Audio feature ranking for sound-based COVID-19 patient detection

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

[1] School of Computing, Engineering and Mathematics, University of Brighton, Brighton BN2 4GJ, United Kingdom
[2] Department of Computer Science, Royal Holloway University of London, Surrey TW20 0EX, United Kingdom
{J.Meister,K.A.Nguyen}@brighton.ac.uk
{Zhiyuan.Luo}@rhul.ac.uk

Abstract. Audio classification using breath and cough samples has recently emerged as a low-cost, non-invasive, and accessible COVID-19 screening method. However, no application has been approved for official use at the time of writing due to the stringent reliability and accuracy requirements of the critical healthcare setting. To support the development of the Machine Learning classification models, we performed an extensive comparative investigation and ranking of 15 audio features, including less well-known ones. The results were verified on two independent COVID-19 sound datasets. By using the identified top-performing features, we have increased the COVID-19 classification accuracy by up to 17% on the Cambridge dataset, and up to 10% on the Coswara dataset, compared to the original baseline accuracy without our feature ranking.

Keywords: COVID-19 classification · Audio event engineering · Sound feature ranking.

1 Introduction

A widely accessible, non-invasive, low-cost testing mechanism is the number one priority to support test-and-trace in most pandemics. The advent of COVID-19 has abruptly brought respiratory audio classification into the spotlight as a viable alternative for mass pre-screening, needing only a smartphone to record a breath or cough sample [3].

Just in the past 12 months, many universities and research institutions have set up audio data collection systems, generally reliant on voluntary submissions, resulting in a variety of smartphone applications based on audio pre-processing and ML classification. However, at the time of writing, none has yet been officially endorsed for medical usage, largely because of the high accuracy and reliability expectations for such a critical healthcare task.

This paper aims to give a holistic overview, evaluation, and ranking of 15 audio features in the context of the binary COVID-19 audio classification task, which, to the best of our knowledge, has not been researched yet.

1.1 The paper's contributions

This paper makes the following contributions to the binary COVID-19 respiratory audio classification task:

i. *Audio feature analysis and ranking.* We perform an extensive comparative analysis and ranking of 15 sound features prevalent in speech and non-speech audio classification. The evaluation is carried out on two independent datasets, allowing the findings to be generalised.
ii. *Highlighting effective features.* We identify ML features with strong discriminative performance that go against common rules of thumb regarding audio feature selection.
iii. *Increasing the COVID-19 detection accuracy.* A natural culmination of the previous points. Compared to the baseline results presented in the datasets' original papers, we increase the classification accuracy by up to 17%, just by incorporating new training data obtained through our feature ranking.

The findings described in this paper are directly relevant to the COVID-19 sound-based classification task and would benefit future implementations using the same approach.

The remainder of the paper is organised into four sections. Section 2 provides a thorough description of the relevant audio features. Section 3 includes information about the implementations and then focuses on the extensive experimental analysis. Section 4 outlines the related work in the COVID-19 classification domain. Finally, Section 5 summarises our findings and outlines further work.

2 Audio features overview

Feature engineering is a vital step in any ML application, as a model's predictive efficiency relies directly on the discriminating information encoded in the input vectors. Before delivering a comprehensive comparison of 15 audio features in the context of binary COVID-19 audio classification, we first provide a detailed overview and intuition of their function. The selected features cover a variety of domains, including those in speech and non-speech audio tasks. A summary is presented in Table 1.

2.1 Time domain

Low-level features extracted directly from the audio signal without requiring a transformation are grouped in the time domain. While such features are often not meaningful to humans, they are commonly included in a larger feature set in audio classification tasks because they are very efficient to calculate. In the context of lung-sound classification, such features can identify explosive and discontinuous sounds (e.g. crackling) that often occur due to a buildup of fluid or secretions in the throat and lungs [21]. The selected features have been previously extracted for COVID-19 classification [3,22].

Table 1: *Audio feature selection.* The following 15 audio features, grouped by domain, will be considered in the paper.

Domain	**Feature category**	**Name**	**Intuition**
Time	Signal energy	RMSE	Loudness of the signal.
	Waveform	ZCR	Percussive vs tonal.
Frequency	Spectral	S-BW	Perceived timbre.
	Spectral	S-CENT	'Brightness' of a sound.
	Spectral	S-CONT	Prevalence of formants[1].
	Spectral	S-FLAT	Similarity to white noise.
	Spectral	S-FLUX	Rate of frequency changes.
	Spectral	S-ROLL	'Skewness' of the energy.
Time-frequency	Cepstral	MFCC	Timbre, tone colour/ quality.
	Cepstral	MFCC-Δ	Velocity of temporal change.
	Cepstral	MFCC-Δ^2	Acceleration of temporal change.
	Tonal	C-ENS	Pitch.
	Tonal	C-CQT	Pitch.
	Tonal	C-STFT	Pitch.
	Tonal	TN	Pitch & pitch height.

[1]Formants can be described as a spectrum's local maxima. In speech audio, they correspond to regularly occurring energy concentrations at certain frequencies in signals produced by the human vocal tract [4].

i. *Root mean square energy (RMSE).* A description of the signal's mean amplitude, calculated by taking the Root Mean Square (RMS) of energy over N frames, see Equation (1). x_n is the average energy per frame n [17].

$$\text{RMS} = \sqrt{\sum_{n=1}^{N} x_n^2} \tag{1}$$

ii. *Zero-crossing rate (ZCR).* The rate of the signal's sign change over time is given by Equation (2). Here x_n is the signal's amplitude at frame with index n (N frames overall), and $sign(a)$ returns 1 if $a > 0$, 0 if $a = 0$, and -1 otherwise [17].

$$\text{ZCR} = \frac{1}{2} \times \sum_{n=2}^{N} |sign(x_n) - sign(x_{n-1})| \tag{2}$$

2.2 Frequency domain

In its original format, digital audio is encoded as a temporal sequence of samples. Decomposing the signal into its constituent frequencies (e.g. with the Fourier Transform) reveals information about the frequency content. Because most frequency-domain features, alternatively spectral features, describe only a small aspect of the audio signal, they are rarely used individually for audio classification tasks. The selected features describe and compare the signal's intensity, which can provide information about the state of the respiratory tract, e.g. identifying abnormal lung sounds if it is affected by a respiratory disease [21]. A subset of the following features has previously been used for COVID-19 detection [3,22].

i. *Spectral bandwidth (S-BW).* Also referred to as spectral spread, S-BW describes the signal's energy concentration around the centroid. Equation (3) defines bandwidth as the variance around the signal's expected frequency E given the energy P_k and corresponding frequency f_k in $1 \leq k \leq K$ subbands [19].

$$\text{S-BW} = \sqrt{\textstyle\sum_{k=1}^{K}(f_k - E^2 \times P_k)} \tag{3}$$

ii. *Spectral centroid (S-CENT).* The centroid identifies a signal's mean frequency, i.e. the band with the highest energy concentration. Equation (4) shows its breakdown into the weighted and unweighted sums of spectral magnitudes P_k in the k-th of K subbands. f_k is the corresponding frequency range [23].

$$\text{S-CENT} = \frac{\sum_{k=1}^{K} P_k \times f_k}{\sum_{k=1}^{K} P_k} \tag{4}$$

iii. *Spectral contrast (S-CONT).* The audio signal's contrast is evaluated by comparing the spectral energy peaks P_k and valleys V_k in each frequency subband k, see Equation (5). N represents the number of frames and $x'_{k,n}$ the FFT vector of the k-th subband in frame n with elements in descending order [7].

