# IMPACT OF DATA-SPLITS ON GENERALIZATION: IDENTIFYING COVID-19 FROM COUGH AND CONTEXT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Rapidly scaling screening, testing and quarantine has shown to be an effective strategy to combat the COVID-19 pandemic. We consider the application of deep learning techniques to distinguish individuals with COVID from non-COVID by using data acquireable from a phone. Using cough and context (symptoms and meta-data) represent such a promising approach. Several independent works in this direction have shown promising results. However, none of them report performance across clinically relevant data-splits. Specifically, the performance where the development and test sets are split in time (retrospective validation) and across sites (broad validation). Although there is meaningful generalization across these splits the performance significantly varies (up to 0.1 AUC score) [Table 1]. In addition, we study the performance on symptomatic and asymptomatic individuals across these three splits. Finally, we show that our model focuses on meaningful features of the input, 'cough' bouts for cough and relevant symptoms for context.

## 1 INTRODUCTION

Screening, testing and quarantining has been an effective strategy employed by public health systems around the world to combat the spread of the COVID-19. There are two main objectives of any public health system, namely, minimize the spread of the virus and save lives. To achieve these objectives, especially during a pandemic, there are limited resources available like testing kits, hospital beds, x-ray machines etc. We consider the approach of using machine learning algorithms to increase the efficiency of resource allocation within the public health system. Furthermore, a non-invasive, widely accessible, scalable, and accurate model could be an impactful solution. Identifying COVID-19 from respiratory acoustics like cough and user-reported symptoms represent two such approaches. Both of these approaches simply require a phone with a microphone. This could be used by the public health system as a screening tool for effective allocation of the lab-tests.

There has been several prior work using cough (Brown et al. (2020); Imran et al. (2020); Bagad et al. (2020)), voice (Bartl-Pokorny et al. (2020); Pinkas et al. (2020)), breathing sound (Faezipour & Abuzneid (2020)), symptoms and their combinations (Schuller et al. (2020); Coppock et al. (2021)) to identify COVID-19. Qian et al. (2020) and Deshpande & Schuller (2020) provide fairly comprehensive surveys on related works. Given this evidence and the approaches' practical advantages, it's surprising that it has seen limited clinical deployment. One of the reasons could be the lack of trust from the healthcare community as verbalized by Topol (2020) in this article. This paper debates highly publicized work claiming 98.5% accuracy for identifying COVID from cough. Their critical review could be summarized as following. 1) Lack of reproducibility 2) Lack of clinically relevant evaluations 3) Interpretation of deep learning models. In this work, we address these limitations to build confidence in this line of work and spur further research.

The main contribution of this work is as following:

1) Dataset and model: We describe our multi-site, lab-tested ground-truth dataset collected for this paper. Detailed description of our modelling approach to identify COVID-19 from cough and contextual-data.

2) Evaluations: We show the model's performance across three clinically meaningful data splits as defined in this paper by Norgeot et al. (2020) and Kleppe et al. (2021). Specifically, the performance where the development and test sets are split randomly, in time (retrospective validation) and across
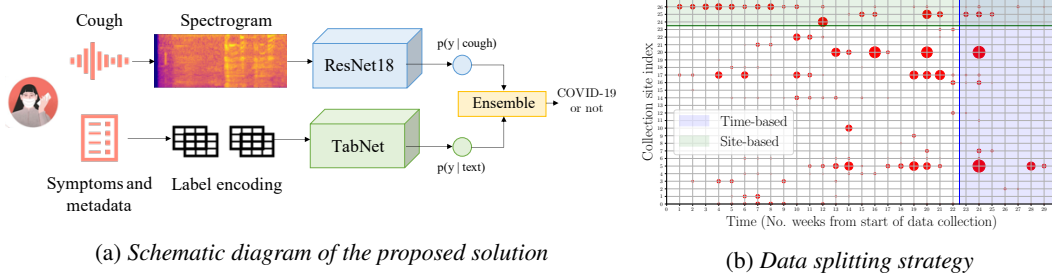
(a) *Schematic diagram of the proposed solution*



(b) *Data splitting strategy*

Figure 1: (a) For every individual, we get predicted probabilities $p(y = 1|\mathbf{x}_{\text{cough}})$ and $p(y = 1|\mathbf{x}_{\text{context}})$ from the individual classifiers. For our final prediction we use a simple ensembling scheme that averages the predictions from the two classifiers. (b) The data has been partitioned into train/test splits in 3 ways: (i) random, (ii) time-based (blue region denotes the test set), (iii) site-based (green). The size of a point denotes the number of samples collected in a given week ($x$-axis) at a given site ($y$-axis) of the data collection drive.

sites (broad validation). We see meaningful generalization across these splits however the performance varies up to 0.1 AUC score [Table 1]. In addition, we report model performance across a special sub-populations: symptomatic and asymptomatic individuals [Table 1].

