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 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 accuracies 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].

It has long been common knowledge that respiratory diseases physically alter the respiratory environment in a way that often induces audible changes [19]. Consequently, manual auscultation[1] is a common method to identify and diagnose respiratory disorders. However, many abnormalities can affect only subtle changes in auditory cues, making the inherently subjective manual auscultation process error-prone even when performed by a trained medical professional [2].

To counteract subjectivity, automated audio classification approaches with promising results have become more and more common in recent years [1,2,8].

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

One of the main limiting factors is the availability of ground truth data which can be difficult to obtain, is prone to representing limited diversity in the underlying population, and requires medical training to label correctly.

Because COVID-19 detection and classification are widespread and critical problems, multiple universities and research institutions have published COVID-19 audio datasets [3,21]. This offers a unique opportunity to develop and verify classification solutions on independently collected samples from a diverse population. The datasets have supported the development of a variety of applications with Machine Learning (ML) audio classification. However, at the time of writing, none have 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 improve COVID-19 audio classification by exploring and optimising the audio signal's representation for ML. This is achieved by giving a holistic overview, evaluation, and ranking of 15 audio features in the context of binary COVID-19 audio classification.

1.1 The paper's contributions

The findings presented in this paper are directly relevant to the binary COVID-19 respiratory audio classification task and can benefit future implementations using the same approach. The following contributions are made:

i. *Audio feature analysis and ranking.* We perform an extensive comparative analysis and ranking of 15 sound features prevalent in speech and non-speech classification to optimise the audio signal's representation. The evaluation is carried out on two independent datasets, allowing the findings to be generalised.
ii. *Highlighting effective features.* We identify sound-based ML features with strong discriminative performance that go against common rules of thumb.
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 see an increase in classification accuracy of up to 17% by incorporating new training features based on our feature ranking.

2 Audio features overview

Feature engineering is a vital step in any ML application, as a model's predictive efficiency relies directly on the discriminative information encoded in the input vectors. We provide a detailed overview and intuition of 15 audio features from a variety of signal domains before delivering a comprehensive comparison in the context of COVID-19 audio classification (see 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

Table 1: *Audio feature selection.* The 15 audio features 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.
	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.

meaningful to humans, they are commonly included in audio feature sets 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 occur due to a buildup of fluid or secretions in the throat and lungs [19]. The selected features have been previously used for COVID-19 classification [3,21].

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 [15].

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

ii. *Zero-crossing rate (ZCR).* The rate of a signal's sign change over time is given by Equation (2). Here x_n is the signal's amplitude at frame n (N frames overall), and $sign(a)$ returns 1 if $a > 0$, 0 if $a = 0$, and -1 otherwise [15].

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

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, or 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 caused by a respiratory disease [19]. A subset of the following features has previously been used for COVID-19 detection [3,21].

i. *Spectral bandwidth (S-BW).* Also referred to as spectral spread, S-BW describes a signal's energy concentration around the centroid. Equation (3) defines bandwidth as the variance around a signal's expected frequency E given energy P_k and corresponding frequency f_k in $1 \leq k \leq K$ subbands [17].

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

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 [22].

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

iii. *Spectral contrast (S-CONT).* An audio signal's contrast is evaluated by comparing spectral energy peaks P_k and valleys V_k in each frequency band 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 [6].

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

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. $1 \leq k \leq K$ [9].

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

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 [22].

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

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 proportion S of total energy, see Equation (8). P_k is spectral energy in one of K frequency subbands [22].

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

2.3 Time-frequency domain

This feature category illustrates a signal's frequency-related information as it transitions 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 non-linear Mel-frequency Cepstrum (MFC), as it is ubiquitous in audio 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 information. MFC features have been previously used for COVID-19 detection [3,12].

i. *Mel-frequency cepstral coefficients (MFCC).* Derived from the MFC power spectrum, a signal is converted into the time-frequency domain by discrete cosine transform in Equation (9). K is the number of coefficients and $s(k)$ calculates the logarithmic energy of the k-th coefficient at frame n [18].

