scientific **data**

DATA DESCRIPTOR

Check for updates

Coswara: A respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection

Debarpan Bhattacharya[1], Neeraj Kumar Sharma[2], Debottam Dutta[1], Srikanth Raj Chetupalli[3], Pravin Mote[1], Sriram Ganapathy[1 ✉], C. Chandrakiran[4], Sahiti Nori[4], K. K. Suhail[4], Sadhana Gonuguntla[5] & Murali Alagesan[6]

This paper presents the Coswara dataset, a dataset containing diverse set of respiratory sounds and rich meta-data, recorded between April-2020 and February-2022 from 2635 individuals (1819 SARS-CoV-2 negative, 674 positive, and 142 recovered subjects). The respiratory sounds contained nine sound categories associated with variants of breathing, cough and speech. The rich metadata contained demographic information associated with age, gender and geographic location, as well as the health information relating to the symptoms, pre-existing respiratory ailments, comorbidity and SARS-CoV-2 test status. Our study is the first of its kind to manually annotate the audio quality of the entire dataset (amounting to 65 hours) through manual listening. The paper summarizes the data collection procedure, demographic, symptoms and audio data information. A COVID-19 classifier based on bi-directional long short-term (BLSTM) architecture, is trained and evaluated on the different population sub-groups contained in the dataset to understand the bias/fairness of the model. This enabled the analysis of the impact of gender, geographic location, date of recording, and language proficiency on the COVID-19 detection performance.

Background & Summary

As of July 2022, with more than 550 million reported COVID-19 cases and a fatality ratio of more than 1%, the COVID-19 pandemic has emerged as the most consequential global health crisis since the influenza pandemic of 1918[1]. The outbreak has largely outpaced global efforts to characterize the infection and contain its spread. The measures such as population surveillance, case identification, contact tracing, vaccination, physical distancing, mask mandates, and lock-downs have helped control the outbreak to a certain extent. Inventing alternative COVID-19 screening methodologies, which can function as point-of-care tests (POCT), and are efficient in terms of time, cost and performance was highlighted as an urgent requirement by the World Health Organization (WHO)[2]. The application of digital technologies for large-scale health-related data collection and analysis can help meet this requirement[3,4] and build infrastructure to tackle future pandemics. The longitudinal studies, recording self-reported symptoms using smartphone-based applications, had explored the development of easily accessible COVID-19 screening tools[5–8]. Besides symptoms, measurement of physiological signals such as respiration rate using wearable electronic sensors had also been explored for COVID-19 detection[9].

The COVID-19 is primarily a respiratory illness[10]. Listening to respiratory sounds, such as deep breathing, with a stethoscope has served as a useful methodology to screen respiratory ailments since 1821[11]. Currently, utilizing digital technologies, respiratory sound samples can be collected via internet connected devices. The computer aided analysis of respiratory sounds can bring new insights for the design of POCT solutions as well as in the monitoring of SARS-CoV-2 infection. Motivated by this, we conceptualized the creation of a respiratory sound and symptom dataset named *Coswara*, (the word Coswara is an amalgamation of "Co" (from COVID-19) and "swara" (sound in Sanskrit)), composed of breathing, cough, and speech sounds along with the health related symptoms, recorded from individuals with, without, and recovered from SARS-CoV-2 infection[12].

[1]Department of Electrical Engineering, Indian Institute of Science, Bangalore, India. [2]Mehta Family School of Data Science and Artificial Intelligence, Indian Institute of Technology Guwahati, Guwahati, India. [3]Fraunhofer Institute of Integrated Circuits, Erlangen, Germany. [4]Ramaiah Medical College Hospital, Bangalore, India. [5]General Hospital, Hoskote, Bangalore, India. [6]PSG Institute of Medical Sciences and Research, Coimbatore, India. ✉e-mail: sriramg@iisc.ac.in

Dataset Name	Data Collected From	Data Access	Meta-data	Sound Data	Sound Samples per Subject	Uncompressed audio files	Audio Quality Validation	Subject Count			COVID-19 Screening Tool Released
								Non-COVID-19	COVID-19	Recovered from COVID-19	
COUGHVID[13]	EU and beyond	Open	Age Gender Country Symptoms Medical history	Cough	1	×	×	26395	1155	×	×
COVID-19 Sounds[33]	EU, USA, and beyond	On request	Gender Symptoms Medical history	Cough Breathing Speech	3	×	Automatic	6450	2106	×	×
Tos COVID-19[15]	Argentina	Open	Age Gender Location Symptoms	Cough	1	×	×	125183	21197	×	×
Coswara (This work)	India and beyond	Open	Age Gender State Country Symptoms Medical history (includes respiratory health, COVID-19 vaccination status, other diseases, and more)	Cough (2 types) Breathing (2 types) Vowel (3 types) Speech (2 types)	9	✓	Manual Listening	1819	674	142	✓

Table 1. List of publicly accessible COVID-19 related respiratory sound datasets.

This paper presents a detailed description of the data collection protocol, a summary of the data records, and results from bias analysis performed with a pre-trained classifier model.

In comparison with other efforts on respiratory sound sample collection[13–15], the Coswara dataset was primarily collected from India. Further, the data records were collected between April-2020 to February-2022, allowing the Coswara dataset to contain records associated with infections due to multiple variants of the SARS-CoV-2 virus. The participant metadata provided in the Coswara dataset spanned a wide range of attributes and included demographic, health, symptom, and COVID-19 status information. For respiratory sound recordings, each participant provided nine sound recordings. These correspond to breathing (breathing deep and breathing shallow), cough (coughing deep and coughing shallow), sustained vowel phonation (three different vowels), and continuous speech (counting from 1–20 at normal and fast pace). These sound recordings were released without application of lossy compression standards. In order to validate the audio files that were collected using the crowd-sourced web-link, the human annotators listened to all the sound recordings and marked the quality in a scale of 0–2. The sound quality annotations obtained from human listening were also released as part of the dataset. A detailed comparison of the Coswara dataset with few other relevant dataset is provided in Table 1. As shown in this Table, the Coswara dataset is unique in terms of containing nine different types of sound samples, augmented with a rich meta-data, audio quality annotation obtained through human listening. The dataset also contains labels associated with COVID-19 recovered subject population. A COVID-19 screening tool designed using this dataset is made publicly available. Furthermore, we report a detailed bias and fairness analysis to understand the confounding factors associated with the COVID-19 screening tool developed using the collected data.

