# Cough Against COVID: Evidence of COVID-19 Signature in Cough Sounds

**Paper ID 7493**


## Abstract

Testing capacity for COVID-19 remains a challenge globally due to the lack of adequate supplies, trained personnel, and sample-processing equipment. These problems are even more acute in rural and underdeveloped regions. We demonstrate that solicited-cough sounds collected over a phone, when analysed by our AI model, have statistically significant signal indicative of COVID-19 status (AUC 0.72, t-test, $p < 0.01$, 95% CI 0.61—0.83). This holds true for asymptomatic patients as well. Towards this, we collect the largest known (to date) dataset of microbiologically confirmed COVID-19 cough sounds from 3,621 individuals. When used in a triaging step within an overall testing protocol, by enabling risk-stratification of individuals before confirmatory tests, our tool can increase the testing capacity of a healthcare system by 43% at disease prevalence of 5%, without additional supplies, trained personnel, or physical infrastructure.

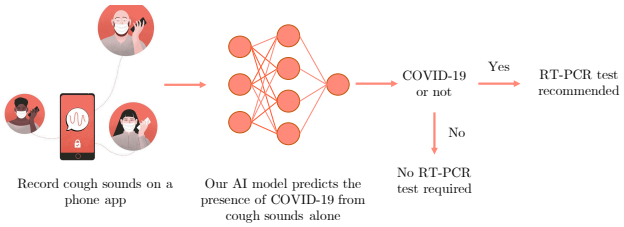




Figure 1: *Cough Against Covid.* An overview of our non-invasive, AI-based pre-screening tool that determines COVID-19 status from solicited-cough sounds. With the AI model set to an operating point of high sensitivity, an individual is referred for gold standard RT-PCR test if they triage +ve for risk of COVID-19. At 5% disease prevalence, this triaging tool would increase the effective testing capacity by 43%.

## 1 Introduction

On 11th March, 2020, the World Health Organisation (WHO) declared COVID-19 (also known as the coronavirus disease, caused by SARS-CoV2) a global pandemic. As of 20th August, 2020, there were more than 22M confirmed cases of COVID-19 globally and over 788K deaths (JHU 2020). Additionally, COVID-19 is still active, with 267K new cases and 6,030 deaths per day world wide. As we eagerly await new drug and vaccine discoveries, a highly effective method to control the spread of the virus is frequent testing and quarantine at scale to reduce transmission (Kucharski et al. 2020). This has led to a desperate need for triaging and diagnostic solutions that can scale globally. While the WHO has identified the key symptoms for COVID-19 – fever, cough, and breathing difficulties, and recently, an expanded list (WHO 2020b), these symptoms are non-specific, and can deluge healthcare systems. Fever, the most common symptom, is indicative of a very wide variety of infections; combining it with a cough reduces the possible etiologies to acute respiratory infections (ARIs), which affect millions at any given time. Additionally, the majority of COVID-19 positive individuals show none of the above symptoms (asymptomatics) but they continue to be contagious (WHO 2020a; Daniel P. Oran 0; Day 2020). To address this challenge, we present an AI-based triaging tool to increase the effective testing capacity of a given public health system. At the current model performance and at a prevalence of 5–30%, our tool can increase testing capacity by 43–33%.

There have been various successful efforts using CT scans and X-rays to classify COVID-19 from other viral infections (Wang and Wong 2020; Hall et al. 2020; Gozes et al. 2020; He et al. 2020). This suggests that COVID-19 affects the respiratory system in a characteristic way (Huang et al. 2020; Imran et al. 2020) (see Section II (B) of (Imran et al. 2020) for a detailed summary). The respiratory system is a key pathway for humans to both cough and produce voice – where air from the lungs passes through and is shaped by the airways, the mouth and nasal cavities. Respiratory diseases can affect the sound of someone's breathing, coughing, and vocal quality – as most readers will be familiar with from having e.g. the common cold. Following this intuition we investigate whether there is a COVID-19 signature in solicited cough sounds and if it can be detected by an AI-model.

The main contributions of this paper are as follows: (i) We demonstrate with statistical significance that solicited-cough sounds have a detectable COVID-19 signature; (ii) Our modelling approach achieves a performance of 72% AUC (area under the ROC curve) on held out subsections of our col-

lected dataset; (iii) We demonstrate with statistical significance that solicited-cough sound has a detectable COVID-19 signature among *only asymptomatic* patients (Fig. 7b); (iv) We collect a large dataset of cough sounds paired with individual metadata and COVID-19 test results. To the best of our knowledge this is currently the largest cough dataset with verified ground truth labels from COVID-19 Reverse Transcription Polymerase Chain Reaction (RT-PCR) test results; and (v) Finally, we describe a triaging use case and demonstrate how our model can increase the testing capacity of the public health system by 43%.

## 2 Motivation and Related Work

Sound has long been used as an indicator for health. Skilled physicians often use stethoscopes to detect the presence of abnormalities by listening to sound from the heart or the lungs. Machine learning (ML), in particular, deep learning, has shown great promise in automated audio interpretation to screen for various diseases like asthma (Oletic and Bilas 2016) and wheezing (Li et al. 2017) using sounds from smartphones and wearables. Open-source datasets like AudioSet (Gemmeke et al. 2017) and Freesound Database (Fonseca et al. 2018) have further boosted research in this domain.

