Supplementary Materials for COVID-19 Sounds: A Large-Scale Audio Dataset for Digital Respiratory Screening

Contents

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

1 Datasheet

The original questions are in bold. The subtext to each question is in italics. The answers are in plain text with no formatting[1].

1.1 Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The COVID-19 Sounds Dataset is a crowd-sourced audio dataset created with three goals in mind:

1. Being at a large scale respiratory sound dataset to enable machine learning model training and evaluation. This means the dataset should have tens of thousands of participants.
2. Being accessible and easy-to-use for researchers to facilitate healthcare model development.
3. Begin to answer questions about the potential of exploring sounds for COVID-19 or other respiratory health status detection.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

This dataset was crowdsourced through the COVID-19 Sounds project, approved by the Ethics Committee of the Department of Computer Science and Technology at the University of Cambridge and partly funded by the European Research Council through Project EAR #833296. Our project website is `https://www.covid-19-sounds.org/en/`.

Any other comments?

No.

1.2 Composition

Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Audio files of crowd-sourced breathing, coughs and voices of anonymous users. Their age, location, sex, medical history and symptom, and if they have tested positive for COVID-19, as declared.

How many instances are there in total (of each type, if appropriate)?

A total of 53,449 audio samples (over 552 hours) from 36,116 participants are included.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

[1]The questions were copied from the paper which introduced this concept: Datasheets for Datasets `https://arxiv.org/abs/1803.09010`

It is a sample from the larger set. We only include up to five sample per user, and a few users have given more samples but we do not include those.

If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

It is representative as most of the data collection has happened. While the app is still functioning only a few users are still contributing data.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The raw data consists of audio waveforms that you can listen to, along with meta-data in the csv file that you can read.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, the meta-data of a particular audio file acts as the "label".

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

No.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

We recommend the data splits for the defined benchmark tasks, which can be found in `https://github.com/cam-mobsys/covid19-sounds-neurips.git`. We created these user-independent splitting with demographics carefully balanced in each set.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

There is acoustic noise in our dataset. We conducted audio quality check, and provided the sound type detection results. We suggest researchers to use the samples that can be recognised as high-quality cough, breathing, or voice recordings.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Yes, the dataset contain participants' personal health information, which is sensitive and should be considered confidential. The participants are anonymous, but we cannot guarantee the user will not be re-identified for the audio samples, especially voice recordings. Thus, to access and to use this dataset for academic research purpose, signing a data transfer agree to restrict the usage is needed.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Yes, the datast contains repeated cough sounds, and thus if viewed directly and continually, it might cause anxiety.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes. All of our data comes from real people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Yes, our dataset covers different age and gender groups. 62% are male, 36% are female and the others prefer not to tell their gender. 9.2% are under 20, 24.1% are aged 20-29, 26.5% are 30-39, 19.8% are 40-49, 11.2% are 50-59, 5.3% 60-69, 1.7% are 70-79, 0.2% are over 80, and the others prefer not to tell their age.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

It is possible, as in our dataset, voice recordings might be used to identified a participant from his or her public media if exists.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

Yes. Our data is health data, containing medical history and smoking status information.

Any other comments?

No.

1.3 Collection process

The answers to questions here may provide information that allow others to reconstruct the dataset without access to it.

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

We crowd-sourced data from volunteers. The data was directly reported by subjects. The data was not clinically verified.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

We developed the data collection app COVID-19 Sounds and launched it to app market, which is free to download. To validate the mechanism, our team members first downloaded and tested the app.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

PhD students and postdocs in our research group were involved in the data collection process. We thank the data contributors and did not pay them.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g. recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The released dataset was collected around one year from April 2020 to April 2021. This timeframe matched the creation timeframe of the data.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Yes, the study was approved by the ethics committee of the Department of Computer Science at the University of Cambridge, with ID #722. Our app displays a consent screen, where we ask the user's permission to participate in the study by using the app. Also note that the legal basis for processing any personal data collected for this work is to perform a task in the public interest, namely academic research. More information is available at `https://covid-19-sounds.org/en/privacy.html`.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

