Virufy: Global Applicability of Crowdsourced and Clinical Datasets for AI Detection of COVID-19 from Cough

Gunvant Chaudhari[1,2], Xinyi Jiang[1], Ahmed Fakhry[1,3], Asriel Han[1,4], Jaclyn Xiao[1,5], Sabrina Shen[1,6], Amil Khanzada[1]

1. Virufy AI Research Group
2. University of California San Francisco, School of Medicine
3. University of Alexandria, Department of Electronics and Communication
4. Stanford University, Department of Humanity Science
5. Duke University, Department of Biomedical Engineering
6. Harvey Mudd College, College of Engineering

***Abstract*— Rapid and affordable methods of testing for COVID-19 infection are essential to manage infection rates and prevent medical facilities from becoming overwhelmed. This study demonstrates that crowdsourced cough audio samples acquired on smartphones across the world and paired with COVID-19 status labels can be used to develop an AI algorithm that accurately predicts COVID-19 infection with an ROC-AUC of 77.1% (75.2%-78.3%). Furthermore, this AI algorithm is able to generalize to crowdsourced samples from Latin America and clinical samples from South Asia, without further training using the specific samples. As more crowdsourced data is collected, further development can be implemented using various respiratory audio samples to create a cough analysis-based AI solution for COVID-19 detection that can likely generalize globally to all demographic groups in both clinical and non-clinical settings.**

1 Introduction

As of November 9th 2020, there were more than 59M cases of COVID-19 worldwide [1]. Widespread testing and isolation of individuals infected with COVID-19 is necessary to control infection rates and optimize healthcare resources [2]. The current gold standard of Reverse Transcription Polymerase Chain Reaction (RT-PCR) testing requires person to person contact to administer, has variable turnaround time, with the longest take days, and is not easily accessible for everyone globally [3]. With cases still increasing and vaccine approval and distribution still on the horizon, accessible and affordable testing is essential.

Artificial Intelligence (AI) algorithms can be a powerful tool for a preliminary indication of a person's COVID-19 status and have been developed to accurately predict COVID-19 infection from smartphone-acquired cough sounds as shown in results from Cambridge and MIT [12,13]. With smartphone usage high and continually rising in developing countries, these devices are an ideal platform for widespread collection of respiratory audio recordings and for implementing audio-based COVID-19 testing.

A variety of COVID-19 cough recording datasets have been collected by various groups and used to train machine learning models for COVID-19 detection. However, each of these models has been trained on data of a variety of formats and recording settings. While some, such as Coswara, collect additional counting and vowel recordings [9], others gather cough recordings exclusively [10]. Furthermore, these datasets come from various sources, such as clinical setting recorded data [10], crowdsourcing [9], and extraction from public media interviews [11]. Recordings can be done in different compression formats as well as no formal standard for gathering COVID-19 cough data has yet been defined. Variations of sources, contents and formats of datasets raise a generality challenge for the AI model developers.

Research groups have explored the prediction of COVID-19 status from cough sounds, but there is still doubt about the generalizability of such models across various environments and collection methods. The largest obstacle to implementation of machine learning algorithms is their reliability on unseen data [8]. Therefore, a globally deployable AI solution needs to be trained and tested on global data to ensure high performance in all locations and situations. We present the first study showing generalizability of a multimodal deep learning model in detecting COVID-19 status on crowdsourced and clinical cough datasets from around the world, using five distinct datasets from multiple sources. An overview of our approach is provided in Figure 1.

Our group, Virufy, is uniting the world to build a global artificial intelligence (AI) database of crowdsourced cough sounds to identify patterns that signify respiratory diseases such as COVID-19. The authors of this paper are current students and alumni from various institutions who have come together to tackle the pandemic.

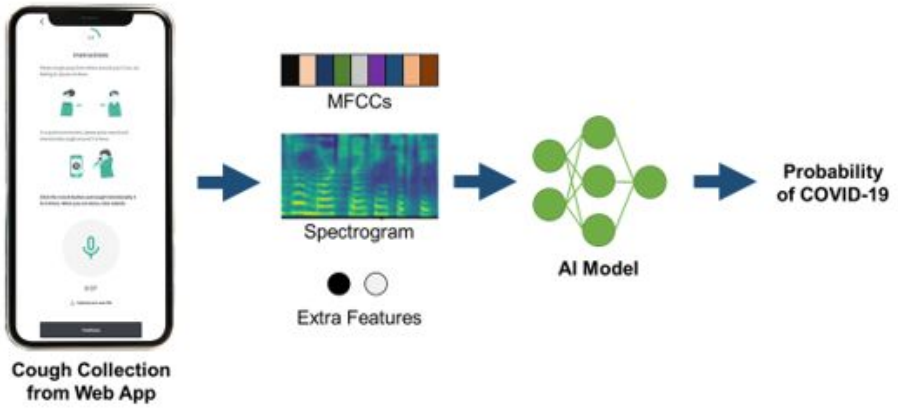


Figure 1: Overview of our proposed cough-based COVID-19 detection system .