Automated reading of chest X-rays and CT scans (Wang and Wong 2020; Hall et al. 2020; Gozes et al. 2020; He et al. 2020) have been widely studied along with typically collected healthcare data (Soltan et al. 2020) to screen for COVID-19. Respiratory sounds have also been explored for diagnosis (see (Deshpande and Schuller 2020) for a nice overview). Some research has explored the use of digital stethoscope data from lung auscultation as a diagnostic signal for COVID-19 (hui Huang et al. 2020). The use of human-generated audio as a biomarker offers enormous potential for early diagnosis, as well as for affordable and accessible solutions which could be rolled out at scale through commodity devices.

Cough is a symptom of many respiratory infections. Triaging solely from cough sounds can be simple operationally and help reduce load on the healthcare system. (Saba 2018; Botha et al. 2018) detect tuberculosis (TB) from cough sounds, while (Larson et al. 2012) track the recovery of TB patients using cough detection. A preliminary study on detecting COVID-19 from coughs uses a cohort of 48 COVID-19 tested patients versus other pathology coughs to train a combination of deep and shallow models (Imran et al. 2020). Other valuable work in this domain investigates a similar problem (Brown et al. 2020), wherein a binary COVID-19 prediction model is trained on a dataset of crowdsourced, unconstrained worldwide coughs and breathing sounds. In (Han et al. 2020) speech recordings from COVID-19 hospital patients are analyzed to automatically categorize the health state of patients. A crowdsourced dataset (Sharma et al. 2020) of cough, breathing and voice sounds was also recently released to enable sound as a medium for point-of-care diagnosis for COVID-19.

Apart from (Imran et al. 2020) and (Brown et al. 2020), none of the previous efforts actually detect COVID-19 from cough sounds alone. (Imran et al. 2020) covers only 48 COVID-19 tested patients, while our dataset consists of 3,621 individuals with 2,001 COVID-19 tested positives. The dataset used in (Brown et al. 2020) was entirely crowdsourced with the COVID-19 status being self-reported, whereas our dataset consists of labels directly received from healthcare authorities. Further, we show that COVID-19 can be detected from the cough sounds of *asymptomatic* patients as well. Unlike previous works, we also demonstrate how label smoothing can help tackle the inherent label noise due to the sensitivity of the RT-PCR test and improve model calibration.

## 3 Data

In this section we outline our data collection pipeline as well as the demographics and properties of the gathered data. We further describe the subset of the data used for the analysis in this paper.

We note here that we use two types of data in this work. First, we describe data collected from testing facilities and isolation wards for COVID-19 in various states of India, constituting the largest dataset of tested COVID-19 cough sounds (to the best of our knowledge). Next, we mention several open-source cough datasets that we use for pretraining our deep networks.

### 3.1 COVID-19 cough dataset

**Data collection** We create a dataset of cough sounds from COVID-19 tested individuals from numerous testing facilities and isolation wards throughout India (collection is ongoing). Testing facilities provide data for both positively and negatively tested individuals, whereas isolation wards are meant only for those who have already tested positive. For isolation wards, we only consider individuals within the first 10 days after an initial positive result through RT-PCR. Our eligibility criteria also requires that individuals should be independently mobile and be able to provide cough samples comfortably. The data collector is required to wear a PPE kit prior to initiating conversation, and maintain a distance of 5 feet at all times from the participant. The participant is required to wear a triple layer mask and provide written consent. For minors, consent is obtained from a legally acceptable representative. Our data collection and study have been approved by a number of local and regional ethics committees [1].

For each individual, our data collection procedure consists of the following 3 key stages:

1. **Subject enrollment:** In the first stage, subjects are enrolled with metadata such as demographic information (including self-reported age and sex), the presence of symptoms such as dyspnea (shortness of breath), cough and fever, recent travel history, contact with known COVID-19 positive individuals, body temperature, and any comorbidities or habits such as smoking that might render them more vulnerable.

[1] The names of the precise committees have been omitted to preserve anonymity, and will be added to any future versions.

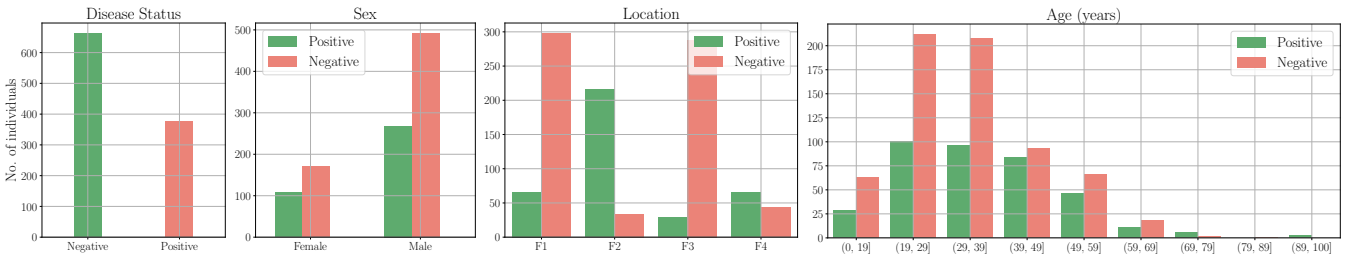



Figure 2: *Dataset demographics.* From left to right – distribution of the number of individuals based on COVID-19 test result, sex, location and age.