India shares about 18% of the global population and features approximately 32% of the global disability adjusted life years (DALY, an estimate of overall disease burden) from respiratory diseases[16]. In addition to the analysis of the health symptoms and respiratory acoustics associated with COVID-19 disease, we foresee the Coswara dataset to be of future relevance to understand respiratory acoustics and aid in the design of cost-effective, scalable, and accurate respiratory health monitoring technologies.

Methods

Design.   The Coswara dataset was based on crowdsourcing via a website (https://coswara.iisc.ac.in/). The application interface was designed with a simple workflow (see Fig. 1). Any individual connected to the internet could access the website using a smartphone (or a computer). The first step was the collection of the participant's consent to record their audio. Subsequently, the participant filled out a short questionnaire to provide demographic information, current health status, and COVID-19 test results. A detailed list of all the metadata recorded is provided in Table 3. The participant recorded nine different kinds of sound samples. A description of the sound samples recorded is provided in Table 2.

Demographics data.   The participants provided demographic information like age, gender, country and the location information in terms of province. Further, they also indicated whether they were proficient in English, and whether they had a face mask during the recording. The smoking status was also recorded.
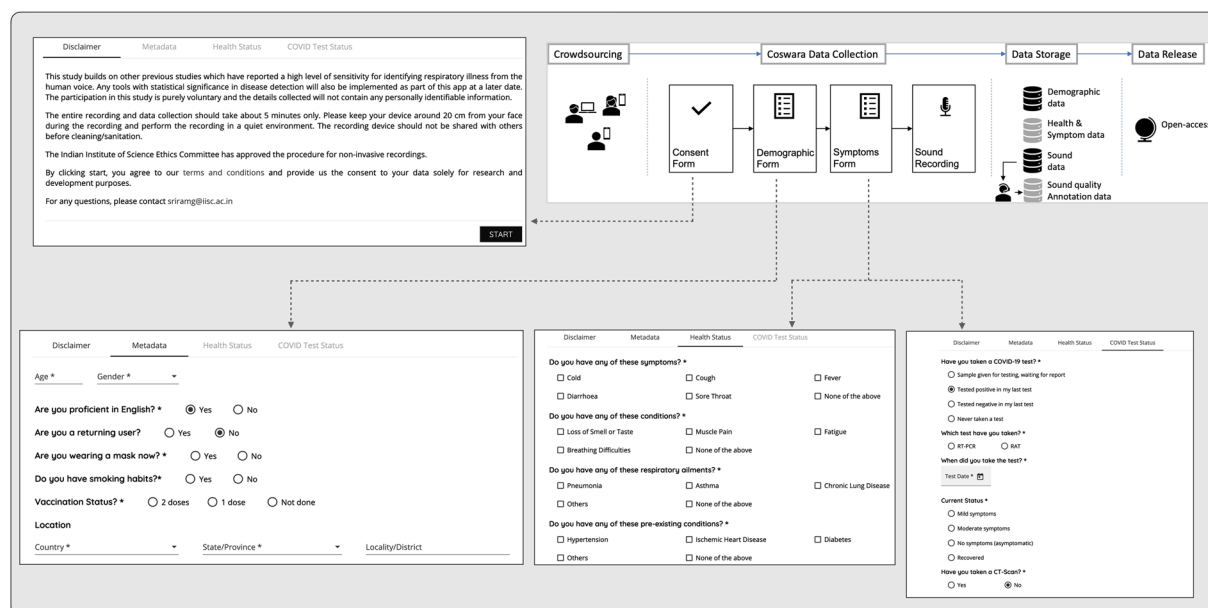
Fig. 1 A block diagram illustration of the steps involved in creating the Coswara dataset.

Symptom data.   The participants provided information about the presence of COVID-19-like symptoms, respiratory ailments, comorbidity, and COVID-19 test status. For each of these, a list of options was provided and the participants chose as many options as were applicable to them. A summary of the list of symptoms provided by the participants is shown in Table 3.

COVID-19 test status.   The participants provided information on their COVID-19 status by classifying themselves into one of the three categories, namely, COVID-19 negative (non-COVID), positive (COVID), and recovered categories. If the participant belonged to a positive or recovered category, they further provided the date of the last COVID-19 test conducted and the kind of test (RAT or RT-PCR).

Respiratory sound data.   The website application provided a sound recording interface. The participants were instructed to record and upload nine sound samples, sequentially as different audio files. An example sound sample for each sound category was also provided. The nine sound samples included two variants of breathing, two variants of cough, three variants of sustained vowel phonation, and two variants of continuous speech. A description of these sound categories is provided in Table 2.

Procedures.   The website URL was shared with the public through various social media platforms and word-of-mouth in the collaborating hospitals and health centers. The crowd-sourcing approach to data collection enabled the recording of data from diverse age groups, geographic locations (within India and to a smaller extent, from outside India), health conditions, sound recording device types and ambient environments. A participant's complete data record comprised of demographic, symptom, COVID status, and sound data. On average, each participant spent approximately 7 minutes to record their data. A significant portion (>95%) of the subjects who were SARS-CoV-2 infected, were inducted through hospitals and health centers. Hence, the label quality of these data records were validated.

