Cough-based COVID-19 detection with audio quality clustering and confidence measure based learning

Alice E. Ashby A.Ashby1@uni.brighton.ac.uk
Julia A. Meister J.Meister@brighton.ac.uk
Khuong An Nguyen K.A.Nguyen@brighton.ac.uk
University of Brighton, East Sussex BN2 4GJ, United Kingdom

Zhiyuan Luo Zhiyuan.Luo@rhul.ac.uk
Royal Holloway University of London, Surrey, TW20 0EX, United Kingdom

Werner Gentzke Werner.Gentzke@cardis.io
Cardisio GmbH, 60549 Frankfurt am Main, Germany

Editor: Ulf Johansson, Henrik Boström, Khuong An Nguyen, Zhiyuan Luo and Lars Carlsson

Abstract

COVID-19 cough classification has rapidly become a promising research avenue as an accessible and low-cost screening alternative, needing only a smartphone to collect and process cough samples. However, audio processing of recordings collected in uncontrolled environments and prediction confidence are key challenges that need to be addressed before cough-screening could be widely accepted as a trusted testing method. Therefore, we propose a novel approach for cough event detection that identifies *cough clusters* instead of individual coughs, significantly reducing onset detection's usual hypersensitivity to energy fluctuations between cough phases. By using this technique to improve training sample quality and quantity by +200%, we improve Machine Learning performance on the minority COVID-19 class by up to 20%, achieving up to +47% precision and +15% recall compared to the dataset baseline. We propose a novel, class-agnostic Conformal Prediction non-conformity measure which takes the *cough sample quality* into account to counteract the variance caused by limiting segmentation to just the training set. Our Conformal Prediction model introduces uncertainty quantification to COVID-19 cough classification and achieves an additional 34% improvement to precision and recall.

Keywords: COVID-19 cough classification, cough event detection, audio segmentation.

1. Introduction

The unanticipated spread of the COVID-19 pandemic has set off a hunt for accessible, resource-light, and easy-to-distribute alternatives to standard medical tests. Audio classification quickly emerged as a potential alternative compared to more personnel and resource intensive tests (e.g., PCR, Lateral Flow Test), requiring only a smartphone to record respiratory sounds while generating no plastic waste. In particular, cough classification achieved promising results in detecting asymptomatic cases (Laguarta et al., 2020) and with forced coughs (Brown et al., 2020).

However, a ubiquitous challenge in time-series classification including cough is how to standardise the input with consistent dimensions for Machine Learning (ML). One of the most sophisticated approaches identifies non-overlapping, isolated cough segments for maximum sample quality with successful prediction results (Mohammed et al., 2021; Andreu-Perez et al., 2021). The difficulty lies in optimising the cough onset detection algorithm to identify the entire cough and avoid hypersensitivity to energy fluctuations within cough phases (Cohen-McFarlane et al., 2020).

Additionally, COVID-19 cough classification suffers under the same limitations as most ML predictive approaches: How certain are we that the model predictions are correct? To answer this question, the Conformal Prediction (CP) framework quantifies the uncertainty of the underlying ML model by statistically guaranteeing performance up to a user-selected error rate (Shafer and Vovk, 2008). Contrary to ML, CP outputs prediction sets, which means that optimisation is centred around reducing the set to 1 label, the most definitive class association (Alvarsson et al., 2021).

Therefore, this paper aims to improve binary COVID-19 classification by proposing and rigorously evaluating a novel cough cluster detection algorithm for audio segmentation, and developing a novel CP non-conformity measure that takes sample quality into account.

1.1. Our contributions

- We improved COVID-19 classification precision and recall by up to +47% and +15% compared to the published dataset baseline respectively, significantly narrowing the class-performance differences (Section 4.4). The results were achieved with a novel cough event detection approach which increases the quality and quantity of training samples (+200%, Section 2.2).
- We demonstrated the valuable confidence and performance benefits that Conformal Prediction provides for COVID-19 cough classification in a high-risk setting (+29% precision and recall of forced predictions, Section 4.5).
- Finally, we developed a novel, class-agnostic non-conformity measure for Conformal Prediction based on the sample quality (Section 3.2). By accounting for background noise, we improved forced predictive efficiency by an additional 6% on average across a range of metrics including precision and recall (Section 4.5).

2. Audio segmentation to isolate cough training samples

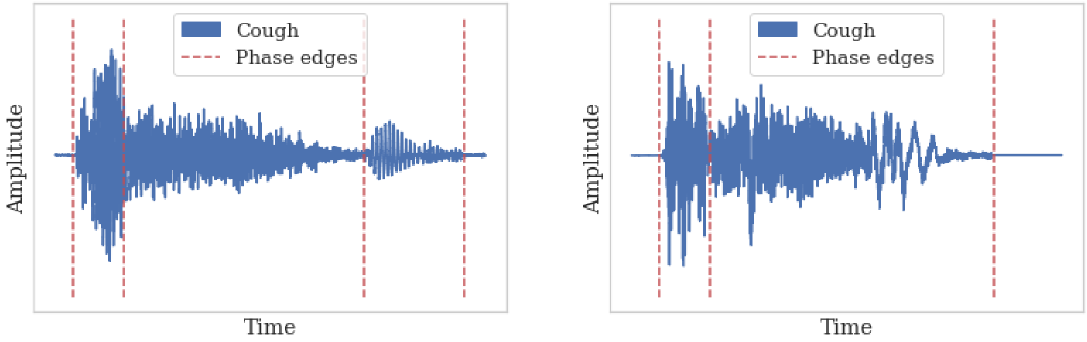
Sample quality is a ubiquitous Machine Learning challenge because noise can significantly skew prediction performance by obscuring class-differentiating information (Gupta and Gupta, 2019). In this paper, we use a COVID-19 cough dataset with volunteer-submitted recordings (Brown et al., 2020). Consequently, the samples have a large variation in duration (0.2–10 seconds), number of coughs per sample (1–15 individual coughs), and background noise levels.

Thus, we propose a procedure that segments the training audio by isolating cough intervals, which leads to a larger quantity of training samples and reduces sample variations caused by uncontrolled recording environments.

2.1. Cough event detection

To segment a time-series in a non-overlapping manner, we need to identify the starting point of events to isolate. The preferred automated approach are a class of algorithms called *Onset Detection*, which locates extreme changes in an audio signal (Cohen-McFarlane et al., 2019).

