Though the potential consequences of false positive cases aren't ideal, the health risk of a false negative far outweigh those of the false positive [50]. This type of decision is used similarly for other medical conditions such as cancer diagnosis [51].

5.1 Limitations of Datasets and process

While we were able to create a working model, our results are far from ideal. A key factor that likely influence the model performance is the fidelity of the data set used. The recordings, upon further inspection, do not have consistent audio quality. In some, there are extraneous noise such as a chair squeaking on the floor or clothes rustling. Additionally,



(a) Sensitivity-Specificity curve

(b) Confusion Matrix using threshold given by Sensitivity-Specificity curve

Figure 20: DenseNet model generalized to Coughvid Dataset

the volume, likely linked to how far or close the person was from the microphone, varies from recording to recording. A variation in the distance from the microphone will change how the microphone picks up audio due to the environment in which the recording takes place [52]. Furthermore, COVID-19 labels in the current datasets are self-reported, and therefore do not necessarily reflect the actual COVID-19 status, especially as many COVID-19 infections are asymptomatic [53]. These, in addition to other inconsistencies, created an inconsistent dataset which likely contributed to the difficulty in distinguishing the differences between coughs.

5.2 Comparison to Previous Literature

Some previous literature that have reported promising results also have several flaws that should be noted. First, almost all past efforts in COVID-19 cough detection have suffered from extremely small sample sizes, with only 16-70 samples in the COVID-19 positive group [19][20][21]. In addition, as COVID-19 datasets usually have a very significant class imbalance with very limited COVID-19 positive data, some past works have addressed this problem by manually creating a subset by downsampling the majority class. However, they did so without reporting confidence intervals for different sampling schemes [20][17]. Furthermore, some works have reported accuracy and AUC_ROC measures even though the testing dataset is imbalanced [20][19], which can be misleading as they are more suitable for balanced datasets [48].

Some additional strategies adopted by past efforts might also contribute to the better performance. Imran and colleagues [21] trained and applied a cough detection CNN in which only samples that were classified as cough were used for COVID-19 classification. This virtually discarded all audios with poor recording quality or extensive environmental noise. Using a XGBoost (eXtreme Gradient Boosting) classifier, Mouawad and collegues achieved performance of mean F_1-score of 91% and 89% for COVID-19 detection from cough and vowel respectively, by employing a novel informative undersampling method based on Information Rate (IR) to tackle the class imbalance problem after finding out that oversampling did not produce desirable result [54].

6 Conclusion

While the methodology of creating a classification model using a CNN on Mel Spectrograms of the cough samples in parallel with an SLP using hand-crafted features was viable in implementation, the quality of the model was severely limited by the variations and inconsistencies in the dataset. Specifically, this was due to extraneous noise, large variation in recording methods and using labels that were self reported rather than a conclusive test for each person. Despite this, we still were able produce a working model with above chance performance using transfer learning from the DenseNet model pre-trained on Imagenet. The current model could be further improved by acquiring a larger dataset with accurate COVID-19 testing results, filtering audio files to discard samples with bad quality, and experimenting additional ways to address the class imbalance problem.

References

[1] M. F. Pullen, C. P. Skipper, K. H. Hullsiek, A. S. Bangdiwala, K. A. Pastick, E. C. Okafor, S. M. Lofgren, R. Rajasingham, N. W. Engen, A. Galdys, D. A. Williams, M. Abassi, and D. R. Boulware, "Symptoms of COVID-19 Outpatients in the United States," *Open Forum Infectious Diseases*, vol. 7, 06 2020. ofaa271.

[2] A. Krishnan, J. P. Hamilton, S. A. Alqahtani, and T. A. Woreta, "Covid-19: An overview and a clinical update," *World Journal of Clinical Cases*, vol. 9, no. 1, p. 8, 2021.

[3] World Health Organization, "WHO Coronavirus Disease (COVID-19) Dashboard." `https://covid19.who.int/`. Accessed on 27 March 2021.

[4] M. J. Mina and K. G. Andersen, "Covid-19 testing: One size does not fit all," *Science*, vol. 371, no. 6525, pp. 126–127, 2021.

[5] O. Albahri, A. Zaidan, A. Albahri, B. Zaidan, K. H. Abdulkareem, Z. Al-Qaysi, A. Alamoodi, A. Aleesa, M. Chyad, R. Alesa, *et al.*, "Systematic review of artificial intelligence techniques in the detection and classification of covid-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects," *Journal of infection and public health*, 2020.