Yes, data was collected from real people.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Data was collected directly from users. Questionnaire can refer to `https://www.covid-19-sounds.org/en/app/`.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Yes. When a user accesses our app for the time time, the app will pop up the consent as shown in Figure 1. If the user agrees, the data recording will start.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Users can also request that their data is deleted at any time by contacting us at covid-19-sounds@cl.cam.ac.uk and quoting the ID that appears in the last screen of the app. We will remove the data from our server, however, any data already shared with researchers at other institutions will not be deleted from their copies. Please refer to our privacy policy `https://www.covid-19-sounds.org/en/privacy.html`.

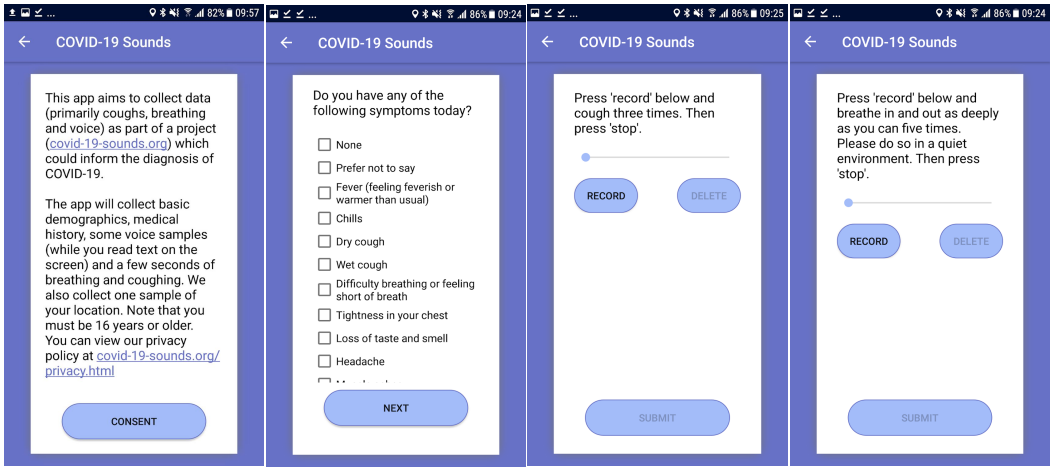


Figure 1: **Screens of the data collection app.** The users are asked to input their symptoms along with medical history, as well as to record breathing, cough, and voice sounds every couple of days.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Yes, we conducted the analysis on our dataset to explore the potential of sounds for health status detection. Two tasks (i.e., respiratory symptom detection, and COVID-19 prediction) have been defined and three baselines have been implemented, yielding ROC-AUCs above 70%. Implementation details can refer to `https://github.com/cam-mobsys/covid19-sounds-neurips.git`.

Any other comments?

No.

1.4 Reprocessing, cleaning, and labelling

The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Sound type detection has been conducted as a prepossess, in order to access the quality of each sample. All the dataset points are remained with an extra quality check list attached. Please refer to section 3.2 of our paper.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Yes. Those who are interested may request access to the Google Cloud Storage drive containing the raw data.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes. The code is open source `https://github.com/cam-mobsys/covid19-sounds-neurips/tree/main/YAMNet`.

Any other comments?

No

1.5 Uses

These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should Not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

Has the dataset been used for any tasks already? If so, please provide a description.

Yes, for training respiratory symptom and COVID-19 detection models. Please refer to Section 4 in our paper.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No.

What (other) tasks could the dataset be used for? Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Based on the dataset, we proposed and implemented two tasks: respiratory symptom and COVID-19 detection. In addition, we reflect on a plethora of applications that can be empowered by our data, including biometric user authentication, demographic prediction, smoking status detection, etc. Please refer to Section 5.2 in our paper.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Data recipients are not allowed to use the data for non-research purposes. They are also not allowed to re-identify individual from the data.

Any other comments?

No.