However, one of the difficulties in cough detection is identifying the event in its entirety rather than its individual phases. As shown in Figure 1(*a*), there may be a second cough energy peak that should not be split off (Cohen-McFarlane et al., 2020). In other cases, a cough may consist of only the first two phases, meaning that every energy peak is an individual cough (Figure 1(*b*)). To address the challenge of handling both 2-phase and 3-phase coughs correctly, we propose detecting *cough clusters*, i.e. a series of coughs with only brief intervals between them. This ensures that samples are not split into too many extremely short single-phase cough segments ($\leq$0.1s), while maintaining the removal of extended non-cough intervals that would skew audio features.



(*a*) 3-phase cough.

(*b*) 2-phase cough.

Figure 1: Different inter-phase energy fluctuations make isolating individual coughs in their entirety challenging. Since coughs may have varying phase configurations with different durations, setting a minimum threshold before the next onset detection performed poorly in our tests.

2.2. Proposed segmentation procedure

For cough onset detection, we chose Root Mean Square Energy (RMSE) as the underlying feature as it may be interpreted as a signal's loudness (Meister et al., 2021), an intuitive indicator for a cough event with low to medium background noise levels. The audio feature describes a signal's energy as a function of its amplitude x over N frames.

$$RMSE = \sqrt{\tfrac{1}{N} \textstyle\sum_{n=1}^{N} x_n^2} \tag{1}$$

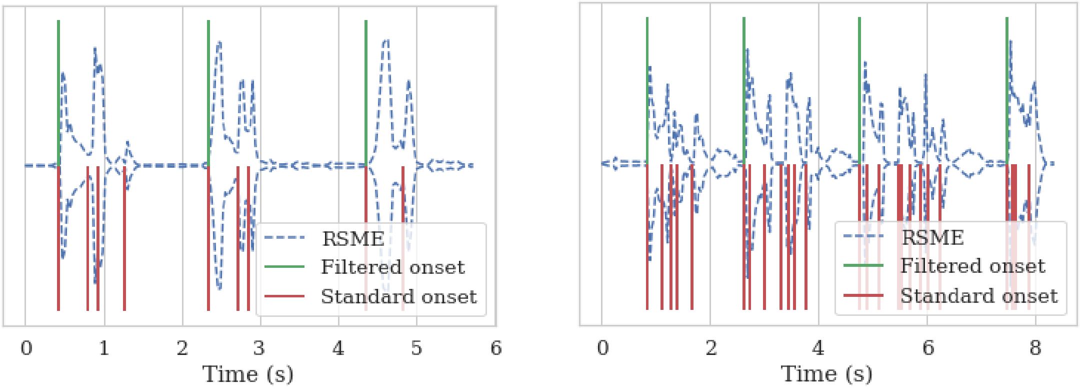
Given RMSE, the Python audio-processing package `librosa` identifies any extreme changes in the signal's energy levels. To reduce the function's hypersensitivity to different cough phases and their energy fluctuations (Figure 1), we propose a novel onset filtering technique that identifies *cough clusters* instead of attempting to find individual coughs. Algorithm 1 gives a pseudo-code overview of the proposed onset filtering which produces meaningful, non-overlapping audio segments.

In contrast to the common method of setting a minimum distance before the next onset is detected (which performed poorly in our test because of the varying duration

Algorithm 1: Cough onset detection and cough cluster filtering.

```
Input: $sample, threshold$
Output: $onset\_frames\_filtered$, the onsets of cough clusters
$rmse\_per\_frame \leftarrow$ calculate_rmse($sample$)
$onset\_frames \leftarrow$ get_energy_onsets($rmse\_per\_frame$)
$onset\_frames\_filtered \leftarrow$ list([$onset\_frames[0]$])
for $i \leftarrow 1$ to $onset\_frames.length$ do
    if $(onset\_frames[i] - onset\_frames[i-1]) \geq threshold$ then
        $onset\_frames\_filtered$.append($onset\_frames[i]$)
    end
end
```

of a cough), our novel approach successfully identifies cough clusters, and avoids over-segmenting audio clips between cough phases. The novelty and improved performance was achieved by the inclusion of a `for-loop` in Algorithm 1, which evaluates whether detected coughs are the start of a cough cluster. By checking whether the current detected cough (`onset_frames[i]`) is a minimum distance from the previous cough (`onset_frames[i-1]`), we effectively filter only for coughs that are the first in a rapid series of coughs, i.e. a cough cluster (`onset_frames_filtered`). The results are presented in Figure 2.



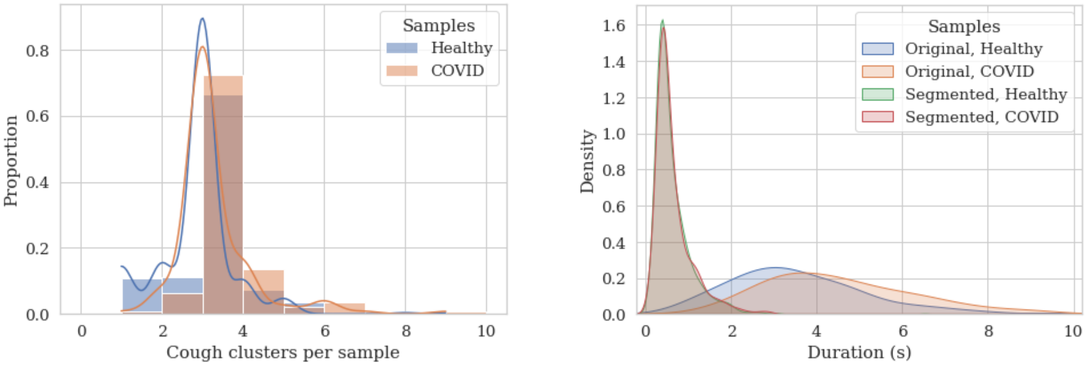
(a) Three cough clusters, low noise.

(b) Four cough clusters, medium noise.

Figure 2: The onsets detected by our proposed filtering method successfully identified meaningful cough clusters. In comparison, standard onset detection was hypersensitive to cough-phase energy fluctuations and noise, over-segmenting the audio.

The segmented audio samples were visually verified and manually spot-checked to ensure the quality of the proposed method. An overview of the final statistics is presented in Figure 3. Three cough clusters were identified in most original samples as expected, and the number of individual coughs within a cluster ranged from 1-5. Assessing the distribution of sample durations, we see that the segmentation has drastically reduced the sample lengths, with the majority just under 1 second.

Promisingly, this has also reduced a cough-external difference (i.e. silent sample recording lengths) between COVID-19 and Healthy samples. Before segmentation, COVID-19 samples tended to have longer recordings while maintaining a similar number of cough clusters. After isolation, the class duration distributions aligned almost perfectly. From these results, we hypothesise an improvement in classification performance and a narrowing of the performance gap between classes because cough isolation has reduced spurious background noise sample characteristics. A detailed results analysis in Section 4.4 confirms the effectiveness of our segmentation method.