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

ii. *MFCC-Δ.* MFCC's first-order derivative, referred to as velocity, represents temporal change [4] and is often included due to its low extraction cost.
iii. *MFCC-Δ^2.* Similarly to MFCC-Δ, the second-order derivative, acceleration, is commonly included because it can improve audio classification [4].

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 [13]. Two feature groups are considered, distinguished by their underlying representation: Chroma features (chromagram) and Tonnetz (lattice graph). While the Tonnetz (tone-network) encodes tone quality and height, chroma omit interval information. A common consequence of respiratory diseases is a narrowing of airways by secretions. The effect is a wheeze because the pitch of in- and expiration is altered [19], which can be heard in COVID-19 audio recordings.

i. *Chroma energy normalised (C-ENS).* A chroma abstraction by considering short-time statistics within the chroma bands. Normalisation of the feature vector makes it resistant to dynamic variations, such as timbre [13].
ii. *Constant-Q chromagram (C-CQT).* Chroma are extracted from a time-frequency representation of audio via a filter bank. The constant-Q transform (CQT) is employed, which has a good resolution of low frequencies [7].
iii. *Short-time Fourier Transform chromagram (C-STFT).* The extraction process is similar to C-CQT. The difference lies in the audio signal's initial transformation, in this case via the Short-time Fourier Transform (STFT) [7].
iv. *Tonnetz (TN).* A lattice graph of harmonic information. Distances between points become meaningful by encoding pitch as geometric areas [5].

[2] Pitch classes of the equal tempered scale, prominent in Western tonal music [13].

3 Experimental method and results

To make the findings generalisable, the selected 15 audio features are ranked based on the empirical results and analysis on two independent datasets. The assumption is that any patterns repeated across both are likely inherent to the COVID-19 respiratory 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?

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 healthy and COVID-positive cough and breath recordings. The data used in this paper is a curated set of 48kHz WAV file samples collected during April and May 2020.

The Indian Institute of Science has collected shallow and deep breath and cough recordings in the *Coswara dataset* [21]. Samples with a compatible format collected between April and December 2020 are considered. For consistency with the Cambridge dataset, we filter for COVID-positive and healthy participants.

3.3 Feature engineering

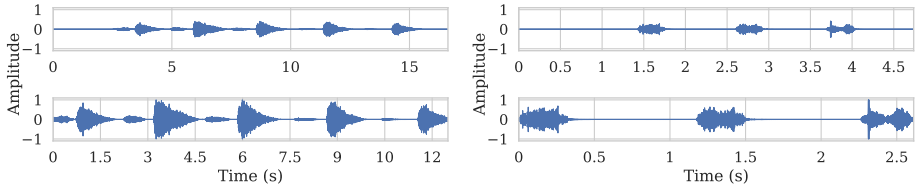
Cleaning the audio data is especially important because the recording devices and environments were not controlled. The pre-processing steps include sampling the audio at 48kHz, converting the signal to mono, trimming the leading/trailing silences, and normalising the amplitude to $[-1, 1]$ (see Figure 1 for the effects). The Python-toolkit `librosa` [10] (version 0.8) was used for the signal processing.

The basis of all of our evaluations are the 15 audio features from three signal domains identified in Section 2 and listed in Table 1. 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 extracted per frame, summary statistics are taken to capture all of the available information. This leads to a feature vector with consistent dimensions, regardless of the underlying sample's length. The extracted statistics 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 feature number is 812, as detailed in Table 3. Large feature dimensions bring a risk of overfitting, however, only a small subset is considered at a time for feature evaluation and ranking.

3.4 Results description and analysis

The paper's main contribution is an extensive analysis and ranking of 15 audio features for COVID-19 classification. We identify informative features by evaluating two datasets in parallel: the Cambridge and the Coswara datasets. Due to their independence, we propose that any recurring patterns in predictive efficiency are likely independent of the underlying dataset, and that the identified features should be strongly considered for future ML COVID-19 audio classification applications. The 15 audio features summarised in Table 3 are analysed over the following configurations to provide a 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.



(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.

