DSCI 400 - COVID DETECTION THROUGH COUGH SAMPLES

INITIAL 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

March 13, 2021

1 Introduction

1.1 Background

The Coronavirus Disease 2019 (COVID-19) is undoubtedly 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. Its primary mode of transmission is through the respiratory tract by indirect means such as droplets and aerosols [1]. According to the most recently available statistics from the World Health Organization, COVID-19 has infected more than 100 million people and caused more than 2,554,000 deaths [2]. 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. 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. We believe that such methods may provide a method of diagnosing COVID-19 using coughing sounds.

1.2 Related Work

For audio classification, various feature extraction methods are used. For instance, Oikarinen et al. developed a method of classifying sounds from the common marmoset using a spectrogram representation of the audio data [3]. Similarly, Emre et al. used audio recordings to classify ten different animal species and were able to obtain a classification accuracy of 75% [4]. 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 [5].

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 [6]. 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 [7].

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 [8]. 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 [9]), 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 [10]. These coefficients are calculated by converting the audio 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 [11].

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 [12] [13]. Specifically, various crowdsourced cough datasets were collected and made available for researchers interested in tackling such problems [14][15][16]. Previous results obtained by researchers attempting to detect COVID-19 using patient cough samples are encouraging. Using a set of handcrafted features including duration, onset, and tempo, Brown et al. achieved an ROC-AUC of 80% with the simplistic logistic regression and 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% [17]. 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 [18]. 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.

1.3 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 accuracy. This model should be able to take multiple inputs, such as the age or sex of the person, and produce a Boolean output along with a measure of certainty for the classification. Our initial objectives are to extract data from a set of already determined audio files and create parameters and sorting methods to manage the important features of the audio in forming our model.

1.4 Data Science Pipeline

Given our objectives, we will follow the data science pipeline to successfully analyze and implement a model that can achieve our desired task.

Data Wrangling

We will begin the process of Data Wrangling by removing unnecessary labels that contain missing values as well as only extracting the cough samples for use in our model.

Exploratory Data Analysis

To begin the process of modeling, we want to understand the data we have. Therefore, we want to visualize the age distributions, gender distributions, and health status distribution among all participants.

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 will be important to distinguish between those who are not diagnosed with COVID-19 but coughed due to unrelated respiratory illnesses, those who are perfectly healthy and coughed, and those who have COVID-19 and coughed as a symptom.

Feature Extraction

For features, we need to determine the important features of the given audio samples. We will primarily use MFCC's from *librosa* [19] to preprocess our data and obtain meaningful features. However, to help reinforce and validate our model, we will extract additional feature vectors.

Handcrafted Features

To add some additional features, we plan to include the following:

- Duration: the duration of the cough recording after trimming silence for each audio sample.
- Onset: The salient change in a sound's pitch, energy or timbre. Hard onsets are characterized by sudden increases in energy.
- Tempo: the acoustic tempo feature for each recording. This measures the rate of beats which occur at regular intervals and will be used for its peak detection capabilities.
- Period: the main frequency of the signal.
- RMS Energy: The power of the signal.
- Spectral Centroid: the mean magnitude of the spectrogram (frame level).
- Zero-crossing: the rate of sign-changes of the audio signal.
- 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.
- Δ-MFCC: the temporal differential of the MFCC.
- Δ2-MFCC: the differential of the delta of the MFCC.

We can create our hand-extracted feature vector for each audio sample by compiling these features as a feature vector. For the RMS Energy, Spectral Centroid, and the variants of MFCCs, we will extract additional statistical features to capture the following distributions: mean, median, root-mean-square, maximum, minimum, 1st and 3rd quartile, interquartile range, standard deviation, skewness, and kurtosis.

Additional Features

In addition to handcrafted features, we plan to employ a VGGish [20] or a ResNet model [21] to extract audio features automatically. These will employ a pre-trained neural network to create a learned feature vector.

Finally, we obtain a handcrafted feature vector and a learned feature vector. Therefore, the final feature vector for an audio sample is the concatenation of the handcrafted features and the neural network-based features. These feature vectors will then be reduced using Principal Component Analysis (PCA) [22]. These feature vectors will be used alongside the MFCC images in our models to help predict the COVID-19 status of an individual.

Modeling

Given the raw MFCC image, our initial model to predict infectious coughing sounds will be a Convolutional Neural Network. This follows as CNNs have proven to be prominent image classifiers. [23]. Although we plan to experiment with our model in terms of the number of layers and types of kernels, we will first attempt to implement several variations of suggested CNNs in the literature regarding cough recognition such as the model in Figure 6 derived from the work done by Bales et al [24].