(a) Detected cough clusters.

(b) Duration distributions.

Figure 3: Statistics of the cough clusters isolated by our proposed onset detection algorithm. The number of isolated clusters varied but was centred on three as expected. Promisingly, cough isolation removed the spurious association that COVID-19 samples have a longer duration with the same number of coughs.

3. A confidence measure for COVID-19 cough classification

The cough isolation pre-processing described above is limited to the training samples to increase the size of the training set. Therefore, we propose a novel Conformal Prediction non-conformity measure that takes the audio sample's quality into account to address the newly introduced differences between the training and test sets.

3.1. Inductive Conformal Prediction

Conformal Prediction (CP) is a Machine Learning framework based on hypothesis testing that could be combined with any underlying algorithm to quantify the certainty of individual predictions. It has been used in high-risk settings such as pharmaceutical compound prediction because it statistically guarantees a user-specified maximum error rate under the minimal i.i.d. assumption (Alvarsson et al., 2021). This section gives a brief overview of the concepts that are relevant for the paper, specifically focusing on the computationally efficient Inductive Conformal Prediction (ICP) variant for classification. Interested readers may refer to Shafer and Vovk (2008) for a more detailed treatment of the framework and its context.

ICP achieves a guaranteed error rate for a given significance level ϵ by outputting a prediction set $\Gamma^{1-\epsilon}$ containing all plausible sample labels. The validity property ensures that the following equation holds:

$$Pr(y_i^* \notin \Gamma_i^{1-\epsilon}) \leq \epsilon \tag{2}$$

In other words, the ICP predictor makes a mistake (sample x_i's true label y_i^* is not included in the ϵ-conditioned prediction set) with maximum probability up to ϵ, subject to statistical fluctuations (Shafer and Vovk, 2008).

An ICP predictor is defined by its non-conformity measure (NCM), which scores a sample-label pair's non-conformity α_i^y compared to the known data. Each test sample x_i is extended to $z_i^y = (x_i, y)$ with every possible label in the label space $y \in Y$, and evaluated against the training proper set $(z_1, ..., z_n)$. Given a calibration set $(z_{n+1}, ..., z_m)$ with NCM scores $(\alpha_{n+1}^{y^*}, ..., \alpha_m^{y^*})$ and a test sample x_{m+1} with NCM scores $\alpha_{m+1}^y, y \in Y$, the test sample's class-conditional p-values p_{m+1} are calculated with:

$$p_{m+1}(y) = \frac{|\{j = n+1, ..., m : \alpha_j^{y^*} \leq \alpha_{m+1}^y\}| + 1}{|\{j = n+1, ..., m : y_j^*\}| + 1} \tag{3}$$

From the p-values, we generate an ϵ-conditioned prediction set $\Gamma^{1-\epsilon}$ containing all probable labels:

$$\Gamma^{1-\epsilon} = \{y \in Y | p_{m+1}(y) > \epsilon\} \tag{4}$$

A variant on ICP is the Mondrian Inductive Conformal Predictor (MICP) which extends the validity guarantee to label-conditional maximum error rates, even in cases where there is a large class performance gap in the underlying algorithm. This stricter validity is achieved by adjusting the p-value calculation in Equation (3) to:

$$p_{m+1}(y) = \frac{|\{j = n+1, ..., m : y_j^* = y, \alpha_j^{y^*} \leq \alpha_{m+1}^y\}| + 1}{|\{j = n+1, ..., m : y_j^* = y\}| + 1} \tag{5}$$

Because the error rates are guaranteed, ICP optimisation is centred on maximising predictive efficiency, measured as a function of set size. Improvements are often achieved by fine-tuning the NCM function (Alvarsson et al., 2021), which will be discussed in more detail in the next section.

3.2. A class-agnostic Conformal Prediction non-conformity measure

Optimising a non-conformity measure (NCM) will significantly improve CP predictive efficiency (i.e. set sizes). We chose Inverse Probability Function (IPF) as our baseline to build on since it is a common and versatile NCM function that could be used in combination with any probabilistic underlying algorithm. IPF is calculated as the inverse of a model's prediction on a sample x with the tentative class y_i.

$$NCM_{IPF} = 1 - P(y_i|x) \tag{6}$$

Improvement of the training set quality and quantity through cough cluster isolation (Section 2.2) has introduced variation in the training and test set distribution. We minimised the effects with a carefully designed NCM that takes sample quality, i.e. background noise, into account. Normalised NCM scores are regularly used to improve Conformal Prediction (CP) efficiency (Wisniewski et al., 2020) but to the best of our knowledge, the normalising factor is generally a Machine Learning model's accuracy estimate of the test sample. This implies that the normalising score will be affected by the same spurious class associations that already decreases CP's predictive performance (e.g., class imbalance).

Instead, we propose Signal-to-Noise Ratio (SNR) as a class-agnostic normalising factor for NCM. SNR measures the average power of the background noise and the average power of the foreground sound (Monge-Álvarez et al., 2019), in our case, coughs. Since the difference introduced between the training and test samples was a shifted cough vs non-cough ratio, SNR provided a convenient metric for sample quality. SNR is often measured on a logarithmic decibel scale because the signals tend to have a wide dynamic range.

The proposed SNR-based NCM is defined in Equation (7). The higher SNR is, the closer the test sample is to the isolated training coughs, and the smaller the NCM will be, i.e., the sample is conforming.

$$NCM_{SNR} = \frac{NCM_{IPF}}{SNR_{dB}} \tag{7}$$

$$SNR_{dB} = 20 \cdot \log_{10} SNR \tag{8}$$

$$SNR = \left(\frac{amplitude_{signal}}{amplitude_{noise}}\right)^2 \tag{9}$$

Figure 4 illustrates the distribution of SNR scores before and after cough isolation. Looking at the original audio, the scores were concentrated near 0, because of a higher ratio of non-cough intervals in the samples. In contrast, the curve became much shallower with a wider range once the samples were segmented. The sample quality was much improved by removing the majority of non-cough background information, and consequently audio features were less skewed. Additionally, COVID-19 and Healthy scores overlapped much more cleanly, confirming that cough isolation diminished the effects of spurious, class-obscuring characteristics such as sample length, which in turn could improve prediction performance (see Section 4.4).

