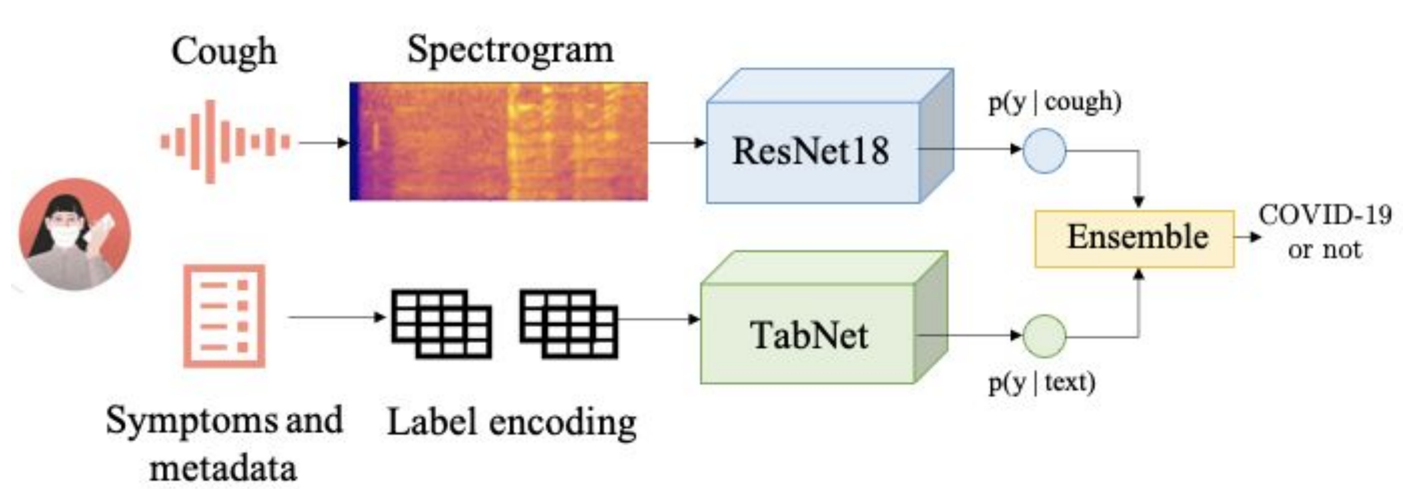
Impact of data-splits on generalization : Identifying COVID-19 from cough and context

Makkunda Sharma*, Nikhil Shenoy*, Jigar Doshi*,
Piyush Bagad *, Aman Dalmia, Saurabh Rane,
Amrita Mahale, Neeraj Agarwal, Rahul Panicker
Wadhwani Institute for Artificial Intelligence

Introduction

- We consider the application of classifying COVID from non-COVID patients using cough data acquireable from a phone
- There have been several independent works in this direction , however, **none of them report** performance across clinically relevant data-splits.
- We compute the performance where the development and test sets are split in time (**retrospective validation**) and across sites/hospitals (**broad validation**) as defined here. Although there is meaningful generalization across these splits the performance significantly varies (up to 10% AUC score)
- We are releasing the code and checkpoints with this paper https://github.com/WadhwaniAI/cough-against-covid

