COSWARA DATA SET

ABSTRACT

Performed an Exploratory Data Analysis on the COSWARA Metadata and Audio files. Conductedstatistical analysis and tested a model to predict Covid.

DANA-4800

Javier

MerinoJosé

Vargas

Manraj

Singh

Table of Contents

1. Objective

1. Perform a comprehensive exploratory data analysis on the metadata to identify patterns,distributions, and potential anomalies
2. Study how the age, health status and gender of individuals influence the acoustic characteristicsof the audio recordings.
3. Evaluate how different features extracted from the audio recordings vary between individualswith different health status.
4. Apply a range of statistical techniques, such as T-tests, ANOVA, and Chi-square tests, to identify significant differences and relationships between groups within the dataset.
5. Use a model to classify the health status of individuals based on the features of their audiorecordings.

2. Data Description

The data comprise information related to health status mainly related to respiratory illnesses and Covid 19 symptoms for 2,745 subjects, who recorded 9 type of audios each. The health information was collected through 41 variables (metadata) and if we consider 9 audio files per each subject, there are approximately 24,705 audio files.

2.1. Metadata

If considering the combined data presented by the COSWARA team we had 36 variables, but after combining all the CSV reports inside all the folders we got 41 variables which are presented in Table 1. Nevertheless, those additional variables “dT”, “fV”, “iF”, “Date”, “test” (variables 37 to 41 in Table 1) had no significant values, and there is not much detail on them, so the research teammight have discarded them on purpose.

Table 1 Variable Description

No	Variable Type Coswara	Variabla Name	Variable Description	Variable Type	Variable Valid Range
1	Demographic	id	Participant identifier	Categorical-Ordinal	character
2	Demographic	record_data	Date of the recording	Date	Dd-mm-yy
3	Demographic	a	age	Numerical - discrete	18-90
4	Demographic	g	gender	Categorical-Ordinal	Male
					Female
					Other

5	Demographic	l_c	country	Categortical-Nominal	List of countries
6	Demographic	l_s	state	Categortical-Nominal	List of states inside thecountries
7	Demographic	vacc	Vaccination status	Categorical-Ordinal	y
					p
					n

No	Variable Type Coswara	Variab la Name	Variable Description	Variable Type	Variable Valid Range
					NA
8	Demographic	ep	Proficient in english	Categorical-Ordinal	TRUE
					FALSE
9	Demographic	smoker	Regular smoker	Categorical-Ordinal	TRUE
					FALSE
					NA
10	Demographic	rU	Returning participant	Categorical-Ordinal	TRUE
					FALSE
					NA
11	Demographic	um	Wearing mask	Categorical-Ordinal	TRUE
					FALSE
					NA
12	Covid-19 like symptoms	cough	Has cough	Categorical-Ordinal	TRUE
					FALSE
					NA
13	Covid-19 like symptoms	cold	has cold	Categorical-Ordinal	TRUE
					FALSE
					NA
14	Covid-19 like symptoms	diarrhoea	has diarrhoea	Categorical-Ordinal	TRUE
					FALSE
					NA
15	Covid-19 like symptoms	bd	has breathing diffculties	Categorical-Ordinal	TRUE
					FALSE
					NA
16	Covid-19 like symptoms	st	has sore throat	Categorical-Ordinal	TRUE
					FALSE
					NA
17	Covid-19 like symptoms	fever	has fever	Categorical-Ordinal	TRUE
					FALSE
					NA
18	Covid-19 like symptoms	ftg	suffering from fatigue	Categorical-Ordinal	TRUE
					FALSE
					NA
19	Covid-19			Categoric	TRUE

19	like symptoms	mp	has muscle pain	al-Ordinal	FALSE NA
20	Covid-19 like symptoms	loss_of_smell	loss of smell and/or taste	Categorical-Ordinal	TRUE FALSE NA
21	Respiratory ailments	asthma	Has asthma relatedissues	Categorical-Ordinal	TRUE FALSE NA
22	Respiratory ailments	cld	Has chronic lung disease	Categorical-Ordinal	TRUE FALSE NA
23	Respiratory ailments	pneumonia	has pneumonia	Categorical-Ordinal	TRUE FALSE NA

No	Variable Type Coswara	Variabla Name	Variable Description	Variable Type	Variable Valid Range
24	Respiratory ailments	others_resp	has other respiratoryillness	Categorical-Ordinal	TRUE
					FALSE
					NA
25	Comorbidity	ht	has hypertension	Categorical-Ordinal	TRUE
					FALSE
					NA
26	Comorbidity	diabetes	has diabetes	Categorical-Ordinal	TRUE
					FALSE
					NA
27	Comorbidity	ihd	has ischemic heartdisease	Categorical-Ordinal	TRUE
					FALSE
					NA
28	Comorbidity	others_preexist	any other pre-existingcomorbidity	Categorical-Ordinal	TRUE
					FALSE
					NA
29	Covid-19 health	test_status	status of covid-19 test	Categorical-Ordinal	p (positive)
					n(negative)
					na(not taken)
					ut(under testing)
					NA
30	Covid-19 health	covid_status	covid related healthstatus	Categorical-Ordinal	healthy
					positive_mild
					no_resp_illness_exposed
					positive_moderate
					resp_illness_not_identified
					recovered_full
					positive_asymp
					under_validation
31	Covid-19 health	testType	type of covid-19 testtaken	Categorical-Ordinal	RTPCR
					RAT
					FALSE
					NA

32	Covid-19 health	test_date	date of covid-19 test	Date	2020-09-01 - 2022-09-01
33	Covid-19 health	ctDate	date of CT-scan	Date	2020-12-31 - 2022-07-22
34	Covid-19 health	ctScore	CT value	Numerical - discrete	number
35	Covid-19 health	ctScan	The participant has a CTScan	Categorical-Ordinal	TRUE
					FALSE
					NA
36	Demographic	l_l	Name of Cities	Categortical-Nominal	List of cities
37	Demographic	dT	Adquisition medium	Categorical-Ordinal	android
					web
					NA
38	Demographic	fV	No available information	Categorical-Ordinal	2
					NA
39	Demographic	iF			TRUE

No	Variable Type Coswara	Variabla Name	Variable Description	Variable Type	Variable Valid Range
			No available information	Categorical-Ordinal	FALSE
40	Demographic	Date	Dates taken the 2020-04-19	Dates	Dates NA
41	Demographic	test	No available information	Categorical-Ordinal	TRUE FALSE

2.2. Audio Files

The audio files in the Coswara dataset consist of 9 different recordings for each subject. Each subject has a unique ID in the metadata which allows to link that information to the audio files. The audio files include:

Sound	Type	Description
Breathing	Deep	This type of audio captures the patient's breathing patterns, distinguishing between deep and shallow breaths.
	Shallow	
Cough	Heavy	This audio records the patient's coughs
	Shallow	
Counting	Fast	Involves the patient counting numbers at different speeds.
	Normal	
Vowel vocalization	"a"	This type of audio captures the patient vocalizing specific vowels
	"e"	
	"o"	

3. Metadata Cleaning

After a heat map analysis, we encounter that most variables have missing values "NA". In figure 1 it is shown the extend of the NA values, so we analyzed the logic of the filling of the data, to tryto eliminate the missing values using our own assumptions.

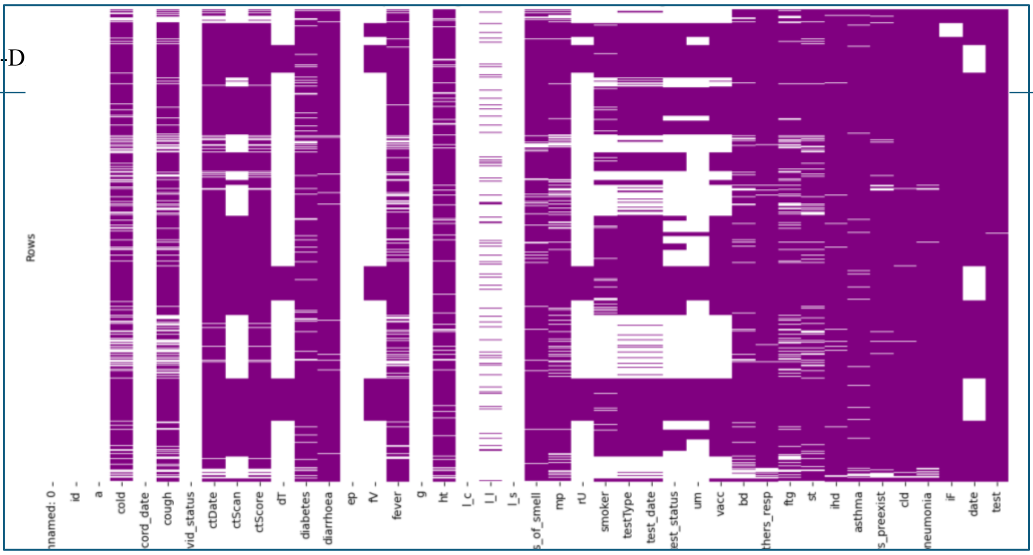


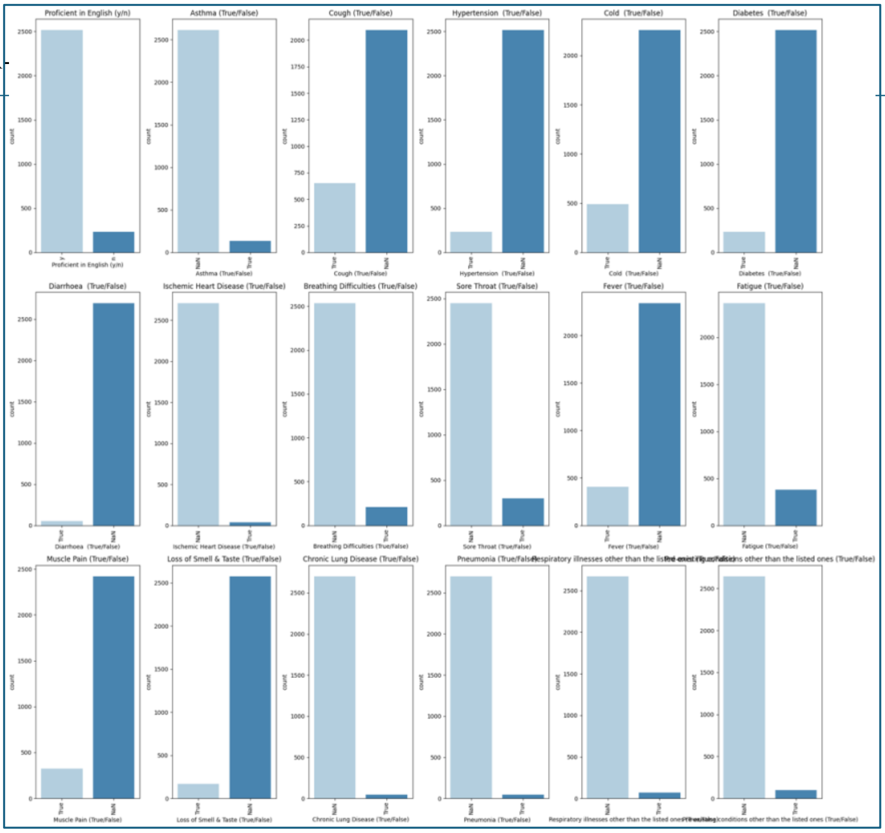
Figure 1 Heatmap of NA values for each variable.