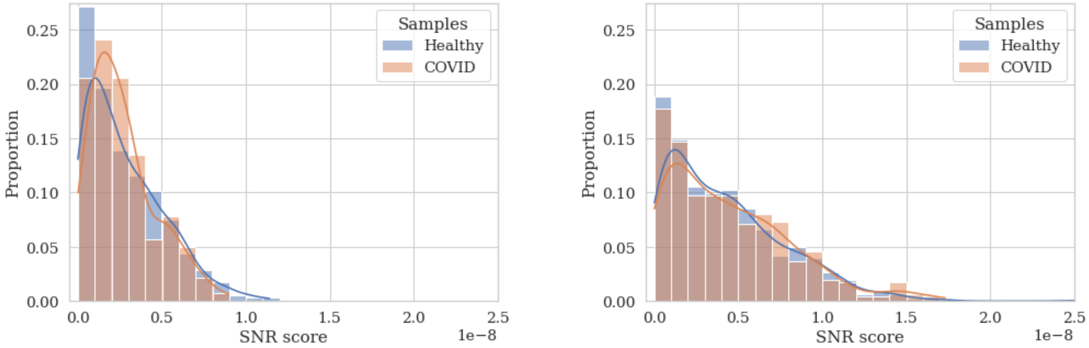
4. Experimental results

This section empirically evaluates the proposed methods in the context of COVID-19 binary cough classification.

4.1. Research questions

The overarching goal was to improve COVID-19 classification performance. We define three research questions to focus our results analysis and discussion in the following sections:

- Does increasing the quality and quantity of training samples through cough event detection and audio segmentation improve COVID-19 classification results?



(a) Original audio.

(b) Cough-segmented audio.

Figure 4: Signal-to-noise ratio (SNR) of the audio samples measures the ratio of cough to noise intervals. The higher the score, the higher the quality of the sample. Cough isolation successfully reduced non-cough, class-confounding effects.

- What valuable new insights does Conformal Prediction give us for COVID-19 cough classification?
- Can we improve Conformal Prediction's performance by quantifying the quality of a particular sample and incorporating this information into the non-conformity measure?

4.2. Data and feature engineering

A curated version of the COVID-19 audio dataset described in Brown et al. (2020) was used for the experiments because of its large sample count compared to other COVID-19 options. The dataset contains cough samples recorded by volunteers in April and May 2020 submitted through an Android app and a website, along with their COVID-19 status (positive or negative). The dataset includes 487 cough samples, of which 29% are COVID-19 positive and 71% are Healthy, i.e. COVID-negative and no other cough symptoms. Once the samples were split into the train and test sets (2:1, stratified by label), we used our proposed method for isolating cough events from Section 2.2 to increase the number of training samples by almost 200% (see Table 1).

The original samples are encoded as WAV files at 22kHz with varying sample lengths and around three cough clusters each. Before feature extraction, samples were standardised by scaling the amplitudes to $(-1, 1)$, and trimming leading and trailing sections under 10dB to remove as much non-cough data as possible. Out of 15 widely-used audio features, previous analysis of the dataset found MFCCs (Mel-frequency cepstral coefficients) to be the most differentiating feature for COVID-19 cough classification, reaching 83.25% ROC-AUC (Meister et al., 2021). MFCCs describe a signal's power spectrum (Brown et al., 2020) and may be interpreted as the audio's tonal quality. Because MFCCs are calculated as a time series, we took the mean to account for different audio lengths. To avoid overfitting models on a small dataset, we identified `MFCC_12` as the most impactful coefficient for both classes

Table 1: COVID-19 cough sample counts and statistics. Samples were stratified and split 70:30% into the training and test sets. Cough segmentation (Section 2.2) increased samples by almost 200%, but maintained the 2:1 imbalanced class ratio.

Dataset	**COVID-19**	**Healthy**	**Total**
Original	141 (29%)	346 (71%)	487 (100%)
Segmented	462 (32%)	986 (68%)	1448 (100%)
Increase	321 (185%)	640 (228%)	961 (197%)

by a wide margin through SHAP analysis, shown in Figure 5. SHAP (SHaply Additive exPlanations) is a model agnostic alternative to the classic feature selection techniques (Marcílio and Eler, 2020).

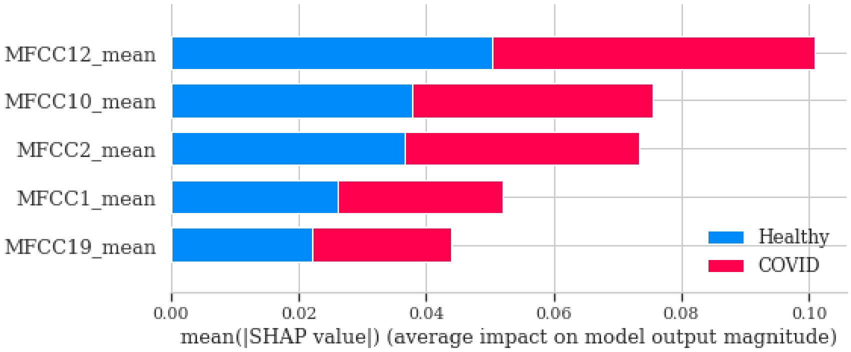


Figure 5: SHAP feature importance for feature selection. We chose MFCC12 because it had the highest impact on both COVID-19 and Healthy classes.

4.3. Evaluation metrics

Standard Machine Learning (ML) and Conformal Prediction (CP) require unique metrics to evaluate their performance because they output predictions in different formats. An overview of the considered metrics is available in Table 2.

For ML, our focus lies on maximising accurate point predictions while taking the imbalanced dataset into account. Consequently, we focus on metrics such as accuracy, precision, recall, and F1-score. In addition, CP has a guaranteed level of error while outputting prediction sets, which encourages measuring efficiency as a function of set sizes. Optimally, each prediction set has exactly one element.

Table 2: Evaluation metrics for Machine Learning (ML) and Conformal Prediction (CP). T and F stand for True and False, P and N for Positive and Negative. n is the number of test samples, Γ the CP prediction set, and y_i^* is sample i's true label.

Metric	**Formula**	**Intuition**
Accuracy	$ACC = \frac{TP+TN}{TP+FP+TN+FN}$	Correct predictions, both classes.
Precision	$PR = \frac{TP}{TP+FP}$	Correct COVID-19 predictions.
Recall	$REC = \frac{TP}{TP+FN}$	COVID-19 samples correctly identified.
F1-score	$F1 = 2 \cdot \frac{PR \cdot REC}{PR+REC}$	Accuracy taking imbalance into account.
Error rate	$(\sum_{i=1}^{n} y_i^* \notin \Gamma_i)/n$	True label not in prediction set.
Empty rate	$(\sum_{i=1}^{n} \lvert\Gamma_i\rvert = 0)/n$	Outlier samples.
Multi rate	$(\sum_{i=1}^{n} \lvert\Gamma_i\rvert > 1)/n$	Samples on the class boundary.
Single rate	$(\sum_{i=1}^{n} \lvert\Gamma_i\rvert = 1)/n$	Efficient predictions.
True single rate	$(\sum_{i=1}^{n} \lvert\Gamma_i\rvert = 1 \wedge y_i^* \in \Gamma_i)/n$	Correct efficient predictions.