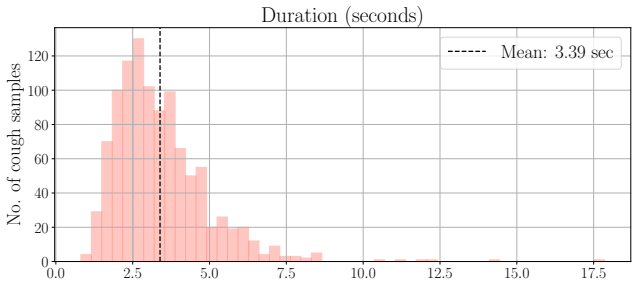



Figure 3: *Duration statistics.* Distribution of the duration of the cough audio recordings.

2. **Cough-sound recording:** Since cough is an aerosol generating procedure, recordings are collected in a designated space which is frequently disinfected as per facility protocol. For each individual, we collect 3 separately recorded audio samples of the individual coughing, an audio recording of the individual reciting the numbers from one to ten and a single recording of the individual breathing deeply. Note here that these are non-spontaneous coughs, i.e. the individual is asked to cough into the microphone in each case, even if they do not naturally have a cough as a symptom.
3. **Testing:** RT-PCR test results are obtained from the respective facility's authorized nodal officers.

For each stage, we utilise a separate application interface. Screenshots for the apps and further details are provided in suppl. material. We note here that the COVID-19 test result is *not* known at the time of audio cough recording – minimising collection bias, and that all data collection is performed in environments in which potential solutions may actually be used.

**Dataset** As of August 16th, 2020 our data collection efforts have yielded a dataset of 3,621 individuals, of which 2,001 have tested positive. In this paper we focus on a curated set of the collected data (until 20 July, 2020). We also restrict our models to use only the cough sounds (and not the voice or breathing samples). Henceforth, all results and statistics will be reported on this data used in our analysis after filtering and manual verification (details of which are provided in the suppl. material). Our curated dataset consists of 3,117 cough sounds from 1,039 individuals. We aim to release some or all of the data publicly to the research community. Figures 2 and 3 show distribution statistics of the data. Out of 1,039 individuals, 376 have a positive RT-PCR test result (Fig. 2, left) and the sex breakdown is 760 male and 279 female. (Fig. 2, center-left). (Fig. 2, center-right) highlights the distribution by the facility from which the data was collected (we use data from 4 facilities, F1-F4). (Fig. 2, right) shows the age distribution, which is skewed towards middle-aged individuals (between 20-40 years of age), while Fig. 3 shows the distribution of the lengths of our cough samples. Fig 4 shows the distribution of symptoms recorded for dyspnea, cough and fever. Interestingly, note that most individuals are asymptomatic. In our dataset, the most common single symptom among COVID-19 positive individuals is fever while that among negatives is cough, followed by an intersection of cough and fever.

### 3.2 Open-source non-COVID cough datasets

In the absence of explicit feature engineering, deep Convolutional Neural Networks (CNNs) are data hungry – relying on thousands of manually annotated training examples. Given the challenges of training deep CNNs from scratch on small datasets, we collect a larger dataset of cough samples from various public datasets (Fonseca et al. 2018; Al Hossain et al. 2020; Sharma et al. 2020) which we use to pretrain our model. In total we obtain 31,909 sounds segments, of which 27,116 are non-cough respiratory sounds (wheezes, crackles or breathing) or human speech, and 4,793 are cough sounds. The various data sources and their statistics are as follows:

**1. FreeSound Database 2018 (Fonseca et al. 2018):** This is an audio dataset consisting of a total of 11,073 audio files annotated with 41 possible labels, of which 273 samples are labelled as 'cough'. We believe the cough sounds correspond to COVID-19 negative individuals as these sounds were recorded well before the COVID-19 pandemic.

**2. Flusense (Al Hossain et al. 2020):** This is a subset of Google's Audioset dataset (Gemmeke et al. 2017), consisting of numerous respiratory sounds. [2] We use 11,687 audio segments of which 2,486 are coughs.

**3. Coswara (Sharma et al. 2020):** This is a curated dataset of coughs collected via worldwide crowd sourcing using

[2] Including speech, coughs, sneezes, sniffles, silence, breathing, gasps, throat-clearing, vomit, hiccups, burps, snores, and wheezes.

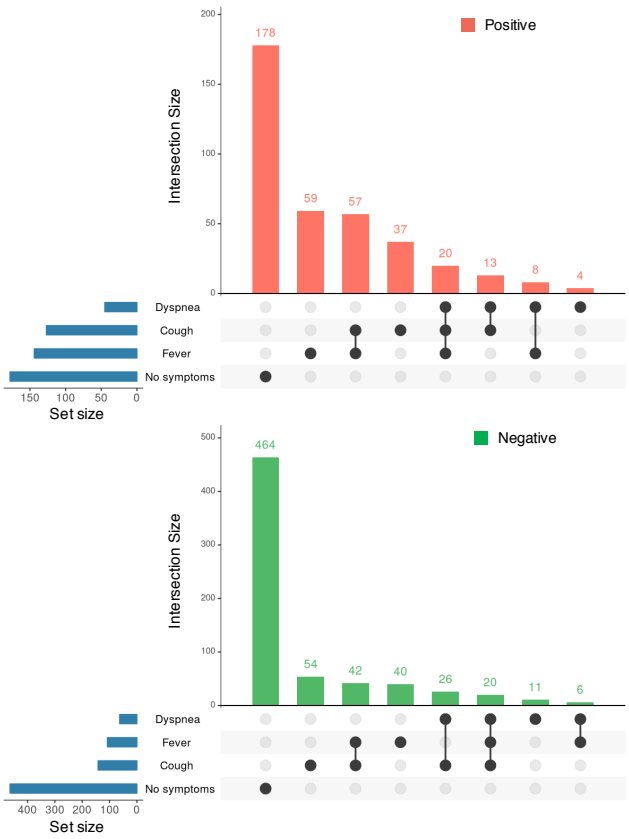



Figure 4: *Symptom co-occurrence statistics.* We show statistics for individuals with an RT-PCR positive (top) and negative (bottom) test for the following symptoms: dyspnea (shortness of breath), cough and fever.