3.1. Categorical variables with True and NA values

We identified that for 18 variables the research team only filled the TRUE value, and left the other possibility as NA, since there can only be one of two choices, we will assume that the NA values will be FALSE. In Figure 2 it is displayed a counting for each of these variables. The variables thatwere considered in this analysis are shown in Table 2.

Table 2 Variables which contain the value True and NA

No.	VARIABLE
1	English Proficiency
2	Cough
3	Cold
4	Diarrhoea
5	Breathing Difficulty
6	Sore Throat
7	Fever
8	Fatigue
9	Muscle Pain
10	Loss of Smell
11	Asthma
12	Chronic Lung Disease
13	Pneumonia
14	Has Other Respiratory illness
15	Hypertension



No.	VARIABLE
16	Diabetes
17	Ischemic heart Disease
18	Other Preexistence Comorbidity

Figure 2 Variables with TRUE and NA values

3.2. Categorical variables filled with True, False and NA values

For the variable Smoker, we found that it was filled with values of only True until June 8 2021, then the research team started to record as N, Y, and some occasional False. We didn't find the logic behind this change, so we had to assume in this case, and consider the Y and True as positive, and N and False as Negative. The variables like "rU" (returning user), "um" (using mask), and "CT-Scan", have only values of N, Y and Blank. We cannot be 100% sure for these particular cases that the blanks are N or False, but we must assume as there is no additional information to think otherwise, so we turned the N and NA values into False. In Figure 3 was presented the counts foreach of these variables.

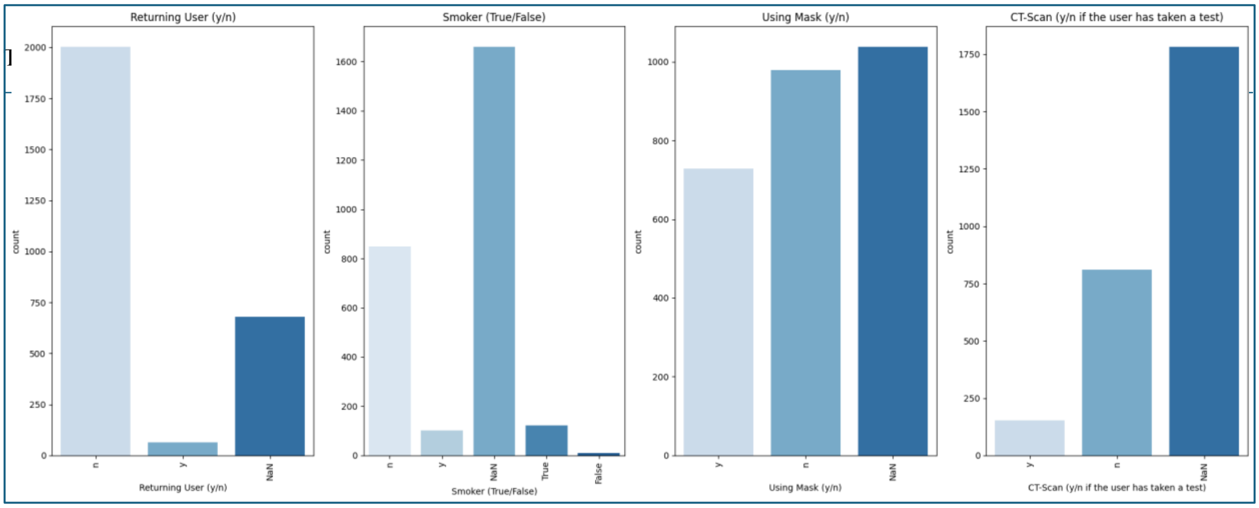


Figure 3 Variables filled with NA, True and False values.

Table 3 Variables filled with NA, True and False Values

No.	VARIABLE
19	Returning User
20	Smoker
21	Wearing Mask
22	CT Scan

3.3. Categorical variables with more than one entry

In Figure 4, it is presented the categorical variables that had other entries beside true or false, so in order to group them we made some assumptions:

- For the Covid Test Type there were four entries, rtpcr, rat, false and NA. We are considering the NA values as false as there is no additional information to consider them differently.
- For Vaccination Status we have recode the entries as “Both Doses”, “One Dose”, “No Dose”, and assigned the NA values as “No Dose”. The NA were above the 1.750 values, and as thereis not other relevant information, we decided to count them as “No Dose”.
- For “Covid Test Status” we have “Positive”, “Negative, “Not Taken”, and “Under testing”. As there is not additional information, we group the NA values as “Not Taken”.
- For “Covid Status”, under the Positive value we grouped: “Positive moderate”, “Positive mild”, “Positive asymptomatic”; under the Negative value we grouped: “Healthy”, “Recoveredfull”, “No resp illness exposed”, and under “Respiratory illness not identified” we grouped: “respiratory illness not identified”, “under validation”.

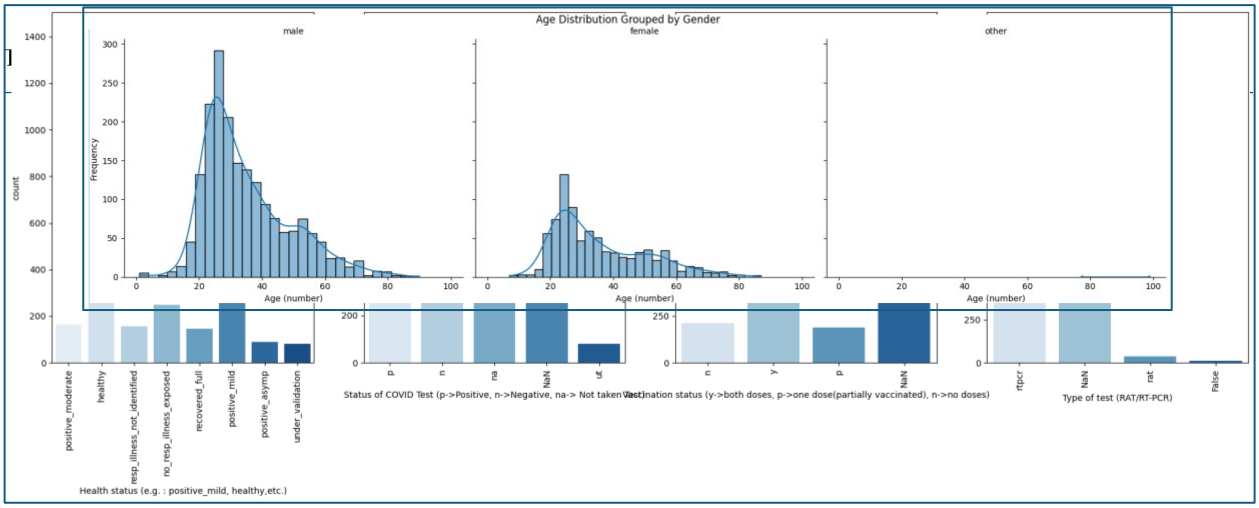


Figure 4 Categorical variables that had other entries beside True and False Values

Table 4 Variables filled with more than one entry

No.	VARIABLE
23	Vaccination Status
24	Covid Status
25	Covid Test Result
26	Covid Test Type

3.4. Other variables to be used in the analysis

We filtered the range of Age from 18 to 80 and turned out that one of the observations filtered (age 99) had the gender “Other”. The only other subject with Gender marked as “Other” was 77 years old, so we decided to eliminate this observation to keep only Male and Female values. The other values to be considered were the CT Score which ranges from 0 to 25, the subject’s Id and the dateof the recording.

Figure 5 Age vs Gender

Table 5 Other variables

No.	VARIABLE
27	Age
28	Gender
29	CT Score
30	Id
31	Recorded date

3.5. Variables that will not be used in the analysis

The variables identified in Table 6 were not considered in the analysis. Variables 37 to 41 lack some details and specific information. Variables 32 to 34 describe the subject's locations, which will not be considered in the analysis. The dates of covid and CT scan from variables number 35 and 36 were not used on any analysis.

Table 6 Variables that will not be used in the Analysis

No.	VARIABLE
32	Country
33	City
34	State
35	Covid Test Date
36	CT Scan Date
37	dT
38	fV
39	iF
40	date
41	test

4. EDA Metadata

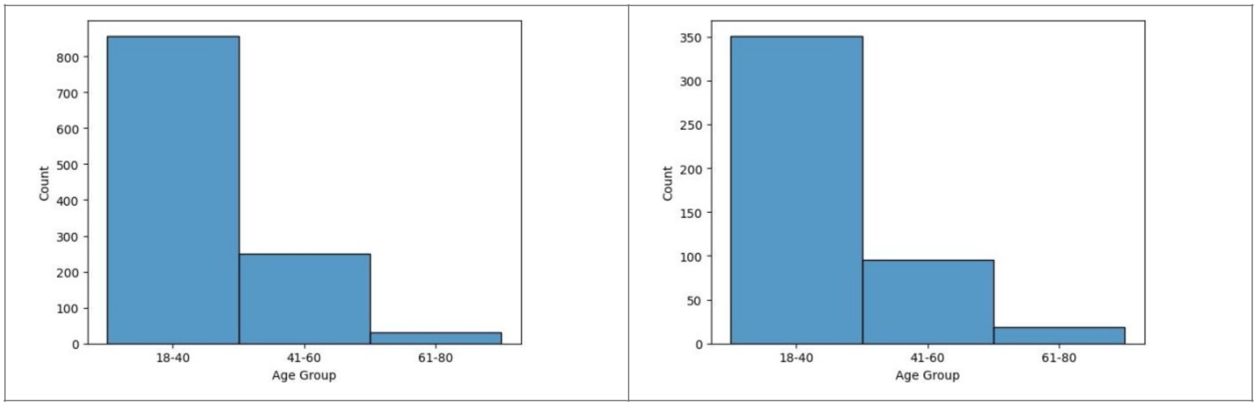
To analyze the information and provide a comparison context, we created tree groups based on age and dividing them by female and male. The analysis we displayed in this section has this groupingmethod.

4.1. Subjects with no respiratory illness

We filtered the dataset to identify subjects with no respiratory illnesses, in order to have a subject basis to whom we could compare the audio files from the ones affected by covid. The results are presented in Figure 6. We also filtered the smoker positive values to discard any possibleaffectation through that.