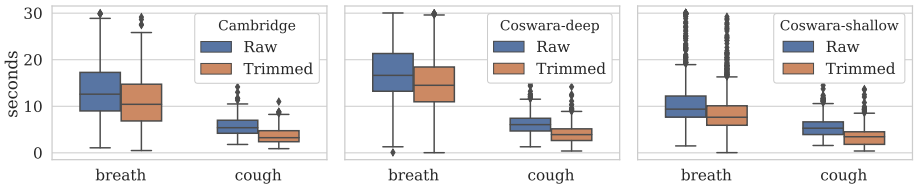


Fig. 2: *Sample lengths before and after pre-processing.* By trimming the leading and trailing silences at 60dB (empirically identified cutoff point) we can remove non-discriminative data. Sample lengths are reduced by 1–3 seconds on average.

Table 3: *Feature dimensions.* 812 features are considered. 7 Summary statistics (min, max, mean, median, var, Q_1, and Q_3) are taken across frames to ensure consistent vector dimensions, regardless of the sample's length (1–30s). To reduce the risk of overfitting, small feature subsets are considered at a time for ranking.

Feature	**Name**	**Count**	**Total** ($\times 7$)
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

iii. 5 ML models, selected for the variety in which they partition the label space. The models are implemented with the `scikit-learn` [16] package version 0.24, and optimised with parameter grid searches[3]: AdaBoost with Random Forest (ADA), K-Nearest Neighbours (KNN), Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM).

Given the datasets' imbalance, 5-fold Cross-Validation (CV) is employed to ensure the results' reliability. We select 3 metrics to compare the features' impact on the audio classification task at hand: *Receiver Operating Characteristic* (ROC), *Precision* (P), and *Recall* (R), see Figure 3 for a brief overview.

[3] ADA: criterion, depth, number of estimators. KNN: K, weights. LR: C, penalty, solver. RF: criterion, depth. SVM: C, γ, kernel.

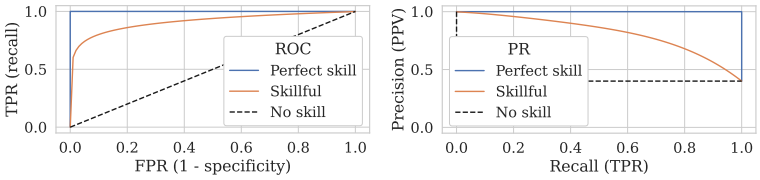
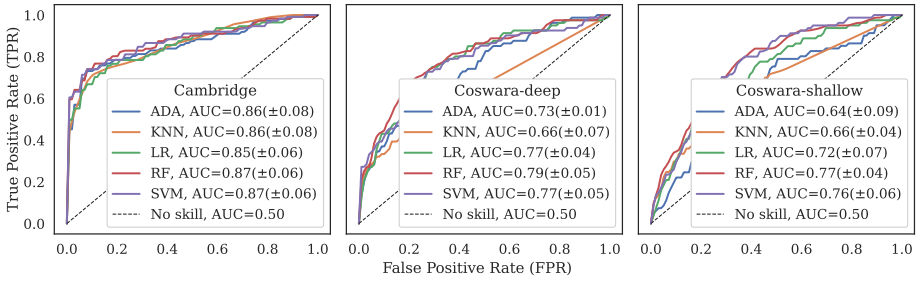
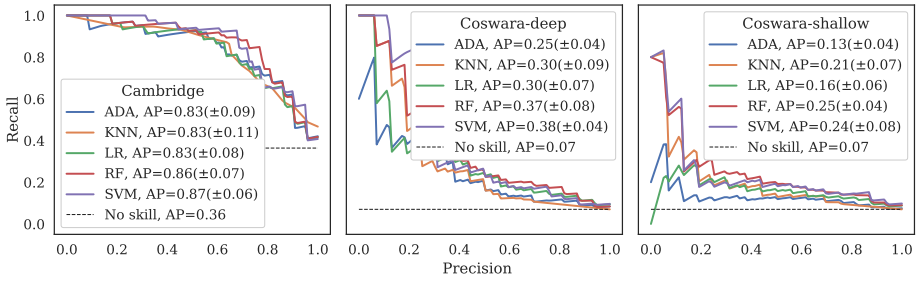


Fig. 3: *Intuition of the considered metrics.* In addition to ROC, PR curves are a valuable tool for evaluating imbalanced dataset because they counteract ROC's optimism by omitting true negatives [20]. PR no-skill classifiers correspond to the dataset's positive sample ratio (i.e. precision at threshold 0.0).