a website application[3]. The dataset contains samples from 570 individuals, with 9 voice samples for each individual, including breathing sounds (fast and slow), cough sounds (heavy and shallow), vowel sounds, and counting (fast and slow). In total the dataset consists of 2,034 cough samples and 7,115 non-cough samples. We are unaware of the COVID-19 status of the coughs in this dataset as it was collected after the pandemic broke out.

# 4 Method

Inspired by the recent success of CNNs applied to audio inputs (Hershey et al. 2016), we develop an end-to-end CNN-based framework that ingests audio samples and directly predicts a binary classification label indicating the probability of the presence of COVID-19. In the following sections, we outline details of the input, model architecture, training strategies employed and inference.

## 4.1 Input

During training we randomly sample a 2-second segment of audio from the entire cough segment. We use short-term magnitude spectrograms as input to our CNN model. All audio is first converted to single-channel, 16-bit streams at a 16kHz sampling rate for consistency. Spectrograms are then generated in a sliding window fashion using a hamming window of width 32ms and hop 10ms with a 512-point FFT. This gives spectrograms of size 257 x 201 for 2 seconds of audio. The resulting spectrogram is integrated into 64 mel-spaced frequency bins with minimum frequency 125Hz and maximum frequency 7.5KHz, and the magnitude of each bin is log transformed. This gives log-melspectrogram patches of 64 x 201 bins that form the input to all classifiers. Finally, the input is rescaled by the largest magnitude over the training set to bring the inputs between -1 and 1.

## 4.2 CNN architecture

An overview of our CNN architecture can be seen in Fig. 5. As a backbone for our CNN model we use the popular ResNet-18 model consisting of residual convolution layers (He et al. 2016), followed by adaptive pooling layer in both the time and frequency dimensions. Finally, the output is passed through 2 linear layers and then a final predictive layer with 2 neurons and a softmax activation function, which is used to predict whether the input cough sample has COVID-19. Dropout (Srivastava et al. 2014) and the ReLU activation function are used after all linear layers.

## 4.3 Training strategies

**Augmentation** Given the medium size of our dataset, we adopt the standard practise of data augmentation, applying transformations to our data to boost performance and increase robustness. We perform data augmentation online, i.e. transformations are applied randomly to segments during training. We perform two types of augmentation: (1) the addition of external background environmental sounds from the ESC-50 dataset (Piczak 2015), and (2) time and frequency masking of the spectrogram input (Park et al. 2019). ESC-50 (Piczak 2015) consists of 2,000 environmental audio recordings from 50 environmental classes. At train time, we randomly select a single noise sample and modulate the amplitude by a random factor between 0.4 and 0.75, before adding it to the input cough spectrogram.

**Pre-training** Our model architecture is first pretrained on the open source cough datasets outlined in Sec. 3.2. We partition the data into train and validation (the validation set consists of 648 cough and 2882 non-cough sounds), and train our model to simply predict the presence of a cough or not (cough detection). Note that this is a proxy task and we use this simply to pretrain our model and learn a good initialisation of weights.

We first initialise the ResNet-18 backbone with weights obtained from pretraining on ImageNet (the additional linear layers after are initialised randomly). Given the highly unbalanced nature of the pretraining data, we upsample the minority class to ensure that each batch has the equal number of cough and non-cough samples. AdamW (Loshchilov and Hutter 2017) is used as the optimizer with a learning rate of 1e-5 and weight decay 1e-4. The model is trained for

[3]https://coswara.iisc.ac.in/

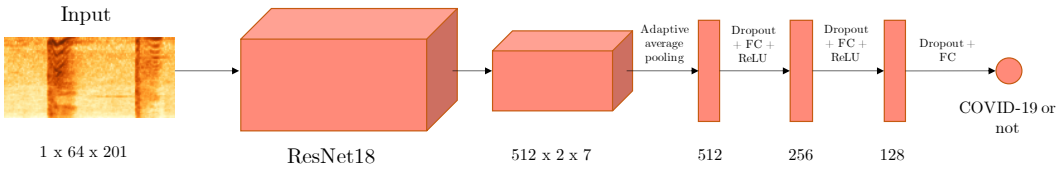



Figure 5: *Network architecture.* An input cough spectrogram goes through a deep CNN to predict the presence of COVID-19.

200 epochs and on the proxy cough vs. non-cough task, we achieve an AUC of 0.98 on the validation set.

**Label smoothing** For our final task of COVID-19 classification, we note here that the ground truth labels come solely from the RT-PCR test for COVID-19. Even though this test is widely used, it is known to make mistakes, i.e. it is estimated to have a sensitivity of almost 70% at a specificity of 95% (Watson, Whiting, and Brush 2020). Hence it is possible that a number of cough samples may have the wrong label, and penalising our model for making mistakes on these samples can harm training. Hence we apply a standard label smoothing technique (Müller, Kornblith, and Hinton 2019) during training for each instance. Label smoothing is also known to improve model calibration (Müller, Kornblith, and Hinton 2019). Results are provided in Sec. 6.2.