$$\text{S-CONT}_k = P_k - V_k = (\log \tfrac{1}{N} \textstyle\sum_{n=1}^{N} x'_{k,n}) - (\log \tfrac{1}{N} \textstyle\sum_{n=1}^{N} x'_{k,N-n+1}) \tag{5}$$

iv. *Spectral flatness (S-FLAT).* Also called a tonality coefficient, flatness measures a signal's similarity to white noise (flat spectrum). It is defined as the ratio between the geometric and arithmetic means as shown in Equation (6), where P_k is the signal's energy at the k-th frequency band s.t. $1 \leq k \leq K$ [10].

$$\text{S-FLAT} = \frac{(\prod_{k=1}^{K} P_k)^{\frac{1}{K}}}{\frac{1}{K}\sum_{k=1}^{K} P_k} \tag{6}$$

v. *Spectral flux (S-FLUX).* A measure of a signal's change in energy between frames, estimated by Equation (7). $E_{n,k}$ represents the k-th normalised DFT (Discrete Fourier Transform) coefficient in frame n across K coefficients [23].

$$\text{S-FLUX}_n = \textstyle\sum_{k=1}^{K} E_{n,k} - E_{n-1,k}^2 \tag{7}$$

vi. *Spectral rolloff (S-ROLL).* A description of the relationship between frequency and energy, rolloff represents the minimum frequency f_R s.t. the energy accumulated below is not less than the specified proportion S of the total energy. P_k is the spectral energy in one of K frequency subbands [23].

$$\text{S-ROLL} = \arg\min f_R \in \{1, \ldots, K\} \textstyle\sum_{k=1}^{f_R} P_k \geq S \sum_{k=1}^{K} P_k \tag{8}$$

2.3 Time-frequency domain

This feature category illustrates a signal's frequency-related information as it varies over time. We consider two types of time-frequency features: cepstral features (encoding timbre or tone colour) and tonal features (describing pitch).

Cepstral features This paper focuses on the Mel-frequency Cepstrum (MFC), as it is by far the most commonly used cepstral feature variant in audio classification tasks. MFC mimics the non-linear human perceptions of sound and is applied ubiquitously in both speech and non-speech classification tasks. While both spectral and cepstral features can facilitate respiratory classification by exploring a signal's frequency content, the latter's benefit is the inclusion of temporal and transitional information. MFC features have been previously used for COVID-19 detection [3,13].

i. *Mel-frequency cepstral coefficients (MFCC).* MFCC features are derived from the MFC power spectrum. In Equation (9) the signal is transformed into the time-frequency domain by a discrete cosine transform. K is the number of coefficients and $s(k)$ calculates the logarithmic energy of the k-th coefficient at frame n [20].

$$\mathrm{MFCC}_n = \textstyle\sum_{k=1}^{K} s(k) \cos \frac{\pi n(k-0.5)}{K} \tag{9}$$

ii. *MFCC-Δ.* As the first-order derivative of MFCC, also referred to as velocity, the feature represents temporal change [5]. It is often included in combination with MFCC, as it has a low extraction cost.

iii. *MFCC-Δ^2.* Acceleration, MFCC's second-order derivative, is commonly included when MFCC is extracted from an audio signal because it is resource-efficient to calculate and can improve audio classification [5].

Tonal features Tonal features primarily encode an audio signal's harmonics information in 12 pitch classes[2] and are based on the human perception of periodic pitch [15]. Two feature groups are considered, distinguished by their underlying representation: Chroma features (chromagram) and Tonnetz (lattice graph). While the Tonnetz encodes tone quality and height, chromagrams omit interval information. A common consequence of respiratory diseases is a narrowing of the airways by secretions. The effect is a wheezing sound because the pitch of in- and expiration is altered [21], which can be heard in COVID-19 lung-sound recordings.

i. *Chroma energy normalised (C-ENS).* A chromagram feature abstraction introduced in [14] by considering short-time statistics over energy distributions within the chroma bands. Normalisation of the feature vectors makes it resistant to dynamic variations, such as timbre and articulation [15].

[2] Pitch classes of the equal tempered scale, prominent in Western tonal music [15].

ii. *Constant-Q chromagram (C-CQT).* Chroma features are extracted from a time-frequency representation of audio via a filter bank. In this case, the initial audio transformation is the constant-Q transform (CQT), which has a good resolution of low frequencies [8].
iii. *Short-time Fourier Transform chromagram (C-STFT).* The feature extraction process is similar to the description for C-CQT. The difference lies in the audio signal's transformation into the time-frequency domain, which in this case is calculated via the Short-time Fourier Transform (STFT) [8].
iv. *Tonnetz (TN).* A Tonnetz (German: tone-network) encodes harmonic data in a lattice graph. The benefit of a graphical representation is that distances between points are musically meaningful, as pitch is encoded as geometric areas in the graph [6].

3 Experimental method and results

The ranking of the above 15 selected audio features will be based on the empirical results and analysis of two datasets to make the findings more generalisable. The assumption is that any distinct patterns repeated across independent datasets are likely inherent to the COVID-19 breath and cough audio recordings, not the underlying datasets.

3.1 Research questions

Exploring the following questions is the focus of this body of work. They are centred on the binary COVID-19 audio classification task and have informed the experimental design and consequent results analysis.

i. What are the most distinguishable ML audio features?
ii. Are the feature rankings comparable across independent datasets?
iii. What is the performance accuracy of the new ML models using the most dominant features?

To answer the above research questions, this section contains a brief description of the datasets underlying the evaluated features, the data preparation and pre-processing steps, and an extensive description and analysis of the results. Finally, we compare our improved results to the baseline ML accuracies presented in the datasets' original papers.

3.2 The datasets

Two independent datasets are considered in parallel throughout the paper to indicate whether identified feature rankings are likely specific to the underlying dataset or generally applicable: the *Cambridge* and the *Coswara* COVID-19 audio datasets. The distribution of sample counts can be found in Table 2.

Table 2: *Sample counts and label stratification of the Cambridge and Coswara datasets.* 'Shallow' and 'deep' refer to to the 'shallow' and 'deep' breath (B), cough (C), and breathcough (BC) recordings available for every participant.

Label	**Cambridge**			**Coswara-deep**			**Cos.-shallow**		
	B	C	BC	B	C	BC	B	C	BC
COVID-19	111	111	111	81	81	81	81	81	81
Healthy	194	194	194	1074	1074	1074	1074	1074	1074
$\sum$	305	305	305	1155	1155	1155	1155	1155	1155

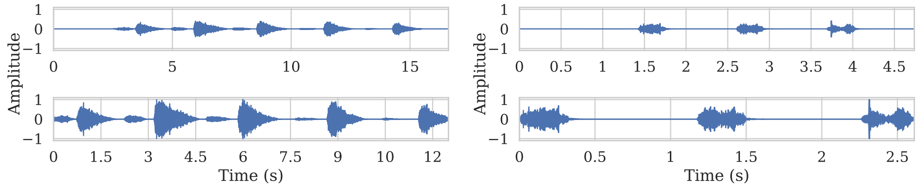
Introduced in [3], the *Cambridge dataset* is a collection of voluntary web and android recordings of coughing and breathing sounds from healthy, COVID-positive, and asthmatic people. Only the first two categories are considered, as the latter only has eight samples. The data available for this paper is a curated set of samples collected during April and May 2020. While the paper describes various metadata statistics over the entire dataset (e.g. age, gender, location, and symptom distribution), such information is not included in the curated dataset. The data comes in 2 to 30-second WAV files with a 48kHz sampling.