4.4. Experiment 1: Training models with isolated coughs

In keeping with the research questions laid out in Section 4.1, we are particularly interested in exploring to what extent training models on isolated coughs could improve COVID-19 cough classification performance. Therefore, we compared two scenarios:

- *Scenario 1*: Models were trained and tested on the original dataset. All samples were full user-submitted recordings.
- *Scenario 2*: Models were trained on the segmented set and tested on the original set. Train samples were isolated cough clusters as identified by our automated audio segmentation procedure (Section 2.2). Test samples were full user-submitted recordings.

In both scenarios, we employed four ML algorithms to evaluate the model-agnostic trends: K-Nearest Neighbours (KNN), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR). Each model was optimised with grid-search hyperparameter tuning, and all results were generated with 5-fold Stratified Cross Validation (CV) to increase their stability. In Scenario 2, folds were assigned on the origin-level to ensure that all isolated coughs from one submitted recording were contained in the same fold.

Scenario 1 results. Table 3(*a*) provides an overview of the ML results in Scenario 1, with models reaching up to 70% accuracy on the original dataset. As expected from class imbalance (2:1, Figure 1), COVID-19 samples were especially difficult to predict correctly. Their accuracy was between -41% (RF) and -78% (SVM) lower than the Healthy-class equivalent for all models, leading to low precision and recall (29% and 16% on average).

Scenario 2 results. The results summary in Table 3(*b*) shows only marginal overall accuracy improvements up to 72% with the isolated training coughs, but it is well-established

Table 3: 5-fold CV test results to measure the benefit of training models on isolated cough clusters (improved sample quality and quantity). Performance on the original training set was poor for the minority COVID-19 samples. A segmented training set significantly narrowed the performance gap between classes.

(*a*) Original cough training set.

Metric (%)	**KNN**	**RF**	**SVC**	**LR**
Acc [COVID]	26.78	31.93	3.81	13.77
Acc [Healthy]	77.71	73.05	81.85	**82.42**
Accuracy	65.92	61.41	69.62	70.84
F1	52.24	52.49	42.83	48.10
Precision	35.62	32.74	5.71	42.22
Recall	22.07	31.26	2.86	8.52

(*b*) Isolated cough training set.

Metric (%)	**KNN**	**RF**	**SVC**	**LR**
Acc [COVID]	36.96	**39.11**	24.88	26.65
Acc [Healthy]	77.55	73.80	82.17	**82.42**
Accuracy	66.95	63.46	71.26	**71.67**
F1	**57.25**	56.45	53.53	54.53
Precision	40.91	38.41	52.86	**57.00**
Recall	34.09	**40.44**	17.02	17.76

that raw accuracy is biased with imbalanced data (Kotsiantis et al., 2006). Assessing better-suited metrics like class-specific accuracy, precision, recall, and F1-score, we see the true effects of segmenting the training set: There was significant improvement in COVID-19 predictions across the board, with all four models. This is demonstrated more clearly in Figure 6, where we plotted the difference in all metrics' means between Scenarios 1 and 2. Not only did we achieve significant improvement in terms of COVID-19 prediction performance, we also confirmed that there was virtually no deterioration in the models' accuracy on Healthy samples.

SVM and LR achieved the largest improvements in Scenario 2 across most metrics, e.g. +47% and +15% precision respectively. However, RF was identified as the best model overall, since we considered good performance on COVID-19 samples as the deciding factor. RF had both the highest COVID-19 accuracy (39%) and recall (40%) of all examined models. Consequently, the F1-score (56%) was also high, second by only $< 1\%$ (KNN).

4.5. Experiment 2: Conformal Prediction with a novel non-conformity measure

After identifying the best combination of audio formatting (isolated training coughs and full-recording test samples) and the underlying ML model (Random Forest) in Section 4.4, we developed and assessed Conformal Prediction (CP) for COVID-19 cough classification. A novel class-agnostic non-conformity measure (NCM) which takes the sample quality into account further improved predictive performance by narrowing the gap between classes. As explained in Section 3.2, we measured audio quality with the Signal-to-Noise Ratio (SNR), which scores a sample's cough vs non-cough proportions. Segmented training samples had much higher SNR scores since they contain an isolated cough cluster. We used the SNR score to normalise the well-known Inverse Probability Function (IPF-NCM), creating SNR-NCM. The CP models were implemented in the `nonconformist` package.

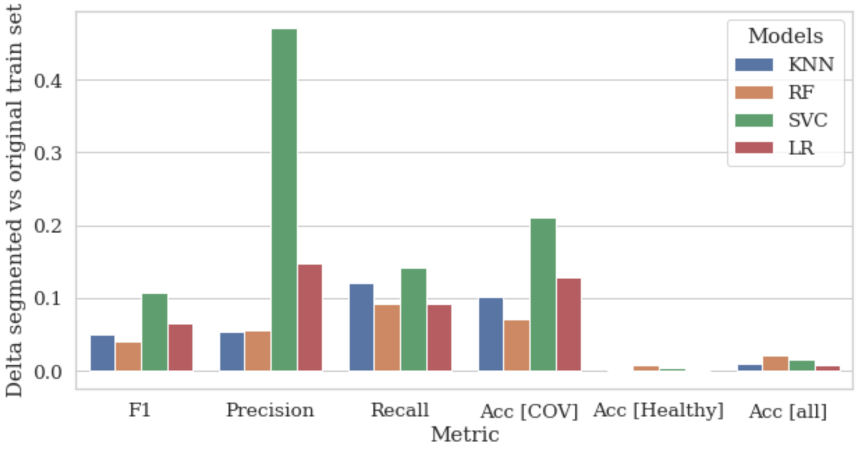


Figure 6: Effects of segmenting the training samples. There were significant improvements across all COVID-19 related metrics, even though the dataset's relative class imbalance was unchanged (2:1). Importantly, minority class benefits did not come at the cost of Healthy prediction performance, which stayed virtually unchanged.

The Inductive (ICP) and Mondrian Inductive (MICP) variants of Conformal Prediction require three datasets: the training proper (45%), calibration (25%), and test sets (30%). To ensure that the validity guarantee is maintained, the calibration and test sets have to be i.i.d. and therefore both sets were comprised of original full-recording samples. Only samples from the training proper set were run through the cough-isolation procedure proposed in Section 2.2, which increased the number of training samples by 200% and ensured that all coughs from the same original recording were contained in one subset.