3) Interpretability: We compute the saliency maps of the cough classifier and show qualitatively it focuses on the 'cough' parts of the input. For context-based classifiers, we show the most predictable features at an instance level.

## 2 DATA

**Open-source non-COVID cough datasets:** Given the challenges of training deep learning model from scratch on a relatively small dataset, we collate a larger dataset of audio samples from various public datasets - FreeSound Database by Fonseca et al. (2018), FluSense by Hossain et al. (2020) and Coswara by Sharma et al. (2020) which we use to pretrain our model for cough detection. 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.

**COVID-19 dataset:** We collect this dataset from individuals who have undergone a COVID-19 test, from numerous testing sites across the country. In addition, contextual data such as symptoms, travel history, contact with confirmed case and demographic information etc is collected. Unlike crowd-sourced datasets such as Brown et al. (2020), Orlandic et al. (2020) and Sharma et al. (2020) that rely on self-reported COVID-19 status, our ground-truth is lab test results from the healthcare facilities. This dataset consists of 12,780 cough sounds from 4,260 individuals from 27 different sites. 1,394 have a positive test result and 2866 remaining are tested negatives. Please see Appendix A.1, A.2 for more details on data collection process and list of contextual features.

**Dataset splits:** A key challenge in developing deep learning models on real-world data is to devise the right developmental (train-validation) and test splits. In order to establish robustness towards deployment of our model, we use 3 different strategies for splitting the data. All samples of an individual are always part of the same set. The details on the data distribution across the splits are shown in Figure 3.

**1. Random** We randomly split the data with train-validation-test splits being 80%:10%:10%.

**2. Time-based**: In order to perform a *retrospective validation* as laid out by Kleppe et al. (2021), we split the data by time - we select two cut-off dates s.t. the validation and test sets are about 10% each and the remaining data forms the training set. This is demonstrated in Figure 1b with $x$-axis denoting the time in weeks from the start of data collection.

**3. Site-based**: In order to perform a *broad validation* as laid out by Kleppe et al. (2021), we split the data by sites. We construct the test set by selecting a set of 3 sites s.t. they constitute about 20% of the dataset. This is demonstrated in Figure 1b with $y$-axis denoting site indices in data collection.

| Model | Task 1 | Task 2 | Task 3 | Task 1 - S/A | Task 2 - S/A | Task 3 - S/A |
|---|---|---|---|---|---|---|
| Cough-based | 0.787 | 0.690 | 0.761 | 0.820 / 0.713 | 0.510 / 0.699 | 0.709 / 0.613 |
| Context-based | 0.718 | 0.650 | 0.669 | 0.610 / 0.730 | 0.449 / 0.560 | 0.559 / 0.645 |
| Ensembling | **0.797** | **0.718** | **0.774** | 0.816 / 0.740 | 0.498 / 0.709 | 0.707 / 0.671 |

Table 1: Given the skewed label distribution, we use a distribution-agnostic metric: Receiver Operating Characteristic - Area Under Curve (AUC) (Left) Results on the test set for each of the 3 splits: Task 1 (random split), Task 2 (time-based split) , and Task 3 (site-based split). (Right) We analyze performance across symptomatics (S)/ asymptomatics (A) on all tasks. We consider an individual symptomatic if they report at least one of the symptoms: cough, fever, or shortness of breath.

## 3 METHOD

### 3.1 COUGH-BASED CLASSIFICATION

Inspired by the recent success of CNNs applied to audio inputs by Hershey et al. (2016), we develop a CNN-based framework that ingests spectrogram representations of audio and directly predicts the probability of the presence of COVID-19. In this sections, we outline details of the input processing, model architecture, training strategies employed and inference for cough-based model.