The *Coswara dataset* [22] is collected and freely distributed by the Indian Institute of Science and receives its voluntary samples through a web application. The samples considered in this paper were collected between April and December 2020. The available categories and recording types are much more varied, but to remain consistent with the Cambridge dataset, we filter the data for COVID-positive and healthy participants that have submitted breathing and coughing recordings. The 'shallow' and 'deep' variants are considered as two separate datasets. Conveniently, the data format is compatible with the same WAV format at 48kHz and 1 to 30-second long recordings.

3.3 Feature engineering

Because the recording devices and environments were not controlled, consistently cleaning the audio data is important to reduce non-discriminatory variance and improve the samples' comparability to each other. The applied pre-processing steps include converting the audio to mono, resampling it to 48kHz, trimming the leading and trailing silences, and normalising the signal's amplitude to $[-1, 1]$. The effects can be seen in Figure 1. The Python-toolkit `librosa` [11] (version 0.8) was used for the signal processing tasks.

The basis of all of our evaluations are the 15 audio features from the time, frequency, and time-frequency domains identified and described in Section 2. In general, ML models require input with a consistent format and dimension. Because the recordings have vastly different lengths (1–30 seconds, see Figure 2) and the selected audio features are calculated over frames, the question of how to represent the feature vector at a constant dimension was a challenge. A range



(a) Raw and pre-processed 'breath' audio. (b) Raw and pre-processed 'cough' audio.

Fig. 1: *The effects of cleaning the raw audio recordings.* Pre-processing steps include converting the audio to mono at 48kHz, trimming, and normalising.

of summary statistics over frames is taken to capture all available data without resorting to padding (infeasible due to the up to 29-second difference). This leads to a feature vector guaranteed to have the same number of dimensions, regardless of the sample length. The statistics we consider are the (i) minimum, (ii) maximum, (iii) mean, (iv) median, (v) variance, (vi) 1st quartile, and (vii) 3rd quartile, giving us a wide range of descriptive information about the features' distribution over frames. The total count of features analysed and ranked individually and by category is 812, as detailed in Table 3. no dimensionality reduction to maintain interpretability of the features

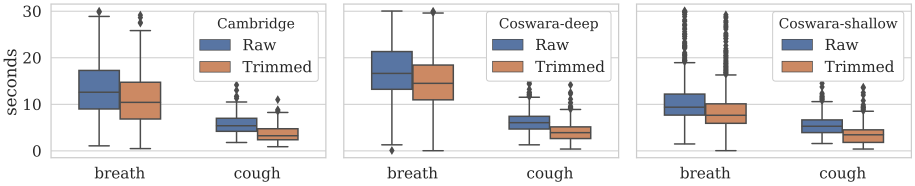


Fig. 2: *Sample lengths before and after pre-processing.* By trimming the leading and trailing silences at 60dB (empirically identified cutoff point) we can remove non-discriminative data. Sample lengths are reduced by 1–3 seconds on average.

3.4 Results description and analysis

The paper's main contribution is an extensive and in-depth analysis and ranking of 15 audio features for the binary COVID-healthy classification task. The goal is to identify particularly informative features and feature categories by carrying out the evaluation on two independent datasets in parallel: the Cambridge and the Coswara (deep and shallow variants) datasets. Due to their independence, we propose that any recurring patterns in predictive efficiency are likely independent of the underlying dataset and should therefore be strongly considered for future ML COVID-19 audio classification applications.

Table 3: *Feature dimensions.* A total of 812 features are considered. Seven summary statistics (min, max, mean, median, var, Q_1, and Q_3) are taken across frames to make the feature vector dimensions compatible regardless of the underlying recording's length (1–30s) and the number of respiratory events (1–10). The large number of features brings a risk of overfitting, however, for the majority of category evaluations only a small subset of features is used at a time.

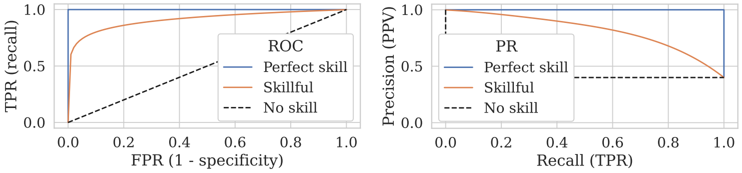
Feature	**Name**	**Count**	**Total** (×7)
RMSE	Root mean square energy	1	7
ZCR	Zero-crossing rate	1	7
S-BW	Spectral bandwidth	1	7
S-CENT	Spectral centroid	1	7
S-CONT	Spectral contrast	7	49
S-FLAT	Spectral flatness	1	7
S-FLUX	Spectral flux	1	7
S-ROLL	Spectral rolloff	1	7
MFCC	Mel-frequency cepstral coefficients	20	140
MFCC-Δ	Mel-frequency cepstral coefficients Δ	20	140
MFCC-Δ^2	Mel-frequency cepstral coefficients Δ^2	20	140
C-ENS	Chroma energy normalised	12	84
C-CQT	Constant-Q chromagram	12	84
C-STFT	Short-time Fourier Transform chromagram	12	84
TN	Tonnetz	6	42

The 15 audio features summarised in Table 3 are analysed over the following configurations to provide a detailed picture of their predictive efficiency:

i. The Cambridge, Coswara-deep, and Coswara-shallow datasets.
ii. 'Breath' (B), 'cough' (C), and 'breathcough' (BC) feature vectors. The latter is a concatenation of the previous two feature vectors, i.e. double the size.
iii. Five ML models, selected for the variety in which they partition the label space. The models are implemented with the `scikit-learn` [18] package version 0.24, and optimised with parameter grid searches: AdaBoost with Random Forest (`nr. estimators`, `criterion`), K-Nearest Neighbours (`K`, `weights`), Logistic Regression (`C`, `penalty`, `solver`), Random Forest (`max. depth`, `criterion`), and Support Vector Machine (`C`, γ, `kernel`), referred to as ADA, KNN, LR, RF, and SVM respectively.

To ensure that the generated results are reliable even on the relatively small and very imbalanced available datasets, 5-fold Cross-Validation (CV) stratified by labels is employed. We select three metrics to compare the features' impact on the audio classification task at hand: *Receiver Operating Characteristic* (ROC), *Precision*, and *Recall*. The latter two counteract ROC's optimism on highly imbalanced datasets, see Figure 3 for a brief overview.

In the following, we provide a detailed analysis and discussion of the previously described audio features' (Section 2) performance on the selected datasets



(a) ROC curve. (b) PR curve.

Fig. 3: *Intuition of the considered metrics.* In addition to ROC, Precision-Recall (PR) curves are a valuable tool when evaluating imbalanced datasets because they counteract the effect of relative imbalances by omitting true negatives (TN or specificity). PR no-skill classifiers correspond to the positive sample ratio in the dataset (i.e. precision at threshold 0.0).