Ethical issues.   The data collection procedure was approved by the Institutional Human Ethics Committee, at the Indian Institute of Science, Bangalore. The informed consent was obtained from all participants who uploaded their data records. All the data collected was anonymized, and excluded any participant identity information.

Data Records

Manual curation.   Approximately 65 hours of respiratory sound samples (23700 recordings coming from 2635 subjects each having 9 categories of recordings) were subjected to manual listening by the human annotators. The listeners classified the recordings based on the sound quality into one of the three categories, namely, excellent (no ambient noise), moderate (slight ambient noise), and poor (significant ambient noise and distortions). The curation process resulted in identifying 78.6% of the samples as excellent, 11.7% as moderate, and 9.7% as poor quality sound recordings, respectively.

The Coswara dataset is publicly available as a Zenodo repository[17].

Sound Category	Collected Sound Sample	Description
Breathing	Breathing-shallow	Few cycles of inspiration and expiration without exertion on the lungs
	Breathing-deep	Few cycles of inspiration and expiration with exertion on the lungs
Cough	Cough-shallow	Few bouts of (induced) cough without exertion on lungs
	Cough-heavy	Few bouts of (induced) cough with exertion on lungs
Vowel Phonation	Vowel-[u]	Sustained phonation of vowel-[u] (as in the word boot)
	Vowel-[i]	Sustained phonation of vowel-[i] (as in the word beet)
	Vowel-[æ]	Sustained phonation of vowel-[æ] (as in the word bat)
Speech	Counting-normal	Counting numbers from 1 to 20 in normal pace
	Counting-fast	Counting numbers from 1 to 20 in fast pace

Table 2. Description of the sound categories in the Coswara dataset.

Metadata Type	Field name	Description	Allowed values
Demographic	id	participant identifier	assigned using a unique random code generator
	record_data	data recording date	dd-mm-yy
	a	age	number
	g	gender	male/female/other
	l_c	country	one from a list of 217 countries
	l_s	state	one from a list of province associated with l_c
	vacc	COVID-19 vaccination status	y (at least two doses)/p (one dose)/n (unvaccinated)
	ep	proficient in english	True/False
	smoker	regular smoker	True/False
	rU	returning participant	True/False
	um	wearing mask	True/False
COVID-19-like symptoms	cough	has cough	True/False
	cold	has cold	True/False
	diarrhoea	has diarrhoea	True/False
	bd	has breathing difficulties	True/False
	st	has sore throat	True/False
	fever	has fever	True/False
	ftg	suffering from fatigue	True/False
	mp	has muscle pain	True/False
	has loss_of_smell	loss of smell and/or taste	True/False
Respiratory ailments	asthma	has asthma related issues	yes/no
	cld	has chronic lung disease	True/False
	pneumonia	has pneumonia	True/False
	others_resp	has other respiratory illness	True/False
Comorbidity	ht	has hypertension	True/False
	diabetes	has diabetes	True/False
	ihd	has ischemic heart disease	True/False
	others_preexist	any other pre-existing comorbidity	True/False
COVID-19 health	test_status	status of COVID-19 test	p (positive)/n (negative)/na (not taken a test)
	covid_status	covid related health status	positive_mild, healthy, positive_moderate, positive_asymptomatic, exposed
	testType	type of COVID-19 test taken	RAT/RT-PCR
	test_date	date of COVID-19 test	dd/mm/yy
	ctDate	date of CT-scan	dd/mm/yy
	ctScore	CT value	number
	ctScan	the participant had a CT-scan	True/False

Table 3. Description of the meta-data collected and released as part of the Coswara dataset.

Technical Validation

COVID-19 representation. From the perspective of SARS-CoV-2 infection, the dataset was comprised of three kinds of subject categories (see Fig. 2a). First, the *non-COVID* category contained data records from 1819 subjects. These subjects were completely healthy (78%), had some respiratory ailments (9%), or had COVID-19 like symptoms (13%). Further, a subset of these participants self-reported themselves as exposed to the virus
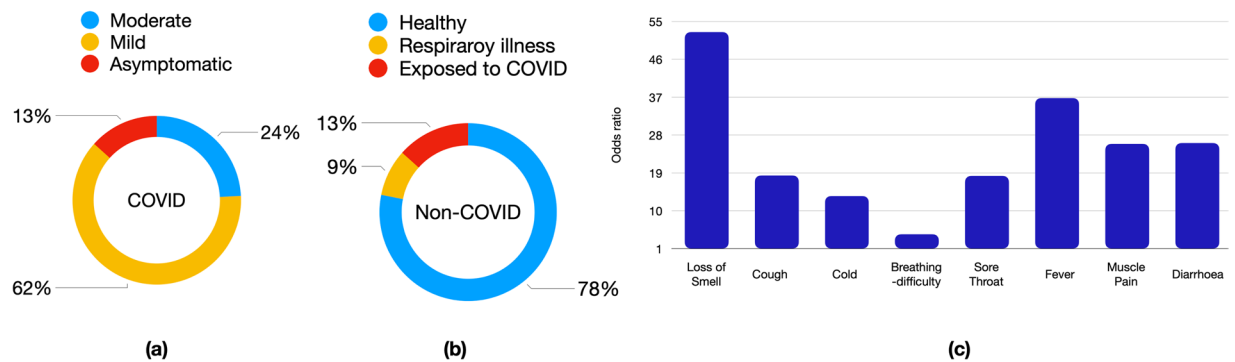
Fig. 2 Distribution of the (**a**) COVID, and (**b**) non-COVID subjects corresponding to different sub-categories. Figure (**c**) shows the odds ratio for the different symptoms.

(13%), but had not tested positive for SARS-CoV-2 infection in the past or at the time of data recording. Second, the *COVID* category contained data records from 674 subjects. These individuals had tested positive for SARS-CoV-2 infection at the time of data recording or in the past 14 days. Using the self-reported health condition, these individuals were further grouped into asymptomatic (14%), mild (62%) and moderate (24%) COVID-19 patients. Third, the *recovered* category contained data records from 142 subjects. These subjects had completed at least 14 days since testing positive for COVID-19.