The number of male subjects double the number of female subjects, and the greatest number is allocated in the group from 18 to 40 years old.

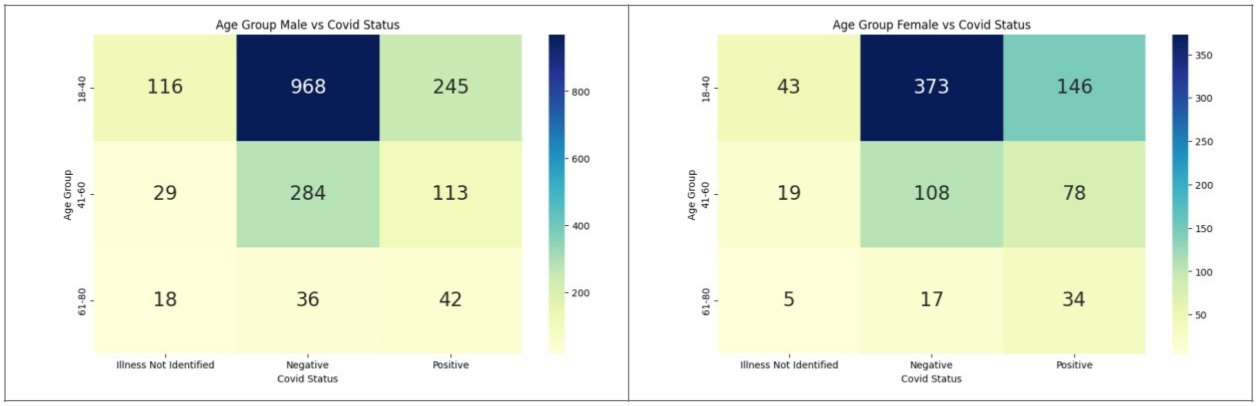
Figure 6 No respiratory illness a) male b) female



4.2. Covid Analysis

We found that the population of male nearly doubles the population of female in the case of Negative Covid.

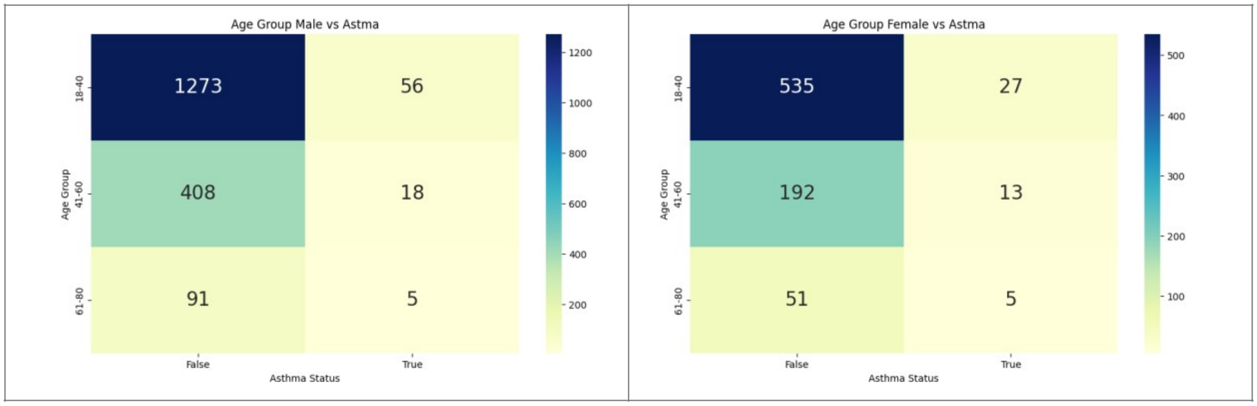
Figure 7 Covid Status per group



4.3. Asthma Analysis

The values for positive asthma are concentrated mainly in the group of Age from 18-40.

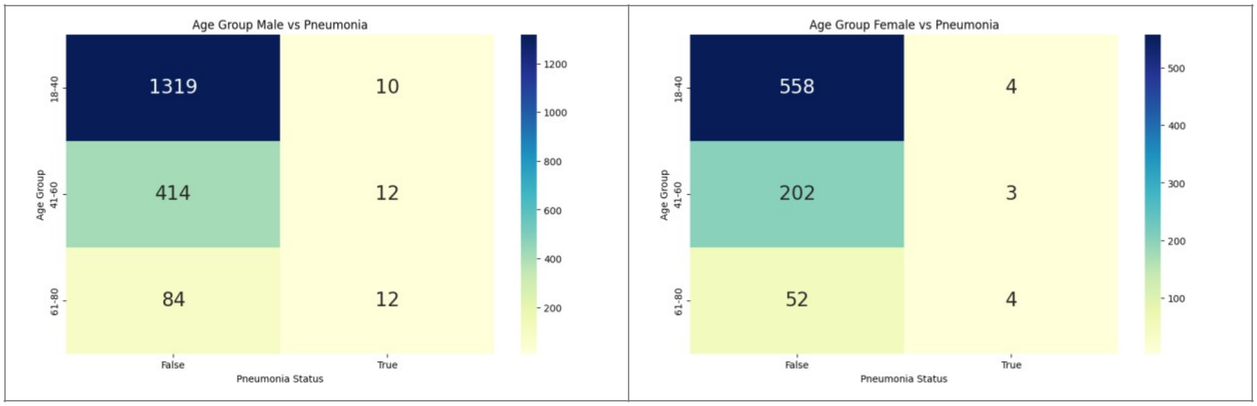
Figure 8 Asthma Status per group



4.4. Pneumonia Analysis

There are a limited number of subjects that reported pneumonia positive.

Figure 9 Pneumonia Status per group



5. EDAAudio recordings

5.1. Audio Duration

We analyzed the audio duration for all sounds for each subject by extracting this information fromall the Coswara folders with a python code. Initially, we started working with the “202201” folderwhere we observed that for breathing deep, the mean duration was 15.7 seconds and for breathingshallow the mean duration was 9.52 seconds.

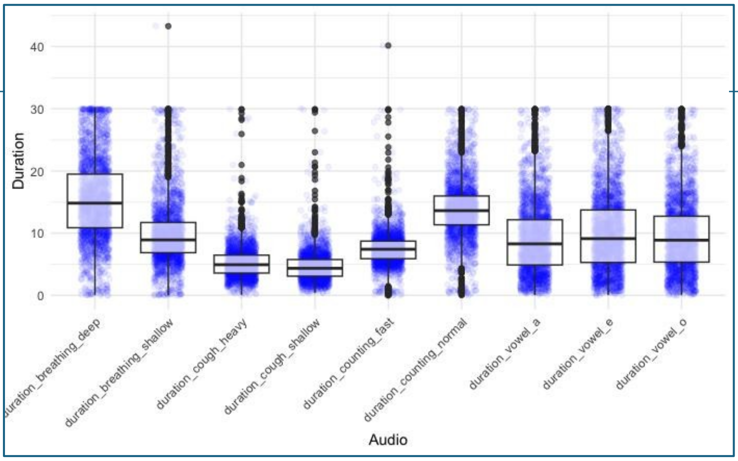


Figure 10 Breathing deep folder 202201

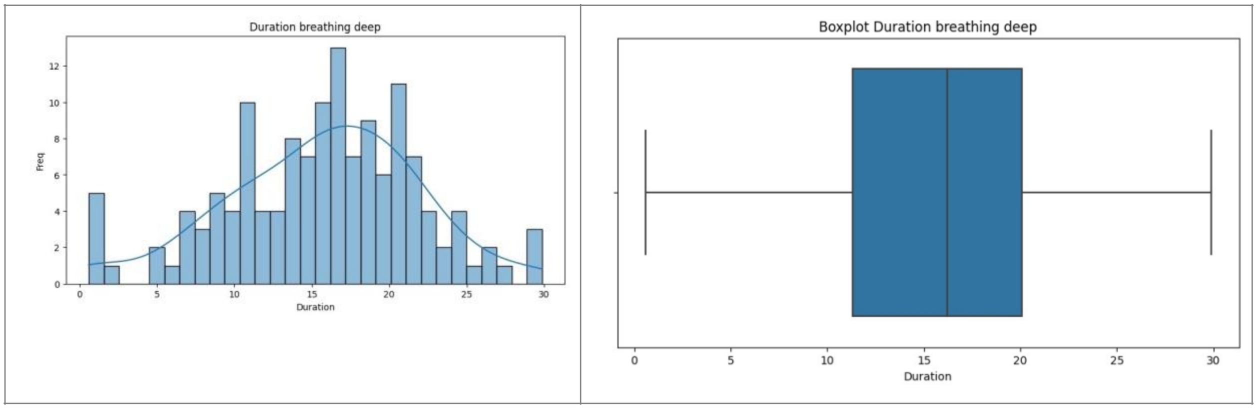
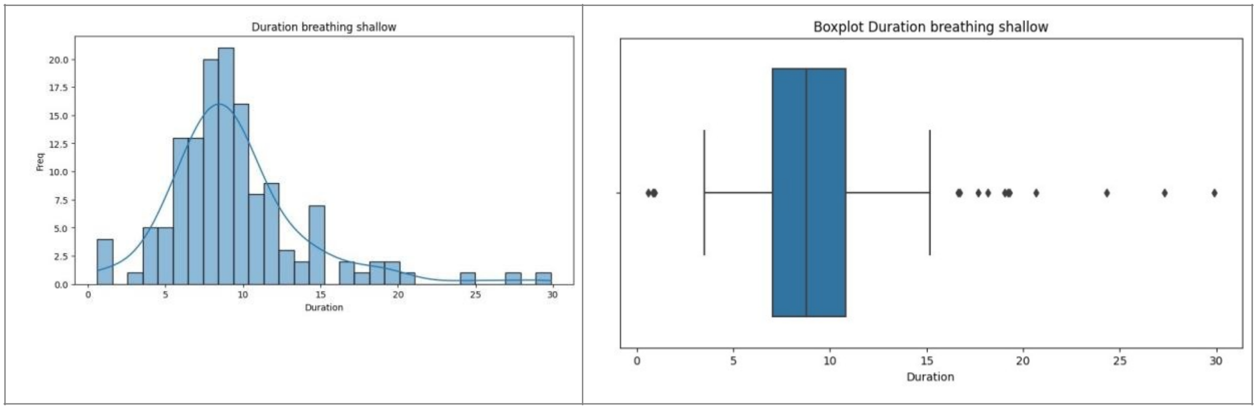
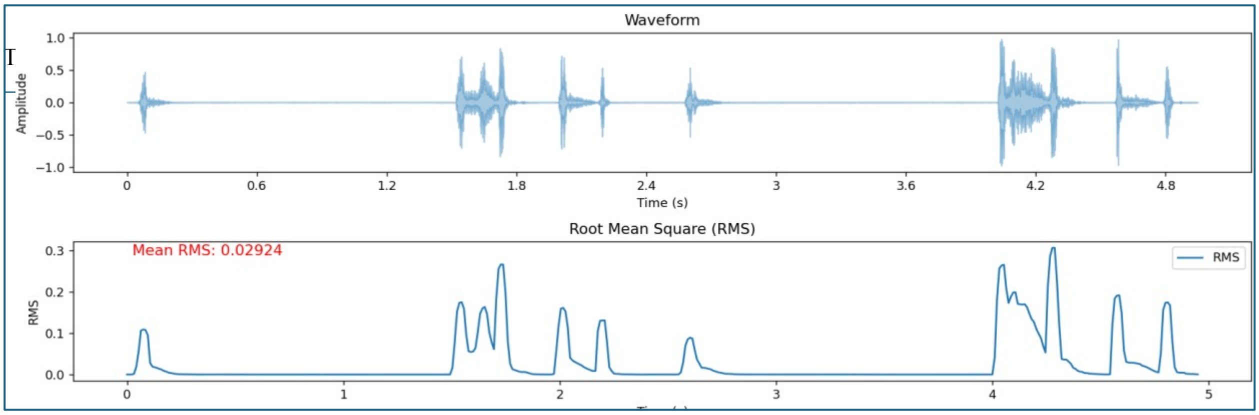