**Implementation details** For cough classification, we use the pretrained weights from the cough non-cough pretraining task to initialize the model. SGD is used as the optimizer, with an initial learning rate of 0.001 and a decay of 0.95 after every 10 epochs. We use a batch size of 32 and train for a total of 110 epochs. Label smoothing is randomly applied between 0.1 and 0.3. Our model is implemented in PyTorch (Paszke et al. 2019) (version 1.6.0) and trained using a single Tesla K80 GPU on the Linux operating system. The same seed has been set for all our experiments (more details can be found in suppl. material). We used Weights & Biases (Biewald 2020) (version 0.9.1) for experiment tracking and visualisation.

### 4.4 Inference

Every cough sample is divided into 2-second segments using a sliding window with a hop length of 500ms. We take the median over the *softmax* outputs for all the segments to obtain the prediction for a single sample. We pad inputs less than 2 seconds with zeros. A comparison of different aggregation methods have been provided in suppl. material.

**Individual-level aggregation** For each individual in the dataset, we have three cough samples. We consider the *max* of the predicted probabilities of the three cough samples to obtain the prediction for a single individual. All performance metrics have been reported at the individual level.

## 5 Experimental Evaluation

### 5.1 Tasks

Although we train our model on the entire dataset once, we focus on three clinically meaningful evaluations:

- **Task 1:** Distinguish individuals tested *positive* from individuals tested *negative* for COVID-19.
- **Task 2:** Distinguish individuals tested *positive*, from individuals tested *negative* for COVID-19, specifically for individuals that do *not report cough as a symptom*. We refer to this set as Asymptomatic (no C).
- **Task 3:** Distinguish individuals tested *positive*, from individuals tested *negative* for COVID-19, specifically for individuals that do *not report cough, fever or breathlessness as a symptom*. We refer to this set as Asymptomatic (no C/F/D).

The number of cough samples in the validation set for each task are provided in Table 1. Fig. 7b shows the comparison in performance across the three tasks.

| Task | Positive | Negative |
|---|---|---|
| (1) | 87-102 | 108-117 |
| (2) | 57-75 | 78-105 |
| (3) | 45-66 | 69-93 |

Table 1: *Dataset statistics per task.* Number of cough samples in the validation set for each task. Since we perform 5-fold validation, we show the range from min-max. Note that the precise number of samples varies across folds as we select 10% of the total dataset but ensure that the validation set is balanced per facility. Note that each individual has three cough samples.

### 5.2 Triple-stratified cross-validation

In order to create a fair evaluation, we (1) create training and validation sets from disjoint individuals, (2) we balance the number of positive and negatives obtained from each facility in the validation set, to ensure that we are not measuring a facility specific bias, and (3) we upsample the minority class samples per facility in the train set (facility-wise class distribution has been shown in Fig. 2). We split our dataset into train and validation sets of approximately 90%:10% ratio, and following standard practise for ML methods on small datasets, perform 5-fold cross-validation.

### 5.3 Evaluation metrics

We report several standard evaluation metrics such as the Receiver Operating Characteristic - Area Under Curve (ROC-AUC), Specificity (1 - False Positive Rate (FPR)), and Sensitivity (also known as True Positive Rate (TPR)). Since

this solution is meant to be used as a triaging tool, high sensitivity is important. Hence, we report the best specificity at 90% sensitivity. We report mean and standard deviation across all 5 cross-validation folds. For fairness, all hyperparameters are set on the first fold and applied, as is, to other folds, including epoch selection.

### 5.4 Comparison to shallow baselines

We also compare our CNN-based model to shallow classifiers using hand-crafted audio features. We experiment with the following classifiers: (1) Logistic Regression (LR), (2) Gradient Boosting Trees (3) Extreme Gradient Boosting (XGBoost) and (4) Support Vector Machines (SVMs). As input to the classifiers, we use a range of features such as the tempo, RMS energy and MFCCs (see Sec. 4.1 from (Brown et al. 2020) for an exhaustive list of the features used.) For all methods, we follow the preprocessing design choices adopted by (Brown et al. 2020). We optimize the hyperparameters following the same procedure outlined in 5.3.

### 5.5 Stacked ensemble

We ensemble the individual-level predictions from ResNet-18 (both with and without label smoothing) and the XGBoost classifier (described in detail in Sec. 5.4) using Stacked Regression (Van der Laan, Polley, and Hubbard 2007). The stacked regressor is a XGBoost classifier using the predicted probabilities from each of the above models as features. The hyperparameters for the regressor are mentioned in the suppl. material. We report performance with and without the ensemble (Fig. 7a).

### 5.6 Ablation analysis

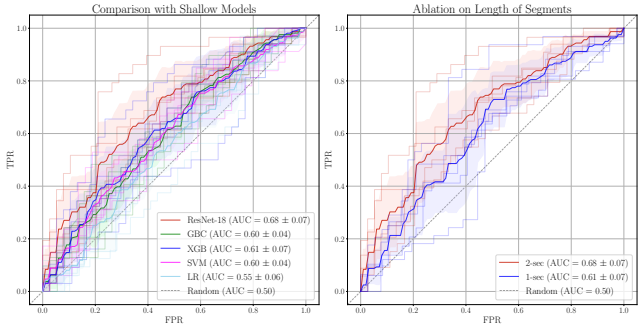
We also quantify the effect of several aspects of our training pipeline, notably - pretraining, label smoothing and the length of the input segment. We experiment with two segment lengths - 1 second and 2 seconds. For the model trained on 1-second input segments, we perform hyperparameter tuning again. Results for all ablation analysis are provided in Sec. 6.

