Data Science Project: AttackonStats

Aryan Chaudhary	Shubham Dattatray Patil	Vimal Kirti Singh
MT22019	MT22125	MT22089

1. Dataset

Link: https://zenodo.org/record/4048312/files/public_dataset.zip?download=1

About dataset:

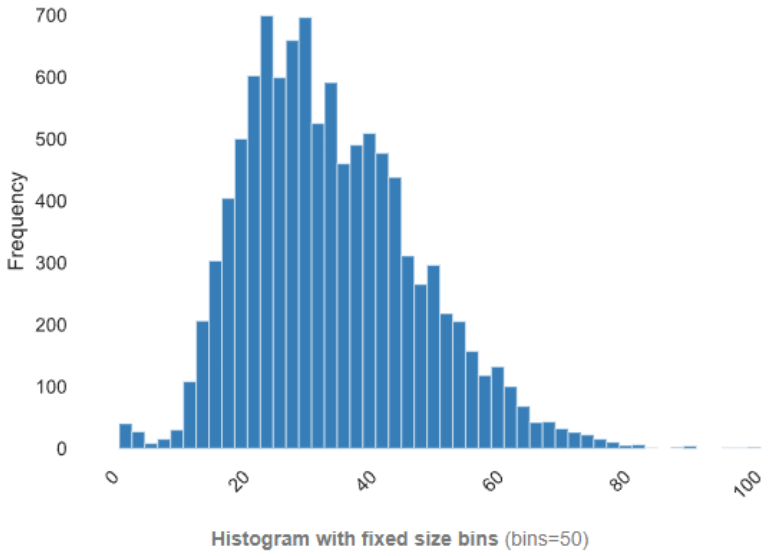
The dataset provides over 20,000 crowdsourced cough recordings representing a wide range of subject ages, genders, geographic locations, and COVID-19 statuses. Furthermore, experienced pulmonologists labeled more than 2,000 recordings to diagnose medical abnormalities present in the coughs, thereby contributing one of the largest expert-labeled cough datasets in existence that can be used for a plethora of cough audio classification tasks.

Along with the cough audio, for every data sample we have the following attributes:

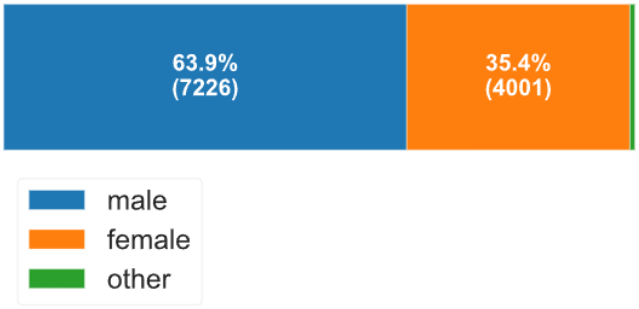
a. DateTime: Date and time when the sample was collected
b. Cough_detected: the probability that the individual has cough
c. Age: age of the individual
d. Gender: Gender of the individual
e. Respiratory_condition: Whether the person has a respiratory condition or not
f. Fever_muscle_pain: Whether the person has muscle and fever pain
g. Status: whether the person has COVID-19 or not
h. Latitude and longitude: Geolocation details of the person.
i. Labeled recording has more features that describe the cough:
    i. Quality
    ii. Cough_type
    iii. Dyspnea
    iv. Wheezing
    v. Stridor
    vi. Choking
    vii. Congestion
    viii. Nothing
    ix. Diagnosis
    x. severity

There were some steps taken to first prepare the data. We will talk about those steps in point
Lets analyze some features of our data:

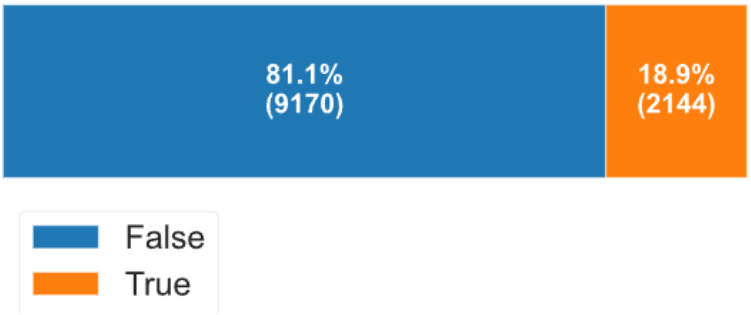
Age



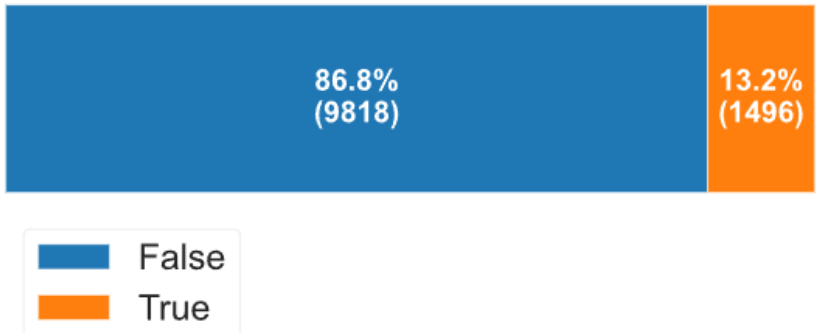
Gender



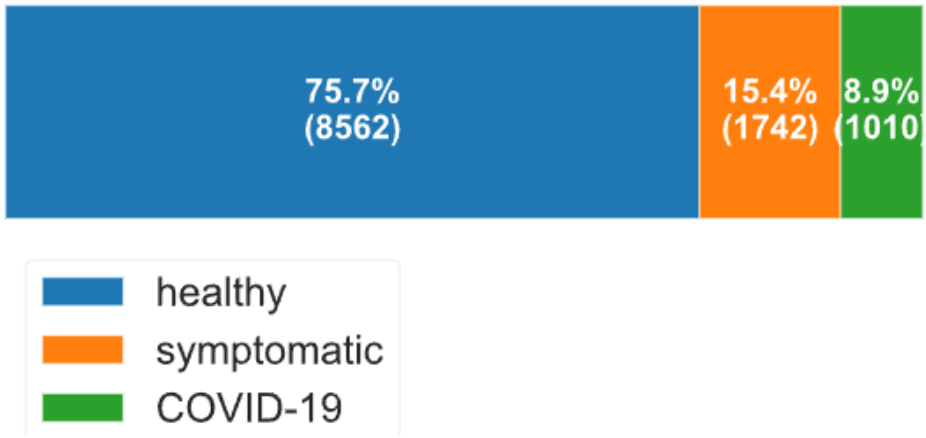
Respiratory condition



Fever_muscle_pain

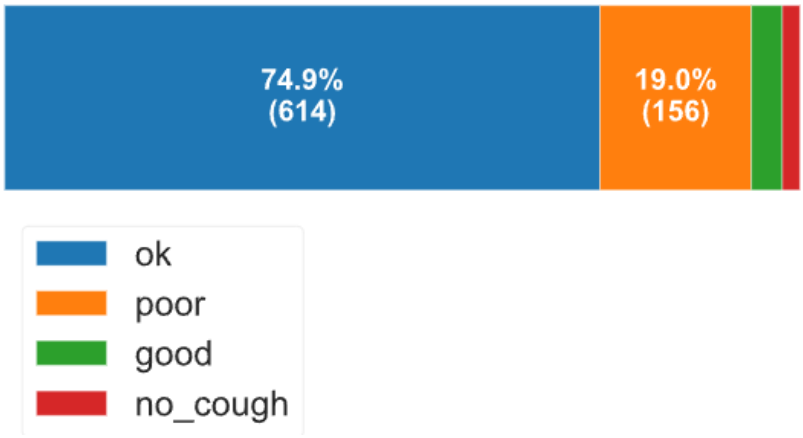


Status

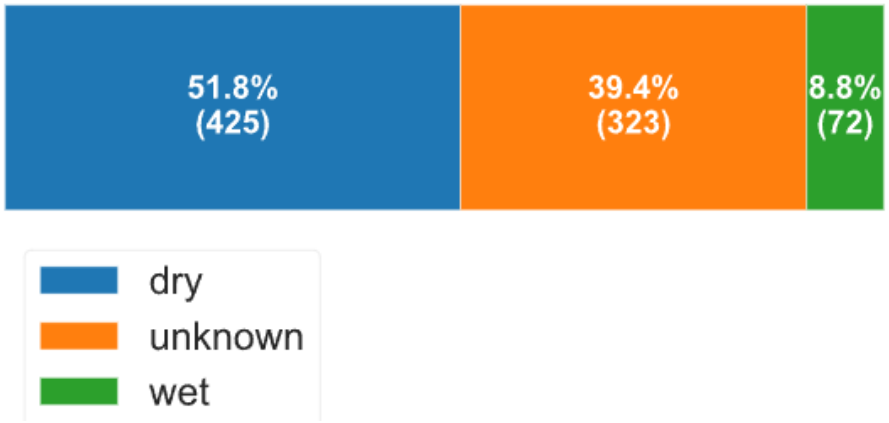


As we know some portion of our dataset is labeled. Let's look at all the labeled features and what are the values in them