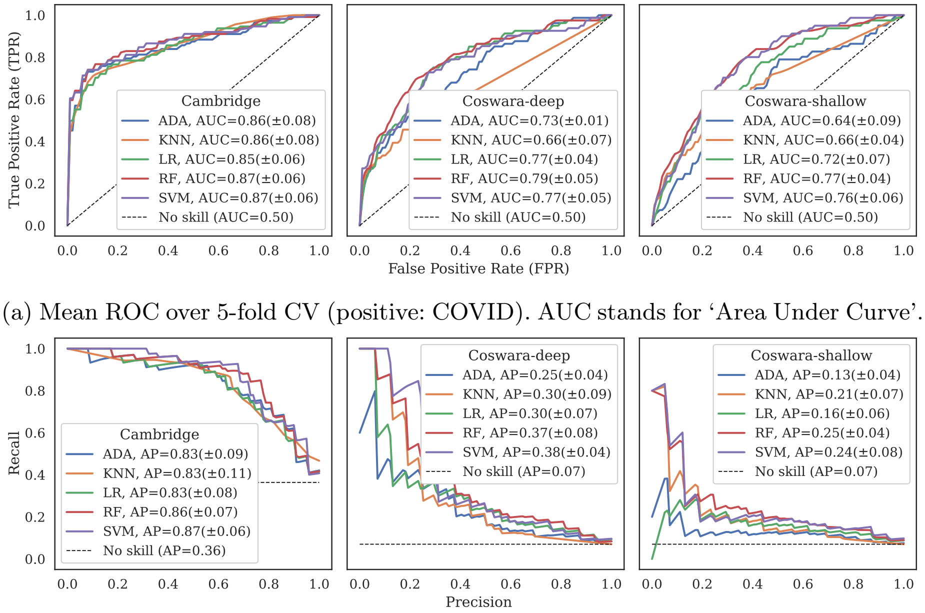
(Section 3.2) for the binary COVID-19 classification task. Finally, we compare our improved results to the baseline ML accuracies presented in the datasets' original papers.

Feature categories. An initial overview of the 'breath', 'cough', and 'breath-cough' full feature vectors' COVID-19 discriminatory efficiency shows promising results, as most models outperform their no-skill equivalent. Figure 4 visualises the mean ROC and Precision-Recall (PR) curves over 5 CV folds on the 'breath-cough' vector for each of the considered models. It clearly establishes SVM and RF outperforming their counterparts across all datasets and metrics, with a similar trend observed for the other two data types.

Even though the Cambridge and Coswara datasets have similarly shaped ROC curves, it becomes immediately apparent that the Cambridge dataset significantly outperforms its counterparts when looking at the PR curves. This illustrates ROC-AUC's optimism when applied to vastly imbalanced datasets, justifying our approach of considering multiple metrics throughout our analysis. An influential factor in Coswara's lower overall accuracies is the greater imbalance of COVID-positive samples compared to healthy ones at 13:1 compared to 2:1 in the Cambridge data (see Table 2 for sample counts). Nonetheless, models trained on the Coswara datasets still perform noticeably better than an entirely unskilled classifier with AP (Average Precision) scores between 13–38% compared to the unskilled 7% (equivalent to the positive sample ratio).

The results contained in Table 4 confirm our selection of SVM and RF as the best-performing models. The table shows the same 'breathcough' feature vector's predictive efficiency, but this time only considering one feature category (time domain, spectral, cepstral, tonal) at a time. Apart from two exceptions, both SVM and RF achieve higher accuracies than the other ML models across the board.

Considering SVM's mean ROC-AUC accuracies on the 'breathcough' vector across all datasets, we noted that the 4 feature categories could be broadly ranked



(a) Mean ROC over 5-fold CV (positive: COVID). AUC stands for 'Area Under Curve'.

(b) Mean PR over 5-fold CV (positive: COVID). AP stands for 'Average Precision'.

Fig. 4: *Model results on the 'breathcough' feature vector.* The graphs show the 5 considered models and an unskilled classifier. For PR-curves, this corresponds to the ratio of positive labels in the dataset. Even though the ROC-curves look similar across datasets, the PR-curves reveal that Cambridge performs better overall. We can also identify SVM and RF as the top-performing models.

in the following order of increasing predictive efficiency (Cambridge, Coswara-deep, Coswara-shallow): *time domain* (78.78%, 63.94%, 55.90%), *tonal* (82.59%, 72.98% 68.81%), *spectral* (84.84%, 74.46%, 72.32%), and *cepstral* (87.15%, 75.62%, 70.62%). As evidenced by the results, the spectral and cepstral categories perform equally well, where spectral slightly outperforms cepstral for Coswara-shallow by about 2%. More noteworthy is that the same ranking pattern is prevalent for all 5 considered ML models, leading to the conclusion that the cepstral and spectral feature categories encode particularly informative data for COVID-19 classification contained in the breathing and coughing signals.

Individual features. Now turning our attention to the analysis of individual features, the initial focus lies on the previously identified best-performing SVM and RF classifiers before broadening again to include all models, letting us identify generally applicable patterns of predictive efficiency. The feature accuracies on which the following descriptions are based are available in Tables 5 to 7 for the Cambridge, Coswara-deep, and Coswara-shallow datasets respectively.

Table 4: *'Breathcough' 5-fold CV ROC-AUC results.* The mean μ and standard deviation σ are reported for four feature categories (i.e. the feature vectors are a concatenation of the category components, see Table 1 for details). SVM and RF perform best overall, and the feature categories can be ranked in the following increasing order: time domain, tonal, spectral, and cepstral.

Dataset	**Category**	**ADA**		**KNN**		**LR**		**RF**		**SVM**	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Cambridge	Time dom.	67.17	0.04	77.96	0.07	76.01	0.07	78.21	0.05	**78.78**	0.07
	Spectral	87.09	0.04	85.34	0.05	84.17	0.06	**87.15**	0.05	84.84	0.07
	Cepstral	83.84	0.05	85.56	0.07	83.27	0.06	87.82	0.07	**87.15**	0.06
	Tonal	**84.74**	0.09	81.04	0.05	81.44	0.04	81.11	0.07	82.59	0.07
Coswara-deep	Time dom.	55.65	0.07	62.34	0.02	54.21	0.09	**64.65**	0.05	63.94	0.07
	Spectral	65.77	0.07	68.18	0.04	72.03	0.05	71.76	0.06	**74.46**	0.06
	Cepstral	70.83	0.06	71.03	0.03	75.01	0.05	**77.55**	0.06	75.62	0.08
	Tonal	69.29	0.06	66.27	0.02	68.02	0.03	72.32	0.06	**72.98**	0.03
Coswara-shallow	Time dom.	**61.63**	0.04	55.05	0.06	56.16	0.09	54.27	0.07	55.90	0.09
	Spectral	66.69	0.04	61.02	0.05	69.85	0.05	69.15	0.05	**72.32**	0.04
	Cepstral	63.13	0.09	68.35	0.04	65.83	0.03	**71.79**	0.06	70.62	0.04
	Tonal	58.37	0.08	63.98	0.05	65.21	0.08	67.17	0.08	**68.81**	0.08

Taking a general look at the results, we see that the majority of the 15 features significantly outperforms random guesses for the binary COVID-19 classification task across datasets and sample types ('breath', 'cough', 'breathcough'), with better accuracy on the Cambridge dataset. The lowest accuracy on average is achieved by Coswara-shallow, matching our previous findings when considering both the entire feature vector and individual feature categories. Looking at which underlying sample type performs best further underlines the similarities between the Cambridge and Coswara-deep datasets compared to Coswara-shallow. When considering the former, 'breathcough' achieves the highest mean ROC-AUC scores on average (except for time domain features), whereas Coswara-shallow is split evenly between 'breath' (time domain, spectral) and 'breathcough' (cepstral, tonal). However, given all considered features in a single feature vector, the Coswara-shallow dataset still shows its highest accuracy on 'breathcough' samples since cepstral and tonal features are very influential overall.

