Confident COVID-19 cough prediction on imbalanced data

Julia A. Meister*[1], Khuong An Nguyen[1], Zhiyuan Luo[2]

[1]University of Brighton, [2]Royal Holloway, University of London

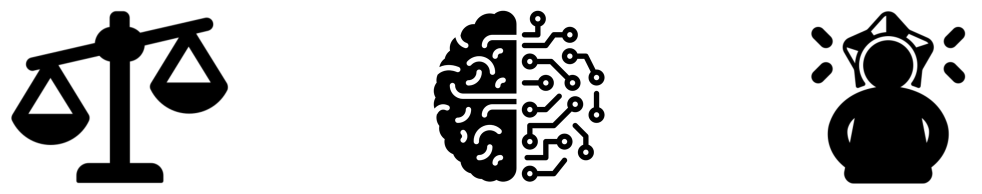
*J.Meister@brighton.ac.uk

Motivation

- COVID cough data is *heavily imbalanced*, and it is challenging to collect more samples.
- Therefore, *models are biased* and their *predictions cannot be trusted*.

→ ***We propose a confidence measure for COVID-19 cough classification.***

Challenges



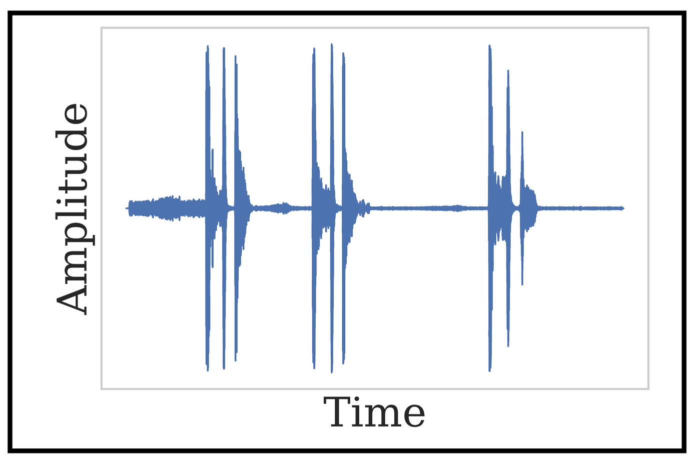
Imbalanced data | Biased models | Prediction confidence

1. COVID-19 cough classification

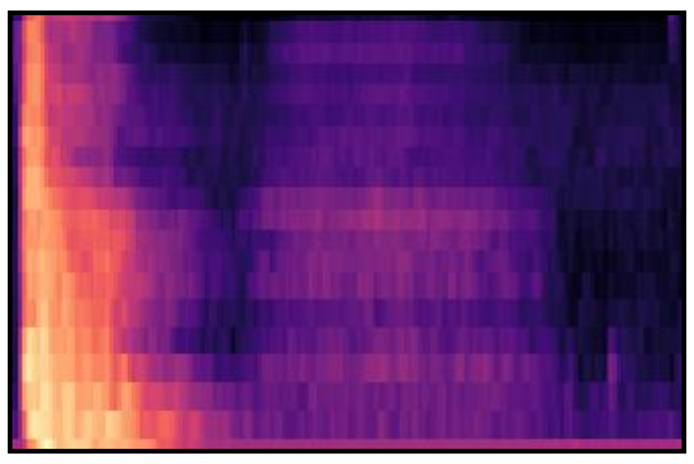
- *Cough classification* is an accessible, low-cost, and environmentally friendly COVID-19 screening alternative.
- Machine Learning models can *successfully identify COVID-19 and healthy coughs* from Mel frequency cepstral audio features [1].
- But we have to make the *risky assumption* that the models will on average perform as well on unseen data as on the training set.

→ ***How trustworthy are our COVID-19 predictions?***

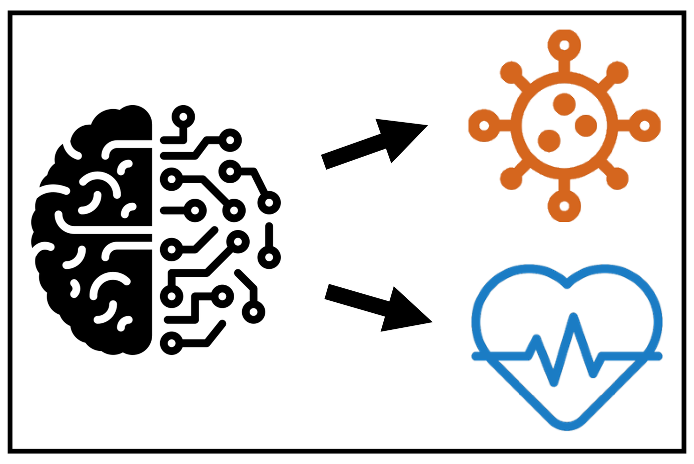
Cough recording



Audio processing



Feature extraction



Machine Learning prediction

2. Confident classification with Conformal Prediction (CP)

- CP acts as a wrapper for any Machine Learning model, transforming point predictions $y_j \in \boldsymbol{Y}$ into prediction sets $\boldsymbol{\Gamma_i^{1-\epsilon}}$ [2].