[6] T. Oikarinen, K. Srinivasan, O. Meisner, J. B. Hyman, S. Parmar, R. Desimone, R. Landman, and G. Feng, "Deep convolutional network for animal sound classification and source attribution using dual audio recordings," *The Journal of the Acoustical Society of America*, 2018.

[7] E. Sasmaz and F. B. Tek, "Animal sound classification using a convolutional neural network," *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, 2018.

[8] I. Mierswa and K. Morik, "Automatic feature extraction for classifying audio data," *Machine Learning*, vol. 58, no. 2-3, p. 127–149, 2005.

[9] H. Chatrzarrin, A. Arcelus, R. Goubran, and F. Knoefel, "Feature extraction for the differentiation of dry and wet cough sounds," *2011 IEEE International Symposium on Medical Measurements and Applications*, 2011.

[10] J. Amoh and K. Odame, "Deep neural networks for identifying cough sounds," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 5, p. 1003–1011, 2016.

[11] R. X. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas, "A cough-based algorithm for automatic diagnosis of pertussis," *PLOS ONE*, vol. 11, no. 9, 2016.

[12] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997.

[13] S. Young, P. Woodland, and W. Byrne, "Spontaneous speech recognition for the credit card corpus using the htk toolkit," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, p. 615–621, 1994.

[14] Y. Wang, M. Hu, Y. Zhou, Q. Li, N. Yao, G. Zhai, X. P. Zhang, and X. Yang, "Unobtrusive and automatic classification of multiple people's abnormal respiratory patterns in real time using deep neural network and depth camera," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8559–8571, 2020.

[15] B. W. Schuller, D. M. Schuller, K. Qian, J. Liu, H. Zheng, and X. Li, "Covid-19 and computer audition: An overview on what speech and sound analysis could contribute in the sars-cov-2 corona crisis," 2020.

[16] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, P. K. Ghosh, S. Ganapathy, *et al.*, "Coswara–a database of breathing, cough, and voice sounds for covid-19 diagnosis," *arXiv preprint arXiv:2005.10548*, 2020.

[17] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data," *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Jul 2020.

[18] G. Chaudhari, X. Jiang, A. Fakhry, A. Han, J. Xiao, S. Shen, and A. Khanzada, "Virufy: Global applicability of crowdsourced and clinical datasets for ai detection of covid-19 from cough," 2021.

[19] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. Schuller, "End-to-end convolutional neural network enables covid-19 detection from breath and cough audio: a pilot study," *BMJ Innovations*, vol. 7, no. 2, 2021.

[20] R. Dunne, T. Morris, S. Harper, and et al, "High accuracy classification of covid-19 coughs using mel-frequency cepstral coefficients and a convolutional neural network with a use case for smart home devices," *PREPRINT (Version 1) available at Research Square*, August 2020. (Accessed on 03/01/2021).

[21] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel, "Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, p. 100378, 2020.

[22] E. S. Olivas, J. D. M. Guerrero, M. Martinez-Sober, J. R. Magdalena-Benedito, L. Serrano, *et al.*, *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques: Algorithms, methods, and techniques*. IGI Global, 2009.

[23] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," 2014.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2015.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[29] M. Huh, P. Agrawal, and A. A. Efros, "What makes imagenet good for transfer learning?," *arXiv preprint arXiv:1608.08614*, 2016.

[30] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2661–2671, 2019.

[31] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[32] H. Ravishankar, P. Sudhakar, R. Venkataramani, S. Thiruvenkadam, P. Annangi, N. Babu, and V. Vaidya, "Understanding the mechanisms of deep transfer learning for medical images," in *Deep learning and data labeling for medical applications*, pp. 188–196, Springer, 2016.

[33] L. Orlandic, T. Teijeiro, and D. Atienza, "The coughvid crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms," *ArXiv*, vol. abs/2009.11644, 2020.

[34] R. Lenain, J. Weston, A. Shivkumar, and E. Fristed, "Surfboard: Audio feature extraction for modern machine learning," 2020.

[35] L. Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE T. Signal. Proces.*, vol. 56, pp. 1830–1839, 06 2008.

[36] P. Thaine and G. Penn, "Extracting mel-frequency and bark-frequency cepstral coefficients from encrypted signals," *Interspeech 2019*, 2019.

[37] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. C-23, no. 1, pp. 90–93, 1974.