1.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes. The dataset will be available under data transfer agreement fro research purpose.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

Data will be distributed through Google Drive Storage. There is no DOI at this time.

When will the dataset be distributed? Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so,

please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset has been prepared and will be accessible to researchers under data transfer agreement. No IP involved.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

No.

1.7 Maintenance

These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

Who is supporting/hosting/maintaining the dataset? How can the owner/curator/manager of the dataset be contacted (e.g., email address)? Is there an erratum? If so, please provide a link or other access point.

The COVID-19 Sounds research group handles hosting and maintenance. Please contact `covid-19-sounds@cl.cam.ac.uk` with questions. Instead of an "erratum", we plan to publish updates to the emails that people use to request the dataset.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes. It will be updated on an as-needed basis, with updates sent to all emails provided by users who request data access.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

We would retain the data indefinitely if no specific deleting request is received.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes, in the case that some data needs to be removed for legal or ethical reasons.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Absolutely. Our data collection app is open source:
Andriod app: https://github.com/cam-mobsys/covid19-sounds-android-app,
iOS app: https://github.com/cam-mobsys/covid19-sounds-ios-app,
based on which, researchers can launch their app to crowd-source data. If someone would like to contribute directly back to The COVID-19 Sounds, we recommend contacting us covid-19-sounds@cl.cam.ac.uk.

Any other comments?

No.

2 Data Access

The data is sensitive as voice sounds can be deanonymised. Anonymised data will be made available for academic research upon requests. Please email covid-19-sounds@cl.cam.ac.uk. Academic institutions will need to sign a Data Transfer Agreement with the University of Cambridge to obtain the data. A copy of the data will be transferred to the institution requesting the data. Documentation and the dataset are available on Google Drive (restricted to invited people only). Once the agreement is signed, full access will be provided.

3 Benchmark Reproducibility

All the code is publicly available. Data splitting, environment setting, running instructions can be found on GitHub https://github.com/cam-mobsys/covid19-sounds-neurips.git.

4 Statement of Responsibility

The Recipient shall include a disclaimer in any publication or presentation, to the effect that Cambridge does not bear any responsibility for the Recipient's analysis or interpretation of the Data, which shall be stated as representing the Recipient's own view.

5 Hosting and Maintenance Plan

The COVID-19 Sounds study hosts the following online assets:

- **Official website**: https://covid-19-sounds.org
- **GitHub repository**: https://github.com/cam-mobsys
- **E-mail**: covid-19-sounds@cl.cam.ac.uk is used for contact.
- **Google Drive**: data is stored here with url only available on request.

We have a team consisting of professors, postdocs, and PhDs, who are responsible to maintain the dataset and response to any issues related to the data in time.

6 Potential Negative Societal Impacts

The data collected is sensitive as it contains voices from participants which could be re-associated back to individuals if cross examined with other datasets. This may potentially lead to linking medical history or symptoms to specific individuals: our data is released with a data sharing agreement to protect against these operations.

A Appendix

A.1 Automating Audio Quality Check

YAMNet [4][2] was trained for audio classification with more than 500 audio event classes including cough, breath, and speech. Given a 16 KHz mono .wav file, it can output per-class score, i. e., the softmax probability of belonging to each class. In our case, after experimentation we came up with the following heuristic algorithm: 1) For a cough recording, if the predicted probability of being cough appears in the Top 5 class probabilities, then label its audio quality as acceptable. 2) For a breath recording, similar to cough, if the predicted probability of being breath appears in the Top 5 probabilities, we likewise assign an acceptable label. 3) For a speech recording, the probability of being speech should be the highest (Top 1) and greater than 0.4. The first two criteria allow some silent and abrupt (like throat clearing) segments in cough and breath recordings, while for voice, considering that background noise is very likely to be included, we require speech segments that take over 40% of a recording. Note that by doing this, we only detect the sound type, while semantic information is not considered.