Demographic representation. The demographic distribution for each subject category, namely, non-COVID, COVID and recovered, is shown in Fig. 3. The participating subjects spanned a wide age group (between 15–90 years), with a majority between 15–45 years. Further, the dataset also contained more male participants. The geographic distribution of the data was concentrated primarily in India (91% from India). Within India, a majority of the data came from Karnataka, a province in southern India. The rest of the data was drawn from various other provinces across India.

Symptom representation. The distribution of the symptoms among the non-COVID and COVID subjects is shown in Fig. 4. As expected, the COVID-like symptoms such as cold, cough, fever and fatigue, were relatively more prevalent in COVID subjects compared to non-COVID. In the non-COVID class, a majority of the subjects were healthy, i.e., without any COVID-19-like symptoms, respiratory ailments, or comorbidity. However, there were considerable number of non-COVID subjects with COVID-19-like symptoms of fever (225 individuals), cough (173 individuals) and cold (90 individuals). Further, there were also several non-COVID subjects with pre-existing respiratory ailments of pneumonia (83 individuals), Asthma (110 individuals) and other respiratory illnesses (120 individuals). Hence, the non-COVID category data broadly represents the population level incidence of different respiratory ailments. Some of the subjects, also indicated a vaccination status of 2 doses.

Figure 2c depicts the odds ratio, a statistic that quantifies the strength of the association between health status and a symptom. It is defined as the ratio of the odds of "symptom x" in the COVID category to the odds of "symptom x" in the non-COVID category. Here, we see that most of the COVID-19 like-symptoms had an odds ratio >1.

Sound recording specification. The sound files were recorded and stored in an uncompressed audio format. Figure 5a depicts the duration distribution of all the sound files after discarding the audio files of quality rating 0 (poor quality, as identified by the human listeners). The maximum duration in the recording setup was limited to 30 s. Each dot in the figure corresponds to the duration of a single audio file. Additionally, a standard box plot representation is overlaid on the distribution to highlight the median, 25th, and 75th percentile etc. Figure 5a indicates that, on average, the cough samples have smaller duration compared to breathing and sustained vowel phonation. Also, as expected, the counting-normal audio files are on average longer in duration than those corresponding to counting-fast.

Within this duration range, the nine sound categories had a different distribution. On average, the cough recordings had a relatively smaller duration (median of 5 s) while the breathing-deep recordings had the longest duration (median of 15 s). The three vowel sounds corresponding to sustained phonation had a similar distribution. Further, a majority of the sound samples were recorded at 48 kHz (90%) (see Fig. 5b)

Sound category classification. The nine sound categories were chosen such that the excitation and physical state of the respiratory system is well captured. We validated the complimentary nature of the acoustic sounds by building a nine-class classifier amongst the different sound samples. The classifier was trained on acoustic features extracted from the sound samples. From each sound file, a 128-dimensional averaged mel-spectrum vector was extracted. This mel-spectrum was computed by averaging the short-time mel-spectrogram obtained with a window of 25 msec duration, 10 msec shift and with 128 mel-spaced spectral filters[18].

A random forest classifier was trained using the acoustic feature data. A 70-15-15% train-val-test split made by (stratified) random sampling was done for training, validation and testing of the classifier, respectively. The "gini" impurity criterion was used to train the classifier and the number of estimators was optimized using the validation dataset. A test set accuracy of 56.5% was obtained. Figure 6 shows the resulting confusion matrix for

Fig. 3 Distribution of the (**a**) non-COVID (1819 subjects), (**b**) COVID-19 positive (674 subjects), and (**c**) recovered from COVID-19 (142 subjects) population in terms of gender, age and location.
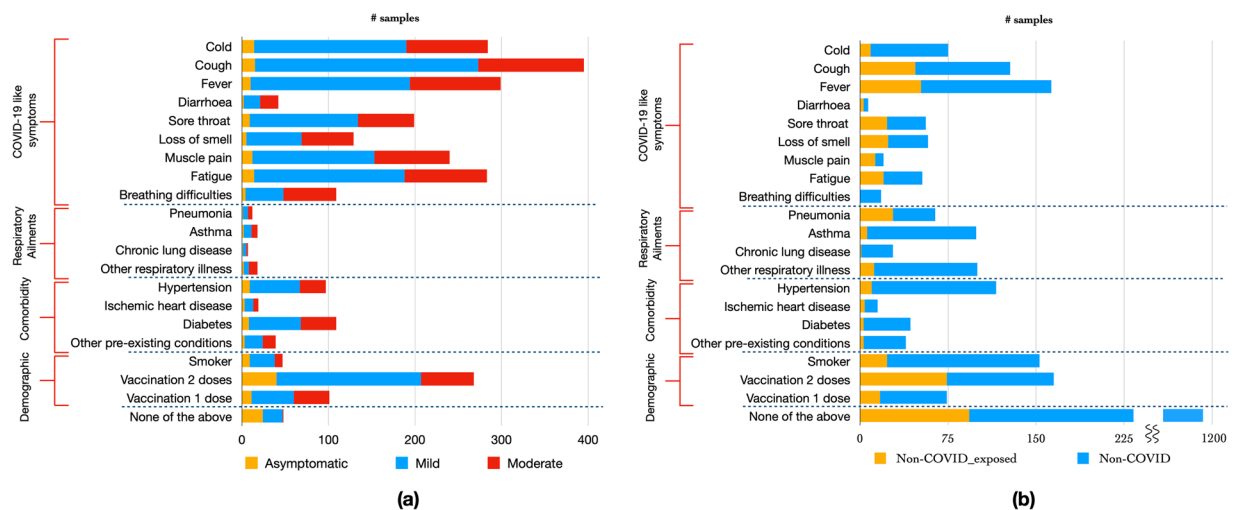