## 6 Discussion

Fig. 6a shows that the CNN-based model outperforms all shallow models by atleast 7% in terms of AUC. We also perform a statistical significance analysis of the results of our model. We conduct a $t$-test with the Null Hypothesis that there is no COVID-19 signature in the cough sounds and the results were found to be statistically significant, $p < 1e-3$, 95% confidence interval (CI) 0.61—0.83.

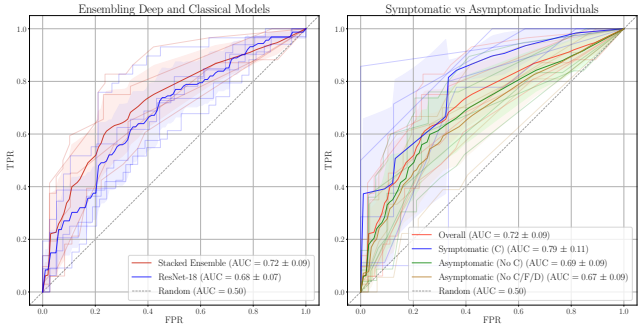
### 6.1 Effect of ensembling

It is widely known that ensembling diverse models can improve performance, even if some models perform worse than others individually (Sagi and Rokach 2018). Fig. 7a empirically validates this for our task by showing that ensembling the deep and shallow models improves performance compared to any of the individual models. This also indicates that there is further room for performance improvement through better ensembling techniques and using more diverse models.




(a) Shallow models vs CNN (b) 1-sec vs 2-sec segments

Figure 6: *Ablation results.* Comparison of ROC curves across (a) different model families - ResNet-18 outperforms other shallow baselines; (b) different segment lengths. 2-second is found to be the optimal segment length




(a) Ensembling (b) Asymptomatic individuals

Figure 7: *Classification results.* Comparison of ROC curves (a) for our best model obtained by ensembling a shallow model with the deep model. (b) for symptomatic and asymptomatic individuals: C - cough, F - fever, D - dyspnea (shortness of breath). Our model is able to identify COVID-19 from the cough sounds of asymptomatic individuals as well.

### 6.2 Effect of label smoothing

The effect of applying label smoothing has been reported in Table 2. Besides improving AUC, label smoothing also improves the specificity at 90% sensitivity. This shows that at the required operating point (threshold on the softmax scores) for a triaging tool, the model is able to classify better with smoothened labels. This suggests that explicitly dealing with label noise can improve performance. We also empirically verify that label smoothing improves model calibration (Müller, Kornblith, and Hinton 2019) as it drives the optimal threshold for COVID-19 classification much closer to 0.5.

### 6.3 Effect of pre-training

Table 3 shows the utility of using pretrained weights. Pretraining improves the mean AUC by 17%, showing it's im-

| Model | AUC | Specificity | Threshold |
|---|---|---|---|
| with LS | $0.68 \pm 0.07$ | $0.31 \pm 0.13$ | $0.422 \pm 0.062$ |
| no LS | $0.65 \pm 0.08$ | $0.27 \pm 0.11$ | $0.002 \pm 0.002$ |

Table 2: *Effect of label smoothing.* Label smoothing improves specificity at 90% sensitivity and model calibration.

portance in dealing with small or medium sized datasets like ours.

| Model | AUC |
|---|---|
| with pretraining | $0.68 \pm 0.07$ |
| no pretraining | $0.51 \pm 0.07$ |

Table 3: *Effect of pre-training.* Pre-training greatly improves model performance.

### 6.4 Optimal segment length

Fig. 6b indicates that using segments of 2-seconds performs better than 1-second segments. We suspect that this happens because our dataset contains several samples with silence at the start and the end, increasing the probability of noisy labels being assigned to random crops during training.

### 6.5 Asymptomatic individuals

Fig. 7b shows the performance for asymptomatics. We see that while our model performs significantly better for symptomatic individuals, performance for asymptomatic individuals is still far above random. A $t$-test was conducted with the Null Hypothesis that there is no COVID-19 signature in the cough sounds of asymptomatic patients and the results were found to be statistically significant, $p < 1e-2$.

### 6.6 Performance across sex and location

While we note that our dataset contains more males than females, there is no obvious bias in COVID-19 test results (Fig. 2), and performance is similar for both male ($0.71 \pm 0.11$) and female ($0.72 \pm 0.11$) individuals.

Samples collected from different *locations* can have different label distributions. For example, testing facilities (F1, F3 and F4) tend to have predominantly COVID-19 negatives while isolation wards (F2) tends to contain COVID-19 positives (Fig. 2). Naively training a classifier on this combined dataset would lead to significantly inflated performance because it could simply learn a *location* classifier instead of a COVID-19 cough classifier. This is a known phenomenon in deep learning and medical imaging (Badgeley et al. 2019) (Wachinger et al. 2019). To address this issue, we carefully constructed our validation set to contain only testing facilities with equal number of positive and negative samples per location. Future work will explore algorithmic mitigation by applying techniques such as (Zhang, Lemoine, and Mitchell 2018).