When comparing the results within the feature categories, MFCC (cepstral), S-CONT (spectral), and C-ENS/ C-STFT (tonal) stand out as the highest-scoring features in their respective categories across datasets and models. In contrast, the time domain features are much more varied in which one performs best. It is worth mentioning that spectral contrast (S-CONT) is the only composite feature (7-D) in the Spectral category, which could be part of the reason it performs better. However, the heat maps in Figure 5 clearly show that individual S-CONT features perform better or on par with other top spectral features in a

Table 5: *5-fold CV results on all audio features extracted from the Cambridge dataset.* The mean μ and standard deviation σ ROC-AUC results are reported. The majority of features provide the most accurate results when considering the 'breathcough' ('BC') vector. We also find that the feature categories can be ranked in the following order of increasing accuracy across both models: time domain, tonal, spectral, and cepstral.

(a) SVM results.

Category	**Feature**	**Breath**		**Cough**		**BC**	
		μ	σ	μ	σ	μ	σ
All	All	85.86	0.07	85.80	0.05	**87.68**	0.06
Time dom.	All	72.77	0.04	74.90	0.08	**78.78**	0.07
	RMSE	72.28	0.05	76.45	0.08	**77.88**	0.08
	ZCR	64.59	0.08	69.73	0.06	**71.40**	0.06
Spectral	All	**85.28**	0.06	84.03	0.07	84.84	0.07
	S-BW	69.24	0.08	71.57	0.04	**75.45**	0.08
	S-CENT	73.45	0.08	70.06	0.08	**78.07**	0.07
	S-CONT	**86.14**	0.06	84.03	0.08	85.98	0.08
	S-FLAT	74.22	0.07	75.44	0.05	**75.87**	0.06
	S-FLUX	79.70	0.08	77.14	0.06	**82.08**	0.06
	S-ROLL	70.70	0.07	67.22	0.04	**71.22**	0.06
Cepstral	All	86.25	0.06	83.98	0.06	**87.15**	0.06
	MFCC	86.56	0.04	83.25	0.05	**87.68**	0.04
	MFCC-Δ	84.21	0.04	79.67	0.08	**85.54**	0.08
	MFCC-Δ^2	84.25	0.09	78.29	0.07	**85.24**	0.09
Tonal	All	79.69	0.07	78.06	0.07	**82.59**	0.07
	C-CQT	76.29	0.06	71.12	0.09	**77.30**	0.06
	C-ENS	77.56	0.07	72.11	0.07	**83.50**	0.03
	C-STFT	77.57	0.05	72.65	0.03	**77.78**	0.07
	TN	74.28	0.04	70.85	0.04	**77.57**	0.05

(b) RF results.

Category	**Feature**	**Breath**		**Cough**		**BC**	
		μ	σ	μ	σ	μ	σ
All	All	86.35	0.07	86.89	0.07	**87.78**	0.06
Time dom.	All	66.07	0.02	74.12	0.09	**78.21**	0.05
	RMSE	61.86	0.05	72.08	0.09	**75.19**	0.06
	ZCR	53.77	0.04	66.80	0.06	**66.41**	0.05
Spectral	All	85.07	0.07	86.23	0.07	**87.15**	0.05
	S-BW	70.69	0.07	69.48	0.05	**76.39**	0.04
	S-CENT	71.40	0.08	70.06	0.05	**74.10**	0.07
	S-CONT	85.61	0.06	85.04	0.07	**85.74**	0.08
	S-FLAT	76.28	0.08	**81.29**	0.05	81.14	0.06
	S-FLUX	79.83	0.07	76.90	0.07	**82.34**	0.06
	S-ROLL	69.82	0.03	67.49	0.04	**71.99**	0.05
Cepstral	All	86.55	0.05	83.70	0.07	**87.82**	0.07
	MFCC	84.27	0.05	83.27	0.05	**85.26**	0.05
	MFCC-Δ	**86.79**	0.06	79.62	0.11	**86.80**	0.08
	MFCC-Δ^2	86.55	0.06	78.02	0.08	**88.31**	0.07
Tonal	All	78.78	0.07	76.75	0.05	**81.11**	0.07
	C-CQT	74.34	0.06	68.78	0.06	**76.11**	0.06
	C-ENS	79.05	0.06	71.08	0.06	**81.57**	0.05
	C-STFT	**77.58**	0.08	69.57	0.05	**77.67**	0.07
	TN	75.34	0.06	63.54	0.06	**76.35**	0.04

Table 6: *5-fold CV results on all audio features extracted from the Coswara-deep dataset.* The mean μ and standard deviation σ ROC-AUC results are reported. Like with the Cambridge dataset, most features show the highest accuracy on the ‘breathcough’ (‘BC’) vector. Promisingly, these results show the exact same feature ranking (increasing efficiency): time domain, tonal, spectral, and cepstral.

(a) SVM results.

Category	**Feature**	**Breath**		**Cough**		**BC**	
		μ	σ	μ	σ	μ	σ
All	All	76.79	0.04	70.85	0.06	**77.15**	0.05
Time dom.	All	61.80	0.04	58.58	0.06	**63.94**	0.07
	RMSE	55.89	0.10	61.14	0.07	**61.81**	0.07
	ZCR	**64.68**	0.03	59.45	0.13	**64.60**	0.04
Spectral	All	**76.34**	0.05	66.74	0.05	74.46	0.06
	S-BW	61.63	0.07	63.51	0.05	**65.46**	0.04
	S-CENT	68.53	0.06	59.91	0.06	**71.95**	0.05
	S-CONT	**74.89**	0.05	63.42	0.08	73.57	0.09
	S-FLAT	61.77	0.08	59.86	0.06	**61.14**	0.03
	S-FLUX	63.79	0.06	62.76	0.07	**67.20**	0.04
	S-ROLL	65.35	0.05	63.16	0.05	**67.58**	0.08
Cepstral	All	74.57	0.03	70.15	0.09	**75.62**	0.08
	MFCC	74.24	0.03	70.74	0.01	**75.38**	0.05
	MFCC-Δ	64.85	0.07	**68.90**	0.05	**68.99**	0.04
	MFCC-Δ^2	66.65	0.08	67.72	0.06	**70.72**	0.07
Tonal	All	71.74	0.05	64.06	0.06	**72.98**	0.03
	C-CQT	**67.87**	0.04	62.78	0.07	61.50	0.05
	C-ENS	**70.03**	0.07	65.14	0.03	65.96	0.05
	C-STFT	67.01	0.05	61.80	0.08	**68.19**	0.10
	TN	60.90	0.04	**62.84**	0.02	61.33	0.03

(b) RF results.