**Input processing:** During training, we randomly sample a 2-second audio-segment of the cough recording and transform it into a log-melspectrogram patch of $64 \times 201$ bins that forms the input to the classifier. We denote this input as $\mathbf{x}_{\text{cough}}$ and the predicted label as $y$. Details in Appendix A.3

**Data augmentations:** In order to increase robustness to noise, we apply the following augmentations (online) while training: (a) addition of external background environmental sounds from ESC-50 dataset by Piczak (2015), and (b) time and frequency masking of the spectrogram input as in Park et al. (2019). At train time, we randomly select a noise sample from ESC-50, modulate the amplitude by a random factor between 0.4 and 0.75 and add it to the input cough sound.

**Training:** We use ResNet-18 by He et al. (2016) as the backbone network, followed by two linear layers and the final output layer with softmax activation that outputs the distribution $p(y|\mathbf{x}_{\text{cough}})$. Dropout with probability $0.4$ and ReLU activation are used after all linear layers.

Our model is first pretrained on the open source cough datasets outlined in Sec. 2. We train the model to classify cough (not covid cough). For the downstream cough classification, we use this pretrained model to train with AdamW, an initial learning rate of $1e-4$ and a decay of 0.95 after every 10 epochs. We use a batch size of 128 and train for a total of 200 epochs. For evaluation, we pick a checkpoint based on the epoch with best AUC on validation set. We set the same seed for all experiments. The same training procedure is followed for all data-splits.

**Inference**: Every cough sample is divided into 2-second segments using a sliding window with a hop length of 500ms. We pad inputs less than 2 seconds with zeros. We take the median over the *softmax* outputs for all the segments to obtain the prediction for a single sample. Since we have three cough samples per individual, we aggregate the predicted probabilities at an individual level using the *max* operator. All performance metrics henceforth have been reported at the individual level.

### 3.2 CONTEXT-BASED CLASSIFICATION

Tree-based models have traditionally worked well when compared to deep neural networks (DNNs) on tabular datasets. However, TabNet, as proposed by Arik & Pfister (2019), combines the benefits of gradient-based learning of DNNs and sparse feature selection of tree-based models. For our context-based classification task, we leverage TabNet as our classifier.

**Input Preprocessing:** The input to the context-based classifier is a feature vector consisting of continuous variables (age, temperature etc) and categorical variables (presence of cough as symptom, travel history etc). For a detailed list of features used, refer to Appendix A.2. Categorical variables are label encoded while continuous variables are normalized to zero mean and unit standard deviation. We denote this input to the context-based classifier as $\mathbf{x}_{\text{context}}$.

(a) GradCAM++ saliency mask

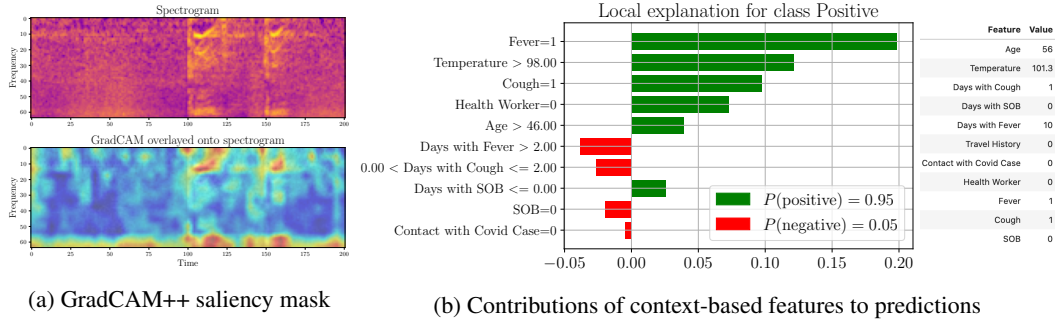(b) Contributions of context-based features to predictions

Figure 2: (a) Interpretability for cough-based model: GradCAM++ saliency map of areas of focus of the model on input spectrogram image. (b) Interpretablity for context-based model: We analyze contributions of context-based features to the prediction score using the LIME method.