2 Background

The clinical presentation of COVID-19 can be highly variable, but multiple common symptoms of the illness impact the human airway and lungs [4]. Dry cough is a distinctive feature of most cases along with more severe symptoms such as pneumonia and ground-glass opacities [5]. Because of the virus's effects on the respiratory system, COVID-19 has been theorized to create unique audio signatures that are distinct from those associated with other respiratory infections. Thus, sounds such as coughing can be analyzed to detect the illness. Though these differences are difficult or impossible to detect with the human ear, AI algorithms have already shown promise in classifying cough sounds to identify respiratory diseases including pertussis, asthma, and pneumonia by using various audio features [6,7].

Various groups across the world have been focused on collecting copious high-quality data to train cough-based machine learning models, including Cambridge University in England, Carnegie Mellon University and the Massachusetts Institute of Technology (MIT) in the US, and Afeka College of Engineering in Israel. Others, such as Coswara, are focused on gathering large datasets of cough and various other human audio sounds [9]. In addition to our open dataset, Coughvid and Coswara have also open-sourced their datasets [9,10].

Work done by Cambridge University [12] and MIT [13] shows the potential of AI to detect COVID-19 status. Cambridge uses mel-frequency cepstral coefficients (MFCC) and other audio statistics, such as duration, onset and zero-crossing and VGG features such as starting features [13]. The features are then further processed by Principal Component Analysis and used as input to a simple binary classifier [13]. The researchers are able to achieve an AUC of 0.82 but their dataset only has a size of 86 coughs [13]. Another approach includes using mel-Spectrogram as input to the model. A pre-trained deep convolutional neural network, ResNet-18 [16], is used and ensembles of both shallow and deep networks are explored [14]. The model is also pretrained on a larger dataset using cough/non-cough labels and techniques, such as noise augmentation, audio segmentation and time and frequency masking, are applied [14]. Both of these models have shown promising preliminary results in the COVID-19 classification task on Cambridge's own dataset [13] and, independently, in the Coswara dataset [9]. MIT researchers, who created a model with significant performance improvement, utilized a biomarker layer with ResNet-50 [16] based models [14]. Their best performing model inputs mel-frequency cepstral coefficients (MFCC) and uses transfer learning to output three separate biomarkers to be used as inputs to the three parallel ResNet-50 convolutional neural networks [14].

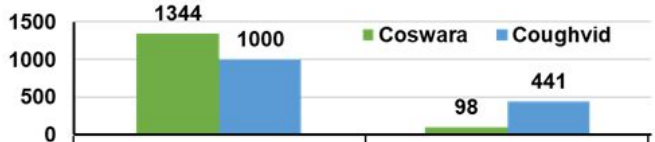
However, research published to date shows results restricted to datasets which are often not open-source, so it is difficult to judge the ability of these models to identify COVID-19 using arbitrary cough recordings. In addition, prior research highlights the importance of using transfer learning and pretraining to enlarge the training data size, because the size of a single dataset may not be big enough to train deep neural networks. Our model is able to adapt to enlarged training and testing datasets from various sources.

3 Methods

3.1 Crowdsourced COVID-19 Training Data

We trained a deep neural network using the openly available Coswara (n=1,543) and Coughvid (n=20,072) datasets of cough sounds. From the Coswara datasets, samples with 'covid-status' of 'positive_mild', 'positive_moderate', and 'positive_asymp' were classified in the positive class (n=1,334), and all other statuses were classified as negative (n=98). Only shallow cough files were used. All were uncompressed audio files.

From the Coughvid dataset, samples with the 'COVID-19 Positive' label were classified in the positive class (n=441) and 1,000 other random samples were selected to be in the negative class to maintain data balance. The distribution of this data is shown in Figure 2.



	Negative Coughs		Positive Coughs	
	Specific Status	**N**	**Specific Status**	**N**
Coswara	Healthy	1,158	COVID-19 Positive, Mild Symptoms	75
	Exposed with No Respiratory Illness	90	COVID-19 Positive, No Symptoms	14
	Unidentified Respiratory Illness	75	COVID-19 Positive, Moderate Symptoms	9
	Fully Recovered from COVID-19	21		
Coughvid	Healthy	557	COVID-19 Positive	441
	Not Provided	340		
	Symptomatic	103		
		2,810		**539**

Figure 2: Distribution of samples in the training dataset. PCR negative labels were not clearly indicated in either crowdsourced dataset.

3.2 Virufy Test Datasets

After verifying our algorithm's performance on Coswara and Coughvid crowdsourced data, we further validated its performance with our own more detailed data. All data had COVID-19 PCR labels and was acquired in conditions that were meant to simulate real-world usage. Audio files were a mixture of compressed and uncompressed files (e.g. wav and mp3 files) depending on the mode of data acquisition. All potential privacy risks and security threats were addressed by our legal and information security teams, who developed localized privacy policies and patient consent forms, along with a Data Protection Impact Assessment (DPIA) and several internal information security policies.