Category	**Feature**	**Breath**		**Cough**		**BC**	
		μ	σ	μ	σ	μ	σ
All	All	78.02	0.04	70.97	0.06	**79.31**	0.05
Time dom.	All	58.98	0.07	58.92	0.04	**64.65**	0.05
	RMSE	50.33	0.06	56.25	0.04	**59.25**	0.06
	ZCR	57.81	0.09	51.63	0.02	**61.39**	0.03
Spectral	All	**74.53**	0.03	65.95	0.07	71.76	0.06
	S-BW	59.58	0.04	62.54	0.10	**66.23**	0.07
	S-CENT	63.35	0.06	58.70	0.07	**64.90**	0.10
	S-CONT	**71.54**	0.03	58.91	0.09	69.94	0.07
	S-FLAT	64.19	0.04	66.23	0.04	**66.12**	0.06
	S-FLUX	60.07	0.04	53.91	0.05	**61.47**	0.04
	S-ROLL	60.57	0.08	59.72	0.09	**66.64**	0.07
Cepstral	All	77.76	0.06	72.88	0.04	**77.55**	0.06
	MFCC	72.99	0.03	67.26	0.03	**75.86**	0.02
	MFCC-Δ	69.07	0.06	70.82	0.07	**73.40**	0.07
	MFCC-Δ^2	67.57	0.07	69.21	0.09	**72.83**	0.07
Tonal	All	71.06	0.06	65.46	0.07	**72.32**	0.06
	C-CQT	**65.55**	0.05	55.87	0.10	63.16	0.06
	C-ENS	**67.51**	0.06	57.66	0.08	63.34	0.05
	C-STFT	65.60	0.07	59.25	0.09	**69.33**	0.08
	TN	63.74	0.06	60.66	0.05	**65.00**	0.05

Table 7: *5-fold CV results on all audio features extracted from the Coswara-shallow dataset.* The mean μ and standard deviation σ ROC-AUC results are reported. While the other two datasets follow very similar patterns, this one is the most different. For example, there is no one sample type that the majority of features perform best on. Nonetheless, the overall category ranking stays the same: time domain, tonal, spectral, and cepstral.

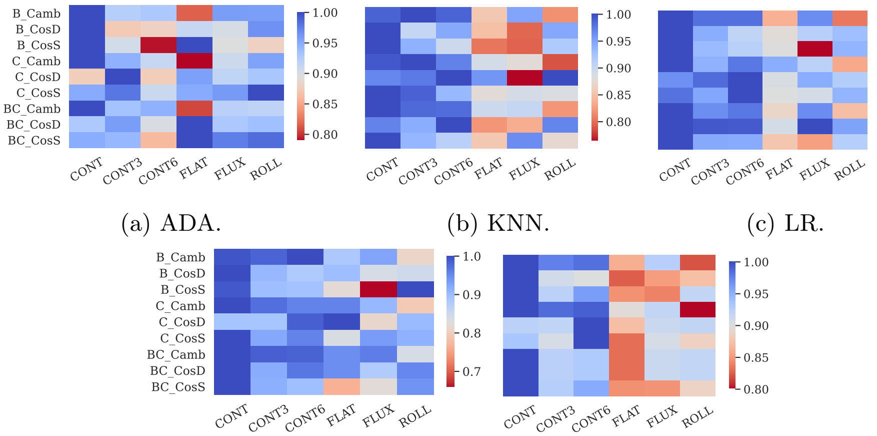
(a) SVM results.

Category	**Feature**	**Breath**		**Cough**		**BC**	
		μ	σ	μ	σ	μ	σ
All	All	72.62	0.06	68.92	0.05	**76.29**	0.06
Time dom.	All	**61.45**	0.08	53.03	0.06	55.90	0.09
	RMSE	**59.92**	0.05	53.74	0.03	**59.84**	0.05
	ZCR	**59.77**	0.03	53.21	0.08	55.67	0.05
Spectral	All	69.29	0.04	65.84	0.04	**72.32**	0.04
	S-BW	**63.10**	0.04	59.81	0.03	61.05	0.05
	S-CENT	**63.40**	0.03	58.86	0.07	62.04	0.04
	S-CONT	70.20	0.05	64.20	0.05	**71.85**	0.05
	S-FLAT	59.42	0.07	57.07	0.04	**60.76**	0.09
	S-FLUX	58.94	0.06	**62.23**	0.04	60.85	0.04
	S-ROLL	**64.48**	0.05	60.84	0.06	63.73	0.04
Cepstral	All	67.80	0.05	66.80	0.05	**70.62**	0.04
	MFCC	**71.27**	0.04	66.88	0.09	**71.17**	0.06
	MFCC-Δ	63.93	0.10	**67.15**	0.06	65.77	0.04
	MFCC-Δ^2	61.34	0.04	63.44	0.04	**65.44**	0.06
Tonal	All	66.78	0.10	63.07	0.03	**68.81**	0.08
	C-CQT	**65.93**	0.06	59.98	0.08	64.97	0.06
	C-ENS	59.79	0.06	63.99	0.04	**65.16**	0.04
	C-STFT	65.82	0.03	63.37	0.03	**68.59**	0.08
	TN	59.85	0.03	60.40	0.03	**60.64**	0.06

(b) RF results.

Category	**Feature**	**Breath**		**Cough**		**BC**	
		μ	σ	μ	σ	μ	σ
All	All	70.54	0.07	71.88	0.04	**76.76**	0.04
Time dom.	All	**61.24**	0.07	59.38	0.04	54.27	0.07
	RMSE	**51.72**	0.07	50.26	0.07	49.25	0.06
	ZCR	58.66	0.05	55.16	0.07	**61.14**	0.05
Spectral	All	68.46	0.05	64.57	0.03	**69.15**	0.05
	S-BW	**58.46**	0.07	46.51	0.05	57.78	0.03
	S-CENT	**59.04**	0.07	57.48	0.07	58.54	0.06
	S-CONT	67.40	0.05	62.84	0.06	**68.80**	0.04
	S-FLAT	**55.92**	0.02	52.57	0.03	52.68	0.05
	S-FLUX	44.77	0.04	**58.70**	0.06	56.65	0.04
	S-ROLL	**68.04**	0.07	57.36	0.05	64.79	0.06
Cepstral	All	68.21	0.06	71.60	0.05	**71.79**	0.06
	MFCC	68.47	0.03	68.23	0.07	**71.92**	0.03
	MFCC-Δ	61.91	0.04	**66.87**	0.05	**66.83**	0.04
	MFCC-Δ^2	60.53	0.05	**69.07**	0.05	65.95	0.05
Tonal	All	65.39	0.08	61.80	0.02	**67.17**	0.08
	C-CQT	**62.54**	0.07	57.99	0.05	61.08	0.04
	C-ENS	62.49	0.05	61.97	0.07	**64.37**	0.01
	C-STFT	64.27	0.06	58.54	0.03	**69.62**	0.07
	TN	56.01	0.07	**56.64**	0.05	54.12	0.04

majority of cases across datasets, sample types, and models, leading to the conclusion that S-CONT's overall positive COVID classification accuracy is in fact based on high-scoring sub-features, rather than just its increased dimensionality.



(a) ADA. (b) KNN. (c) LR.

(d) RF. (e) SVM.

Fig. 5: *Normalised ROC-AUC scores of top spectral features for breath ('B'), cough ('C'), and breathcough ('BC').* The graphs show that S-CONT's high performance is achieved because individual features consistently outperform other spectral features, not just because it is the only composite spectral feature (7-D).