**Modelling:** TabNet, a deep network for tabular data that uses sequential attention to choose which features to reason from at each decision step. Input is a preprocessed context-based features and outputs two logits with softmax activation that produce the distribution $p(y|\mathbf{x}_{context})$. We train with the standard cross-entropy loss and use SGD optimizer with a learning rate of $1e - 2$ and batch size of 128. The validation set is used for early stopping, with AUC as the performance measure. All hyperparameters are set to default values used in this implementation. [1]

## 4 EXPERIMENTAL EVALUATION AND DISCUSSION

**Performance across data splits**: We train three models using the same hyper parameters as defined in Methods for each of the three tasks/split. As evident from Table 1, we observe that it is easiest to generalize on the *random* split followed by *time-based* and *site-based*. We hypothesize that this difference arises from the shift in label distribution across the splits. This also highlights the importance of selecting splits carefully since high-performant models on a randomized split may not generalize well in a deployment setting. Additionally, note that even though the performance on the other two splits is lower, they closely mimic the deployment scenario and thus it may induce confidence and are in-line with recommendations of Norgeot et al. (2020) and Kleppe et al. (2021).

**Performance across (a)symptomatics**: On the *random* split, we report model performance across special sub-populations: symptomatic and asymptomatic groups in Table 1. The cough-based model performs well for the two sub-groups, importantly, even for completely asymptomatic individuals. Contrary to intuition, the context-based model performs much worse on symptomatics than on asymptomatics. On further investigation, we find that the context-based model naively assigns lower probability scores for asymptomatic patients. Note that our dataset is skewed towards asymptomatics and the label distribution among them favors COVID-19 negative set. We suspect that the context-based model picks this association of asymptomatics being negative which disproportionately hurts its performance on symptomatics. The similar performance of cough-based model on these sub-populations is desirable since it is less prone to this data distribution skew.

**Interpretability**: Given the clinical uncertainty of this task and the use of deep learning for it, it is essential for clinicians to qualitatively understand the model behaviour and its predictions. As a sanity check, we employ GradCAM++ by Chattopadhyay et al. (2017) to compute these saliency map. We consistently observe that the focus-areas are on and around the cough bouts as shown in Figure 2a. This reinforces the belief that the model is indeed making predictions based on cough signal and otherwise.

For the context-based classifier, we use Local Interpretable Model-agnostic Explanations (LIME by Ribeiro et al. (2016)) to understand which specific features help the model differentiate between COVID+ and COVID- patients at an instance level. For example in Figure 2b, note that presence of fever/cough, temperature > 98 contribute to positive score whereas no shortness-of-breath contributes to a negative score. For more examples refer to A.4.

---

[1]https://github.com/dreamquark-ai/TabNet

REFERENCES

Sercan O Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *arXiv preprint arXiv:1908.07442*, 2019.

Piyush Bagad, Aman Dalmia, J Doshi, Arsha Nagrani, Parag Bhamare, Amrita Mahale, Saurabh Rane, Neeraj Agarwal, and Rahul Panicker. Cough against COVID: Evidence of COVID-19 signature in cough sounds. *ArXiv*, 2020.

Katrin D Bartl-Pokorny, Florian B Pokorny, Anton Batliner, Shahin Amiriparian, Anastasia Semertzidou, Florian Eyben, Elena Kramer, Florian Schmidt, Rainer Schönweiler, Markus Wehler, and Björn W Schuller. The voice of COVID-19: Acoustic correlates of infection. December 2020.

Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data. *arXiv preprint arXiv:2006.05919v2*, 2020.

Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Gradcam++: Generalized gradient-based visual explanations for deep convolutional networks. *CoRR*, abs/1710.11063, 2017. URL http://arxiv.org/abs/1710.11063.

Harry Coppock, Alexander Gaskell, Panagiotis Tzirakis, Alice Baird, Lyn Jones, and Björn W Schuller. End-2-End COVID-19 detection from breath & cough audio. January 2021.

Gauri Deshpande and Björn Schuller. An overview on audio, signal, speech, & language processing for COVID-19. May 2020.

Miad Faezipour and Abdelshakour Abuzneid. Smartphone-Based Self-Testing of COVID-19 using breathing sounds. *Telemed. J. E. Health.*, 26(10):1202–1205, October 2020. ISSN 1530-5627, 1556-3669. doi: 10.1089/tmj.2020.0114.

Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel PW Ellis, Xavier Favory, Jordi Pons, and Xavier Serra. General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline. *arXiv preprint arXiv:1807.09902*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification, 2016.

Forsad Al Hossain, A. Lover, and George A. Corey. Flusense: A contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas. *ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2020.

Ali Imran, Iryna Posokhova, Haneya N Qureshi, Usama Masood, Muhammad Sajid Riaz, Kamran Ali, Charles N John, Md Iftikhar Hussain, and Muhammad Nabeel. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Inform Med Unlocked*, 20: 100378, June 2020. ISSN 2352-9148. doi: 10.1016/j.imu.2020.100378.