## 7 Use Case: COVID-19 Triaging Tool

In India alone, as of the 21st of August, 2020, there have been over 33M COVID19 RT-PCR tests performed (ICMR 2020). While the current testing capacity is 800k/day, the test positivity rate (TPR) has been increasing at a steady pace, indicating that there is an urgent need for testing to be ramped up even further. The ability to ramp up tests, however, is significantly hindered by the limited supply of testing kits and other operational requirements such as trained staff and lab infrastructure. This has led to an increased urgency for accurate, quick and non-invasive triaging, where individuals most likely to be determined positive for COVID19 are tested as a priority.

To address this, we propose a triaging tool that could be used by both individuals and health care officials. We pick the threshold of the model such that we have a high sensitivity of 90% which is desirable for a triaging tool. At this sensitivity our best model has a specificity of 31%. As shown in Fig. 1, such a model can be used to reliably detect *COVID-19 negative individuals* while we refer the positives for a confirmatory RT-PCR test. In this way, we increase the testing capacity by 43% (a 1.43x lift) when we assume a disease prevalence of 5%. In Table 4, we also show the relative gains at different prevalence levels. Precise calculations can be found in the suppl. material.

| Prevalence | Testing Capacity |
|---|---|
| 1% | +44% |
| 5% | +43% |
| 10% | +41% |
| 30% | +33% |

Table 4: *Utility of our triaging tool.* We show the increase in the effective testing capacity of a system at different disease prevalence levels.

## 8 Conclusion and Future Work

In this paper, we describe a non-invasive, machine learning based triaging tool for COVID-19. We collect and curate a large dataset of cough sounds with RT-PCR test results for thousands of individuals, and show with statistical evidence that our model can detect COVID-19 in the cough sounds from our dataset, even for patients that are entirely *asymptomatic*. At current model performance, our tool can improve the testing capacity of a healthcare system by 43%. Future work will involve incorporating other inputs from our dataset to the model, including breathing sounds, voice samples and symptoms. Our data collection is ongoing, and subsequent models will be trained on individuals beyond the subset in this study. We will also explore fast and computationally efficient inference, to enable COVID-19 testing on smartphones. This will enable large sections of the population to self-screen, support proactive testing and allow continuous monitoring.

## References

Al Hossain, F.; Lover, A. A.; Corey, G. A.; Reich, N. G.; and Rahman, T. 2020. FluSense: a contactless syndromic surveillance platform for influenza-like illness in hospital

waiting areas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4(1): 1–28.

Badgeley, M. A.; Zech, J. R.; Oakden-Rayner, L.; Glicksberg, B. S.; Liu, M.; Gale, W.; McConnell, M. V.; Percha, B.; Snyder, T. M.; and Dudley, J. T. 2019. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ digital medicine* 2(1): 1–10.

Biewald, L. 2020. Experiment Tracking with Weights and Biases. URL https://www.wandb.com/. Software available from wandb.com.

Botha, G.; Theron, G.; Warren, R.; Klopper, M.; Dheda, K.; Van Helden, P.; and Niesler, T. 2018. Detection of tuberculosis by automatic cough sound analysis. *Physiological measurement* 39(4): 045005.

Brown, C.; Chauhan, J.; Grammenos, A.; Han, J.; Hasthanasombat, A.; Spathis, D.; Xia, T.; Cicuta, P.; and Mascolo, C. 2020. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. *arXiv preprint arXiv:2006.05919* .

Daniel P. Oran, E. J. T. 0. Prevalence of Asymptomatic SARS-CoV-2 Infection. *Annals of Internal Medicine* 0(0): null. doi:10.7326/M20-3012. URL https://doi.org/10.7326/M20-3012. PMID: 32491919.

Day, M. 2020. Covid-19: four fifths of cases are asymptomatic, China figures indicate. *BMJ* 369. doi:10.1136/bmj.m1375. URL https://www.bmj.com/content/369/bmj.m1375.

Deshpande, G.; and Schuller, B. 2020. An Overview on Audio, Signal, Speech, & Language Processing for COVID-19. *arXiv preprint arXiv:2005.08579* .

Fonseca, E.; Plakal, M.; Font, F.; Ellis, D. P.; Favory, X.; Pons, J.; and Serra, X. 2018. General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline. *arXiv preprint arXiv:1807.09902* .

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. IEEE.

Gozes, O.; Frid-Adar, M.; Greenspan, H.; Browning, P. D.; Zhang, H.; Ji, W.; Bernheim, A.; and Siegel, E. 2020. Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis. *arXiv preprint arXiv:2003.05037* .

Hall, L. O.; Paul, R.; Goldgof, D. B.; and Goldgof, G. M. 2020. Finding Covid-19 from Chest X-rays using Deep Learning on a Small Dataset. *arXiv e-prints* arXiv:2004.02060.

Han, J.; Qian, K.; Song, M.; Yang, Z.; Ren, Z.; Liu, S.; Liu, J.; Zheng, H.; Ji, W.; Koike, T.; et al. 2020. An Early Study on Intelligent Analysis of Speech under COVID-19: Severity, Sleep Quality, Fatigue, and Anxiety. *arXiv preprint arXiv:2005.00096* .