Lastly, we note a surprising trend regarding MFCC and its derivative features. A prevalent rule of thumb concerning the number of MFCC features that should be included for audio classification tasks is 12 or 13 [3,7,20,22]. However, Figure 6 shows that higher-order features actually provide remarkable discriminative information for the identification of COVID-19 respiratory sounds either on par with (Coswara-deep) or significantly outperforming (Cambridge) the first 13 features. This phenomenon is most noticeable in the 'breathcough' and 'breath' features and MFCC's derivatives. The intuition for MFCCs is that the lower-order features provide information about the signal's energy distribution between high and low frequencies, and the higher-order features contain information about finer details such as pitch and tone quality [12]. From this, we can extrapolate that timbral information is very relevant to COVID audio classification.

Discussion. We have found, described, and analysed in the extensive comparison and ranking of 15 audio features in the previous section that there are distinct efficiency patterns that reoccur on multiple independent datasets.

Starting with the encompassing audio feature categories, there is a distinct order of predictive efficiency that is consistent across different datasets, mod-

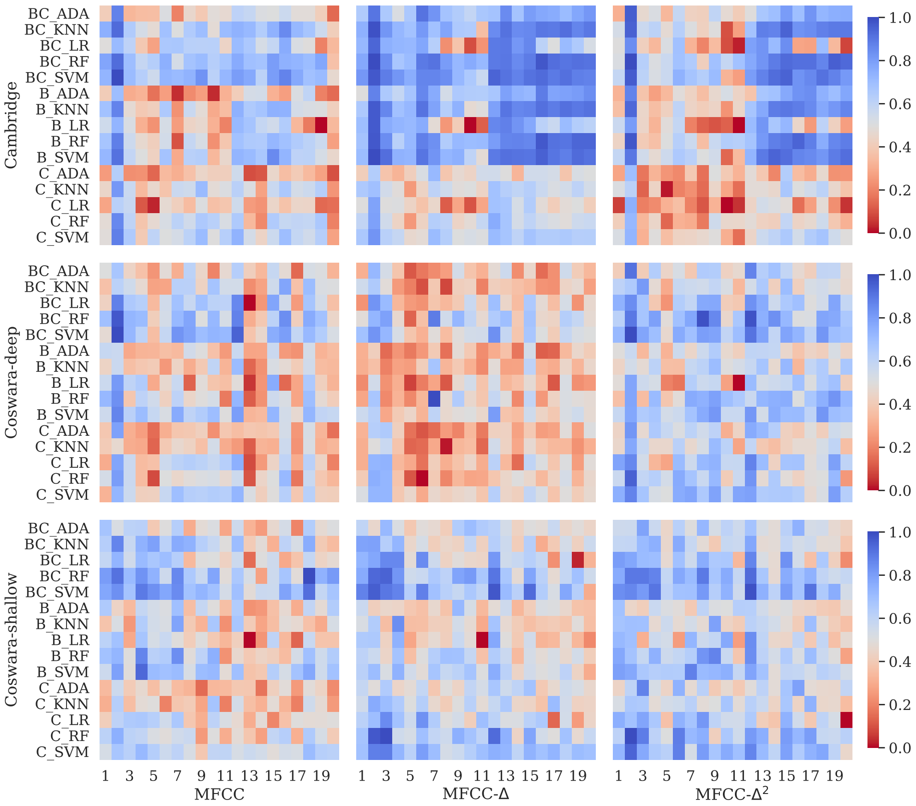


Fig. 6: *Normalised mean ROC-AUC heat map for MFCC and its derivative features.* Surprisingly and contrary to a common audio feature selection rule of thumb [3,7,20,22], higher-order MFFC features (13+) provide significant discriminatory efficiency for COVID-19 classification higher than or on par with lower-order features. This shows that pitch and timbral information is especially relevant to COVID respiratory classification. 'BC', 'B', and 'C' stand for the 'breathcough', 'breath', and 'cough' sample variants.

els, and sample types (increasing): time domain, tonal, spectral, and cepstral. This does not quite follow the intuitive expectation that more complex features provide more discriminative information (e.g. tonal vs spectral features). On the other hand, it can be justified when considering that tonal features describe pitch and so are more suited to tasks with melodic content. The ranking also underlines the significance of frequency-based features by elevating the spectral and cepstral categories. Features in these categories encode an audio signal's frequency content and describe timbral aspects and tone quality or colour. In addition to the feature rankings, we have also shown that the common audio feature selection rule of thumb of using only the first 13 MFCC features [3,7,20,22] is not applicable in this case. Indeed, the higher-order (describing timbre) features' predictive efficiency provides significantly more discriminatory information, especially for the 'breathcough' and 'breath' feature vectors.

Taking a step back from the individual features, we note that the most prevailing pattern across all of the previous descriptions is that the concatenated 'breathcough' feature vector outperforms the individual 'breath' and 'cough' vectors in most cases.

Given our insights, it is interesting to compare our ML results to the baselines presented when the datasets were published, summarised in Table 8. The evaluated models are of similar type and complexity; The major difference is our introduction of new training features. We can see that, in fact, our improved feature vectors significantly outperform both the Cambridge and Coswara baseline accuracies by over 10%, validating our feature selection.

Table 8: *Comparison to baseline results in the Cambridge and Coswara/ DiCOVA challenge papers.* Both papers present multiple results with different configurations. We select the most comparable in terms of feature pre-processing (i.e. no dimensionality reduction or DL embedding) and classification model (simple ML). All results are the average over 5-fold CV.

Origin	**Dataset**	**Sample**	**Model**	**ROC-AUC**		**Precision**		**Recall**	
				μ	σ	μ	σ	μ	σ
This paper	Cambridge	BC	SVM	**87.68**	0.06	87.61	0.07	81.39	0.07
[3]	Cambridge	BC	LR	71.00	0.08	69.00	0.09	66.00	0.14
This paper	Cos-deep	BC	SVM	**77.15**	0.05	76.7	0.05	53.09	0.03
[13]	Cos-Unknown	C	RF	67.59	—	—	—	—	—

4 Related work

While it seems that we are constantly surrounded by speech recognition in our day-to-day lives, when is the last time a digital assistant said 'bless you' after hearing and recognising a sneeze? The ubiquity of speech recognition is at least partially driven by commercial value. In contrast, non-speech sound classification, especially body sound (e.g. sneeze, cough, breathing) classification, has only recently gained traction over the past few years. The sudden emergence of the COVID-19 respiratory disease and the continual lack of testing availability have given the subfield a significant boost.

COVID-19 is not the first application of respiratory classification. It has long been common knowledge that respiratory diseases and disorders affect breathing and coughing by physically altering the respiratory environment. Because many disease-related abnormalities can affect only subtle changes in auditory cues, the inherently subjective manual auscultation[3] process can be error-prone even when performed by a trained medical professional [2]. However, a literature review of existing implementations shows that ML can reliably pick up on those subtle signals for a variety of diseases.

While the following is by no means a comprehensive list of existing implementations, it provides an overview of the current state of research. Smartwatches and small wearable devices have made audio monitoring for healthcare purposes feasible. Nguyen et al. apply a dynamic activated respiratory event detection mechanism to non-intrusively detect coughing and sneezing events [16]. When it comes to the diagnosis of respiratory events, Amrulloh et al. present classifiers trained on audio features such as MFCC to distinguish between asthma and pneumonia for pediatric patients, which are commonly misdiagnosed without proper diagnostic tools in third-world countries, leading to unnecessary antibiotic prescriptions [1]. Lastly, a method of non-binary classification is presented in [2]. Interestingly the audio classification task is transformed into image classification by using a spectrogram as input and achieves comparable results.