Andreas Kleppe, Ole-Johan Skrede, Sepp De Raedt, Knut Liestøl, David J Kerr, and Håvard E Danielsen. Designing deep learning studies in cancer diagnostics. *Nature Reviews Cancer*, 21(3): 199–211, 2021. ISSN 1474-1768. doi: 10.1038/s41568-020-00327-9. URL https://doi.org/10.1038/s41568-020-00327-9.

Beau Norgeot, Giorgio Quer, Brett K Beaulieu-Jones, Ali Torkamani, Raquel Dias, Milena Gianfrancesco, Rima Arnaout, Isaac S Kohane, Suchi Saria, Eric Topol, Ziad Obermeyer, Bin Yu, and Atul J Butte. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature Medicine*, 26(9):1320–1324, 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-1041-y. URL https://doi.org/10.1038/s41591-020-1041-y.

Lara Orlandic, Tomás Teijeiro, and David Atienza. The coughvid crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms. 09 2020.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

Karol J Piczak. ESC: dataset for environmental sound classification. In *ACM Multimedia*, 2015.

Gadi Pinkas, Yarden Karny, Aviad Malachi, Galia Barkai, Gideon Bachar, and Vered Aharonson. SARS-CoV-2 detection from voice. *IEEE Open J. Eng. Med. Biol.*, 1:268–274, 2020. ISSN 2644-1276. doi: 10.1109/ojemb.2020.3026468.

K Qian, B Schuller, and Y Yamamoto. Recent advances in computer audition for diagnosing COVID-19: An overview. *ArXiv*, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Björn W Schuller, Harry Coppock, and Alexander Gaskell. Detecting COVID-19 from breathing and coughing sounds using deep neural networks. December 2020.

Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, Nirmala R., Prasanta Kumar Ghosh, and Sriram Ganapathy. Coswara - a database of breathing, cough, and voice sounds for covid-19 diagnosis, 2020.

Eric J Topol. Is my cough COVID-19? *The Lancet*, 396(10266):1874, dec 2020. ISSN 0140-6736. doi: 10.1016/S0140-6736(20)32589-7. URL https://doi.org/10.1016/S0140-6736(20)32589-7.

## A APPENDIX

### A.1 DATA COLLECTION DETAILS

Our data collection pipeline consists of the following stages: (i) collection of individual specific metadata, (ii) recording of audio samples and finally (iii) obtaining the results of the COVID-19 RT-PCR test. We manually clean and verify all collected data. We achieve this through three separate application interfaces. The details of data collected through these apps are enlisted below:

- **Personal and Demographic information**: We collect the individual's name, mobile number, age, location (facility) and self-reported biological sex.
- **Health-related information**: We collect the COVID-19 RT-PCR test result, body temperature and respiratory rate. We also note the presence of symptoms like fever, cough, shortness of breath and number of days since these symptoms first appear, and any measures undertaken specifically for cough relief. Finally, we also ask individuals if they have any co-morbidities.
- **Additional metadata**: Additional data collected includes location (name of the facility, City and State), travel history of the individual, information about contact with confirmed COVID-19 cases, whether they are a health worker, and information about habits such as smoking, tobacco.

The distribution of positives and negatives in each of the splits are shown in Figure 3. The $x$-axis denotes the set (train or test) and $y$-axis denotes the number of individuals. Note that each individual has three cough samples thus the overall dataset size for cough-based classification triples.

### A.2 CONTEXTUAL FEATURES

The list of all features used in the context-based classification is provided in Table 2

Figure 3: Distribution of positives and negatives in the train and test sets for each of the data splits.

| Feature | Type | Other Details |
|---|---|---|
| Patient Age | Continuous | - |
| Patient Temperature | Continuous | - |
| Days with Cough | Continuous | - |
| Days with Shortness of Breath | Continuous | - |
| Days with Fever | Continuous | - |
| Presence of Cough | Discrete | Yes/No |
| Presence of Shortness of Breath | Discrete | Yes/No |
| Presence of Fever | Discrete | Yes/No |
| Contact with Covid Confirmed Case | Discrete | Yes/No |
| Is Patient a Health Worker | Discrete | Yes/No |
| Travel History | Discrete | No / Inter-district / Inter-state / Inter-country |