(i) $$\boldsymbol{\Gamma_i^{1-\epsilon}} = \{y_j \in \boldsymbol{Y} \mid \boldsymbol{p_i(y_j)} > \boldsymbol{\epsilon}\}$$

- CP *statistically guarantees total error rates* (i.e. the true label y_i^* is not included) up to a maximum, user-selected significance level $\boldsymbol{\epsilon}$.

(ii) $$\mathbb{P}\left(y_i^* \notin \boldsymbol{\Gamma_i^{1-\epsilon}}\right) \leq \boldsymbol{\epsilon}$$

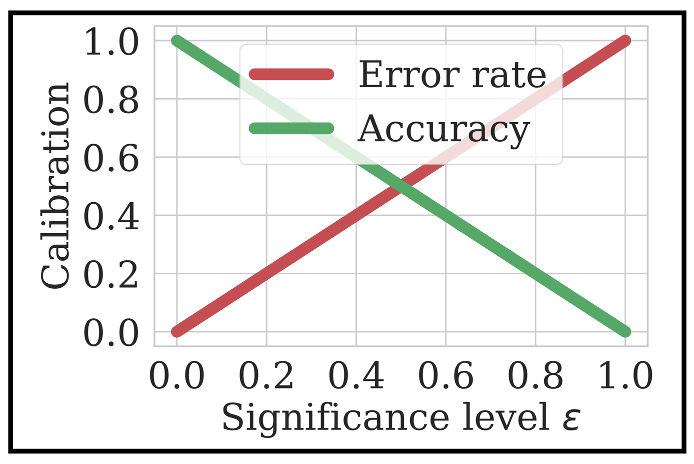
- P-values $\boldsymbol{p_{n+1}(y_j)}$ are derived from the sample $n+1$'s strangeness $\boldsymbol{\alpha_{n+1}^{y_j}}$ compared to n training samples for all possible labels $y_j \in \boldsymbol{Y}$.

(iii) $$\boldsymbol{p_{n+1}(y_j)} = \frac{\left|\left\{ i = 1, \dots, n : \alpha_i^{y_i^*} \leq \boldsymbol{\alpha_{n+1}^{y_j}} \right\}\right|}{n+1}$$

→ ***CP statistically measures the likelihood of a sample and its postulated label given the training data with guaranteed validity.***

Sample	Prediction	Confidence
#1	{COVID}	Certain
#2	{COVID, Healthy}	Uncertain
#3	{}	Outlier
#4	{Healthy}	Certain

CP prediction sets $\boldsymbol{\Gamma_i^{1-\epsilon}}$



Guaranteed validity

3. Adjusting for imbalanced data

- Mondrian CP extends guarantees to class-conditional validity.
- Mondrian CP assesses a sample $n+1$'s strangeness $\alpha_{n+1}^{y_j}$ compared to a class-conditioned training subset for each possible label $y_j \in \boldsymbol{Y}$.
- Adjusted Equations (ii) and (iii):

(ii*) $$\mathbb{P}\left(y_i^* \notin \Gamma_i^{1-\epsilon}\right) \leq \epsilon : \boldsymbol{y_i^* = y_j},\ y_j \in \boldsymbol{Y}$$