Over the past year, there has been an explosion of COVID-related datasets and promising pre-screening implementations, utilising a wide range of sample types. One of the first was [3], which collected and classified breath, cough, and breathcough samples to identify their suitability for COVID-19 classification with a small selection of common audio features. [22] considers further recording types, including vowel intonation and sequence counting. Laguarta et al. propose a different approach, instead applying classification to four biomarkers (muscular degradation, changes in vocal cords, changes in sentiment/ mood, and changes in the lungs/ respiratory tract) that have previously been used to identify the progress of Alzheimer's disease. Intriguingly, this approach has a very high success rate at identifying asymptomatic COVID-carriers [9].

While there are many promising applications available already, the novelty of COVID audio classification means there are still many aspects that need to be explored, partially because only limited and highly imbalanced datasets are publicly available at the time of writing. Many improvements still have to be made before it is reliable enough to use as a pre-screening and diagnosis tool.

5 Conclusion and further work

Our extensive comparative analysis of 15 audio features from different domains has provided significant insights into ML feature selection in the context of COVID-19 respiratory sound classification and addressed the research questions laid out in Section 3.1. As the analysis found recurring patterns of predictive efficiency across two completely independent datasets, we have identified a feature ranking and salient feature characteristics that are likely inherent to COVID-19 respiratory signals rather than the underlying datasets. These findings could be beneficial for future sound-based COVID-19 classification applications.

[3] The diagnostic process of listening to internal body sounds, often with a stethoscope.

Throughout our analysis, we have introduced new training features that were not considered in the baseline evaluations presented in the datasets' papers. Consequently, we have improved the classification results by almost 17% and 10% on the Cambridge and Coswara datasets, without significant discrepancies or differences in the evaluated ML models.

Although this paper has provided a starting point for the holistic evaluation of respiratory audio features, there are still other opportunities to further analyse other relevant aspects. For example, a comprehensive strategy to regularise different sample lengths and preserve temporal information could benefit COVID-19 classification. Additionally, advanced models s.a. Deep Learning should be used as a basis for further feature ranking analysis, as the more complex architecture could reveal thus far hidden relevance of the evaluated audio features.

Although sound-based COVID-19 detection is the primary purpose of this research, many other respiratory diseases and disorders could also benefit from the development and improvement of automatic audio detection systems for diagnosis, treatment, and management purposes. Therefore, the approach described in this paper could be generalised for the detection of other respiratory diseases.

Acknowledgements

We would like to thank Chris Watkins for the stimulating discussions, and University of Cambridge for access to the COVID-19 sound dataset. This research is funded by Brighton's Connected Futures and Radical Futures' initiatives.

References

1. Amrulloh, Y., Abeyratne, U., Swarnkar, V., Triasih, R.: Cough sound analysis for pneumonia and asthma classification in pediatric population. In: 2015 6th International Conference on Intelligent Systems, Modelling and Simulation. pp. 127–131. IEEE (2015)
2. Aykanat, M., Kılıç, Ö., Kurt, B., Saryal, S.: Classification of lung sounds using convolutional neural networks. EURASIP Journal on Image and Video Processing **2017**(1), 1–9 (2017)
3. Brown, C., Chauhan, J., Grammenos, A., Han, J., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P., Mascolo, C.: 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. pp. 3474–3484 (2020)
4. Eyben, F., Wöllmer, M., Schuller, B.: openSMILE: The Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 1459–1462 (2010)
5. Hossan, M.A., Memon, S., Gregory, M.A.: A novel approach for MFCC feature extraction. In: 2010 4th International Conference on Signal Processing and Communication Systems. pp. 1–5. IEEE (2010)

6. Humphrey, E.J., Cho, T., Bello, J.P.: Learning a robust tonnetz-space transform for automatic chord recognition. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 453–456. IEEE (2012)
7. Jiang, D.N., Lu, L., Zhang, H.J., Tao, J.H., Cai, L.H.: Music type classification by spectral contrast feature. In: Proceedings. IEEE International Conference on Multimedia and Expo. vol. 1, pp. 113–116. IEEE (2002)
8. Korzeniowski, F., Widmer, G.: Feature learning for chord recognition: The deep chroma extractor. In: Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR). pp. 37–43. International Society for Music Information Retrieval (ISMIR), New York City, USA (2016)
9. Laguarta, J., Hueto, F., Subirana, B.: COVID-19 artificial intelligence diagnosis using only cough recordings. IEEE Open Journal of Engineering in Medicine and Biology **1**, 275–281 (2020)
10. Madhu, N.: Note on measures for spectral flatness. Electronics letters **45**(23), 1195–1196 (2009)
11. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in Python. In: Proceedings of the 14th Python in science conference. vol. 8, pp. 18–25. Citeseer (2015)
12. Mitrović, D., Zeppelzauer, M., Breiteneder, C.: Chapter 3 - Features for content-based audio retrieval. In: Advances in Computers: Improving the Web, Advances in Computers: Improving the Web, vol. 78, pp. 71–150. Elsevier (2010). https://doi.org/https://doi.org/10.1016/S0065-2458(10)78003-7
13. Muguli, A., Pinto, L., Sharma, N., Krishnan, P., Ghosh, P.K., Kumar, R., Ramoji, S., Bhat, S., Chetupalli, S.R., Ganapathy, S., et al.: DiCOVA challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics. arXiv preprint arXiv:2103.09148 (2021)
14. Müller, M., Ewert, S.: Chroma toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In: Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), 2011. hal-00727791, version 2-22 Oct 2012. Citeseer (2011)
15. Müller, M., Kurth, F., Clausen, M.: Audio matching via chroma-based statistical features. In: ISMIR. vol. 2005, p. 6 (2005)
16. Nguyen, K.A., Luo, Z.: Cover your cough: Detection of respiratory events with confidence using a smartwatch. In: Conformal and Probabilistic Prediction and Applications. pp. 114–131. PMLR (2018)
17. Panagiotakis, C., Tziritas, G.: A speech/ music discriminator based on RMS and zero-crossings. IEEE Transactions on multimedia **7**(1), 155–166 (2005)
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python. the Journal of machine Learning research **12**, 2825–2830 (2011)
19. Peeters, G., Giordano, B.L., Susini, P., Misdariis, N., McAdams, S.: The timbre toolbox: Extracting audio descriptors from musical signals. The Journal of the Acoustical Society of America **130**(5), 2902–2916 (2011)
20. Peng, P., He, Z., Wang, L.: Automatic classification of microseismic signals based on MFCC and GMM-HMM in underground mines. Shock and Vibration **2019** (2019)
21. Rizal, A., Hidayat, R., Nugroho, H.A.: Signal domain in respiratory sound analysis: Methods, application and future development. Journal of Computer Science **11**(10), 1005 (2015)

22. Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S.R., Ghosh, P.K., Ganapathy, S., et al.: Coswara–a database of breathing, cough, and voice sounds for COVID-19 diagnosis. arXiv preprint arXiv:2005.10548 (2020)
23. Stolar, M.N., Lech, M., Stolar, S.J., Allen, N.B.: Detection of adolescent depression from speech using optimised spectral roll-off parameters. Biomedical Journal **2**, 10 (2018)