Figure 11 Breathing shallow folder 202201



After running a python code, we extracted the duration of all the audios for all subjects, nevertheless, we encountered 390 NA values for the duration which perhaps was the case due to some alterations in those audio files. Also, we encountered that there was an audio that lasted nearly 300 seconds (5minutes) for vocal “o”, so after discarding these cases, there were 23,684 audio files. In the next figure are the boxplots for the duration of each sound.

Figure 12 Audio Duration for all sound and subjects.



The results obtained for the duration are like those published by the Coswara team in their article. The breathing-deep audio is the one with the lower number of outliers and is the most normally distributed.

5.2. Audio Features

After exploring different methods to work with audio records, we discovered that extracting features from the audio is essential for thorough analysis. We decided to use the Librosa python library, a great tool for analyzing and processing audio files. Librosa provides functions to extractaudio features and allows us to convert audio files into a numerical format.

We identified important features such as RMS, ZCR, and spectral centroid, among others. However, we focused mainly on two time-domain features: RMS (Root Mean Square) and ZCR (Zero Crossing Rate), we investigated these features further to understand their results and functionality.

For our audio analysis, we chose to use those features because they are highly effective for evaluating the energy and tonal characteristics of audio signals. RMS measures the continuous power of an audio signal, giving an accurate sense of perceived loudness. This is important for assessing the intensity of sounds in patients with respiratory conditions, such as those with COVID-19, and comparing them to healthy individuals.

According to (Parekh 2011), ZCR measures how often the audio signal crosses the zero line. Thishelps us identify the sharpness and tonality of sounds, which is useful for detecting irregular breathing or coughing patterns in patients. The ZCR is valuable in distinguishing between voicedand unvoiced sounds, helping to identify patterns that could indicate respiratory issues.

We selected these methods because they provide clear, quantifiable data that are easy to interpret,which is essential for differentiating the health status of patients in our study.

Figure 13 Waveform and RMS graph for cough-heavy audio COVID negative

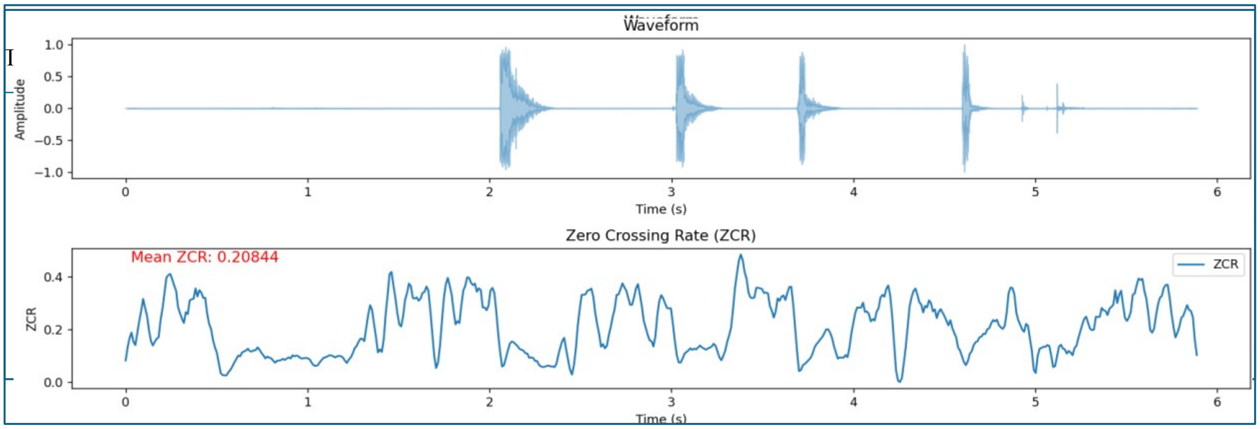


Figure 14 Waveform and RMS graph for cough-heavy audio COVID positive

Figure 15 Waveform and ZCR graph for cough-heavy audio COVID positive

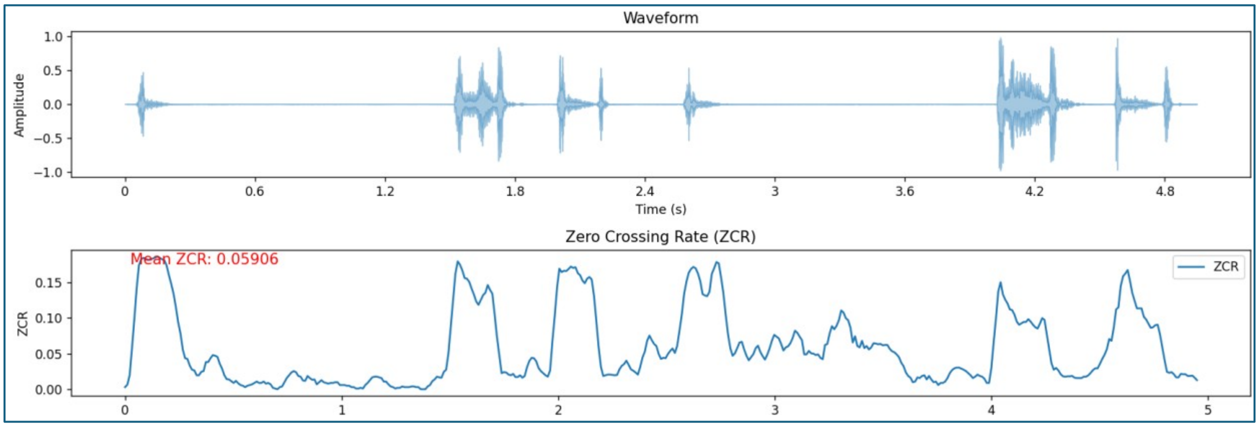


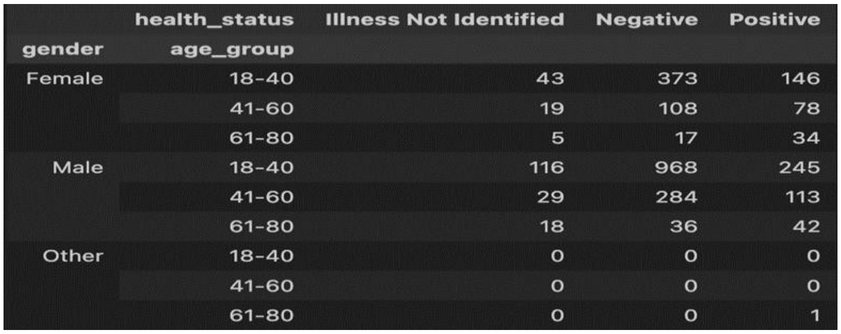
Figure 16 Waveform and ZCR graph for cough-heavy audio COVID negative

5.3. Groups to Analyze

To ensure our analysis is accurate, we divided our data into groups based on gender, age, and health status. The presence of COVID-19 can significantly alter vocal characteristics, so we included health status as a key grouping factor.

Males and females have different vocal characteristics, so we grouped the data by gender to capturethese differences. Additionally, as people age, naturally, natural aging of the vocal cords. We categorized the data into age groups to account for these changes.

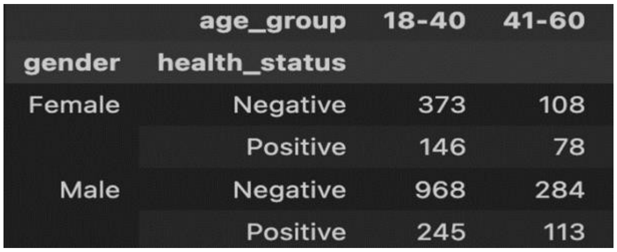
Figure 17 Grouping by Gender, Age Group, and Health Status



	health_status	Illness Not Identified	Negative	Positive
gender	age_group			
Female	18-40	43	373	146
	41-60	19	108	78
	61-80	5	17	34
Male	18-40	116	968	245
	41-60	29	284	113
	61-80	18	36	42
Other	18-40	0	0	0
	41-60	0	0	0
	61-80	0	0	1

Since many of the data is missing in the groups, we omitted the gender "other", age group 61-80 and health status "Illness not identified". In the following table the final grouping is shown:

Figure 18 Final Grouping by Gender, Age Group, and Health Status



	age_group	18-40	41-60
gender	health_status		
Female	Negative	373	108
	Positive	146	78
Male	Negative	968	284
	Positive	245	113

To determine which feature would be the most appropriate considering the 9 audios in the dataset,we conducted descriptive statistics. Firstly, we created histograms based on the groupings shownin Figure 18.