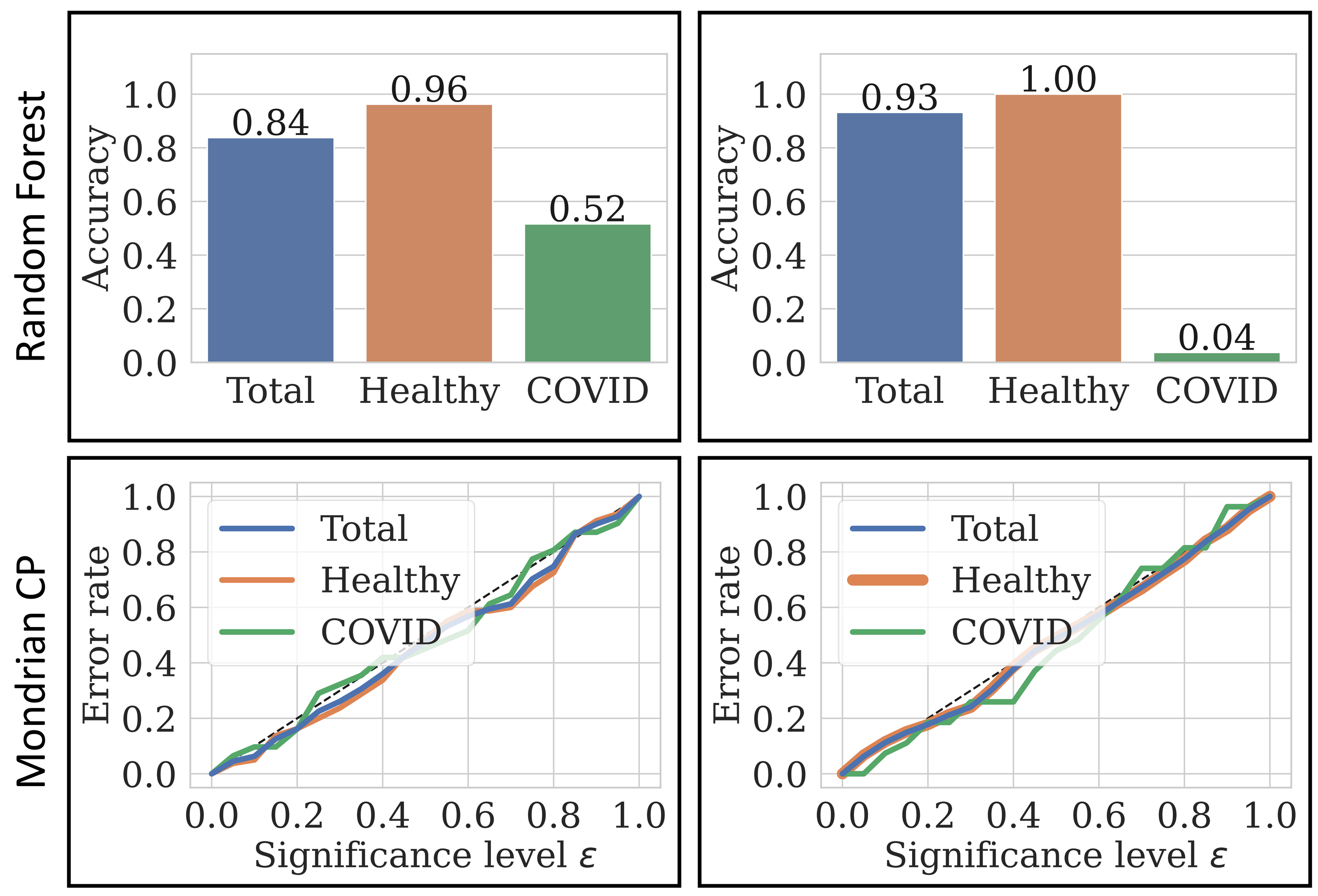
(iii*) $$p_{n+1}(y_j) = \frac{\left|\left\{ i = 1, \dots, n : \boldsymbol{y_i^* = y_j},\ \alpha_i^{y_i^*} \leq \alpha_{n+1}^{y_j} \right\}\right|}{\left|\left\{ i = 1, \dots, n : \boldsymbol{y_i^* = y_j} \right\}\right| + 1}$$

→ ***Mondrian CP guarantees that all class error rates are capped at ϵ regardless of imbalances, e.g. as in the selected two datasets.***

- *COVID-19 cough datasets*: Cambridge [3] and Coswara [4]

Samples	COVID	Healthy	Ratio	Σ
Cambridge	141 (29%)	346 (71%)	1:2.5	487 (100%)
Coswara	81 (7%)	1074 (93%)	1:13	1155 (100%)

4. Empirical results



→ ***CP guarantees a maximum error rate for each cough prediction.***
→ ***Mondrian CP has class-conditional validity and removes bias against the minority class, making each prediction trustworthy.***
→ ***Our proposed method is successful for COVID-19 cough classification, and may be transferred to other problems.***

References

[1] JA. Meister, et al. "*Audio feature ranking for sound-based COVID-19 patient detection.*" EPIA Conference. 2022.
[2] AE. Ashby, et al. "*Cough-based COVID-19 detection with audio quality clustering and confidence measure based learning.*" COPA conference. 2022.
[3] C. Brown, et al. "*Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data.*" SIGKDD. 2020.
[4] N. Sharma, et al. "*Coswara - A database of breathing, cough, and voice sounds for COVID-19 diagnosis.*" 2020.

Acknowledgements

This work is funded by Santander UK's Global Challenges Research grant.