3.2.1 Virufy Latin American Crowdsourced Test Dataset

To mimic one potential use case of COVID-19 detection from cough by smartphone users in the general public, samples used within the models were crowdsourced using the Virufy mobile data collection app[1] (Figure 1). For our analysis, we considered n=178 smartphone-recorded coughs from Peru, Brazil, and Colombia. Each cough sample was accompanied with specific labels of COVID-19 status based on PCR and antibody testing, demographics, past medical history, current symptoms, and a corresponding speech sample of counting from 1 to 10 (A.2). After excluding untested individuals and samples with poor audio quality, n=32 samples of PCR-negative and -positive individuals were aggregated to evaluate our algorithm's generalizability (Table 1).

[1] The app can be accessed here: https://virufy.org/app.

3.2.2 Virufy South Asian Clinical Test Datasets

To determine performance of a COVID-19 detection algorithm in a busy clinical setting, we also collected samples in hospital clinics using smartphones. The explicit patient consent forms electronically accepted by all patients were originally drafted by Virufy clinical researchers and reviewed by our medical advisors. The data was captured directly from patients under the hospitals' Institutional Review Board (IRB) approved clinical research study protocols.

Clinical dataset 1 was collected from a South Asian hospital clinic from 04/13/2020 to 05/21/2020. It consists of cough recordings and labels from a total of 362 unique patients. Personal protective equipment and strict sanitation procedures were used to prevent disease transmission during this data acquisition. All patients were also simultaneously PCR tested for COVID-19 infection. Other demographics, medical history, and current symptoms information were also recorded (A.2).

As this data was acquired in a busy clinic, many samples had background noise, including clinician conversations, equipment sounds, ambient environmental noise, and vehicular traffic. In a manual survey of 50 random samples, 41 (82%) samples had at least one instance of distinctive noise.

Clinical dataset 2 was collected at a second hospital in South Asia from patients who had been PCR tested for COVID-19 with the same data points as in dataset 1. The data was collected from patients being screened at the fever clinic as well as from COVID-19 general and ICU wards.

These datasets were collected from a variety of smartphone types with different audio sampling frequencies, compression artifacts, and background noises.

PCR Result	Virufy Crowdsourced	Clinical Dataset 1	Clinical Dataset 2
Positive	7 (22.6%)	89 (24.6%)	47 (74.6%)
Negative	24 (77.4%)	273 (75.4%)	16 (25.4%)
Total	**31**	**362**	**63**

Table 1: Composition of Test Datasets

3.3 Audio and Clinical Features

We used multiple features from the crowdsourced datasets to train our network. The first feature was mel-frequency cepstral coefficients (MFCCs), a commonly used audio feature derived from the short-term power spectrum [17]. Each cough audio file was resampled to 22.5 kHz and the first 39 MFCCs were extracted using the librosa package [18], with a sampling

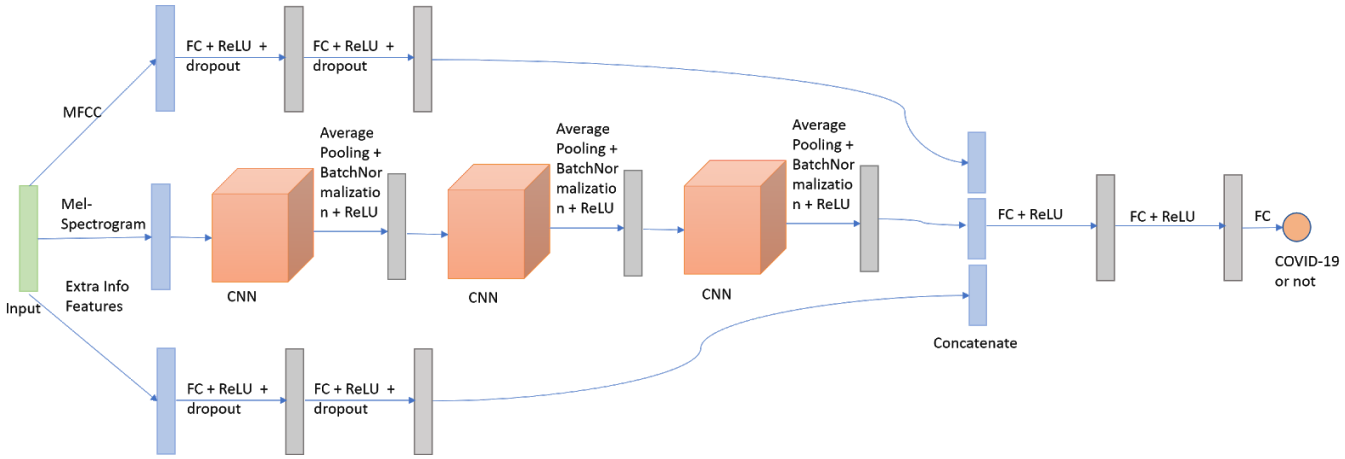


Figure 5: Ensemble Model Structure

rate of 22.5 kHz, hop length of 23ms, window length of 93ms, and a Hann window type. Outputs were averaged across the time-axis to yield mean 39 MFCCs features for each audio file.