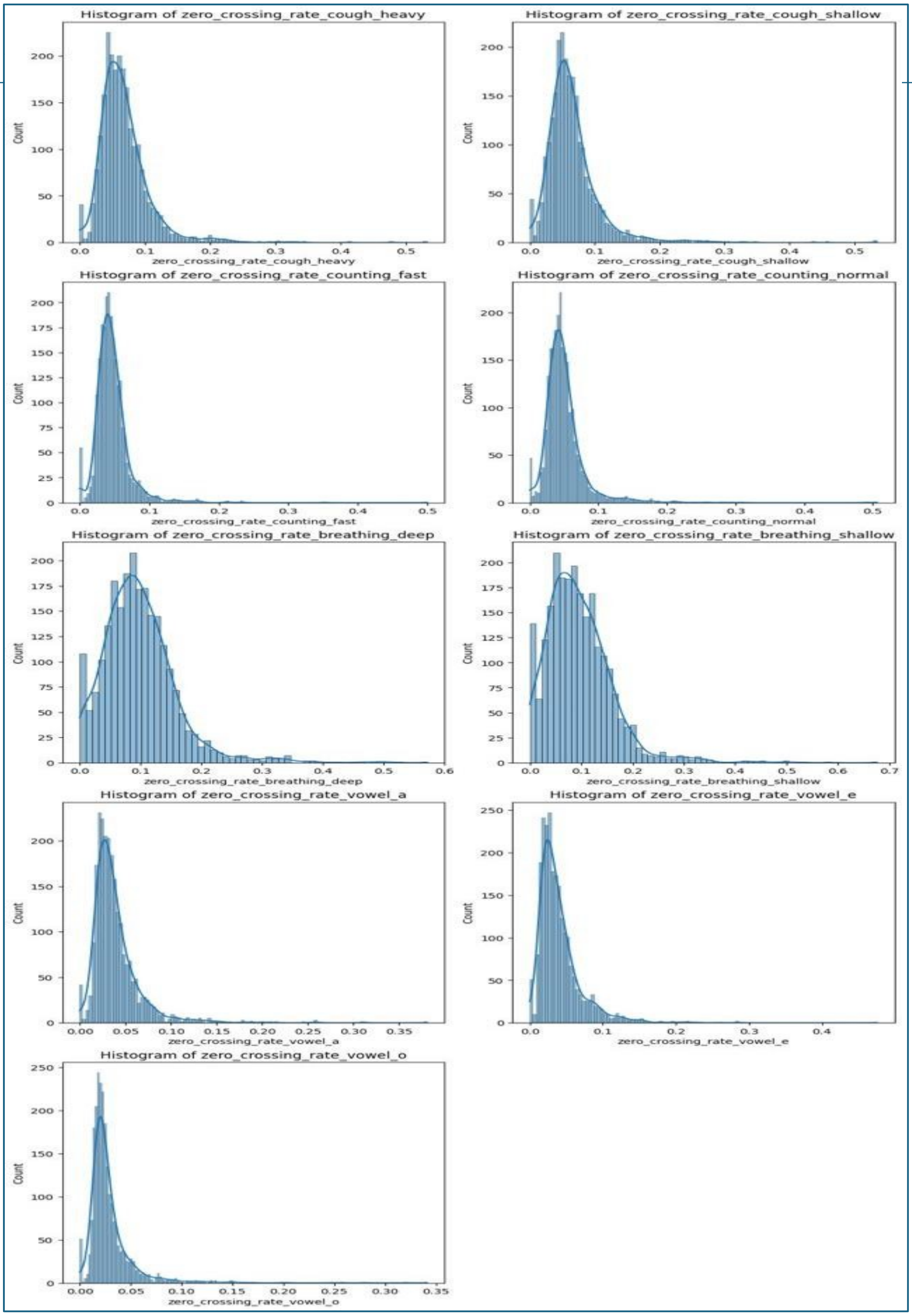


Figure 19 Histograms for Zero crossing rate to all 9 audios

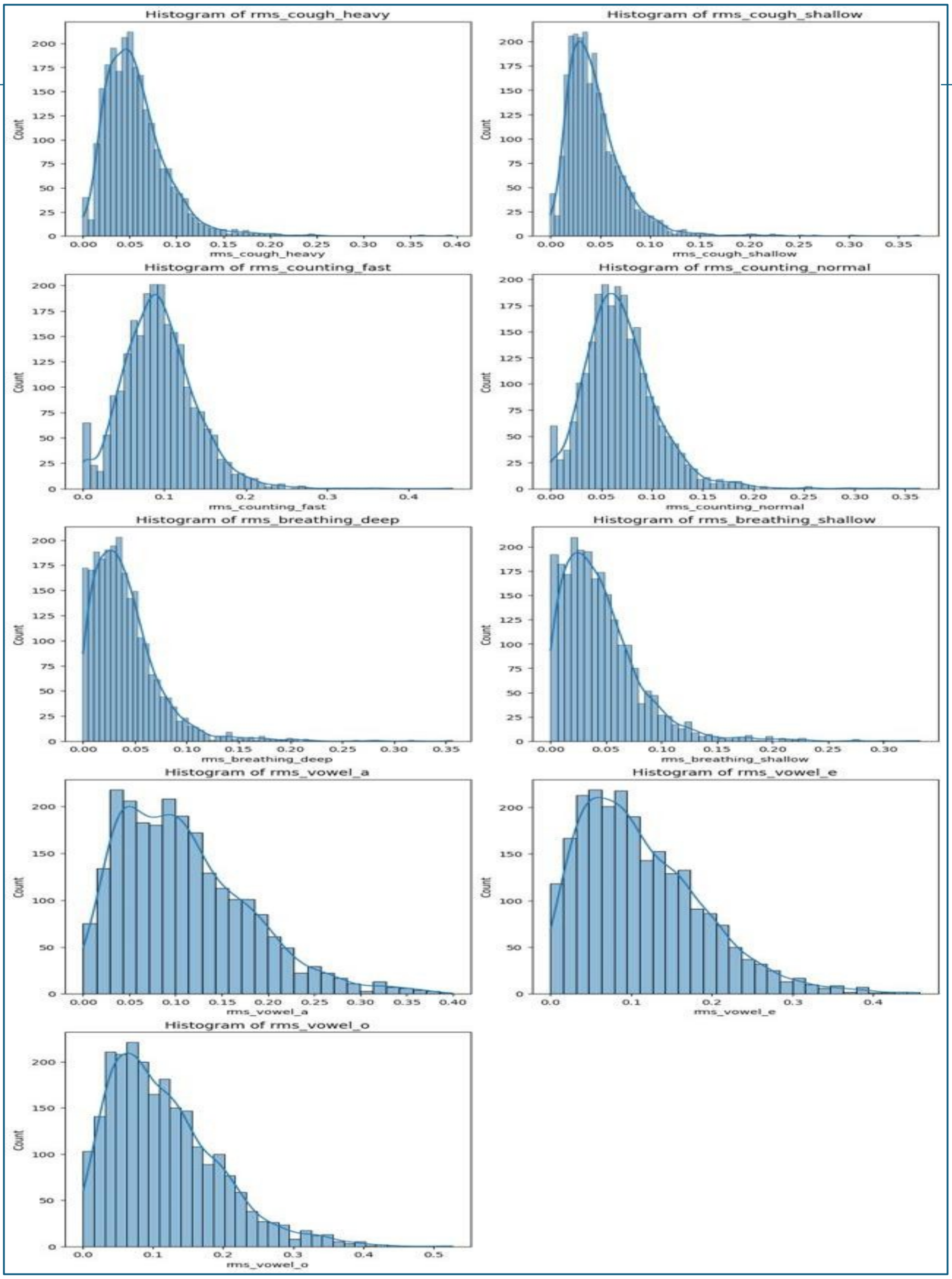
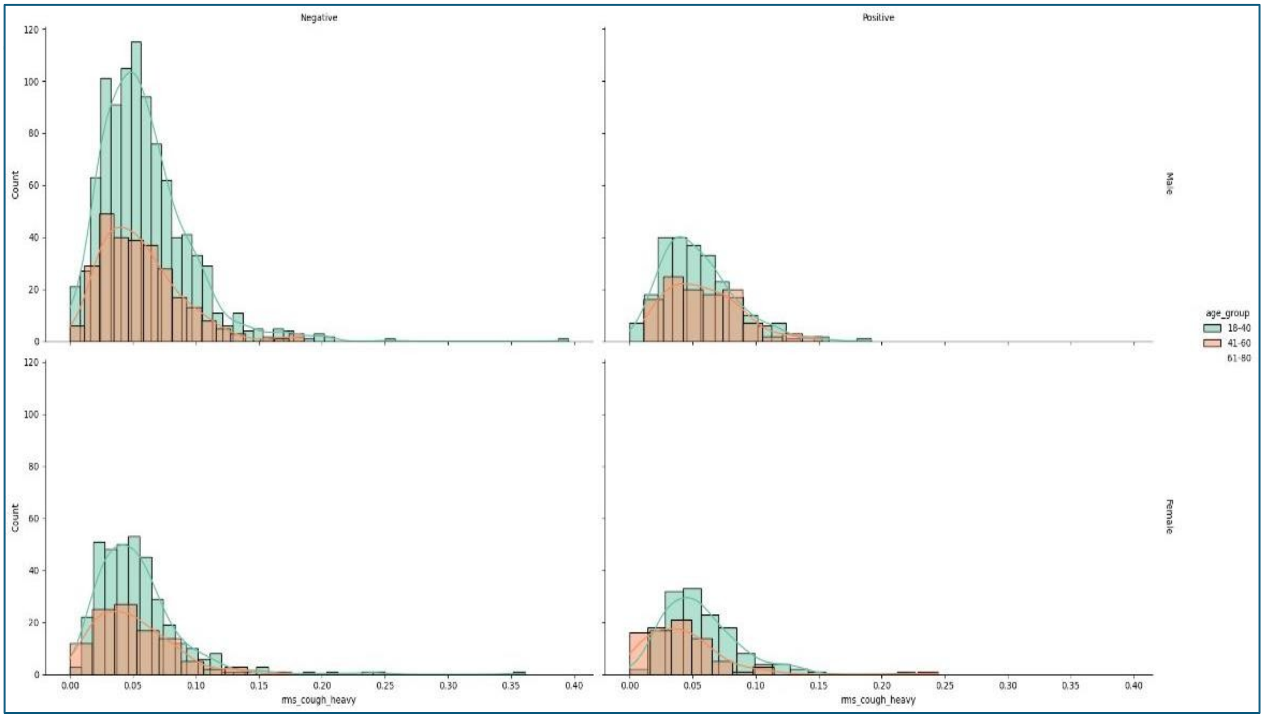


Figure 20 Histograms for RMS to all 9 audios

With these histograms, we proceeded to observe and evaluate the data distribution for each type of audio using the numerical variables of the features extracted using Librosa. This graphical representation is useful for summarizing and describing the main characteristics of the data, such as shape, dispersion and potential anomalies.