Our Proposed Solution

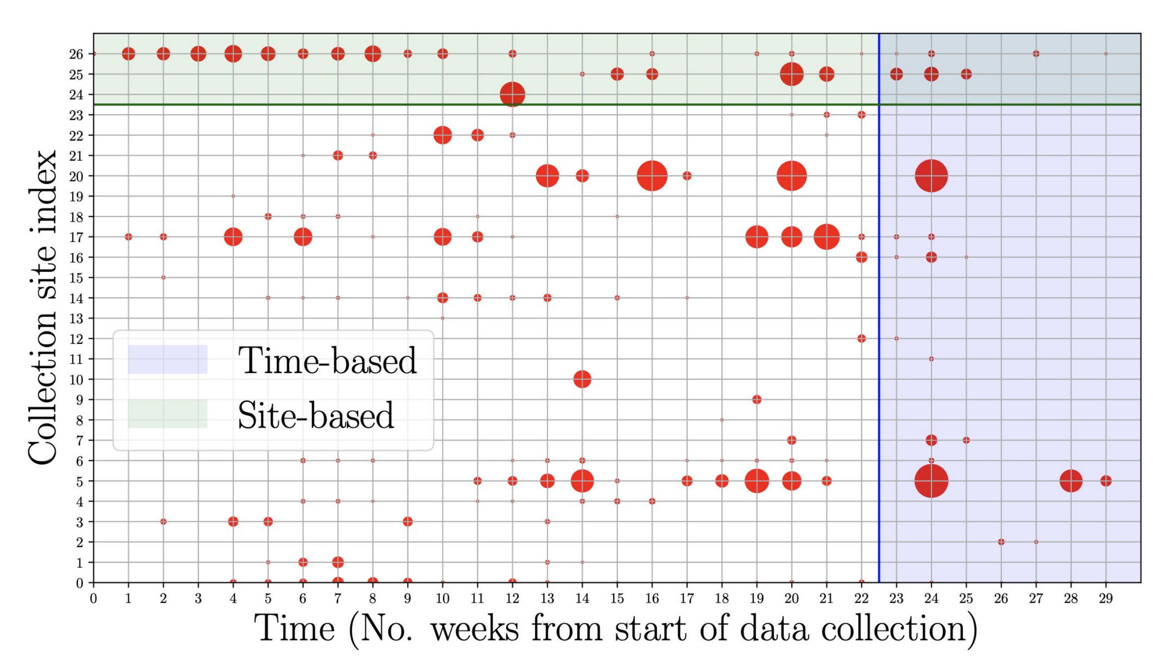


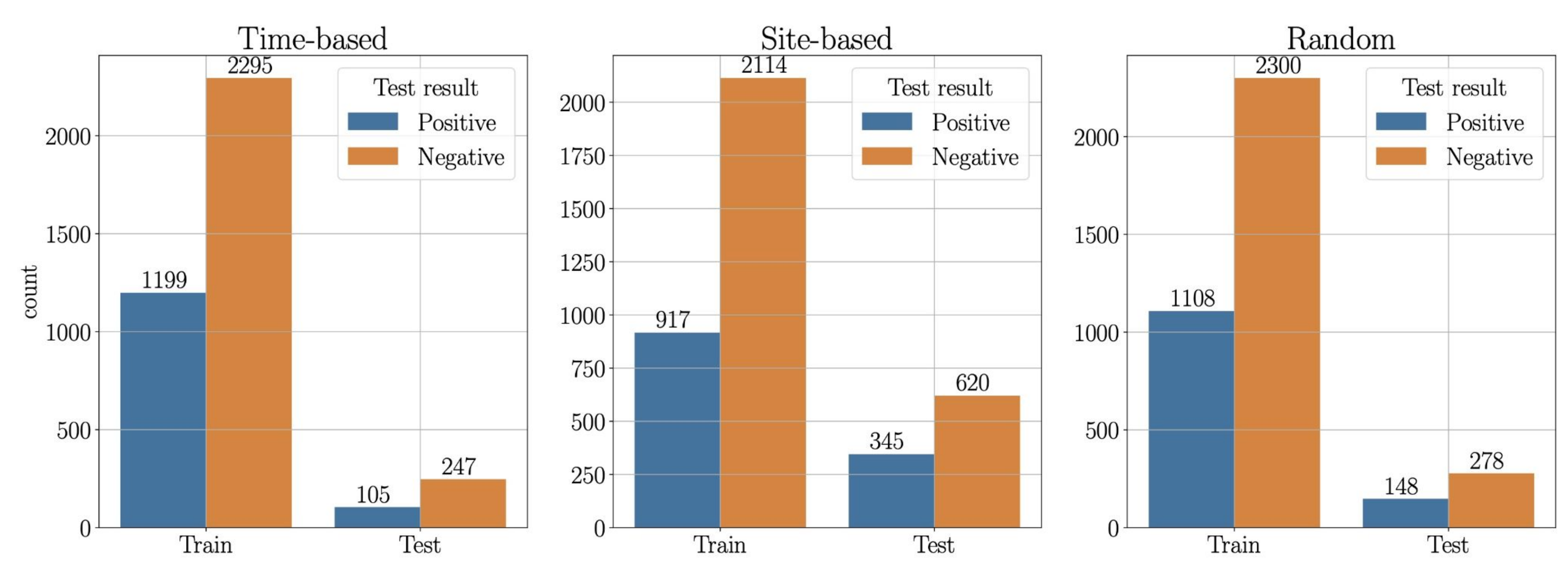
- We use Resnet++ as the classifier that ingests spectrogram representations of audio as input and predicts probability of the presence of COVID-19
- For our context-based classification task, we leverage TabNet as our classifier.
- For our final prediction we use a simple ensembling scheme that averages the predictions from the two classifiers.