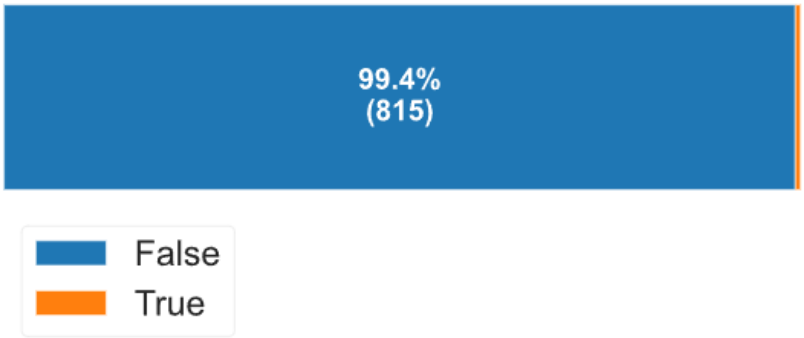
Quality ===> indicates cough quality



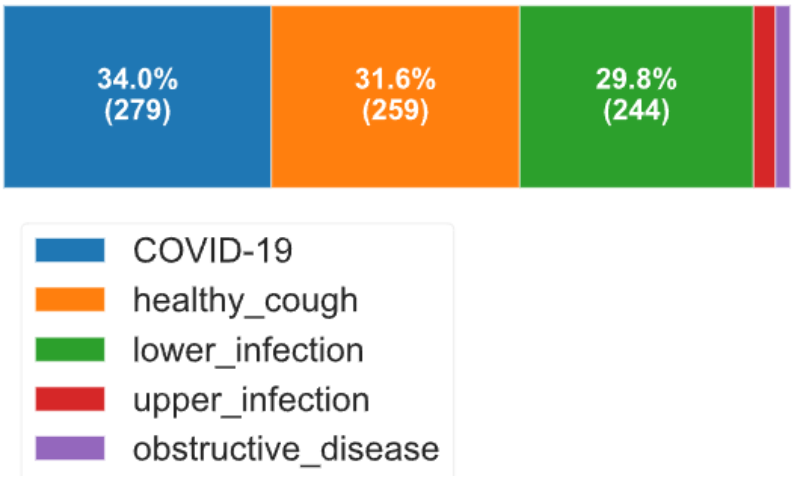
Cough type



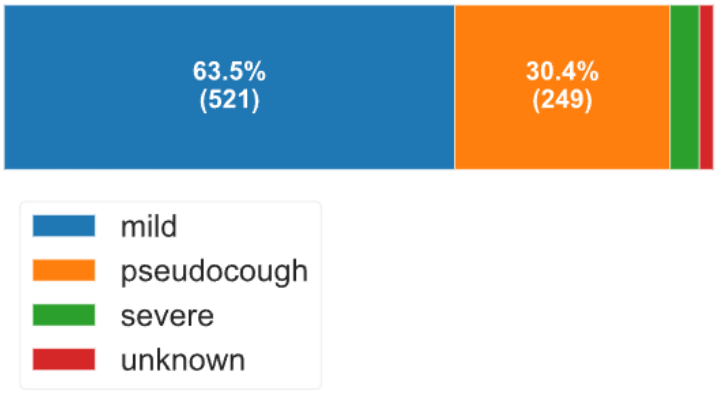
Dyspnea ===> difficulty in breathing



Diagnosis



Severity



2. **Work done / Existing Analysis**
    a. https://www.kaggle.com/code/nasrulhakim86/covid-19-screening-from-audio-part-2
    Given the size of the dataset and its varied quality, it was initially filtered as follows.

    - Only data that has been observed by physicians
    - Remove data without status
    - Select only cough_detected > 0.8
    - Select only data that has been reviewed as good quality by physicians

    Before feeding the audio data into the model for training, a few transformations were made to it for feature extraction.