To validate the above procedure, we manually listened to and labelled over 800 samples. The confusion matrix comparing the annotation by humans and that of YAMNet is presented in Fig.2. Evidently, YAMNet achieves quite high recall of 87.0%, 82.3%, 97.4% for breathing, cough and voice recordings, respectively. It can be observed that YAMNet is very rigorous as around 10% of the recordings are regarded acceptable by human labelling but are predicted as noise by YAMNet, indicating that the rest is of high quality across all modalities.

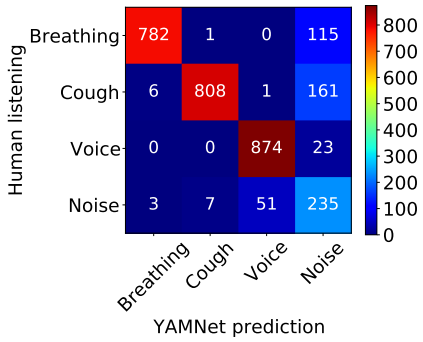


Figure 2: **Confusion matrix for the prediction of YAMNet.** We filtered 3,067 recordings and retained 2,464 usable recordings from the initial dataset for the two principal tasks. YAMNet achieves an overall accuracy of 88.0%.

A.2 Benchmark Implementation

A.2.1 Task Data Selection

For Task 1 of respiratory symptom detection, the symptomatic group is defined as participants who reported cough (dry cough or wet cough) and at least one of other symptoms (e.g., fever, sore throat, shortness of breath, runny nose, headache, dizziness, tightness), while the healthy group consists of participants who did not show symptoms. As for Task 2 of COVID-19 detection, considering there are many asymptomatic COVID-19 cases, symptoms are not used as a filter. Instead, the positive group consists of participants who were tested COVID-19 positive within two weeks before the recording, while the negative group includes participants who were confirmed as negative by a recent COVID-19 test and never tested positive before.

Following the definition of the investigated tasks, two balanced subsets consisting of qualified and complete audio samples (i.e., recordings of good audio quality and each sample with complete breath, cough, and speech) were selected. To minimise the impact of possible biases, for each task, we control for demographics (i. e., gender and age) in the training, validation and testing sets. As COVID-19

[2]Copyright: The TensorFlow Authors, Apache License, Version 2.0

prevalence varies in countries, to avoid using language as a confounding variable, we only retain English-speaking samples for the final splits. Detailed statistics are presented in Table 1 and 2.

	Symptomatic			Asymptomatic		
	Training	Validation	Testing	Training	Validation	Testing
Male	1,305(48.4%)	185(50.4%)	370(48.3%)	1,036(49.6%)	140(47.5%)	284(48.1%)
Female	1,367(50.6%)	180(49.0%)	390(50.9%)	1,014(48.6%)	147(49.8%)	296(50.2%)
16-29	913(33.8%)	125(34.1%)	251(32.8%)	562(26.9%)	70(23.7%)	146(24.7%)
30-39	877(32.5%)	107(29.2%)	260(33.9%)	524(25.1%)	75(25.4%)	142(24.1%)
40-49	564(20.9%)	89(24.3%)	147(19.2%)	442(21.2%)	66(2.4%)	154(26.1%)
50-59	206(7.6%)	33(9.0%)	72(9.4%)	293(14.0%)	34(11.5%)	86(14.6%)
60-69	78(2.9%)	4(1.1%)	19(02.5%)	157(7.5%)	26(8.8%)	31(5.3%)
70-	24(0.9%)	4(1.1%)	6(0.8%)	49(2.3%)	12(4.0%)	12(2.0%)
Total	2,690/3,225	376/461	766/951	2,087/3,423	295/433	590/963

Table 1: **Demographics distribution and data statistics for Task 1**. In each demographic group, #participants with proportion are shown. Only a small proportion of participants preferred not to say their demographics. For the total number, #participants/#samples are presented. The sum of number of patients from the six splits is slightly lagers than 6,623, because some users with more than one samples in different classes can be divided into two splits.