Dataset

- We collect this dataset from individuals the day they undergone a COVID-19 swab test, from **27 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 that rely on self-reported COVID-19 status, our **ground-truth is lab test** (RT-PCR) results from the healthcare facilities
- This dataset consists of **12,780 cough sounds** from **4,260** individuals. 1,394 have a positive test result and 2866 remaining are tested negatives.

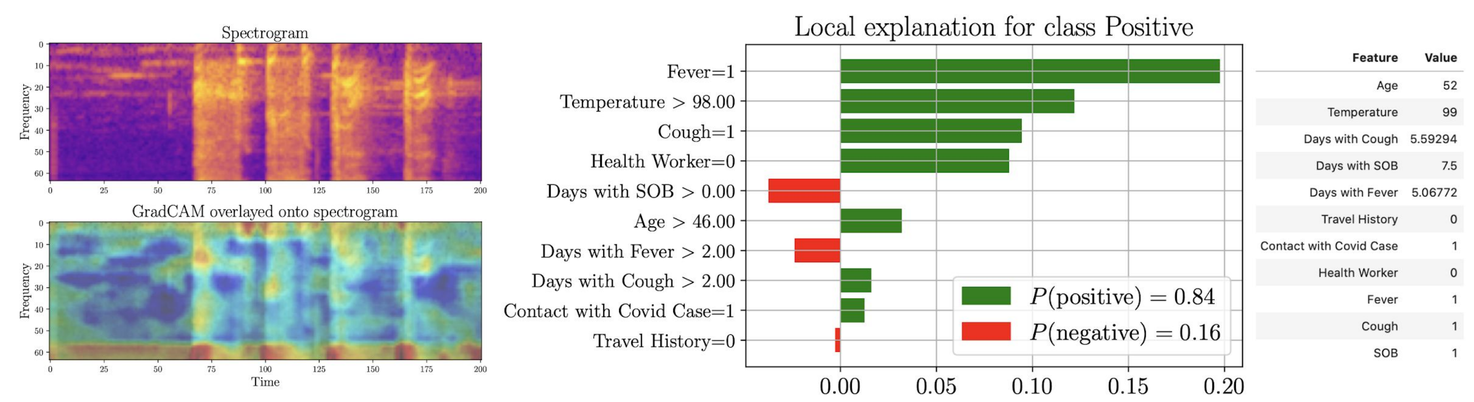
Data Splitting Strategies





Data Visualization

- 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++ to compute these saliency map. We consistently observe that the focus-areas are on and around the cough bouts.
- For the context-based classifier, we use Local Interpretable Model-agnostic Explanations (LIME) to understand which specific features help the model differentiate between COVID+ and COVID- patients at an instance level.



(a) GradCAM++ saliency mask

(b) Contributions of context-based features to predictions

Results

Model	Task 1	Task 2	Task 3	Task 1 - Symptomatic	Task 1 - Asymptomatic
Cough-based	0.787	0.690	0.761	**0.820**	0.713
Context-based	0.718	0.650	0.669	0.610	0.730
Ensembling	**0.797**	**0.718**	**0.774**	0.816	**0.740**

- Data Splitting Tasks:
  - Task 1: Random data split (Typical in ML)
  - Task 2: Time-based data split (Simulating deployment)
  - Task 3: Site-wise data split (Using in new environment)
- Symptomatic: individual has **at least one** of fever, cough or dyspnea.
- 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 highlights the importance of selecting splits carefully since high-performant models on a randomized split may not generalize well in a deployment setting.**

ICLR
USAID FROM THE AMERICAN PEOPLE
BILL & MELINDA GATES foundation
IGIB INSTITUTE OF GENOMICS & INTEGRATIVE BIOLOGY Genomics Knowledge Partner
fondation BOTNAR
WADHWANI AI