Fig. 4 (**a**) COVID-19 positive and (**b**) non-COVID subject meta data distribution.

the test data. Across all sound categories the accuracy was significantly greater than the chance performance of 11.1%. The confusion matrix also indicates a block diagonal like structure indicating that the class confusions mainly reside within the broad sound categories such as breathing, cough, and speech counting.
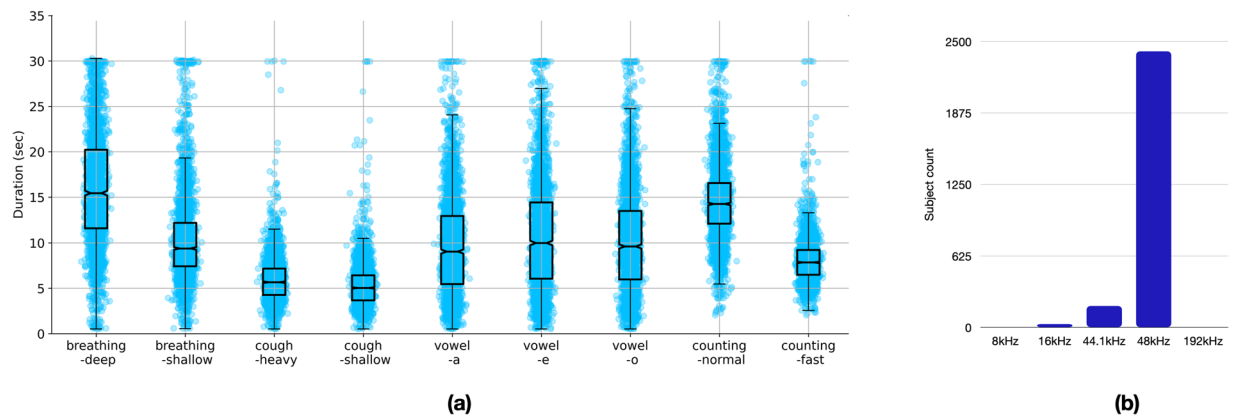
Fig. 5 Distribution of the (**a**) sound recording duration corresponding to the 9 categories, (**b**) sampling rate of the devices used by the subjects.
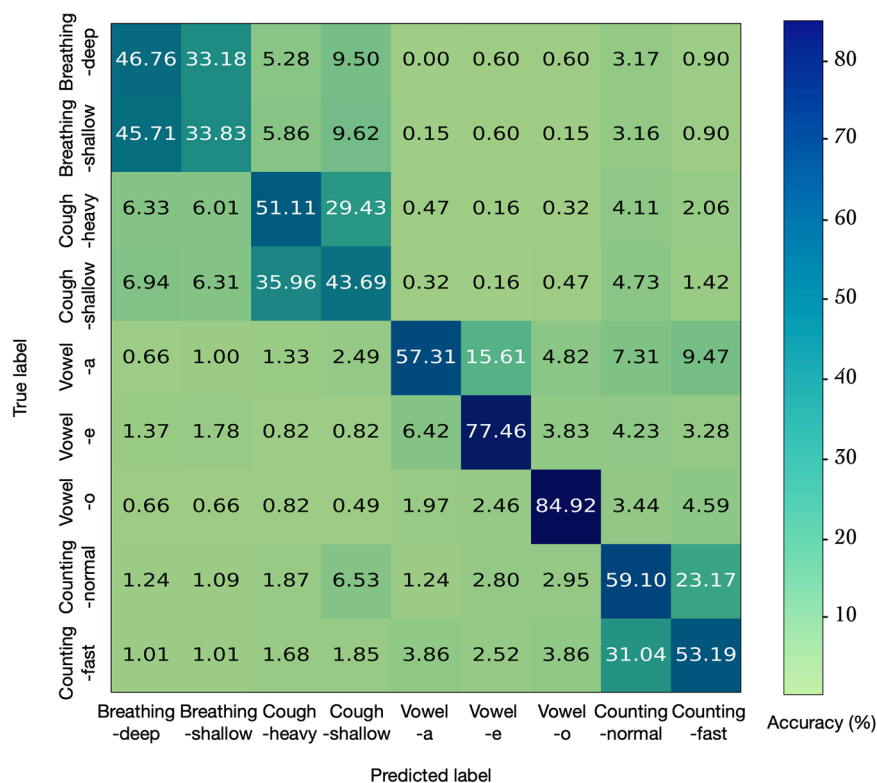

Fig. 6 Confusion matrix for the audio category classification.

Gender classification from sound recordings. According to speech production literature, gender prediction with speech sound recordings is a relatively easy task. In our dataset containing diverse sound samples, we validate the sound recordings for the presence of information related to the participant's gender. We performed a binary classification task, that is, male versus female, across all the sound categories. A stratified random sampling based 70-15-15% train-val-test split was used for training, validation and test dataset creation, respectively. A random forest classifier with "gini" impurity criterion was used. The area-under-receiver-operating-characteristic curve (AUC-ROC) was used as the performance measure. Figure 7 shows the classifier performance on the test set across different sound categories. The AUC was ≥90% for speech sound categories such as vowels and counting. Compared to speech, the performance degrades significantly for cough sound categories and for breathing sound categories.