	Positive			Negative		
	Training	Validation	Testing	Training	Validation	Testing
Male	192(54.9%)	29(58.0%)	58(58.0%))	188(53.7%)	31(62.0%)	52(52.0%)
Female	154(44.0%)	20(40.0%)	42(42.0%)	159(45.4%)	19(38.0%)	46(46.0%)
16-29	92(26.3%)	10(20.0%)	22(22.0%)	73(20.9%)	16(32.0%)	21(21.0%)
30-39	94(26.9%)	16(32.0%)	33(33.0%)	97(27.7%)	13(26.0%)	33(33.0%)
40-49	86(24.6%)	13(26.0%)	23(23.0%)	86(24.6%)	10(20.0%)	24(24.0%)
50-59	39(11.1%)	5(10.0%)	13(13.0%)	50(14.3%)	4(8.0%)	10(10.0%)
60-69	17(4.9%)	1(2.0%)	3(3.0%)	18(5.1%)	4(8.0%)	4(4.0%)
70-	6(1.7%)	2(4.0%)	1(1.0%)	7(2.0%)	1(2.0%)	2(2.0%)
Total	350/490	50/82	100/162	350/530	50/60	100/162

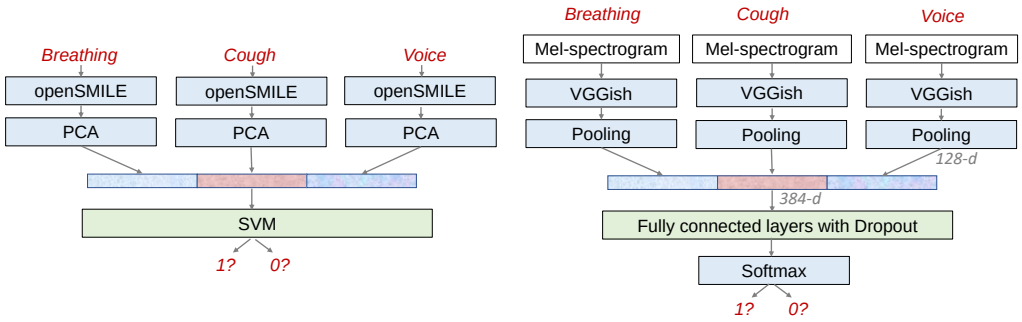
Table 2: **Demographics distribution and data statistics for Task 2**. In each demographic group, #participants with proportion are shown. Only a small proportion of participants preferred not to say their demographics. For the total number, #participants/#samples are presented.

A.2.2 Model Details

Model architectures are illustrated in Fig. 3. Here, we discuss implementation details to reproduce the results.

- **OpenSMILE+SVM.** Following [2, 5], we apply an established acoustic feature set, namely the INTERSPEECH 09 Computational Paralinguistics Challenge (COMPARE) set [6], extracted by the open-source openSMILE toolkit [1]. For each audio file, 12 *functionals* are applied on 16 frame-level descriptors and their corresponding delta coefficients, resulting in a total of 384 features. In particular, the 16 frame-level descriptors chosen are Zero-Crossing-Rate (ZCR), Root Mean Square (RMS) frame energy, pitch frequency (F0), Harmonics to-Noise Ratio (HNR), and Mel-Frequency Cepstral Coefficients (MFCCs) 1-12, covering prosodic, spectral, and voice quality features. Features of a single sound type or concatenation of three sounds were then fed into an SVM for classification. Principal component analysis (PCA) is applied to reduce the feature dimension by retaining 99% of the explained variance, and a linear kernel with $C = 0.001$ is found through the validation set. We use Python 3.8 and Scikit-learn 0.24.1 for these set of experiments.
- **Pre-trained VGGish.** A pre-trained VGGish is employed to extract audio features automatically [4][3]. VGGish is a convolutional neural network that was proposed for audio

[3]Copyright: The TensorFlow Authors, Apache License, Version 2.0



(a) Shallow model structure

(b) Deep model structure

Figure 3: **Model architectures**. Shallow and deep models along with data modalities and pre-processing steps (where applicable).