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, X.; Yang, X.; Zhang, S.; Zhao, J.; Zhang, Y.; Xing, E.; and Xie, P. 2020. Sample-Efficient Deep Learning for COVID-19 Diagnosis Based on CT Scans. *medRxiv* doi:10.1101/2020.04.13.20063941. URL https://www.medrxiv.org/content/early/2020/04/17/2020.04.13.20063941.

Hershey, S.; Chaudhuri, S.; Ellis, D. P. W.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; Slaney, M.; Weiss, R. J.; and Wilson, K. 2016. CNN Architectures for Large-Scale Audio Classification.

Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; Cheng, Z.; Yu, T.; Xia, J.; Wei, Y.; Wu, W.; Xie, X.; Yin, W.; Li, H.; Liu, M.; Xiao, Y.; Gao, H.; Guo, L.; Xie, J.; Wang, G.; Jiang, R.; Gao, Z.; Jin, Q.; Wang, J.; and Cao, B. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* 395(10223): 497 – 506. ISSN 0140-6736. doi:https://doi.org/10.1016/S0140-6736(20)30183-5. URL http://www.sciencedirect.com/science/article/pii/S0140673620301835.

hui Huang, Y.; jun Meng, S.; Zhang, Y.; sheng Wu, S.; Zhang, Y.; wei Zhang, Y.; xiang Ye, Y.; feng Wei, Q.; gui Zhao, N.; ping Jiang, J.; et al. 2020. The respiratory sound features of COVID-19 patients fill gaps between clinical data and screening methods. *medRxiv* .

ICMR. 2020. ICMR: SARS-CoV-2 (COVID-19) Testing Status. https://www.icmr.gov.in/. Accessed: 2020-08-21.

Imran, A.; Posokhova, I.; Qureshi, H. N.; Masood, U.; Riaz, S.; Ali, K.; John, C. N.; and Nabeel, M. 2020. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *arXiv preprint arXiv:2004.01275* .

JHU. 2020. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). https://coronavirus.jhu.edu/map.html. Accessed: 2020-08-21.

Kucharski, A. J.; Klepac, P.; Conlan, A.; Kissler, S. M.; Tang, M.; Fry, H.; Gog, J.; and Edmunds, J. 2020. Effectiveness of isolation, testing, contact tracing and physical distancing on reducing transmission of SARS-CoV-2 in different settings. *medRxiv* doi:10.1101/2020.04.23.20077024. URL https://www.medrxiv.org/content/early/2020/04/29/2020.04.23.20077024.

Larson, S.; Comina, G.; Gilman, R. H.; Tracey, B. H.; Bravard, M.; and López, J. W. 2012. Validation of an automated cough detection algorithm for tracking recovery of pulmonary tuberculosis patients. *PloS one* 7(10): e46229.

Li, S.-H.; Lin, B.-S.; Tsai, C.-H.; Yang, C.-T.; and Lin, B.-S. 2017. Design of wearable breathing sound monitoring system for real-time wheeze detection. *Sensors* 17(1): 171.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* .

Müller, R.; Kornblith, S.; and Hinton, G. 2019. When Does Label Smoothing Help?

Oletic, D.; and Bilas, V. 2016. Energy-efficient respiratory sounds sensing for personal mobile asthma monitoring. *Ieee sensors journal* 16(23): 8295–8303.

Park, D. S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E. D.; and Le, Q. V. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779* .

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, 8026–8037.

Piczak, K. J. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, 1015–1018.

Saba, E. 2018. *Techniques for Cough Sound Analysis*. Ph.D. thesis, University of Washington.

Sagi, O.; and Rokach, L. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8(4): e1249.

Sharma, N.; Krishnan, P.; Kumar, R.; Ramoji, S.; Chetupalli, S. R.; R., N.; Ghosh, P. K.; and Ganapathy, S. 2020. Coswara – A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis.

Soltan, A. A.; Kouchaki, S.; Zhu, T.; Kiyasseh, D.; Taylor, T.; Hussain, Z. B.; Peto, T.; Brent, A. J.; Eyre, D. W.; and Clifton, D. 2020. Artificial intelligence driven assessment of routinely collected healthcare data is an effective screening test for COVID-19 in patients presenting to hospital. *medRxiv* doi:10.1101/2020.07.07.20148361. URL https://www.medrxiv.org/content/early/2020/07/08/2020.07.07.20148361.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15(1): 1929–1958.

Van der Laan, M. J.; Polley, E. C.; and Hubbard, A. E. 2007. Super learner. *Statistical applications in genetics and molecular biology* 6(1).

Wachinger, C.; Becker, B. G.; Rieckmann, A.; and Pölsterl, S. 2019. Quantifying confounding bias in neuroimaging datasets with causal inference. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 484–492. Springer.

Wang, L.; and Wong, A. 2020. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. In *arXiv:2003.09871*.

Watson, J.; Whiting, P. F.; and Brush, J. E. 2020. Interpreting a covid-19 test result. *BMJ* 369. doi:10.1136/bmj.m1808. URL https://www.bmj.com/content/369/bmj.m1808.

WHO. 2020a. Coronavirus disease 2019 (COVID-19) Situation Report – 46. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200306-sitrep-46-covid-19.pdf?sfvrsn=96b04adf_4#:~:text=For\%20COVID\%2D19\%2C,infections\%2C\%20requiring\%20ventilation. Accessed: 2020-08-21.

WHO. 2020b. Q&A on coronaviruses (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses. Accessed: 2020-08-21.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.