Forced predictions. While prediction sets are usually the CP output of choice because they guarantee maximum error rates, we may force point predictions (class with the highest p-value) for a more direct equivalent to ML results. Comparing the performance of the simple Random Forest (RF) algorithm with ICP results in Figure 7, it becomes immediately clear that ICP is the superior COVID-19 cough classification model across a range of metrics, most of which take the dataset's 2:1 imbalance into account. ICP-IPF already improved on RF significantly by 14% on average, and ICP-SNR added a further 2-8% across all metrics, including class-specific accuracy. ICP-IPF/ICP-SNR particularly improved on precision and recall (+29/34%), high-priority metrics in disease classification problems. The clear performance improvements of the point predictions were very promising, but one of the main draws of CP is that the prediction set approach allows for a more nuanced interpretation of the predictions and quantify their uncertainty, which is what we will focus on in the following paragraphs.

ICP results. Inspecting ICP's prediction set results in Figure 8, we see that the overall error rates were well-calibrated for both the IPF- and SNR-NCMs (non-conformity measures), subject to statistical fluctuations. While the class imbalance remained prominent

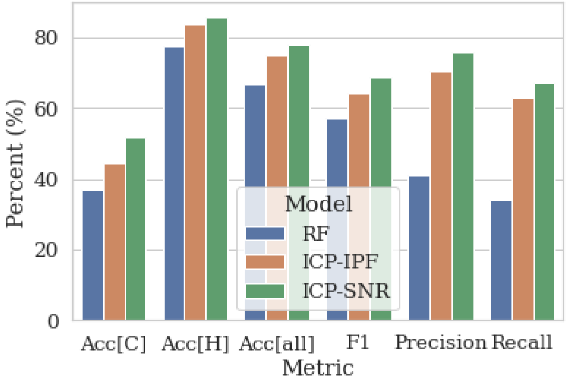
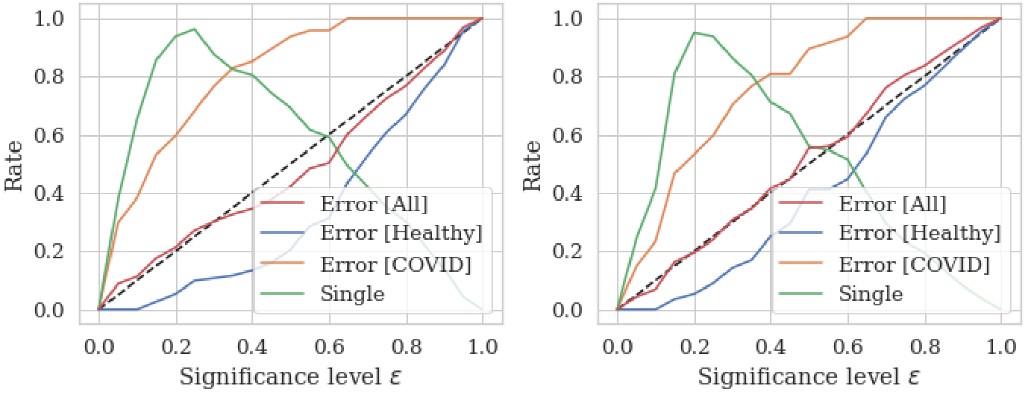


Figure 7: ICP forced predictions compared against RF for COVID-19 classification. ICP-IPF significantly improved all metrics (COVID-19 and Healthy), especially high-priority measures precision and recall up to 29%. ICP-SNR improved all metrics by an additional 6% on average.



(a) ICP-IPF test results.

(b) ICP-SNR test results.

Figure 8: ICP test results showcase the benefit of normalising the NCM with Signal-to-Noise Ratio (SNR). The class-discrepancy gap between COVID-19 and Healthy was visibly narrowed by incorporating sample quality into the NCM, even though training cough isolation had already significantly reduced non-cough class differences (Section 4.4).

in both cases (COVID-19 error rate was higher and Healthy error rate lower than the calibration line), normalising the NCM score with SNR significantly decreased COVID-19 error, narrowing the gap between the classes' error curves. These observations confirm that there were still noticeable non-cough class-discrepancies that could be addressed by incorporating sample quality into the NCM, even though training cough isolation had already significantly reduced spurious feature skewing factors like the background segments between coughs (Section 4.4).

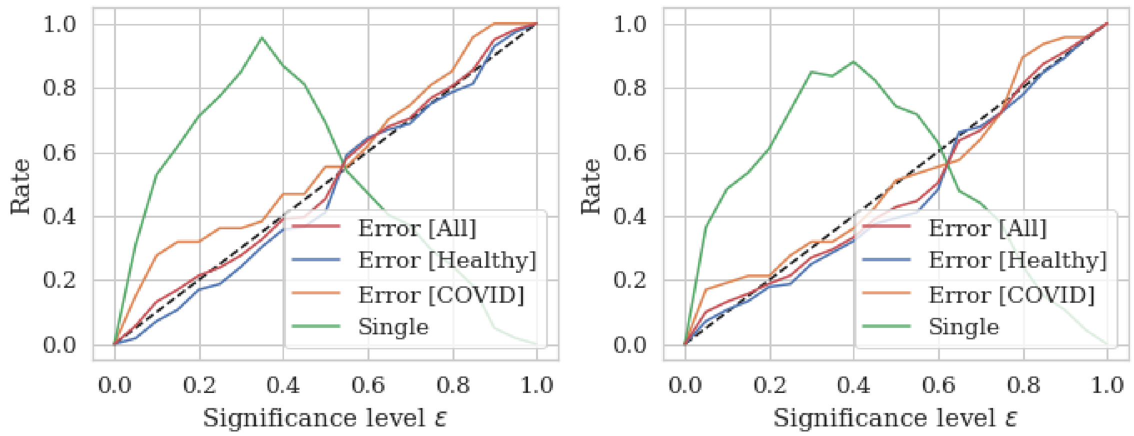
Analysing more precise results in Table 4 for the 0.1-0.3 significance range, we observe that IPC-SNR maintained the single-set rate overall. While the rate of multi-sets increased with lower significance, this was offset by an increase in the COVID-19 true single set rate (i.e. single sets containing the true label, optimal prediction) by 6% at the $\epsilon = 0.2, 0.3$ significance levels. On the other hand, the rate of empty sets stayed virtually unchanged, from which we conclude that ICP-SNR successfully identified and penalised difficult samples. In particular for the minority COVID-19 class, predictions tended to become multi-sets rather than an erroneous single-set prediction.

Table 4: ϵ-conditioned ICP results highlight the benefits of SNR normalisation on the minority COVID-19 class in particular. SNR successfully identified and penalised difficult samples to improve ICP performance.