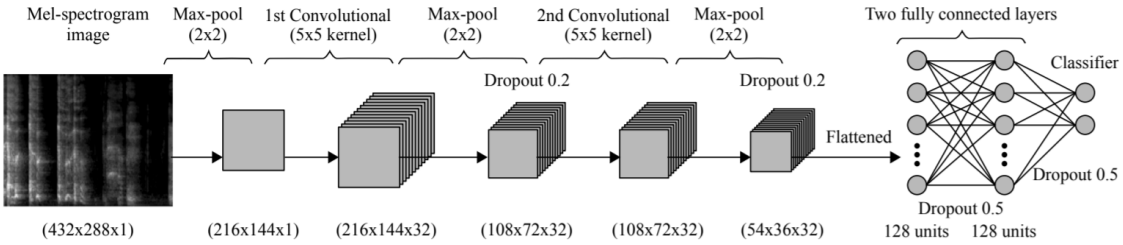

Figure 1: Example Convolutional Neural Network

We first reduce the complexity with a 2x2 max-pooing kernel then pass this into two convolutional layers with another max-pooling kernel after each. After flattening the learned features from a matrix to a vector, we pass them into one or more fully connected layers to then use a binary classifier to distinguish between infected or not infected audio samples.

Validation

Using the dimensionally reduced feature vectors, we will use a classification neural network model that can help reinforce the output of the CNN. By weighting both outputs, this can help predict the health status of the individual.

2 Data & Data Exploration

2.1 Data Set Description

We plan to primarily test and verify our COVID-19 detection model on the Project Coswara data set created by the Indian Institute of Science (IISc) Bangalore. This data set was created via the use of the Project Coswara website [14] 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. The dataset includes different descriptions of the individual's circumstances at the time they submitted their audio samples:

- Core demographic information: age, gender, residence area
- COVID status: healthy, unidentified respiratory illness, COVID exposed, COVID positive (asymptomatic, mild, moderate), COVID recovered
- COVID test status (missing in many subjects)
- Using mask
- Symptoms: cough, diarrhea, breathing difficulties, sore throat, fever, fatigue, muscle pain, loss of smell & taste
- Other diseases or conditions: asthma, smoker, hypertension, cold, diabetes, Ischemic Heart Disease, lung disease, pneumonia
- Other information: English proficiency, returning user

Because there are many missing values in the demographic and health data, as well as missing sounds in the audio data, we will need to clean the data and decide which sounds to use. However, our initial process will involve only using the cough samples to build our model as we will reserve the other sounds and labels for later testing and for improving our model.

Given the option, we also hope to use the COVID-19 Sounds App data provided by the University of Cambridge [15] to help improve and test our model with more samples. This dataset consists of 459 cough and breathing samples from 378 users. Within the dataset 62 users said they had tested positive for COVID-19 and contributed 141 cough and breathing samples to the dataset. The framework used to collect this data consists of a web-based app and an Android app. For each platform, the user was asked to provide their age and gender, as well as some brief medical history. Additionally, users were asked to provide a description of their symptoms (if any) and record respiratory sounds: coughing, breathing and the reading of a sentence provided. Finally, users were asked if they had been tested for COVID-19.

2.2 Metadata Description

The current dataset (released as of February 6, 2021) includes 1503 participants. The subjects' ages ranged between 1 and 87 years old (mean 33.16), and the subjects consisted of predominantly male participants (1137; 75.6%). Figure 2 shows the age distribution of the subjects, where the age distribution for both genders are right skewed.

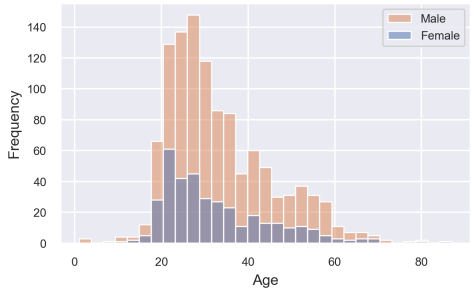

Figure 2: Age Density in Coswara Dataset By Gender

The dataset is also imbalanced across different health statuses (Figure 3a). The majority of the subjects (1198; 79.7%) are healthy individuals, and the rest included subjects who tested positive for COVID, those who were exposed but not yet diagnosed, those who recovered from COVID, and those with non-COVID respiratory diseases. Although the original report of the dataset used binary categorization of healthy vs. unhealthy [14], for this project, we grouped them based on the COVID status. Both healthy and recovered subjects were grouped into COVID 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, were grouped into COVID unsure category and discarded for the current study (Figure 3b). The remaining 1329 subjects included 1221 COVID negative records and 108 COVID positive ones.

(a) Health Status (b) COVID Status

Figure 3: Distribution of Health and COVID Status

2.3 Exploratory Data Analysis

First, we want to examine whether certain symptoms would be a good predictor of the COVID positive status. Figure 4 shows the distribution of different symptoms in COVID negative vs. positive population. All symptoms, except for diarrhea, have a higher rate of appearing in COVID 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 positive status.