The next extracted feature is the mel-frequency spectrogram, another common audio feature. Though MFCCs are derived from the spectrogram, the spectrogram encodes raw power information without any transformations. Spectrograms were extracted using the librosa package with the same parameters as for the MFCCs and interpolated to size (64,64).

Beyond audio files, each sample also contained additional rich information that has potential to enhance prediction accuracy. We chose to add two additional features for each cough file that reflect the clinical picture of the patient. Detectable changes in cough sounds have been shown to occur with diseases other than COVID-19 [19]. Therefore, a binary label about the presence or absence of current respiratory diseases was aggregated to feed into our algorithm as one extra feature. Next, COVID-19 also presents with other symptoms than cough, with some of the most common being fever and myalgia (muscle pain) [20]. The presence or absence of these symptoms may also impact the probability of having COVID-19. To develop as accurate a model as possible, a second binary label of fever or myalgia status was also aggregated from all datasets and fed into the model as a second extra feature. The distribution of each of these two extra features is shown in Figure 4.

3.4 Ensembled Deep Neural Networks

After experimentation with 1D and 2D CNNs, LSTM, and CRNN architectures, the best performing network was an ensemble of 3 separate networks whose structure and hyperparameters were fine-tuned using grid search to minimize overfitting. Outputs from each network were aggregated to predict the probability of having COVID-19 (Figure 5).

The first network is for the MFCCs with input size of (39,) and consists of two hidden layers with ReLU activation, each followed by a dropout layer. The second network is a convolutional neural network with the mel-spectrogram image as input of size (64,64,1). It consists of three 2D convolution layers, with a kernel size of 3 and a stride size of 2 for the first convolution layer and a kernel size of 3 and a stride size of 1 for the rest two convolution

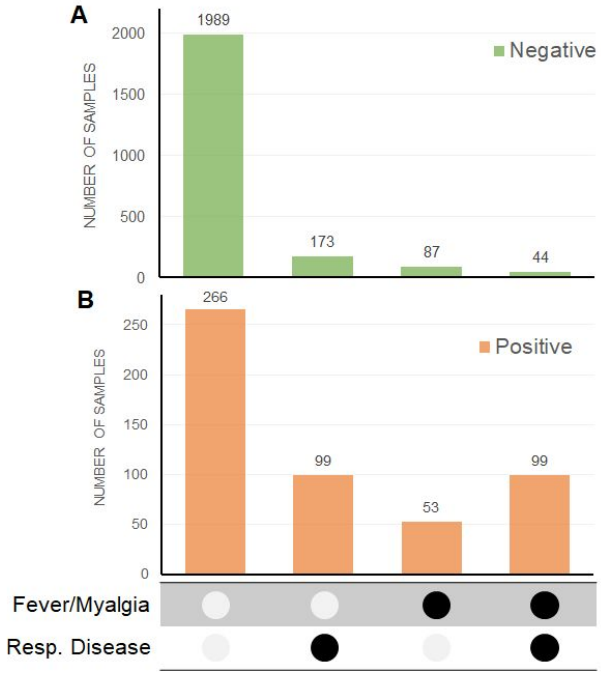


Figure 4: Distribution of extra features in training set (n=3349). Fever/myalgia and medical history of respiratory disease were aggregated into two extra features. 3A shows the distribution for negative samples and 3B shows distribution for positive samples for all patterns of these features.

layer, each followed by a 2D average pooling, a batch normalization, and a ReLU activation. The last network is for each sample's two extra features of fever/myalgia and respiratory conditions. Similar to the first network, it consists of two hidden layers with ReLU activations, each followed by a dropout layer. Outputs from each network were aggregated, fed through two additional hidden layers, each followed by a ReLU activation function, and combined into a final sigmoid output decision layer.

The ensemble network was trained using cross entropy loss, an Adam optimizer, and learning rate of 0.001. The training data was randomly split into train-validation-test datasets using a 70-15-15 split. Every experiment was repeated five times, each with a different random data split. The mean statistical values and 95% confidence intervals are reported, unless otherwise specified.

4 Results

Table 1 contains the test results on the four test datasets that are discussed above. We used both accuracy and Area under the ROC Curve (AUC) as evaluation metrics. As the data is unbalanced, we believe that AUC would be a better presentation of how the model is working. The model is robust enough that the test results on the Virufy Crowdsourced dataset, Clinical Dataset 1 and Clinical Dataset 2 are not significantly impacted by the change of the dataset. As the datasets are a mixture of mp3 and wav files and the mp3 compression downgrades the audio quality, a decrease in performance was expected. However, the Virufy Crowdsourced dataset and the Clinical Dataset 2 have both shown AUC results higher than 0.7 and the Clinical Dataset 1 demonstrates an AUC of 0.59, indicating that the model is generalizable to all four test datasets. The decrease of performance in Clinical Dataset 1 also correlates with the fact that the samples are fairly noisy compared to the others. Figure 6 illustrates the ROC curves of results on the four test datasets, which further confirms our statement that our model can be generalized to different datasets.