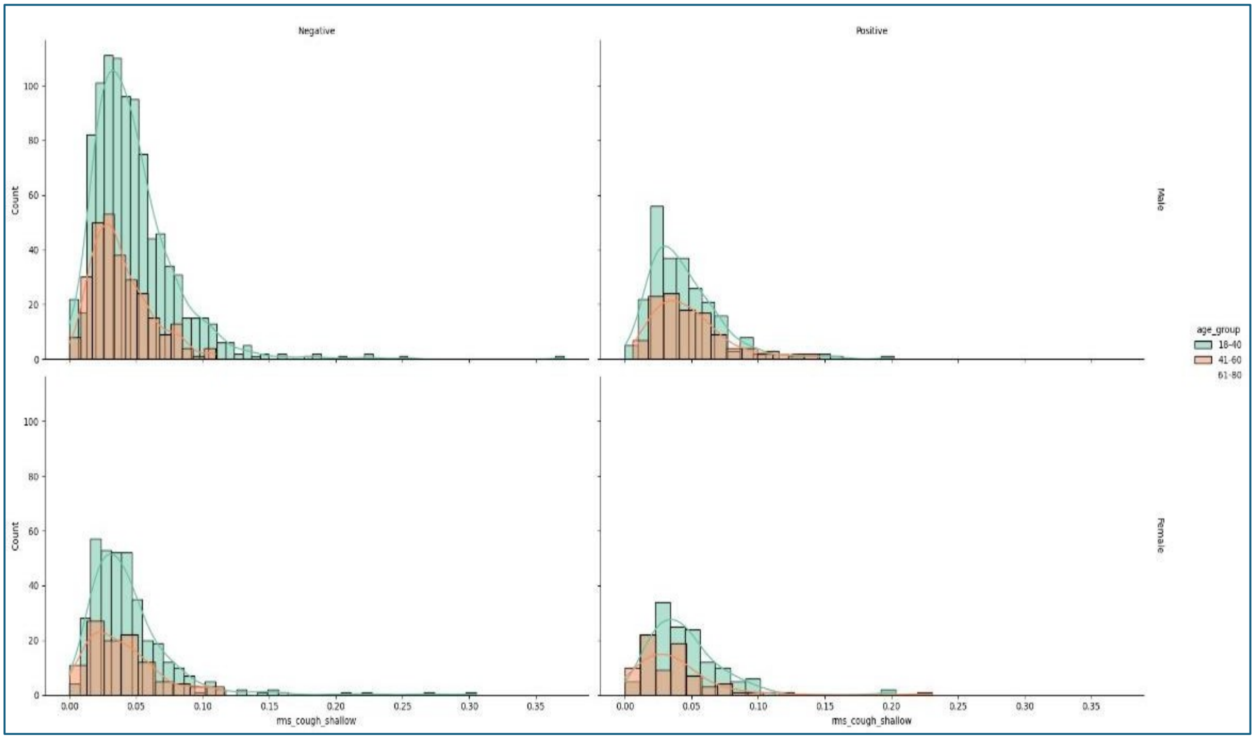
Looking at the histograms in figures 19 and 20 for both features RMS and ZCR we decided to choose those that best visualized a normal distribution. Later, we will proceed to conduct more



statistical tests that require a normal data distribution, as well as other tests to evaluate this distribution.

Figure 21 RMS – Cough Heavy by AGE GROUP, GENDER AND HEALTH STATUS

Figure 22 RMS – Cough shallow by AGE GROUP, GENDER AND HEALTH STATUS



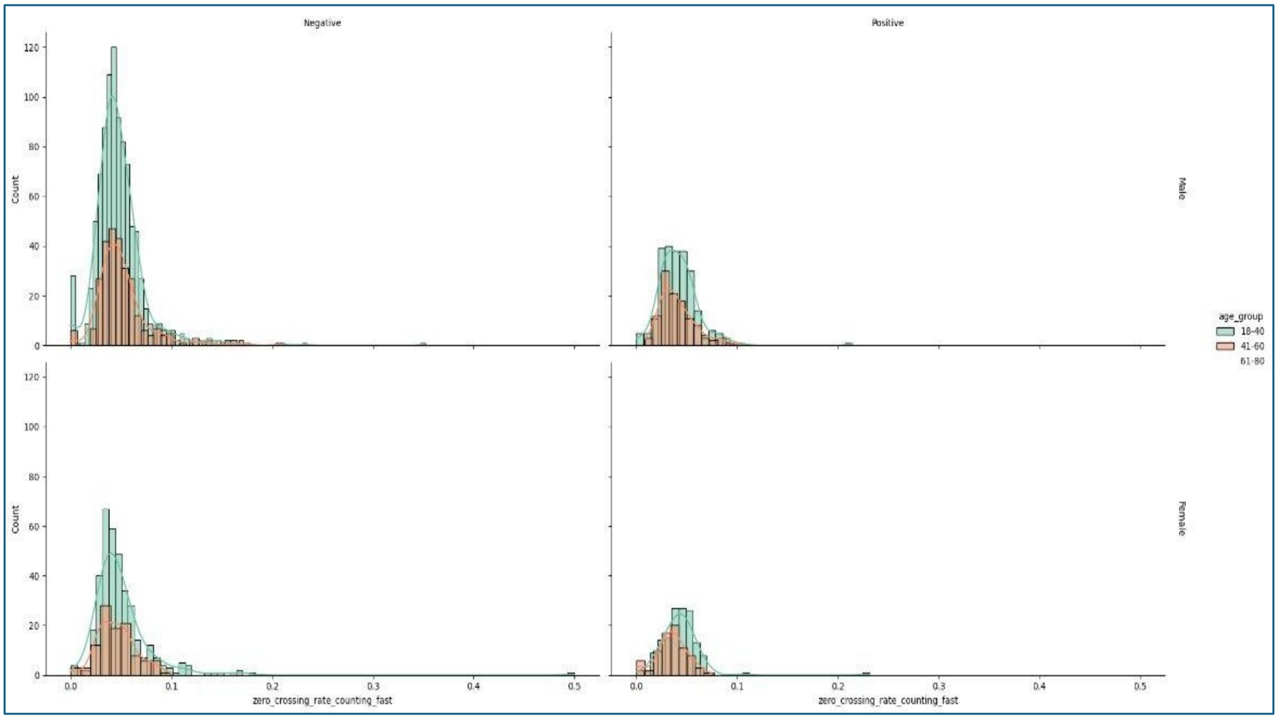
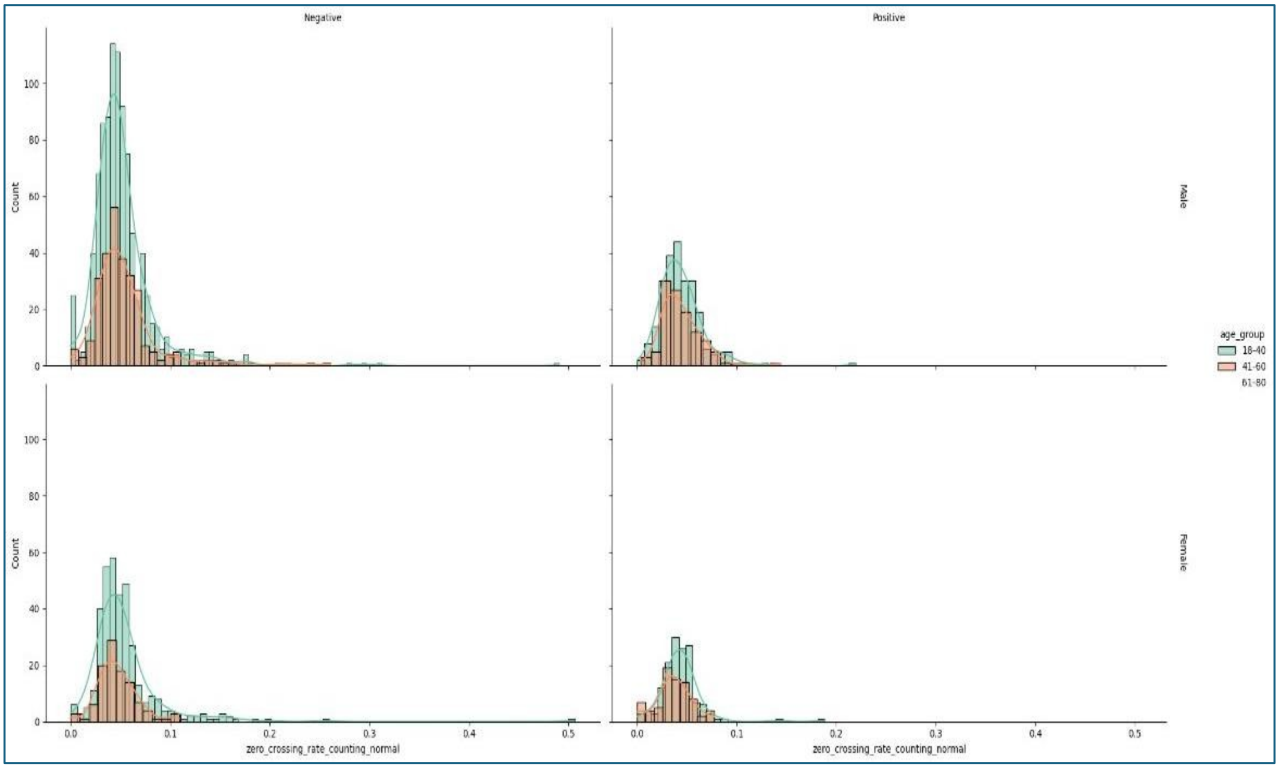


Figure 23 ZCR – Counting fast by AGE GROUP, GENDER AND HEALTH STATUS

Figure 24 ZCR – Counting normal by AGE GROUP, GENDER AND HEALTH STATUS

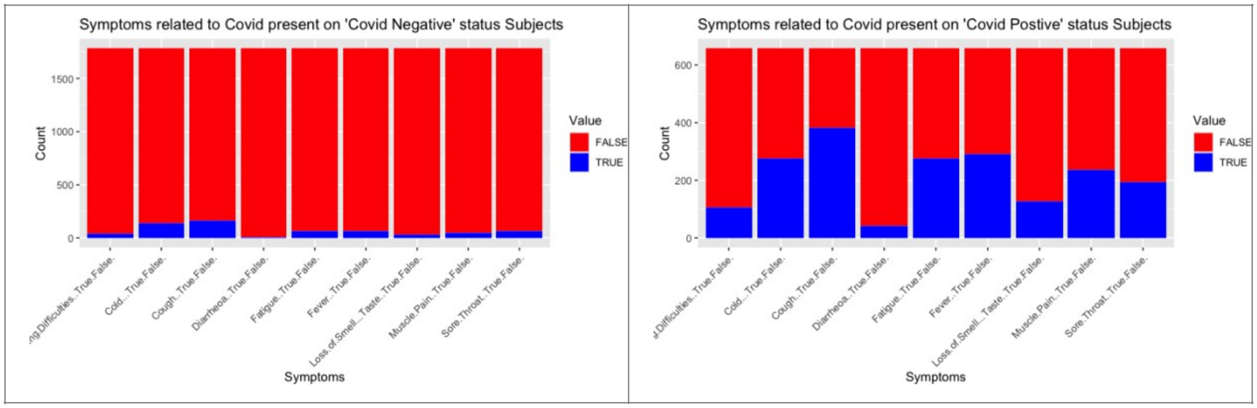


6. Statistical Tests

6.1. Chi-Square

As described in Table 1, there are several symptoms that appear when a subject has Covid. In Figure 12 we display those symptoms in subjects when they reported a "Positive" and a "Negative" covid status. Consequently, among the subjects with a negative covid status most of the symptoms were not prominently present as shown in Figure 12 a); on the other hand, the symptoms were indeed present among the covid positive status subjects, being cough the most frequent symptompresent slightly above the 50% of the covid infected subjects.