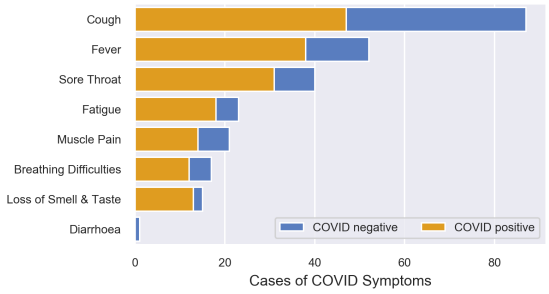

Figure 4: Distribution of COVID Related Symptoms

Furthermore, we analyzed the effect of mask usage on the relationship between age and COVID status using logistic regression. As shown in Figure 5, for the population without regular mask usage, age has a marginally significant ($p = 0.058$) negative effect on the predicted COVID positive cases, with a 2% decrease in the predicted COVID 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 positive. On the contrary, for the population that uses masks consistently, age has a marginally significant positive effect ($p = 0.084$) on the prediction of COVID positive status. For each 10 year increase in age, the possibility of being detected as COVID positive increases 7%. This pattern more closely resembles the common knowledge that the aging population is more susceptible to the virus.

To examine the audio in our data, we have decided to look at the sound samples in the frequency spectrum using a spectrogram. From this, we can examine features as previously mentioned. As an example, a mel spectrogram of the cough of a 57-year-old man from India who had tested positive for COVID-19 from our data set is shown below in Figure 6. This builds on the methods of Dunne et al's analysis [17]. For comparison, a spectrogram of the cough of a healthy 56-year-old female (who is also from India) is shown below in Figure 7.

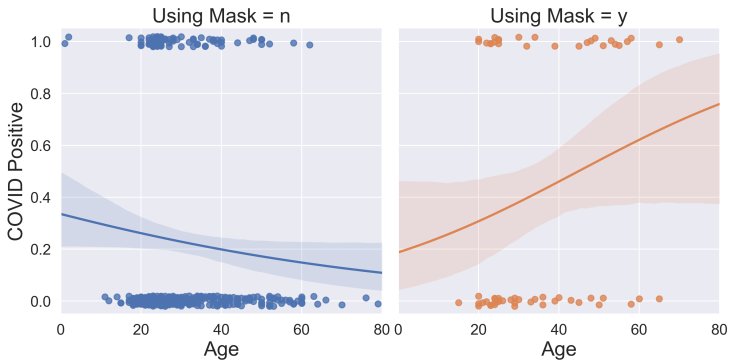

Figure 5: Logistic Regression between Age and Health By Mask Use

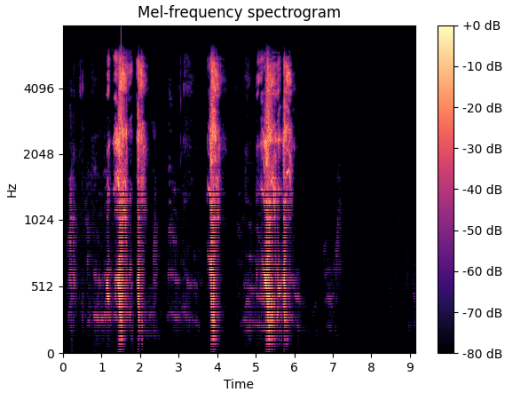

Figure 6: Cough of a 57 year old man from India who had tested positive for Covid-19

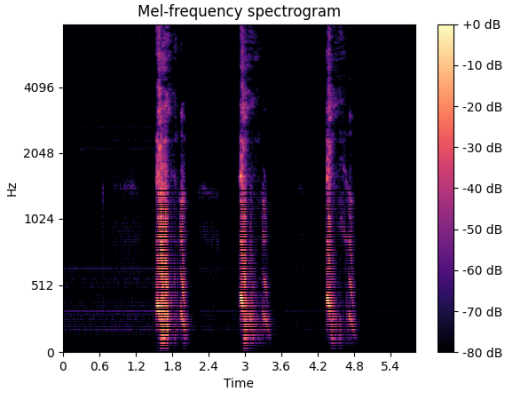

Figure 7: Cough of a healthy 56 year old female from India

References

[1] 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.

[2] World Health Organization, "WHO Coronavirus Disease (COVID-19) Dashboard." `https://covid19.who.int/`. Accessed on 04 March 2021.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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).

[18] 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.

[19] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," *Proceedings of the 14th Python in Science Conference*, p. 18–24, 2015. Austin, TX.

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

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

[22] 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).

[23] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *Journal of Big Data*, vol. 6, Dec 2019.

[24] C. Bales, M. Nabeel, C. N. John, U. Masood, H. N. Qureshi, H. Farooq, I. Posokhova, and A. Imran, "Can machine learning be used to recognize and diagnose coughs?," 2020.

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).