		(*a*) ICP-IPF test results.			(*b*) ICP-SNR test results.		
Samples	**Rate (%)**	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$
All	Error	11.32	21.38	30.19	6.92	19.5	30.82
	Empty	0.00	0.00	12.58	0.00	0.00	13.84
	Multi	34.59	6.29	0.00	58.49	5.03	0.00
	Single	65.41	93.71	87.42	41.51	94.97	86.16
	True single	54.09	72.33	69.81	34.59	75.47	69.18
COVID	Error	38.3	59.57	76.60	23.40	53.19	70.21
	Empty	0.00	0.00	23.40	0.00	0.00	23.40
	Multi	55.32	10.64	0.00	76.60	10.64	0.00
	Single	44.68	89.36	76.60	23.40	89.36	76.60
	True single	6.38	29.79	23.40	0.00	36.17	29.79

MICP results. Turning our attention to the MICP results in Figure 9, we note that both models were close to well-calibrated on a class-level, although COVID-19 samples still had a noticeable disadvantage at low significance ($\epsilon \leq 0.5$). Somewhat surprisingly given MICP's label-conditional validity, ICP-SNR still improved on and once again visibly narrowed the gap between the class error rates. One of MICPs inherent downsides is larger average set sizes as a consequence of the stricter error guarantees. We see this implicitly in the much more shallow and right-shifted single-rate curves in our results.

Even though the single rates tended to drop with MICP-SNR compared to MICP-IPF at low significance, COVID-19 error rates tended to decrease by about 10% as well with an



(*a*) MICP-IPF test results.

(*b*) MICP-SNR test results.

Figure 9: MICP results confirm that normalising IPF with SNR still narrowed class performance discrepancies, even with MCIP's stricter label-conditional validity.

inverse increase in multi rates (Table 5). Interestingly though, the rate of true single sets remained the same or slightly increased, which implies that the MICP-IPF single sets that were changed to multi-sets by MICP-SNR were errors with the true class not included. We infer that MICP-SNR successfully identified difficult, regularly misclassified samples and increased their prediction set size to reduce errors, especially for the minority COVID-19 class. This confirms our ICP-SNR results, and underlines that incorporating sample quality via SNR improved performance both overall and in particular for COVID-19 samples.

4.6. Summary of results

Overall, rigorous empirical evaluation of our proposed methods for COVID-19 cough classification showed a significant performance boost for both ML and CP, which narrowed the performance gap between the 2:1 imbalanced classes.

Experimental results in Section 4.4 confirmed that training ML models on isolated cough clusters (Scenario 2) showed significantly higher performance and improved model generalisation capacity for the COVID-19 class, while maintaining high performance on the majority Healthy class. Performance gains from our segmentation procedure reached up to +47% recall and +15% precision compared to the dataset baseline, with an overall accuracy of up to 73%. The improvements could be traced back to an number of causes, including:

- Training sample quality improvement by removing inter-cough segments for better model generalisation,
- Boosted training set numbers (200% increase), even though the relative ratio of classes remains the same (2:1),

Table 5: ϵ-conditioned MICP results confirm previous observations that SNR normalisation improves results overall and in particular for COVID-19 samples. Difficult samples were penalised, increasing their set sizes to decrease error rates.

		(a) MICP-IPF test results.			(b) MICP-SNR test results.		
Samples	**Rate (%)**	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$
All	Error	13.21	21.38	27.67	13.21	18.87	27.04
	Empty	0.00	0.00	0.00	0.00	0.00	0.00
	Multi	47.17	28.93	15.09	51.57	38.99	15.09
	Single	52.83	71.07	84.91	48.43	61.01	84.91
	True single	39.62	49.69	57.23	35.22	42.14	57.86
COVID	Error	27.66	31.91	36.17	19.15	21.28	31.91
	Empty	0.00	0.00	0.00	0.00	0.00	0.00
	Multi	25.53	14.89	8.51	31.91	23.40	6.38
	Single	74.47	85.11	92.49	68.09	76.60	93.62
	True single	46.81	53.19	55.32	48.94	55.32	6.17

- And the removal of class differences encoded in the sample outside of coughs, e.g. recording length and number of coughs.

The downside of training cough isolation is that we introduced variance in the sample distribution between the train and test sets, which we addressed with a novel Conformal Prediction (CP) non-conformity measure (NCM) normalised with Signal-to-Noise Ratio (SNR). SNR describes the audio quality (cough vs non-cough ratio) and served as an estimate for sample difficulty.

Both Inductive (ICP) and Mondrian Inductive Conformal Prediction (MICP) improved classification performance drastically compared to the simple ML models with both the Inverse Probability Function and SNR (IPF: +29%, SNR: +35% precision and recall with forced predictions, Section 4.5). Prediction sets allowed a much more nuanced interpretation of the results, which confirmed that ICP- and MICP-SNR both drastically reduced the class-performance gap by penalising difficult samples, changing their predictions from (false) single sets to multi-sets, limiting errors.

In summary, our results confirm that both training cough isolation and NCM scores normalised by sample quality (cough vs non-cough ratio) significantly improved COVID-19 classification performance by reducing the effects of confounding factors like sample length and number of coughs. The best results were achieved when the two methods were combined, especially for the COVID-19 minority class.

5. Related work

Computerised respiratory sound classification has a long history supported by the development of sophisticated signal capture and processing technology. Compared to assessing lung sounds manually, automated analysis and prediction techniques introduce less subjectivity (Rizal et al., 2015). Since the onset of the COVID-19 pandemic in 2019, many parties have collected and published COVID-19 audio datasets containing a range of respiratory signals such as breath, cough, and speech recordings (Brown et al., 2020; Sharma et al., 2020; Schuller et al., 2021). The intuition to gather a variety of lung sounds could be drawn back to COVID-19 as a disease that physically affects the lungs, and therefore causes distinct turbulence in all air flows, even in asymptomatic cases (Laguarta et al., 2020). The relevance of COVID-19 respiratory classification is underlined by the 2021 prestigious INTERSPEECH challenge to identify COVID-19 infection from cough and breath samples (Deshpande and Schuller, 2021).

Because coughing is an extremely common symptom for many diseases, a rich literature is available for the classification of non-COVID respiratory diseases. Knowledge transfer through Transfer Learning has been very promising for COVID-19 classification (Pahar et al., 2022), especially because it mitigates the lack of large, high-quality COVID-19 datasets. Additionally, adjusting non-respiratory disease classification research focused on identifying early Alzheimer's from speech has also proven successful (Laguarta et al., 2020).