COVID-19 detection from sound recordings. Since the first release in April-2020, the Coswara dataset has gradually increased in size. The partitions of the dataset have been used by different research groups[19–28] for evaluating the possibility of COVID-19 detection using respiratory sound recordings. Further, two data
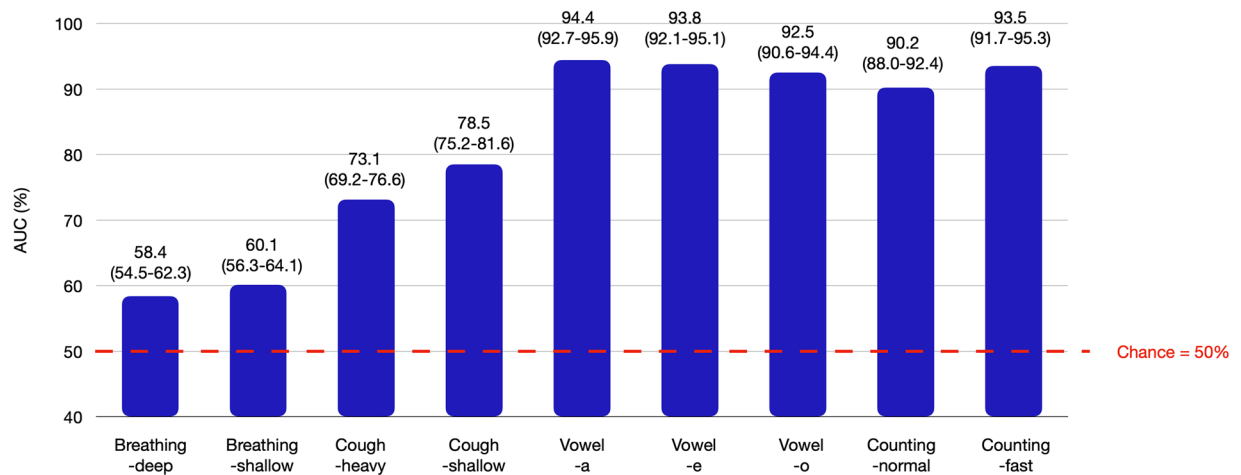
Fig. 7 Test-set area under the curve (AUC) (%) performance for the gender classification task from 9 categories of respiratory sounds.

challenges named, diagnosis of COVID-19 using acoustics (DiCOVA)[29], were conducted using a subset of the data. These challenges invited participants to build machine learning systems for COVID-19 detection using the different sound categories[26,30]. These efforts demonstrated that the dataset can be used for developing machine learning algorithms for COVID-19 screening. Further, an effort was undertaken to understand the impact of different variants of COVID-19 on the audio data[31].

In this paper, we trained the bi-directional long short term memory (BLSTM) based classifier[30] on the proposed Coswara dataset for the binary task of COVID/non-COVID classification. The classifier consisted of two BLSTM layers followed by a fully connected classification head. The classifier was trained on the segments extracted from the audio samples. The log-mel spectrogram was used as feature for training the classifier. The classifier was trained using the weighted binary cross-entropy (BCE) loss. Let, N_c and N_{nc} be the count of COVID and non-COVID subjects used in training, respectively. Let $r = N_c/N_{nc}$ be the class ratio. Then, the total loss is,

$$L = -\sum_{i \in c} log(x_i) - r \sum_{i \in nc} log(1 - x_i)$$

(1)

where, x_i denotes the probability of COVID-19 predicted by the model, and c (nc) denotes the set of COVID (non-COVID) samples.

The participants having a moderate or excellent quality for the audio recordings, for all the nine categories, were selected. The final COVID probability score was obtained after linearly combining scores from the classifiers trained on the nine different audio modalities and the symptoms. The filtered dataset was randomly split into train, validation and test sets consisting of 238, 58 and 96 COVID positive subjects, respectively and 707, 178 and 291 non-COVID subjects, respectively. The area-under-the-receiver-operating-characteristics (AUC-ROC) was used as the evaluation metric as the dataset was imbalanced. The BLSTM classifier achieved an AUC of 91.5%[95% CI 88.5%–94.1%] and sensitivity of 64.6%[95% CI 48.3%–72.6%] calculated at 95% specificity on the test set.

Analyzing bias in COVID-19 detection. The diverse set of metadata, as described in Table 3, also allowed us to analyze the bias and fairness of the trained BLSTM based classifier. In order to analyze the dependence of the model performance on the metadata information, we divided the test set into different population subgroups based on gender, age, vaccination status, mask usage, English proficiency, date of data collection, and province. The COVID-19 detection performance on these subgroups is reported in Table 4. For all the AUC and sensitivity values reported, the two-sided 95% confidence intervals (CIs) were calculated using bootstrap re-sampling with 1000 bootstrap samples (sampled with replacement)[32]. The p-values are obtained by using a t-test for comparing the scores obtained for the full test set and the scores for the population subgroups.

The subject count varied across different subgroups. In the gender category, similar performance was observed for both male and female subgroups. The scores when compared with those obtained for the full test set were not significantly different.

The geographic location was split based on participants from within the Indian province of Karnataka and the rest of India. The resulting performance was not significantly different from the original full test set. In the original test-set, the COVID prevalence was 25%. We experimented with different subsets of the test set by randomly selecting non-COVID samples to match a COVID prevalence level of 2.5%, 5%, 10% and 20%. As seen in Table 4, the test sets with different prevalence levels did not generate statistically significant differences when compared with the original test set. We analyzed the performance attained on two test sets created based on the time period of data collection, that is, chosen as a period before and after the onset of the Omicron variant in India. Note that the training data comprised of data samples collected before December, 2021. The model performance showed similar performance on samples collected after December, 2021 (see Table 4). This indicates