Figure 25 Symptoms related to Covid in a) negative covid status and b) positive covid status

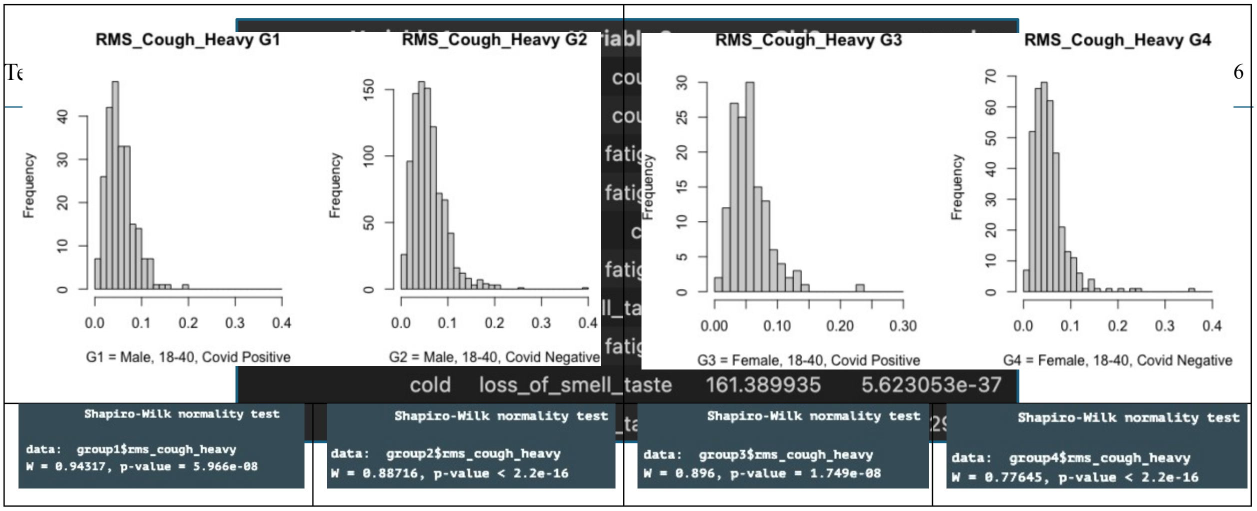


Furthermore, we performed a Chi-Square test to analyze the relationship between the categorical variables: "cough" and "health status" which registered if the subject had covid or not. In table 7 the results are shown.

Table 7 Chi Square Test between Health Status and Cough variable

Hypothesis	Contingency Table	Result	Interpretation
H0: There is no statistical association between the selected variables			Given the result we can reject the null hypothesis and say that there is a statistical relationship between the selected variables.

Additionally, we ran Chi-Square tests between all categorical variables through a python code. Forthis, we considered combinations when there were more than 20 counts in each group inside the contingency table. In the next Figure we present the ten lowest *p-values,* which represent the most significant statistical relationships between the categorical variables. The top ten results comprise the symptoms variables related to Covid, which would be logical as Covid 19 is a trigger for the presence of these symptoms.



```
cold    loss_of_smell_taste    161.389935    5.623053e-37
```

```
        Shapiro-Wilk normality test

data:  group1$rms_cough_heavy
W = 0.94317, p-value = 5.966e-08
```

```
        Shapiro-Wilk normality test

data:  group2$rms_cough_heavy
W = 0.88716, p-value < 2.2e-16
```

```
        Shapiro-Wilk normality test

data:  group3$rms_cough_heavy
W = 0.896, p-value = 1.749e-08
```

```
        Shapiro-Wilk normality test

data:  group4$rms_cough_heavy
W = 0.77645, p-value < 2.2e-16
```

Figure 26 Chi-Square performed with the 10 most relevant p-values

6.2. T-Test

6.2.1. Groups to compare

As shown in section 5.3, the RMS for cough heavy had one of the most normal distribution shapes. That is why we used it to compare differences between the covid status individuals. Also, we usedthe most numerous group in age range which was the 18 to 40. These were the groups:

Feature	Audio File	Group Age	Gender	Covid Status	Group Number
RMS	Cough Heavy	18-40	Male	Positive	1
				Negative	2
			Female	Positive	3
				Negative	4

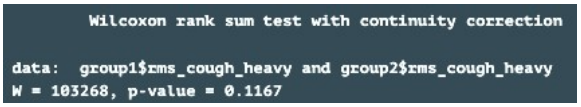
6.2.2. Normality Test

Figure 27 RMS Cough Heavy for the 4 groups

As shown in the results above, the Shapiro-Wilk test to evaluate normality concluded that none ofthe RMS distribution for cough heavy in the four groups had a normal shape, as the p-values whereunder the significance level (0.05). Next, we are going to analyze the groups with a non-parametrictest, which is not affected by the distribution of the data.

6.2.3. Mann Whitney Test

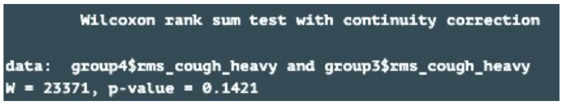
Results for Group 1 and 2



```
         Wilcoxon rank sum test with continuity correction

data:  group1$rms_cough_heavy and group2$rms_cough_heavy
W = 103268, p-value = 0.1167
```

Results for Group 3 and 4



```
         Wilcoxon rank sum test with continuity correction

data:  group4$rms_cough_heavy and group3$rms_cough_heavy
W = 23371, p-value = 0.1421
```

As the p values were greater than the significance level (0.05) we cannot reject the null hypothesis,and therefore there is not a statistical difference between groups 1 and 2, and between groups 3 and 4.

6.3. Anova test

We analyzed various numerical columns extracted from audio features to understand how gender,age, and health status affect these features. Due to the non-normal distribution of the data, we employed multiple statistical tests including ANOVA, Kruskal-Wallis, and Welch's ANOVA to ensure robust results.

6.3.1. Normality and Variance Analysis

Shapiro-Wilk Test Results:

The Shapiro-Wilk test confirmed that none of the columns followed a normal distribution:

Column	Shapiro-Wilk Statistic	p-value	Normal Distribution
rms_cough_heavy	0.881845	0.0	FALSE
rms_cough_shallow	0.827275	0.0	FALSE
zero_crossing_rate_counting_normal	0.708694	0.0	FALSE
zero_crossing_rate_counting_fast	0.746231	0.0	FALSE
rms_vowel_a	0.944696	0.0	FALSE
zero_crossing_rate_vowel_a	0.711047	0.0	FALSE

6.3.2. Homogeneity of Variances

Levene’s and Bartlett’s tests were used to assess variance homogeneity:

Column	Levene’s Test Statistic	Levene’s p-value
rms_cough_heavy	0.943097	0.471750
rms_cough_shallow	1.461245	0.176606
zero_crossing_rate_counting_normal	2.751183	0.007588
zero_crossing_rate_counting_fast	1.532748	0.151506
rms_vowel_a	2.230948	0.029240
zero_crossing_rate_vowel_a	1.326951	0.233179

Given the non-normality and variance issues, we proceeded with non-paramet ric and robust tests.

6.3.3. ANOVA

We conducted ANOVA to explore how different factors affect the numerical features. Significant results were obtained, highlighting the impact of age and interactions among variables:

Variable	Source of Variation	Sum of Squares	df	F	PR(>F)
rms_cough_heavy	(age_group)	0.010655	2.0	4.600946	0.032060
	(gender) (health_status) (age_group)	0.014111	2.0	6.092949	0.013645
rms_cough_shallow	(age_group)	0.033025	2.0	18.373837	0.000019
	(health_status) (age_group)	0.008472	2.0	4.713547	0.030029
	(gender) (health_status) (age_group)	0.009343	2.0	5.198386	0.022700

zero_crossing_rate_counting_normal	(gender) (age_group)	0.012714	2.0	6.380052	0.011608
zero_crossing_rate_counting_fast	(gender) (age_group)	0.009609	2.0	7.112086	0.007711
rms_vowel_a	(age_group)	0.068002	2.0	6.904765	0.008654

6.3.4. Kruskal-Wallis Test Results

Given the non-normal distribution, Kruskal-Wallis tests were conducted to verify the findings from ANOVA:

Column	Group	Kruskal-Wallis Statistic	p-value
rms_cough_heavy	gender	21.497572	3.542769e-06
rms_cough_shallow	gender	10.660557	1.094444e-03
zero_crossing_rate_counting_normal	health_status	46.201839	1.066768e-11
zero_crossing_rate_counting_fast	health_status	55.681593	8.521242e-14
rms_vowel_a	gender	6.455922	1.105832e-02
zero_crossing_rate_vowel_a	health_status	17.777298	2.483287e-05

6.3.5. Welch's ANOVA Results

To address issues with unequal variances, Welch's ANOVA was applied. This test confirmed significant effects and interactions:

Column	Source of Variation	Sum of Squares	df	F	PR(>F)
rms_cough_heavy	(gender) (health_status) (age_group)	0.022736	7.0	2.805012	0.024433
rms_cough_shallow	(age_group)	0.057532	2.0	32.008663	1.727683E-08
	(gender) (health_status) (age_group)	0.019680	7.0	3.128327	0.01408945
zero_crossing_rate_counting_normal	(age_group)	0.008644	2.0	4.337709	0.037388
	(gender) (health_status) (age_group)	0.018977	7.0	2.720816	0.028158
zero_crossing_rate_counting_fast	(gender) (health_status) (age_group)	0.017256	7.0	3.649073	0.005716

7. Model

7.1. Considering the Original Purpose of the Data

Despite the thorough numerical analysis described in this report, the data acquired by the COSWARA team was meant for image-based analysis. Therefore, the conclusions drawn from numerical data alone may not fully capture the nuances of the dataset. To address this issue, we asked ChatGPT to develop a model, where multiple prompts were given to achieve the desired results. ChatGPT developed a Convolutional Neural Network (CNN) model to analyze mel spectrogram images derived from the audio data.

7.2. CNN Model for Image

Analysis**Data Preprocessing**

We preprocessed the mel spectrogram images, resizing them to 224x224 pixels and normalizing the pixel values. The images were then paired with labels indicating whether the corresponding audio data belonged to COVID-19 positive or negative cases.

Model Architecture

The CNN model was built using the following layers:

Convolutional Layers: Extract features from the input images using filters.

Pooling Layers: Reduce the spatial dimensions of the feature maps.

Dense Layers: Perform classification based on the extracted features.