classification based on the Mel-spectrogram of the raw audio input; the VGGish model was trained using a large-scale YouTube dataset and the learned model parameters were released publicly. We deploy it as a feature extractor to transform the raw audio waveforms into embeddings (features). The VGGish pre-trained model first splits the data into 0.96-sec non-overlapping windows, and for every 0.96 window, it returns a 128-dimensional feature vector. Average pooling is applied to handle the varying audio length before feeding these features into the classifier consisting of two fully connected layers and a softmax layer [3, 7]. We used the offical implementation of VGGish[4] in Python 3.8 and Tensorflow 1.18. A learning rate of 1e-4 was used to update the fully connected layers.

- **Fine-tuned VGGish.** Different from the pre-trained VGGish baseline which fixes the parameters of VGGish, in this approach, we jointly fine-tune VGGish and the subsequent fully connected layers. Last, we assigned a lower learning rate to VGGish layers (i.e., 1e-5) and higher learning rate to fully connected layers (i.e., 1e-4).

As for pre-processing, we resample all the recordings to 16 kHz mono audio[5], and then remove the silence period at the beginning and the end of the recording. Finally, audio normalisation by calibrating the peak amplitude to 1 is applied to eliminate the discrepancy across recording devices.

For both pre-trained and fine-tuned VGGish, to avoid over-fitting, we utilised learning rate decay (factor = 0.9) and cross-entropy loss with L2-regularisation (penalty coefficient = 1e-6). Also, Dropout layers (prob. = 0.5) were added after the final two fully connected layers during training. Training was done (batch size = 1) on a single GPU with 64 GB memory. All code will be publicly available in `https://github.com/cam-mobsys/covid19-sounds-neurips`.

References

[1] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462, 2010.

[2] J. Han, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo. Exploring automatic covid-19 diagnosis via voice and symptoms from crowdsourced data. In *Proceedings of the 46th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8328–8332, 2021.

[3] J. Han, T. Xia, D. Spathis, E. Bondareva, C. Brown, J. Chauhan, T. Dang, A. Grammenos, A. Hasthanasombat, A. Floto, et al. Sounds of covid-19: exploring realistic performance of audio-based digital testing. *arXiv preprint arXiv:2106.15523*, 2021.

[4] `https://github.com/tensorflow/models/tree/master/research/audioset`

[5] Sampling rate for several samples used for evaluation (2.6% in Task 1 and 1.8% in Task 2) is lower than 16KHz, and excluding those will not significantly impact the results. We encourage future works to explore the impact of re-sampling rate.

[4] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. CNN architectures for large-scale audio classification. In *Proceedings of the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017.

[5] B. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, et al. The INTERSPEECH 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates. *arXiv preprint arXiv:2102.13468*, 2021.

[6] B. Schuller, S. Steidl, and A. Batliner. The INTERSPEECH 2009 emotion challenge. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2009.

[7] T. Xia, J. Han, L. Qendro, T. Dang, and C. Mascolo. Uncertainty-aware covid-19 detection from imbalanced sound data. *arXiv preprint arXiv:2104.02005*, 2021.

B Meta-data Description

The deception of all fields in the meta-data files we provide is attached as below,

This table shows how the meta data is orginased. Information for all participants are summarised in csv file, with ';' used for field splits.

Each line presents one sample with 18 fields being divided into three major categories:

Demographic:		
Field	**Explanation**	**Value**
Uid	participant's ID	a string with mixed characters and numbers, length of 10: Andriod user, length of 12: iOS user, 'form-app-users': Web user
Age	participant's age group	'18-19': aged betwen 18 and 20, '0-19': aged betwen 18 and 20 (we requst participants aged over 18), '16-19': aged between 18 and 20, '20-29': in twenties, '30-29': in thirties, '40-49': in forties '50-59': in fifties, '60-69': in sixties, '70-79': in seventies, '80-89': in eighties, '90-': aged over 90, 'pnts': prefer not to say
Sex	participant's gender group	'female', 'male', 'other', not female or male, 'pnts': prefer not to say
Medhistory	medical history	'angina':Angina, 'asthma': Asthma, 'cancer':Cancer, 'copd': COPD/Emphysema, 'cystic': Cystic fibrosis, 'diabetes': Diabetes, 'hbp': High Blood Pressure, 'heart': Previous heart attack, 'hiv': HIV or impaired immune system, 'long': Other long-term condition, 'lung':Other lung disease, 'otherHeart': Other heart disease, 'organ': Previous organ transplant, 'pulmonary':Pulmonary fibrosis, 'stroke': Previous stroke or Transient ischaemic attack, 'valvular': Valvular heart disease, 'None': none of the above, 'pnts': prefer not to say
Smoking	smoking history	ex': ex-smokers, 'never': never smoked, 'ltOnce': current smoker (less than once per day), '1to10': current smoker (1-10 cigarettes per day), '11to20': current smoker (11-20 cigarettes per day), '21+': current smoker (20+ cigarettes per day), 'ecig': current smoker (e-cigarettes only), 'pnts': prefer not to say
Language	language of the mobile phone system	'en': English, 'it':Italian, 'es':Spanish, 'de':German, 'pt':Portuguese, 'el':Greek, 'fr':Frech, 'ru':Russian, 'ro':Romanian, 'zh': Chinese, 'hi':Hindi, 'None': not provided or recongized
Daily health status:		
Field	**Explanation**	**Value**
Date	the exact time when the recordings were created	Timestamp
Folder Name	the path of recordings	Unix timestamp string
Symptoms	participant's Covid19 relevant symptoms	'drycough': Dry cough, 'wetcough': Wet cough, 'sorethroat': Sore throat, 'runnyblockednose': Runny or blocked nose, 'tightness': Tightness in your chest, 'smelltasteloss': Loss of taste and smell, 'fever': Fever (feeling feverish or warmer than usual), 'chills': Chills, 'shortbreath': Difficulty breathing or feeling short of breath (Note that sorethroat and runnyblockednose are combined as one option in the app), 'dizziness': Dizziness, confusion or vertigo, 'headache': Headache, 'muscleache': Muscle aches, 'None': None, 'pnts': prefer not to say

Covid-Tested	participant's Covid19 testing results	'positiveLast14': tested positive, in the last 14 days, 'last14': tested positive, in the last 14 days, 'positiveOver14': tested positive, over 14 days ago, 'over14': tested positive, over 14 days ago, 'yes': tested positive, but no time information, 'negativeLast14': tested negative in the last test, but was tested positive before in the last 14 days, 'negativeOver14': tested negative in the last test, but was tested positive before over 14 days ago, 'negativeNever': tested negative and never tested positive before, 'never': never tested before, 'no': never tested before, 'neverThinkHadCOVIDNever': never tested and thought never getting COVID-19, 'neverThinkHadCOVIDNow': never tested but thought getting COVID-19 now, 'neverThinkHadCOVIDLast14': never tested but thought getting COVID-19 before in the last 14 days, 'neverThinkHadCOVIDOver14': never tested but thought getting COVID-19 before over 14 days ago, 'pnts': prefer not to say
Hospitalized	whether the participant was in hospital or not	'yes': is hospitalized, 'no': is not hospitalized, 'pnts': prefer not to say
Location	participant's geolocation	We hold this
Voice filename	the name of voice/speech recording	A string
Cough filename	the name of cough recording	A string
Breath filename	the name of breathing recording	A string
Audio quality:		
Field	**Explanation**	**Value**
Voice check	the modality identified by Yamnet for voice recording	'v': voice, 'n': no voice included or not very good quality
Cough check	the modality identified by Yamnet for coughrecording	'c': cough, 'n': no cough included or not very good quality
Breath check	the modality identified by Yamnet for breathing recordi	'b': breathing, 'n': no breathing included or not very good quality
Sampling rate	sampling rate of the original unprocessed recordings	number in *Hz*