Subject sub-population	Count COVID/Non-COVID	AUC (%)	Sensitivity at 95% specificity	p-value
Gender				
Male	51/209	92.7 (89.3–95.4)	64.7 (37.3–76.7)	0.901
Female	45/82	90.3 (85.1–95.1)	64.4 (48.9–80.0)	0.814
Age				
15–30	39/143	90.0 (85.1–94.3)	59.0 (41.7–76.5)	0.179
30–45	20/86	89.8 (80.8–96.4)	55.0 (33.3–76.5)	0.242
45–60	22/35	94.4 (89.6–98.6)	68.2 (32.1–95.5)	0.590
60–90	8/6	91.7 (75.0–100.0)	50.0 (25.0–100.0)	0.129
Vaccination				
With	66/57	86.3 (80.8–91.3)	31.8 (14.1–70.4)	p < 0.050
Without	30/234	88.8 (82.3–93.9)	53.3 (31.2–78.9)	0.260
Mask				
With	45/36	83.8 (76.1–90.6)	46.7 (12.2–70.2)	p < 0.050
Without	51/255	90.9 (86.4–94.9)	64.7 (50.0–80.7)	0.151
English Proficiency				
With	87/280	91.4 (88.4–94.2)	58.6 (48.8–75.6)	0.899
Without	9/11	86.9 (71.0–100.0)	44.4 (16.7–100.0)	0.379
Date of Collection				
Before Dec. 21	47/147	91.1 (86.6–94.8)	55.3 (44.7–75.0)	0.114
After Dec. 21	49/144	92.3 (88.5–95.8)	63.3 (38.6–78.2)	0.123
Collection Province				
Karnataka	50/95	95.4 (91.6–98.3)	82.0 (42.6–97.3)	0.444
Others	46/196	88.5 (83.9–92.7)	54.3 (38.6–74.4)	0.568
Prevalence				
20%	72/288	90.9 (87.5–94.0)	63.6 (45.6–72.7)	0.814
10%	32/288	91.2 (86.4–95.4)	63.1 (40.3–77.0)	0.783
5%	15/285	88.9 (80.4–95.7)	56.0 (29.1–75.2)	0.574
2.5%	7/273	88.8 (82.8–98.0)	51.4 (34.0–87.6)	0.542

Table 4. Description of the subject categories used in analysis, the AUC performance and the sensitivity at 95% specificity obtained using the BLSTM classifier. The p-values correspond to the t-test performed between scores from the full test set and the ones from the population sub-group considered.

that the timeline of data collection did not influence the AUC performance. We created and analyzed population subgroups based on the vaccination status, and the use of mask during sound recording. The subgroup of subjects wearing a mask gave a performance that was significantly different compared to the full test set. This indicates that presence of a facial mask during the recording modified the acoustic properties of the sound samples that were collected. Further, the subgroup of subjects who were vaccinated also had a significantly different performance. This can be attributed to the relatively milder symptoms observed in COVID subjects who were vaccinated.

Impact and usability. In this work, we have detailed the data collection and modeling efforts done as part of the development of a point-of-care testing (POCT) tool for screening of COVID-19. The data collection efforts spanned a period of 22 months and involved collaborative efforts from multiple hospitals, health centers, and the general public via crowd-sourcing. The rich set of acoustic stimuli, with 9 different categories of sounds consisting of breathing, cough, and speech, along with extensive metadata, allows the development of a COVID-19 screening tool. The designed website tool, using a BLSTM architecture, was also made publicly available. The overall time spent by the participant to record the data ranged from 5–7 mins. Thus, the proposed framework can provide rapid screening results without any additional sophisticated equipment. Further, the entire tool is developed for a smartphone based application, using models built on data collected from wide-range of cellphones. Hence, we hypothesize that the approach may generalize well for a massive population level use case. To the best of our knowledge, this study is a first of its kind to provide a publicly available web-tool for COVID-19 screening. Furthermore, we report a detailed bias and fairness analysis to understand the confounding factors associated with the COVID-19 screening tool developed using the collected data. As future extensions of this work, a collaborative effort can be taken to enlarge the subject count and also increase the diversity (and population size) associated with different demographics, gender, age groups, and health conditions.

Respiratory related diseases are a significant contributor to the leading causes of deaths[16]. In this context, automatic monitoring of the respiratory health can help recommend early interventions to an individual and hence, avoid deterioration in the respiratory health. Automatic monitoring by analyzing the respiratory sounds is particularly interesting owing to the ease in recording and analysis of the sound samples using mobile phones. It can also provide a scalable and cost-effective means for monitoring respiratory health. Providing a means to explore this direction, the Coswara dataset documents the respiratory sound samples drawn majorly from

the Indian population. The subject population spans a wide age group (between 15–90 years) and the dataset contains an extensive meta-data for each subject. The sound samples contain two variants of breathing, two variants of coughs, three kinds of sustained vowel phonation, and two styles of continuous speech. A manual validation of the sound samples via human listening is also provided. The open access release of the dataset has drawn interest in the research community, and multiple studies have analyzed its use to explore COVID-19 detection[19–28]. Complementing the few other COVID-19 respiratory sound sample datasets (comparison provided in Table 1), the Coswara dataset provides scope for development and detailed evaluation of AI-based screening tools. At a broader level, this will benefit design of explainable AI methodologies for respiratory health monitoring.

Code availability

The code used for technical validation is available at https://github.com/iiscleap/Coswara-Data/tree/master/technical_validation.