One of the primary challenges of respiratory classification with Machine Learning (ML) is how to approach the varying length of time-series samples. There are multiple ways to automatically standardise the sample lengths, for example by trimming/padding samples, although this could add uninformative data or even remove important segments (Mohammed et al., 2021). In comparison, using a sliding window to create overlapping segments retains all of the information, but could create uninformative chunks containing only partial information if the parameters are set incorrectly (Andreu-Perez et al., 2021). Finally, the most sophisticated but also most challenging approach is to identify informative, non-overlapping segments that each contain one audio event, e.g. a single cough (Mohammed et al., 2021; Andreu-Perez et al., 2021). To avoid hyper-segmentation of the audio, the respiratory onset detection algorithm has to account for energy fluctuations in the respiratory signals, e.g. different phases in a cough (Cohen-McFarlane et al., 2020).

As in other high-risk settings, the uptake of ML COVID-19 screening techniques could benefit from integrated uncertainty quantification. Conformal Prediction (CP) is a confidence framework rooted in hypothesis testing which provides guaranteed performance for each individual prediction (Shafer and Vovk, 2008). CP for respiratory classification has shown success in identifying cough events in audio recordings (Nguyen and Luo, 2018; Meister, 2020). While CP has been used to forecast daily COVID-19 infections (Stankeviciute et al., 2021), to the best of our knowledge, it has not been used to predict COVID-19 infection from cough samples.

6. Conclusion and future directions

We have proposed and shown the effectiveness of three methods to improve COVID-19 cough classification, reduce the performance gap between the heavily imbalanced classes, and automatically incorporate uncertainty quantification. A new approach to segmenting training samples into cough clusters improved sample quality and quantity by 200%. In combination with a novel, class-agnostic Conformal Prediction confidence measure that takes the sample quality into account, we have achieved +35% precision and recall compared to standard Machine Learning.

Our proposed methods are versatile because they are not restricted to the COVID-19 classification problem. Future works could extended our proposals to any audio cough classification task, optionally with a more problem-specific sample quality measure. One of the paper's limitations is the relatively small dataset size, but cross-validation and experimentation with multiple ML models show stable results.

Future directions of interest include an examination of how background noise within a cough interval might be addressed with a more sophisticated sample quality measure, and evaluating how the choice of audio features alters the effectiveness of the Signal-to-Noise ratio quality measure.

Acknowledgments

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

References

Jonathan Alvarsson, Staffan Arvidsson McShane, Ulf Norinder, and Ola Spjuth. Predicting with confidence: using conformal prediction in drug discovery. *Journal of Pharmaceutical Sciences*, 110(1):42–49, 2021.

Javier Andreu-Perez, Humberto Pérez-Espinosa, Eva Timonet, Mehrin Kiani, Manuel Ivan Giron-Perez, Alma B Benitez-Trinidad, Delaram Jarchi, Alejandro Rosales, Nick Gkatzoulis, Orion F Reyes-Galaviz, et al. A generic deep learning based cough analysis system from clinically validated samples for point-of-need COVID-19 test and severity levels. *IEEE Transactions on Services Computing*, 2021.

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. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 3474–3484. Association for Computing Machinery, 2020.

Madison Cohen-McFarlane, Rafik Goubran, and Frank Knoefel. Comparison of silence removal methods for the identification of audio cough events. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1263–1268. IEEE, 2019.

Madison Cohen-McFarlane, Rafik Goubran, and Frank Knoefel. Novel coronavirus cough database: NoCoCoDa. *IEEE Access*, 8:154087–154094, 2020.

Gauri Deshpande and Björn W Schuller. The DiCOVA 2021 challenge–an encoder-decoder approach for COVID-19 recognition from coughing audio. In *Proc. Interspeech*, volume 2021, pages 931–5, 2021.

Shivani Gupta and Atul Gupta. Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Computer Science*, 161:466–474, 2019.

Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. Handling imbalanced datasets: A review. *GESTS international transactions on computer science and engineering*, 30(1):25–36, 2006.

Jordi Laguarta, Ferran Hueto, and Brian Subirana. COVID-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open Journal of Engineering in Medicine and Biology*, 1:275–281, 2020.

Wilson E Marcílio and Danilo M Eler. From explanations to feature selection: assessing SHAP values as feature selection mechanism. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 340–347. IEEE, 2020.

Julia A Meister. Conformal predictors for detecting harmful respiratory events. Master's thesis, Royal Holloway, University of London, 2020. Supervised by Khuong An Nguyen.

Julia A Meister, Khuong An Nguyen, and Zhiyuan Luo. Audio feature ranking for sound-based COVID-19 patient detection. *arXiv preprint arXiv:2104.07128*, 2021.

Emad A Mohammed, Mohammad Keyhani, Amir Sanati-Nezhad, S Hossein Hejazi, and Behrouz H Far. An ensemble learning approach to digital corona virus preliminary screening from cough sounds. *Scientific Reports*, 11(1):1–11, 2021.

Jesús Monge-Álvarez, Carlos Hoyos-Barceló, Paul Lesso, and Pablo Casaseca-de-la Higuera. Robust detection of audio-cough events using local hu moments. *IEEE Journal of Biomedical and Health Informatics*, 23(1):184–196, 2019. doi: 10.1109/JBHI.2018.2800741.

Khuong An Nguyen and Zhiyuan Luo. Cover your cough: Detection of respiratory events with confidence using a smartwatch. In *Conformal and Probabilistic Prediction and Applications*, pages 114–131. PMLR, 2018.

Madhurananda Pahar, Marisa Klopper, Robin Warren, and Thomas Niesler. Covid-19 detection in cough, breath and speech using deep transfer learning and bottleneck features. *Computers in biology and medicine*, 141:105153, 2022.

Achmad Rizal, Risanuri Hidayat, and Hanung Adi Nugroho. Signal domain in respiratory sound analysis: methods, application and future development. *Journal of Computer Science*, 11(10):1005, 2015.

Björn W Schuller, Anton Batliner, Christian Bergler, Cecilia Mascolo, Jing Han, Iulia Lefter, Heysem Kaya, Shahin Amiriparian, Alice Baird, Lukas Stappen, et al. The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates. *arXiv preprint arXiv:2102.13468*, 2021.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, Prasanta Kumar Ghosh, Sriram Ganapathy, et al. Coswara–a database of breathing, cough, and voice sounds for COVID-19 diagnosis. *arXiv preprint arXiv:2005.10548*, 2020.

Kamile Stankeviciute, Ahmed M Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. *Advances in Neural Information Processing Systems*, 34, 2021.

Wojciech Wisniewski, David Lindsay, and Sian Lindsay. Application of conformal prediction interval estimations to market makers' net positions. In *Conformal and Probabilistic Prediction and Applications*, pages 285–301. PMLR, 2020.