    - Normalize, lowpass filter, and downsample cough samples
    - Select only the cough portion in the audio
    - Remove short segments
    - Make all audio segments the same size.
    - Rescale the data into [-1,1]

    b. https://www.kaggle.com/code/nasrulhakim86/covid-19-screening-from-audio-part-1
    Features extracted from audio:
    - Spectrogram
    - Short-Time Fourier Transform (STFT)
    - Mel-spectrogram
    - Mel-Frequency Cepstral Coefficients
    - Chroma Features

    c. https://www.nature.com/articles/s41597-021-00937-4

68 audio features commonly used for cough classification were extracted from each recording. The details of these state-of-the-art features are listed in Table here: https://www.nature.com/articles/s41597-021-00937-4/tables/3

Analysis was done on Demographic representativeness and Geographic representativeness

3. Problem Statement

Challenges

1. The dataset is available in JSON format. Where there are almost 20000+ JSON files for every data sample collected, we have to club all the data available in JSON files and create one main data frame to work on. Post that we need to also create separate features for the labeled data available so that we can use them in our analysis.
2. We need to deal with the missing values by finding suitable ways to replace them. The imputation of the missing values will not require domain expertise but also it will need complex and data sensitive models. Also the missing values are very scattered; we can not simply just impute or drop the columns.
3. Along with JSON data we have more than 20000+ audio files available. We need to extract multiple features for our analysis.
4. The sheer scale of the Audio data makes it very challenging to preprocess and extract features. On top of that, the data is available at 22,500 Hz sampling rate rather than an 8000 Hz sampling rate which relies on a lot of computation to process. For example just 1 second an audio clip sampled at 22,500 Hz will have 22,500 frames to process.
5. All the existing pre-processing work blindly extracts the features from the Audio data with little or no explainability. Again these new found extracted features have little or no analysis on why and how it works with ML/DL models.
6. We have a plethora of exciting and insightful hypotheses to investigate. Nevertheless, the extensive presence of scattered null values within the dataset significantly hinders our ability to achieve confident results through hypothesis testing.

New Features to Learn:

To enhance our dataset and potentially improve our analysis and model performance, we can consider creating new features based on the existing attributes. Here are some new features (ideas) we can derive from the given dataset:

1. **Day of the Week:** Extract the day of the week from the "DateTime" attribute. This can help you investigate if there are any patterns related to the day when cough samples were collected.

2. **Season:** Create a feature that categorizes the sample collection date into seasons (e.g., spring, summer, fall, winter). Seasonality might impact respiratory conditions and COVID-19 cases.

3. **Age Group:** Categorize age into groups (e.g., children, adults, seniors) to explore if different age groups have varying cough characteristics or COVID-19 prevalence.

4. **Geographical Clusters:** Use clustering algorithms (e.g., K-means) on latitude and longitude data to group individuals into geographical clusters. Analyze if individuals within the same cluster have similar COVID-19 outcomes or cough characteristics.

5. **Geographical Region:** Group latitude and longitude data into geographical regions or zones (e.g., by city or state) to analyze regional variations in cough audio and COVID-19 cases.

6. **Time of Day:** Divide the day into time intervals (e.g., morning, afternoon, evening) based on the "DateTime" attribute. Investigate if cough characteristics or COVID-19 cases vary by time of day.

7. **Interaction Terms:** Create interaction terms between attributes, such as "Age x Gender," "Age x Respiratory_condition," or "Severity x Cough_type," to explore potential combined effects on COVID-19 status.

9. **Cough Duration:** Calculate the duration between the onset of cough and the sample collection date. Investigate if the duration of cough is related to COVID-19 severity.

10. **Age-Related Respiratory Condition:** Create a binary feature indicating whether an individual's age is within a range commonly associated with certain respiratory conditions (e.g., pediatric respiratory conditions).

11. **Frequency of Cough Types:** Calculate the frequency of each "Cough_type" and "Diagnosis" to determine which types are most prevalent in COVID-19 cases.

12. **Severity of Symptoms:** Combine information from "Severity" and "Diagnosis" to create a comprehensive severity score, which may provide a more nuanced view of COVID-19 cases.

13. **Distance to COVID-19 Hotspots:** If you have access to COVID-19 hotspot data, calculate the distance between each sample's geolocation and the nearest hotspot. This can help assess proximity to outbreak areas.

14. **Age-Gender Interaction:** Explore the interaction between age and gender to determine if certain age-gender groups are more susceptible to COVID-19 or exhibit distinct cough characteristics.

15.**Temporal Trends:** Analyze temporal trends by calculating moving averages or trends in cough probabilities over time to identify potential shifts in COVID-19 prevalence.

16. **Weather Data:** If available, integrate weather data (e.g., temperature, humidity) based on the geolocation and "DateTime" attributes to examine if weather conditions influence COVID-19 cases or cough characteristics.