References

1. Cascella, M., Rajnik, M., Aleem, A., Dulebohn, S. C. & Di Napoli, R. Features, evaluation, and treatment of coronavirus (COVID-19). *Statpearls [internet]* (2022).
2. Target product profiles for priority diagnostics to support response to the COVID-19 pandemic v.1.0 (WHO). https://www.who.int/docs/default-source/blue-print/who-rd-blueprint-diagnostics-tpp-final-v1-0-28-09-jc-ppc-final-cmp92616a80172344e4be0edf315b582021.pdf?sfvrsn=e3747f20_1&download=true (2020). [Online; accessed 20-May-2021].
3. Budd, J. *et al*. Digital technologies in the public-health response to COVID-19. *Nature medicine* **26**, 1183–1192 (2020).
4. Lipsitch, M., Swerdlow, D. L. & Finelli, L. Defining the epidemiology of COVID-19–studies needed. *New England journal of medicine* **382**, 1194–1196 (2020).
5. Drew, D. A. *et al*. Rapid implementation of mobile technology for real-time epidemiology of COVID-19. *Science* **368**, 1362–1367 (2020).
6. Menni, C. *et al*. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature medicine* **26**, 1037–1040 (2020).
7. Zoabi, Y., Deri-Rozov, S. & Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digital Medicine* **4**, 1–5 (2021).
8. Natarajan, A., Su, H.-W. & Heneghan, C. Assessment of physiological signs associated with COVID-19 measured using wearable devices. *NPJ digital medicine* **3**, 1–8 (2020).
9. Natarajan, A. *et al*. Measurement of respiratory rate using wearable devices and applications to COVID-19 detection. *NPJ digital medicine* **4**, 1–10 (2021).
10. Hu, B., Guo, H., Zhou, P. & Shi, Z.-L. Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology* 1–14 (2020).
11. Laennec, R. T. H. & Forbes, J. *A Treatise on the Diseases of the Chest, and on Mediate Auscultation* (Samuel S. and William Wood, 1838).
12. Sharma, N. *et al*. Coswara – A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis. In *Proc. Interspeech 2020*, 4811–4815, https://doi.org/10.21437/Interspeech.2020-2768 (2020).
13. Orlandic, L., Teijeiro, T. & Atienza, D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data* **8**, 1–10 (2021).
14. Han, J. *et al*. Sounds of covid-19: exploring realistic performance of audio-based digital testing. *NPJ digital medicine* **5**, 1–9 (2022).
15. Pizzo, D. T., Esteban, S. & Scetta, M. Iatos: Ai-powered pre-screening tool for COVID-19 from cough audio samples. *arXiv preprint arXiv:2104.13247* (2021).
16. Salvi, S. *et al*. The burden of chronic respiratory diseases and their heterogeneity across the states of india: the global burden of disease study 1990–2016. *The Lancet Global Health* **6**, e1363–e1374 (2018).
17. Bhattacharya, D. *et al*. Coswara: A respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection. *Zenodo* https://doi.org/10.5281/zenodo.7188627 (2022).
18. McFee, B. *et al*. librosa/librosa: 0.8.0. Zenodo https://doi.org/10.5281/zenodo.3955228 (2020).
19. Andreu-Perez, J. *et al*. A generic deep learning based cough analysis system from clinically validated samples for point-of-need COVID-19 test and severity levels. *IEEE Trans. Services Computing* 1–1, https://doi.org/10.1109/TSC.2021.3061402 (2021).
20. Coppock, H. *et al*. End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study. *BMJ Innovations* **7** (2021).
21. Pal, A. & Sankarasubbu, M. Pay attention to the cough: Early diagnosis of COVID-19 using interpretable symptoms embeddings with cough sound signal processing. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, 620–628 (2021).
22. Alkhodari, M. & Khandoker, A. H. Detection of COVID-19 in smartphone-based breathing recordings: A pre-screening deep learning tool. *PloS one* **17**, e0262448 (2022).
23. Pahar, M., Klopper, M., Warren, R. & Niesler, T. COVID-19 detection in cough, breath and speech using deep transfer learning and bottleneck features. *Computers in Biology and Medicine* **141**, 105153 (2022).
24. Ponomarchuk, A. *et al*. Project achoo: A practical model and application for COVID-19 detection from recordings of breath, voice, and cough. *IEEE Journal of Selected Topics in Signal Processing* **16**, 175–187, https://doi.org/10.1109/JSTSP.2022.3142514 (2022).
25. Mohammed, E. A., Keyhani, M., Sanati-Nezhad, A., Hejazi, S. H. & Far, B. H. An ensemble learning approach to digital corona virus preliminary screening from cough sounds. *Scientific Reports* **11**, 1–11 (2021).
26. Sharma, N. K. *et al*. Towards sound based testing of COVID-19–Summary of the first diagnostics of COVID-19 using acoustics (DiCOVA) challenge. *Computer Speech & Language* **73**, 101320 (2022).
27. Lella, K. K. & Pja, A. Automatic diagnosis of covid-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: cough, voice, and breath. *Alexandria Engineering Journal* **61**, 1319–1334 (2022).
28. Kranthi Kumar, L. & Alphonse, P. Covid-19 disease diagnosis with light-weight cnn using modified mfcc and enhanced gfcc from human respiratory sounds. *The European Physical Journal Special Topics* **231**, 3329–3346 (2022).
29. Muguli, A. *et al*. DiCOVA Challenge: Dataset, Task, and Baseline System for COVID-19 Diagnosis Using Acoustics. In *Proc. Interspeech 2021*, 901–905, https://doi.org/10.21437/Interspeech.2021-74 (2021).
30. Sharma, N. K. *et al*. The second DICOVA challenge: Dataset and performance analysis for diagnosis of COVID-19 using acoustics. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 556–560, https://doi.org/10.1109/ICASSP43922.2022.9747188 (2022).

31. Bhattacharya, D. *et al*. Analyzing the impact of SARS-CoV-2 variants on respiratory sound signals. *Proc. Interspeech 2022* 2473–2477, https://doi.org/10.21437/Interspeech.2022-10389.
32. Carpenter, J. & Bithell, J. Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in medicine* **19**, 1141–1164 (2000).
33. Xia, T. *et al*. Covid-19 sounds: A large-scale audio dataset for digital respiratory screening. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2021).

Acknowledgements

Author contributions

S.G., S.R.C. and N.K.S. contributed in conceptualizing the study. S.G., S.R.C., N.K.S. and D.B. contributed to the study design. C.C., S.N., S.K.K., S.G., M.A. and P.M. contributed to data acquisition. D.B., S.R.C., S.G., N.K.S. and D.D. contributed to data analysis and verified the data records. D.B. and N.K.S. contributed to the initial drafting of the manuscript. All authors contributed to data interpretation and critical revision of the manuscript. All authors had full access to all the data in the study and took responsibility for the decision to submit this draft for publication.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.