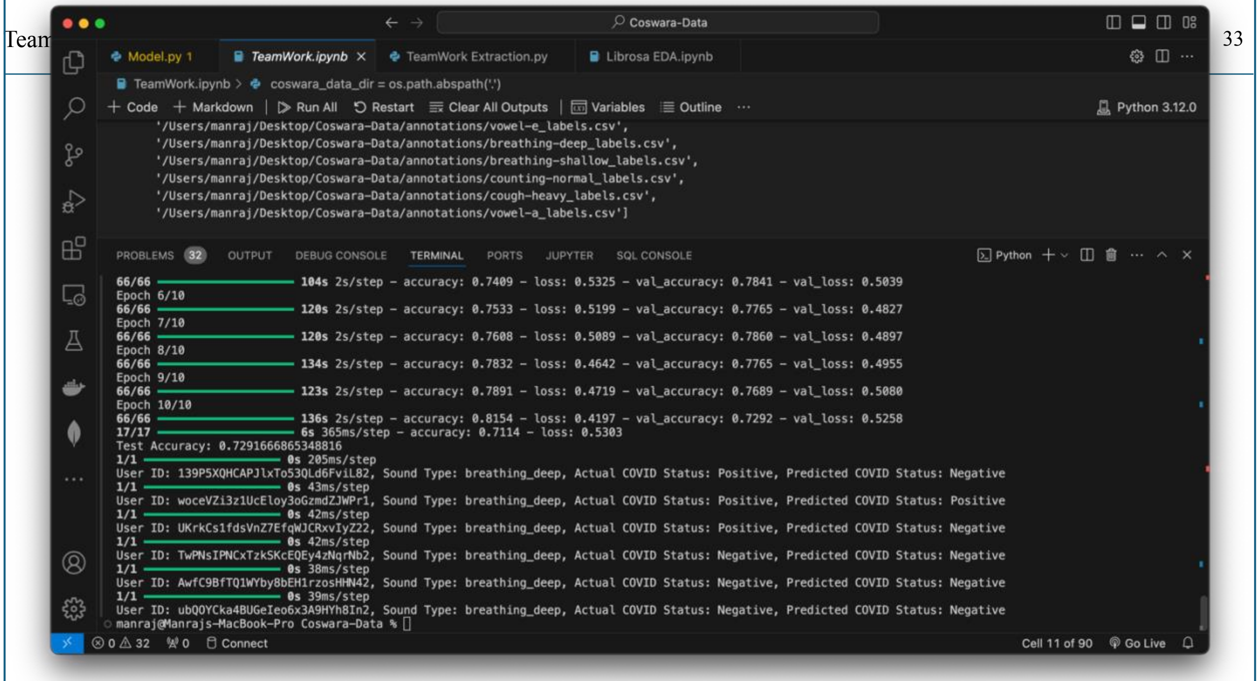
7.3. Training and Evaluation

The model was trained on preprocessed images, with a validation split to assess its performance. The model achieved satisfactory accuracy, demonstrating its effectiveness in distinguishing between COVID-19 positive and negative cases based on mel spectrogram images.

Here is a summary of the training process:

Training Accuracy: Increased consistently across epochs, reaching a final accuracy of 81.54%.

Validation Accuracy: Varied slightly but remained above 72%, indicating a good generalization capability of the model.



```
    '/Users/manraj/Desktop/Coswara-Data/annotations/vowel-e_labels.csv',
    '/Users/manraj/Desktop/Coswara-Data/annotations/breathing-deep_labels.csv',
    '/Users/manraj/Desktop/Coswara-Data/annotations/breathing-shallow_labels.csv',
    '/Users/manraj/Desktop/Coswara-Data/annotations/counting-normal_labels.csv',
    '/Users/manraj/Desktop/Coswara-Data/annotations/cough-heavy_labels.csv',
    '/Users/manraj/Desktop/Coswara-Data/annotations/vowel-a_labels.csv']
```

```
66/66 ━━━━━━━━━━━━━━━━━━━━ 104s 2s/step - accuracy: 0.7409 - loss: 0.5325 - val_accuracy: 0.7841 - val_loss: 0.5039
Epoch 6/10
66/66 ━━━━━━━━━━━━━━━━━━━━ 120s 2s/step - accuracy: 0.7533 - loss: 0.5199 - val_accuracy: 0.7765 - val_loss: 0.4827
Epoch 7/10
66/66 ━━━━━━━━━━━━━━━━━━━━ 120s 2s/step - accuracy: 0.7608 - loss: 0.5089 - val_accuracy: 0.7860 - val_loss: 0.4897
Epoch 8/10
66/66 ━━━━━━━━━━━━━━━━━━━━ 134s 2s/step - accuracy: 0.7832 - loss: 0.4642 - val_accuracy: 0.7765 - val_loss: 0.4955
Epoch 9/10
66/66 ━━━━━━━━━━━━━━━━━━━━ 123s 2s/step - accuracy: 0.7891 - loss: 0.4719 - val_accuracy: 0.7689 - val_loss: 0.5080
Epoch 10/10
66/66 ━━━━━━━━━━━━━━━━━━━━ 136s 2s/step - accuracy: 0.8154 - loss: 0.4197 - val_accuracy: 0.7292 - val_loss: 0.5258
17/17 ━━━━━━━━━━━━━━━━━━━━ 6s 365ms/step - accuracy: 0.7114 - loss: 0.5303
Test Accuracy: 0.7291666865348816
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 205ms/step
User ID: 139P5XQHCAPJlxTo53QLd6FviL82, Sound Type: breathing_deep, Actual COVID Status: Positive, Predicted COVID Status: Negative
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 43ms/step
User ID: woceVZi3z1UcEloy3oGzmdZJWPr1, Sound Type: breathing_deep, Actual COVID Status: Positive, Predicted COVID Status: Positive
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 42ms/step
User ID: UKrkCs1fdsVnZ7EfqWJCRxvIyZ22, Sound Type: breathing_deep, Actual COVID Status: Positive, Predicted COVID Status: Negative
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 42ms/step
User ID: TwPNsIPNCxTzkSKcEQEy4zNqrNb2, Sound Type: breathing_deep, Actual COVID Status: Negative, Predicted COVID Status: Negative
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 38ms/step
User ID: AwfC9BfTQ1WYby8bEH1rzosHHN42, Sound Type: breathing_deep, Actual COVID Status: Negative, Predicted COVID Status: Negative
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 39ms/step
User ID: ubQOYCka4BUGeIeo6x3A9HYh8In2, Sound Type: breathing_deep, Actual COVID Status: Negative, Predicted COVID Status: Negative
manraj@Manrajs-MacBook-Pro Coswara-Data % 
```

7.4. Predicting COVID-19 Status for Random Users

To illustrate the model's capability, we randomly selected user IDs and predicted their COVID-19 status using the trained CNN model. The predictions were compared with the actual test results to evaluate the model's accuracy in real-world scenarios. Here are some example results:

User ID	Sound Type	Actual COVID Status	Predicted COVID Status
139P5XQHCAPJLT0s3Oq	breathing_deep	Positive	Negative
woceVZ1z3iuELOyg3o4	breathing_deep	Positive	Positive
UKrCk1sfdsVnZ7EF4vM	breathing_deep	Negative	Negative
TwPNSlPNCXTKzSKcoE4	breathing_deep	Negative	Negative
AwfC9BfTQlYYbBfHZr4	breathing_deep	Negative	Negative

Figure 28 Model's Output

Results:

In summary, our analysis and model development efforts have provided valuable insights into the factors influencing audio features in COVID-19 cases. The CNN model demonstrates promising accuracy in predicting COVID-19 status based on mel spectrogram images, complementing the numerical analysis. This dual approach allows for a comprehensive understanding of the dataset, leveraging both numerical and image-based analyses to enhance our conclusions.

This combination of statistical analysis and machine learning provides a robust framework for understanding and predicting COVID-19 status from audio data. Moving forward, further refinement of the model and validation on larger datasets will be crucial to improve its accuracy and reliability.

8. Findings

The numerical variables we extracted from the audios with the Librosa library in Python had a high variance and were not normally distributed. This might be caused by the way the audios were recorded, proving that there were not defined parameters or guidelines to record the samples. Nevertheless, we did not deal with these inconsistencies and proceeded to work with the data; therefore, the statistical tests that we conducted might not be accurate but are a start for future analysis. Next are the key findings:

- **Impact of Age on Audio Features:** The age group consistently shows significant effects on several audio features, such as 'rms_cough_heavy', 'rms_cough_shallow', 'zero_crossing_rate_counting_normal', and 'zero_crossing_rate_counting_fast'. This suggests that different age groups might exhibit distinct patterns in cough intensity and frequency characteristics, which could be related to physiological differences or varying responses to respiratory conditions.
- **Role of Gender, Age and Health Status:** The interaction between gender, health status, and age significantly affects features like 'rms_cough_heavy' and 'rms_cough_shallow'. This implies that the combined effect of these factors can influence cough characteristics differently, possibly due to gender-specific health conditions or how different health statuses interact with age.
- **Relationship between covid symptoms:** After running a chi-square test for all the categorical variables, we found that among the most significant p-values were the tests between the covid related symptoms such as cold, cough, fatigue, loss of smell and the covid status variable.
- **Relationship between the RMS feature in the cough heavy audio for the selected groups:** Since the RMS for cough heavy for the four groups was not normally distributed, we used a non-parametric test. The Mann Whitney test performed using R (Wilcoxon test) told that the p values where higher than the significance level, so we concluded that there was no statistical difference between the individuals that had Covid and those who did not have Covid, for the groups chosen (age 18-40, gender, health status) and the feature and audio selected.
- **Model by ChatGPT:** We feed the model with the mel spectrogram extracted from the breathing deep audio, and the model's accuracy was 72.9 %.

9. Recommendations

- We recommend that anyone working with these audio files ensure proper processing to manage inconsistencies in duration and noise. For example, apply filtering to reduce noiseand remove silence. Additionally, standardize the duration of the audio files or use techniques like windowing to handle varying lengths. These steps are essential as inconsistencies can significantly affect statistical tests on features such as RMS and ZCR.
- Establish and follow clear guidelines for audio recording conditions in future studies. This includes maintaining the same distance from the microphone, using high-quality recording equipment, and conducting recordings in a controlled environment to reduce variations andimprove data comparability.
- Before conducting statistical tests that assume normality, assess the distribution of your data using normality tests such as Shapiro-Wilk or Kolmogorov-Smirnov. If the data are not normally distributed, consider data transformations or non-parametric tests.
- For the model we only used the mel spectrogram for breathing deep, but we had some otherimages from the other audios, so in order to increase credibility and accuracy, the other images could be used together to predict the covid status.

10. Bibliography

Parekh, Ranjan. 2011. ""Automated Discrimination of Digital Audio."." *International Journal of Engineering Research and Applications (IJERA)*. Accessed July 30, 2024. https://www.ijera.com.