Table 2: Input features to context-based classifier

## A.3 DETAILS OF INPUT PRE-PROCESSING

During training, we randomly sample a 2-second audio-segment of the input sample and 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 $257 \times 201$ for 2 seconds of audio. and is further 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 $64 \times 201$ 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.
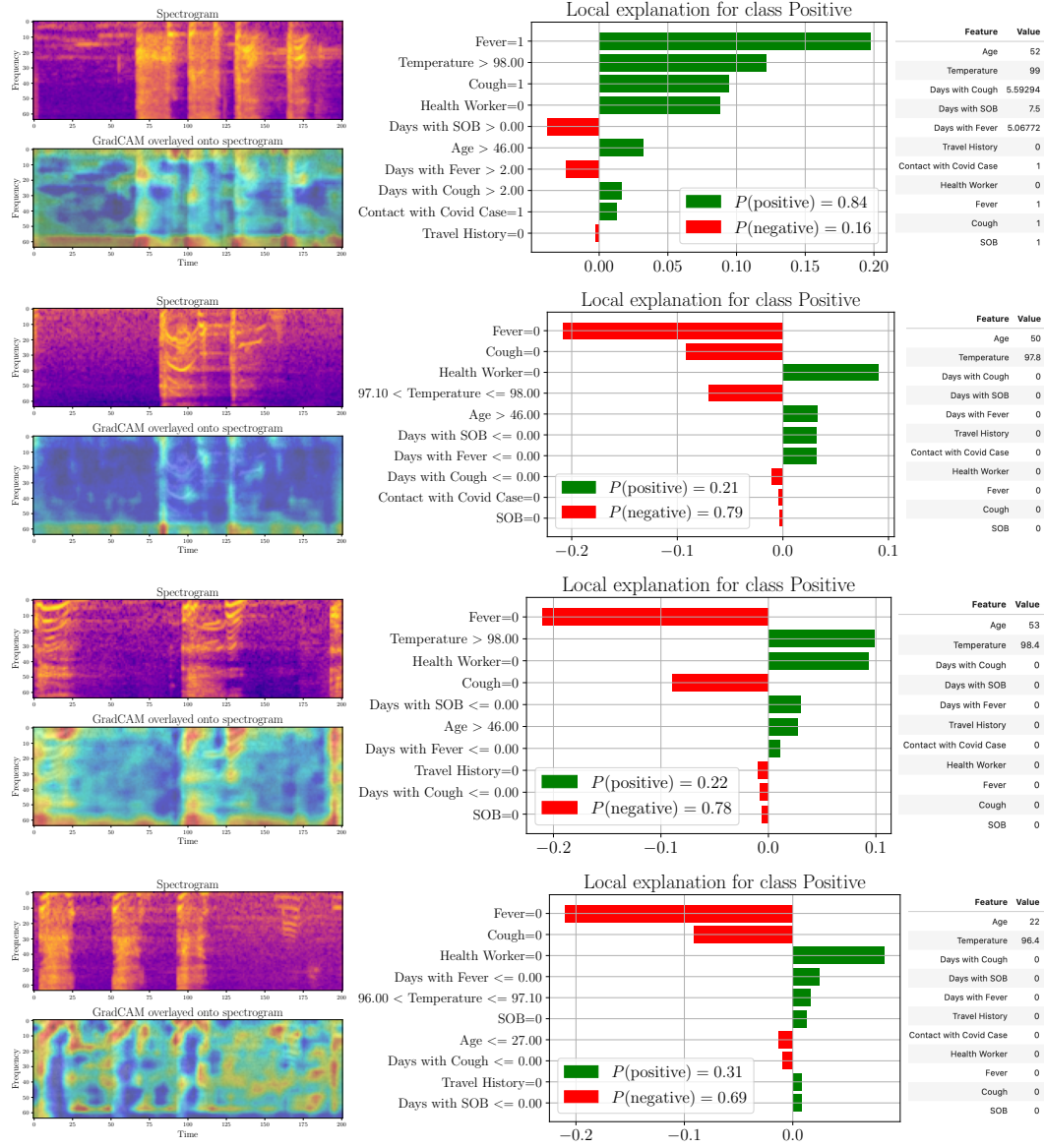
## A.4 MORE SAMPLES FOR INTERPRETABILITY

Figure 4: More examples of interpretability analysis: (Left) GradCAM++ saliency maps overlaid onto spectrograms. (Right) Contribution of contextual feature to prediction scores based on LIME method. The left and right illustrations are for the same individuals in each row.