17. **Temporal Clusters:** Apply clustering algorithms on the "DateTime" attribute to identify temporal clusters or patterns in sample collection times. This can help identify time periods with distinct cough characteristics or COVID-19 prevalence.

Hypothesis:

1. **Hypothesis:** Cough Detection and COVID-19 Status
   - **Null Hypothesis (H0):** There is no statistically significant difference in the "Cough_detected" probability between individuals with COVID-19 and those without COVID-19.
   - **Alternative Hypothesis (H1):** There is a statistically significant difference in the "Cough_detected" probability between individuals with COVID-19 and those without COVID-19.

2. **Hypothesis:** Age and COVID-19 Status
   - Null Hypothesis (H0): Age does not significantly differ between individuals with COVID-19 and those without COVID-19.
   - Alternative Hypothesis (H1): There is a statistically significant difference in age between individuals with COVID-19 and those without COVID-19.

3. **Hypothesis:** Gender and COVID-19 Status
   - Null Hypothesis (H0): Gender is not associated with COVID-19 status.
   - Alternative Hypothesis (H1): There is a statistically significant association between gender and COVID-19 status.

4. **Hypothesis:** Respiratory Condition and COVID-19 Status
   - Null Hypothesis (H0): Having a respiratory condition is not significantly associated with COVID-19 status.
   - Alternative Hypothesis (H1): There is a statistically significant association between having a respiratory condition and COVID-19 status.

5. **Hypothesis:** Fever and Muscle Pain and COVID-19 Status

- Null Hypothesis (H0): Having fever and muscle pain is not significantly associated with COVID-19 status.
- Alternative Hypothesis (H1): There is a statistically significant association between having fever and muscle pain and COVID-19 status.

6. **Hypothesis:** Geolocation and COVID-19 Cases
   - Null Hypothesis (H0): Geolocation (latitude and longitude) is not associated with the prevalence of COVID-19 cases.
   - Alternative Hypothesis (H1): There is a statistically significant association between geolocation and the prevalence of COVID-19 cases.

7. **Hypothesis**: Cough Type and COVID-19 Status
   - Null Hypothesis (H0): There is no significant association between the type of cough (e.g., dry, wet) and COVID-19 status.
   - Alternative Hypothesis (H1): The type of cough is significantly associated with COVID-19 status.

8. **Hypothesis**: Severity of Diagnosis and COVID-19 Status
   - Null Hypothesis (H0): The severity of the diagnosis (e.g., mild, moderate, severe) is not significantly associated with COVID-19 status.
   - Alternative Hypothesis (H1): The severity of the diagnosis is significantly associated with COVID-19 status.

9. **Hypothesis:** Dyspnea and COVID-19 Status
   - Null Hypothesis (H0): The presence of dyspnea (shortness of breath) is not significantly associated with COVID-19 status.
   - Alternative Hypothesis (H1): The presence of dyspnea is significantly associated with COVID-19 status.

10. **Hypothesis**: Wheezing and COVID-19 Status
    - Null Hypothesis (H0): Wheezing is not significantly associated with COVID-19 status.
    - Alternative Hypothesis (H1): Wheezing is significantly associated with COVID-19 status.

We have numerous intriguing and valuable hypotheses to explore. However, the presence of many scattered null values in the dataset poses a significant challenge in obtaining reliable results from hypothesis testing.

Mention any features or labels that you want to learn from the data using some ML/DL models:

We use predictive modeling in two stages. First we will try to learn whether a given audio sample is cough or not. In the second stage we will combine all the extracted features to predict the status of the Covid 19 as Positive, Negative and Symptomatic.

How your problem statements are unique from any other closely related existing analysis.

- We will be using numerous existing techniques to extract features from the cough audio data along with their explainability and use case. We will also extract various features from the existing data we have apart from cough recordings.
- The modeling part for the predictive analysis is always an open challenge for these types of dataset. We will go forward with ML and DL-based techniques to solve the problem and also might try ensembling techniques for the final model.
- All this work will be done on the lines of insights and ideas we develop while understanding the data and all the informative hypothesis testing we do. The novelty lies within the way we approach to solve this problem. The way we approach feature extraction and modeling.
- All the hypothesis testing ideas which we propose (and will execute a few of them) are being done for the first time for this dataset. Various features which we will extract and different feature extraction techniques which we will be using are also quite novel. And the Classification task at hand will again require novel cutting-edge research methods to reach state-of-the-art accuracy.