[38] R. Büssow, "An algorithm for the continuous morlet wavelet transform," *Mechanical Systems and Signal Processing*, vol. 21, no. 8, pp. 2970–2979, 2007.

[39] C. Torrence and G. P. Compo, "A Practical Guide to Wavelet Analysis.," *Bulletin of the American Meteorological Society*, vol. 79, pp. 61–78, Jan. 1998.

[40] A. Meghanani and A. Ramakrishnan, "An exploration of log-mel spectrogram and mfcc features for alzheimer's dementia recognition from spontaneous speech," 01 2021.

[41] "Spectrum: Get skewness," 2002.

[42] S. Tjoa, "Spectral features."

[43] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Phil. Trans. R. Soc.*, April 2016. (Accessed on 03/04/2021).

[44] R. Hyder, S. Ghaffarzadegan, Z. Feng, J. H. Hansen, and T. Hasan, "Acoustic scene classification using a cnn-supervector system trained with auditory and spectrogram image features.," in *Interspeech*, pp. 3073–3077, 2017.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[46] E. Gordon-Rodriguez, G. Loaiza-Ganem, G. Pleiss, and J. P. Cunningham, "Uses and abuses of the cross-entropy loss: case studies in modern deep learning," *arXiv preprint arXiv:2011.05231*, 2020.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[48] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.

[49] M. F. Luce and B. E. Kahn, "Avoidance or vigilance? the psychology of false-positive test results," *Journal of Consumer Research*, vol. 26, no. 3, pp. 242–259, 1999.

[50] B. Healy, A. Khan, H. Metezai, I. Blyth, and H. Asad, "The impact of false positive covid-19 results in an area of low prevalence," *Clinical Medicine*, vol. 21, no. 1, p. e54, 2021.

[51] . N. M. S. Dr. L. Daniel Maxim, Everest Consulting Associates, "Screening tests: a review with examples," *Taylor & Francis Open Select*, vol. 1, no. 3, pp. 3–6, 2014.

[52] I. R. Titze and W. S. Winholtz, "Effect of microphone type and placement on voice perturbation measurements," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 6, pp. 1177–1190, 1993.

[53] Z. Gao, Y. Xu, C. Sun, X. Wang, Y. Guo, S. Qiu, and K. Ma, "A systematic review of asymptomatic infections with covid-19," *Journal of Microbiology, Immunology and Infection*, 2020.

[54] P. Mouawad, T. Dubnov, and S. Dubnov, "Robust detection of covid-19 in cough sounds: Using recurrence dynamics and variable markov model," *Sn Computer Science*, vol. 2, no. 1, 2021.

Appendix

Fields in the Coswara Dataset

- "***id***": Id given to an individual when they submit their audio samples.
- "***a***": Age of the individual.
- "***covid_status***": The health status (e.g. : positive_mild, healthy, etc.) stated by the individual.
- "***ep***": Proficiency in English (y/n).
- "***g***": Gender stated by the individual.
- "***l_c***": Country the individual resides in.
- "***l_l***": City the individual resides in.
- "***l_s***": State the individual resides in.
- "***rU***": Whether or not the individual is a returning user (y/n).
- "***asthma***": Whether or not the individual has asthma (True/False).
- "***cough***": Whether or not the individual has a cough (True/False).
- "***smoker***": Whether or not the individual is a smoker (True/False).
- "***test_status***": Status of the individual's COVID Test (p: Positive, n: Negative, na: Not taken Test)
- "***ht***": Whether or not the individual is experiencing hypertension (True/False).
- "***cold***": Whether or not the individual is experiencing a cold (True/False).
- "***diabetes***": Whether or not the individual has been diagnosed with diabetes (True/False).
- "***diarrhoea***": Whether or not the individual has diarrheoa (True/False).
- "***um***": Whether or not the individual is using a mask (y/n).
- "***ihd***": Whether or not the individual is experiencing Ischemic Heart Disease (True/False).
- "***bd***": Whether or not the individual is experiencing breathing difficulties (True/False).
- "***st***": Whether or not the the individual has a sore throat (True/False).
- "***fever***": Whether or not the individual has a fever (True/False).
- "***ftg***": Whether or not the individual is experiencing fatigue (True/False).
- "***mp***": Whether or not the individual is experiencing muscle pain (True/False).
- "***loss_of_smell***": Whether or not the individual has a loss of smell & taste (True/False).
- "***cld***": Whether or not the individual has chronic lung disease (True/False).
- "***pneumonia***": Whether or not the individual has pneumonia (True/False).