We conducted a Student's t-test on each AUC score with the null hypothesis that the model has no ability to distinguish the difference between coughs from COVID-19 patients and normal coughs at a confidence level of 95%. Table 2 shows all the P-Values calculated with the four test datasets and all of them show a result smaller than $\alpha = 0.05$. Our tests prove that there is a significant statistical difference for all of our datasets.

	Coswara /Coughvid	Virufy Crowd sourced	Clinical Dataset 1	Clinical Dataset 2
AUC	0.771	0.721	0.586	0.718
P-Value	0.001	0.002	0.018	0.0003
CI	0.739-0.802	0.678-0.764	0.614-0.768	0.674-0.763

Table 2: Experiment Results, including AUC scores, P-Values, Confidence Intervals (CI) on various datasets: Coswara/Coughvid subset, Virufy Crowdsourced, Clinical Dataset 1 and Clinical Dataset 2.

5 Conclusion and Discussion

Using relevant audio features, the ensemble deep learning model was successful in identifying COVID-19 positive patients. We verified that a crowdsourced approach to collecting data can yield an accurate COVID-19 detection algorithm from cough. Since crowdsourced data is publicly available, we anticipate future work to build on our deep learning approaches and results.

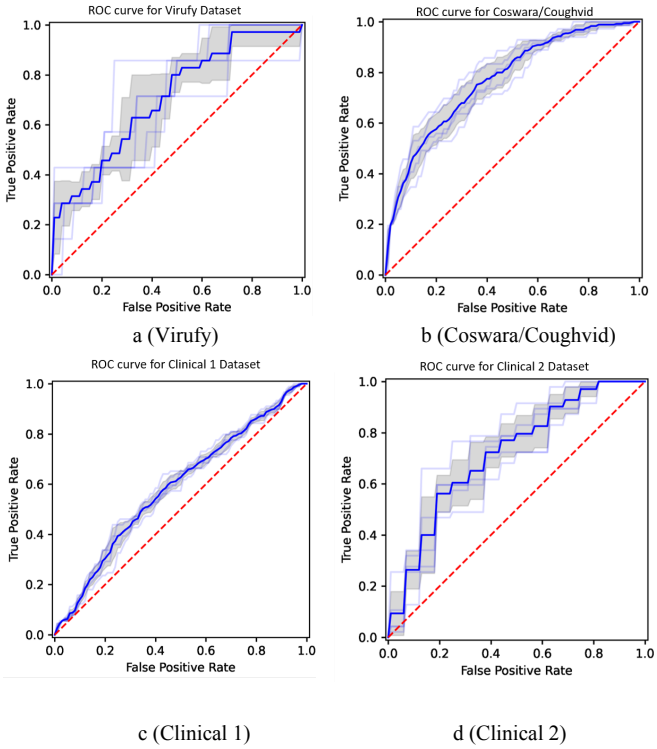


Figure 6: Comparisons of ROC curves between the results from Virufy dataset, the Coswara/Coughvid subset, Clinical Dataset and Clinical Dataset 2.

We also demonstrate that our detection algorithm maintains its performance on external crowdsourced and clinical datasets which were collected using slightly different instructions in environments that were less than ideal and at different stages of infection. This demonstration gives credence to the hypothesis that COVID-19 can be reliably detected from cough sounds, as the virus signature appears to generalize.

Recording cough samples with no audio pollution in a clinical setting, especially in underserved countries, is not easily accomplished. In this study, we demonstrate that a cough-based COVID-19 detection algorithm can perform adequately well on noisy clinical data like our Clinical Dataset 2 (AUC=0.72), but has some limitations as demonstrated by its performance on Clinical Dataset 1 (AUC=0.59). We aim to continue technical work on improving performance on noisy data, with the hope that the eventual algorithm will work on cough data with often unavoidable background noise from hospital wards.

A limitation of this study is sample size: the datasets with detailed labels we selected for this study were not large enough to train the best performing algorithm or to conduct thorough subgroup and longitudinal analyses. To improve the performance of future machine learning algorithms, Virufy intends to collect larger quantities of high-quality

cough data around the world to represent a wide range of ethnicities and community-specific phonological differences across populations. Most existing crowdsourced COVID-19 audio data collection efforts have been focused in Asia, Europe, and the United States. Although Latin America currently has among the highest rates of COVID-19 globally and accounts for 17.3% of current COVID-19 infections at the time of writing this manuscript [21], less than 5% of openly available cough samples are from Latin America [9,10]. We aim to continue our efforts for targeted data collection in Latin America to prevent the exclusion of this community from an AI-based solution.

A well-validated COVID-19 detection algorithm from cough has broad global applicability and can be instrumental in controlling the spread of the disease. However, most current approaches of data collection do not adequately account for variability, such as various settings, COVID-19 status labels, past medical history, and stage and severity of COVID disease. As COVID-19 has a highly variable presentation, including various combinations of anosmia, fever, asymptomatic low oxygen saturation, pneumonia, conjunctivitis, and heart injury [22,23], it is unclear whether any cough-based machine learning algorithm will be equally accurate for the entire spectrum of manifestations.

6 Future work

A key challenge of clinical COVID-19 diagnosis is that its symptoms mimic those of other common respiratory, pulmonary, and cardiac conditions [24]. Therefore, further sub-analyses testing is necessary to determine the ability of machine learning algorithms to distinguish COVID-19 from other illnesses, such as non-COVID-19 pneumonia, upper/lower respiratory infections, asthma, and chronic lung disease exacerbations [19].

Virufy is currently conducting longitudinal crowdsourced studies and clinical studies across various countries. Our goal is to train a machine learning algorithm with more information about human respiratory sound features, including cough and speech, both before symptom onset and over the course of COVID-19 infection. After gathering more audio data in association with PCR and evolving in vitro COVID-19 diagnostics, demographics, and disease course labels, we intend to conduct thorough sub-analyses that can validate an AI solution's performance in a multitude of conditions and demographic groups. As our current models are relatively shallow, we plan to develop deeper models as we collect more data from various contexts.

Although many groups worldwide are collecting various respiratory audio data with COVID-19 labels [9-15], at the time of writing of this manuscript, most data is not openly available. To our knowledge, only Coswara [9] and Coughvid [10] have released open crowdsourced datasets, while only Virufy has released open clinical data[2]. This confidentiality among groups working on the same technical challenge has hampered collaboration, reproducibility of results, and development of a widely-available solution. Moving forward, broad collaboration and data sharing could promote the rapid development of AI-based COVID-19 detection tools.

To facilitate that effort, , Virufy aims to establish a global consortium, bringing together research groups from around the world to share knowledge and datasets to build and rapidly refine AI algorithms to address the COVID-19 pandemic[3].

7 Acknowledgements

We are very grateful to Mary L. Dunne, M.D., Stanford University Distinguished Career Institute Fellow, for her guidance on our data collection procedures and analysis.

We appreciate Siddhi Hedge and Shreya Sriram for their amazing enthusiasm and hard work in facilitating clinical cough data collection from COVID-19 tested patients.

Furthermore, we thank the Coswara and Coughvid groups for open sourcing their COVID-19 datasets. We also thank everyone who has contributed their cough data to Virufy.

References

[1] "Coronavirus (COVID-19)." Google News, Google, news.google.com/covid19/map.

[2] "COVID-19 Overview and Infection Prevention and Control Priorities in Non-US Healthcare Settings." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, www.cdc.gov/coronavirus/2019-ncov/hcp/non-us-settings/overview/index.html.

[3] Udugama, Buddhisha et al. "Diagnosing COVID-19: The Disease and Tools for Detection." ACS nano vol. 14,4 (2020): 3822-3835. doi:10.1021/acsnano.0c02624.

[4] Grant, Michael C et al. "The prevalence of symptoms in 24,410 adults infected by the novel coronavirus (SARS-CoV-2; COVID-19): A systematic review and meta-analysis of 148 studies from 9 countries." PloS one vol. 15,6 e0234765. 23 Jun. 2020, doi:10.1371/journal.pone.0234765.

[5] Mahashur, Ashok. "Chronic dry cough: Diagnostic and management approaches." Lung India : official organ of Indian Chest Society vol. 32,1 (2015): 44-9. doi:10.4103/0970-2113.148450.

[6] Pramono, Renard Xaviero Adhi et al. "A Cough-Based Algorithm for Automatic Diagnosis of Pertussis." PloS

[2] The data can be accessed from: https://virufy.org/data.

[3] The consortium details: https://virufy.org/community.

one vol. 11,9 e0162128. 1 Sep. 2016, doi:10.1371/journal.pone.0162128.

[7] Y. Amrulloh, U. Abeyratne, V. Swarnkar, and R. Triasih, “Cough sound analysis for pneumonia and asthma classification in pediatric population,” 2015 6th International Conference on Intelligent Systems, Modelling andSimulation, pp. 127–131, 2015.

[8] Najafabadi, M. M. et al. “Deep learning applications and challenges in big data analytics.” Journal of Big Data 2 (2014): 1-21.

[9] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. Chetupalli, N. R.,P. Ghosh, and S. Ganapathy, “Coswara – a database of breathing, cough,and voice sounds for covid-19 diagnosis,” 05 2020.

[10] L. Orlandic, T. Teijeiro, and D. Atienza, “The coughvid crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms,” 2020.

[11] M. Cohen-McFarlane, R. Goubran, and F. Knoefel, “Novel coronavirus cough database: Nococoda,” IEEE Access, vol. 8, pp. 154087–154094, 2020.

[12] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat,D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, “Exploring automatic diag-nosis of covid-19 from crowdsourced respiratory sound data,” in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discov-ery amp; Data Mining, KDD ’20, (New York, NY, USA), p. 3474–3484, Association for Computing Machinery, 2020.

[13] J. Laguarta, F. Hueto, and B. Subirana, “Covid-19 artificial intelligence diagnosis using only cough recordings,” IEEE Open Journal of Engineeringin Medicine and Biology, pp. 1–1, 2020.

[14] P. Bagad, A. Dalmia, J. Doshi, A. Nagrani, P. Bhamare, A. Mahale, S. Rane,N. Agarwal, and R. Panicker, “Cough against covid: Evidence of covid-19 signature in cough sounds,” 2020.

[15] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. 2017. CNN architectures for large-scale audio classification. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 131–135.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[17] Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing 28, 4 (1980), 357–366.

[18] McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. "librosa: Audio and music signal analysis in python." In Proceedings of the 14th python in science conference, pp. 18-25. 2015.

[19] J. Korpáš, J. Sadloňová, and M. Vrabec, “Analysis of the Cough Sound: an Overview,” Pulmonary Pharmacology, vol. 9, no. 5-6, pp. 261–268, 1996.

[20] Gómez-Ochoa, S., Franco, O., Rojas, L., Raguindin, P., Roa-Díaz, Z., Wyssmann, B., Guevara, S., Echeverría, L., Glisic, M. and Muka, T., 2020. COVID-19 in Health-Care Workers: A Living Systematic Review and Meta-Analysis of Prevalence, Risk Factors, Clinical Characteristics, and Outcomes. American Journal of Epidemiology.

[21] Dong, E., Du, H. and Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases, 20(5), pp.533-534.

[22] Zaim, S., Chong, J., Sankaranarayanan, V. and Harky, A., 2020. COVID-19 and Multiorgan Response. Current Problems in Cardiology, 45(8), p.100618.

[23] Loffredo, L., Pacella, F., Pacella, E., Tiscione, G., Oliva, A. and Violi, F., 2020. Conjunctivitis and COVID‐19: A meta‐analysis. Journal of Medical Virology, 92(9), pp.1413-1414.

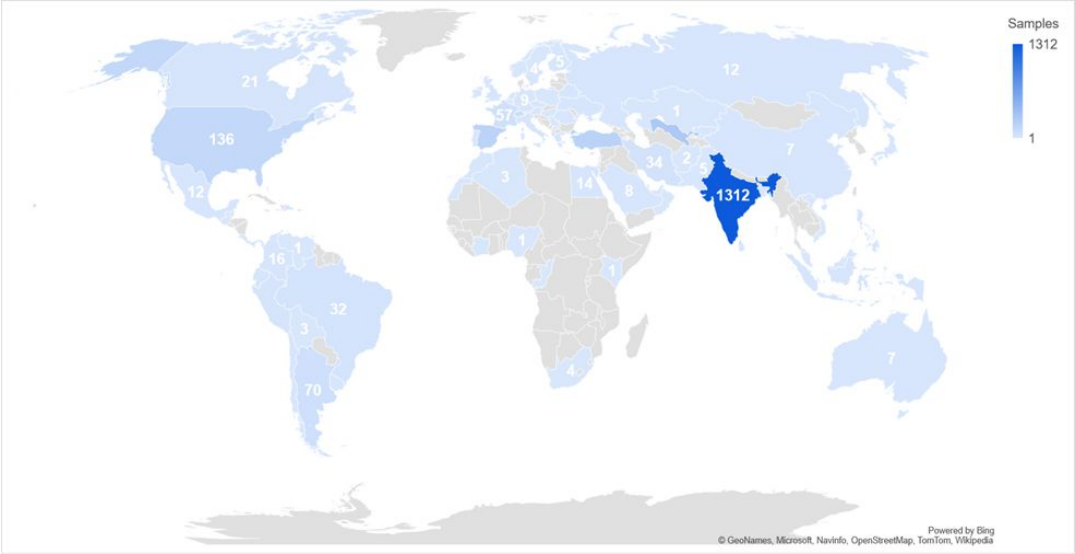
[24] Sahu, K., Mishra, A., Martin, K. and Chastain, I., 2020. COVID-19 and clinical mimics. Correct diagnosis is the key to appropriate therapy. Monaldi Archives for Chest Disease, 90(2).

A Supplementary Material

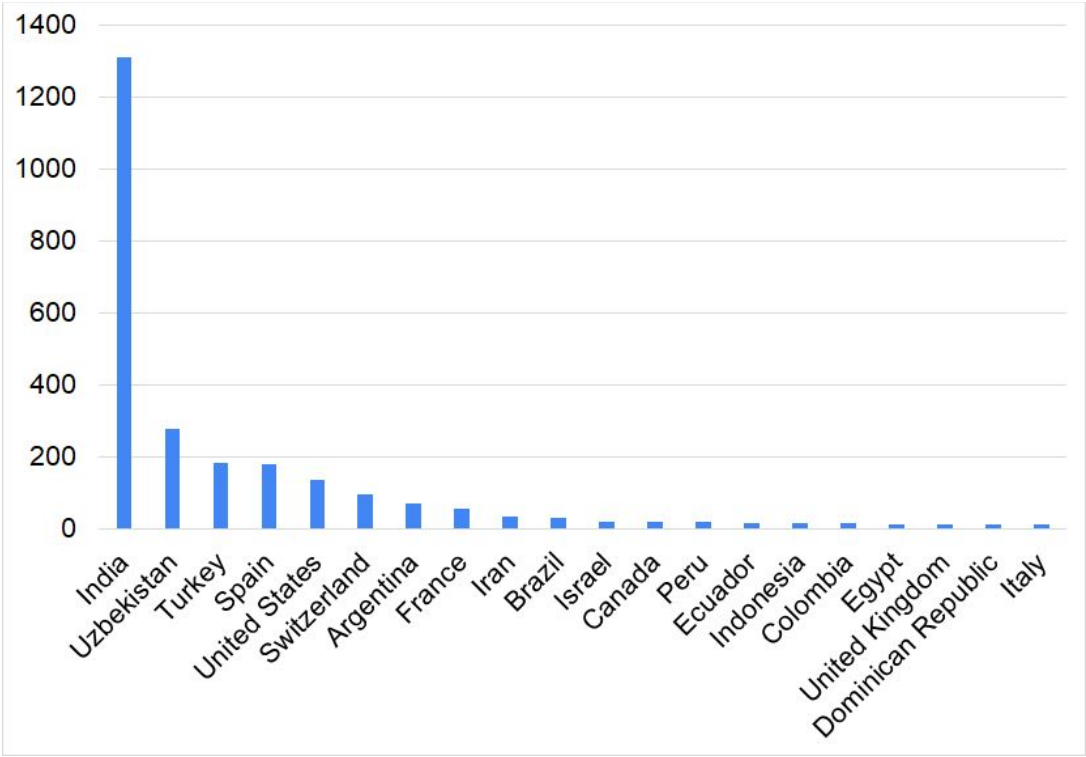
A. 1 Global distribution of training data (n=2748)

The crowdsourced datasets we used had cough samples from all around the world. Graph A below depicts this distribution for the subset of our training data that had location labels. The vast majority of Coswara dataset samples were from India, and most Coughvid dataset ones were from Europe and the United States. Notably, fewer samples were from Latin American and African countries. Graph B shows the number of samples from the top 20 countries.

A)



B)



A. 2 Detailed Characteristics of Test Datasets

Below are several demographic and medical characteristics for the three test datasets used to show model generalizability. People may have several symptoms and medical conditions. (n.c. = not collected)

	Virufy Crowdsourced (n=31)	Clinical Dataset 1 (n=362)	Clinical Dataset 2 (n=63)
PCR Test Result (%)			
Positive	7 (22.6)	89 (24.6)	47 (74.6)
Negative	24 (77.4)	273 (75.4)	16 (25.4)
Sex (%)			
Female	11 (35.4)	117 (32.3)	22 (34.9)
Male	19 (61.2)	244 (67.4)	41 (65.1)
Other	1 (0.03)	1 (0.3)	0 (0)
Age (%)			
0-9	0 (0)	1 (0.3)	0 (0)
10-19	2 (6.5)	15 (4.1)	5 (7.9)
20-29	15 (48.4)	123 (34)	30 (47.6)
30-39	4 (12.9)	86 (23.8)	12 (19)
40-49	2 (6.5)	62 (17.1)	5 (7.9)
50-59	5 (16.1)	56 (15.5)	6 (9.5)
60-69	2 (6.5)	18 (5)	4 (6.3)
70-79	1 (3.2)	1 (0.3)	1 (1.6)
Smoker (%)			
Yes	6 (19.4)	69 (19.1)	17 (27)
No	25 (80.6)	293 (80.9)	46 (73)
Symptoms (%)			
None	20 (64.5)	188 (51.9)	11 (17.5)
New or worsening cough	3 (9.7)	92 (25.4)	24 (38.1)
Sore throat	3 (9.7)	84 (23.2)	28 (44.4)
Fever and/or chills	2 (6.5)	83 (22.9)	40 (63.5)
Dyspnea	3 (9.7)	73 (20.2)	2 (3.2)
Anosmia	1 (3.2)	38 (10.5)	5 (7.9)
Myalgia	3 (9.7)	22 (6.1)	27 (42.9)
Emesis and/or diarrhea	1 (3.2)	4 (1.1)	4 (6.3)
Headaches	5 (16.1)	n.c.	22 (34.9)
Medical Conditions (%)			
None	27 (87.1)	302 (83.4)	57 (90.5)
Asthma	3 (9.7)	18 (5)	1 (1.6)
Chronic Tuberculosis	1 (3.2)	0 (0)	0 (0)
Thrombophilia	1 (3.2)	0 (0)	0 (0)
Extreme Obesity	1 (3.2)	0 (0)	1 (1.6)
Congestive Heart Failure	0 (0)	12 (3.3)	0 (0)
Diabetes	0 (0)	31 (8.6)	5 (7.9)
Hypertension	0 (0)	n.c.	8 (12.7)