Feature categories. An initial overview of the full feature vectors shows promising results, as most models outperform their no-skill equivalent. Figure 4 visualises the mean ROC and PR curves on the 'breathcough' vector for each of the models. It clearly establishes SVM and RF outperforming their counterparts across all configurations, with a similar trend observed for 'breath' and 'cough'.



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



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

Fig. 4: *'Breathcough' results.* 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 PR-curves, the unskilled classifier corresponds to the dataset's positive label ratio.

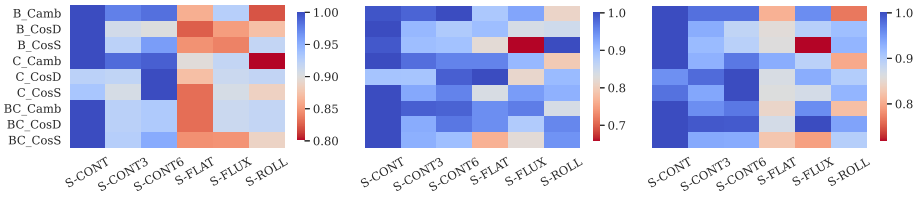
Table 4: *'Breathcough' 5-fold CV ROC-AUC results.* The mean μ and standard deviation σ are reported for four signal domains (see Table 1 for details). SVM and RF consistently achieve the highest accuracies. The feature categories can be ranked in the following increasing order: time domain, tonal, spectral, 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

Even though the Cambridge and Coswara datasets have similarly shaped ROC curves, the former has the best PR curves (i.e. best Average Precision or AP). This illustrates ROC's optimism on imbalanced datasets, justifying our choice of metrics. An influential factor in Coswara's lower overall accuracies is the greater imbalance of COVID-positive samples at 13:1 vs 2:1 in the Cambridge data (see Table 2). Nonetheless, models trained on the Coswara datasets perform noticeably better than an unskilled classifier with AP scores between 13–38% compared to the unskilled 7% (positive sample ratio), see Figure 4b.

Table 4 confirms our selection of SVM and RF as the best-performing models. It shows the same 'breathcough' feature vector's predictive efficiency, but this time considering one signal domain at a time. Apart from two exceptions, 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 note that the 4 feature categories can be broadly ranked in the following order of increasing predictive efficiency (Cambridge, Coswara-deep, Coswara-shallow): *time domain* (79%, 64%, 56%), *tonal* (83%, 73%, 69%), *spectral* (85%, 74%, 72%), and *cepstral* (87%, 76%, 71%). As evidenced by the results, the spectral and cepstral categories achieve similarly high accuracies. More noteworthy is that the same ranking is prevalent for all 5 considered ML models, leading to the conclusion that the cepstral and spectral feature categories encode particularly informative data contained in breathing and coughing signals for COVID-19 classification.



(a) SVM. (b) RF. (c) LR.

Fig. 5: *Normalised ROC-AUC scores of top spectral features for breath ('B'), cough ('C'), and breathcough ('BC').* S-CONT's high performance is achieved because its sub-features (e.g. 3, 6) consistently outperform other spectral features, not just because it is the only composite spectral feature (7-D).

Individual features. Turning our attention to individual features, we focus initially on the best-performing SVM classifier before broadening to include all models, letting us identify general patterns of predictive efficiency. The results forming the basis of our analysis are available in Table 5.

The majority of the 15 features significantly outperforms random guesses for the COVID-19 classification task across all datasets and sample types 'breath' ('B'), 'cough' ('C'), 'breathcough' ('BC'). The lowest accuracies are achieved by Coswara-shallow, matching previous findings, both overall and in individual feature categories. Comparing the underlying sample types further underlines the similarities between the Cambridge and Coswara-deep datasets: 'BC' achieves the highest mean ROC-AUC scores on average (except for time domain features), whereas Coswara-shallow is split evenly between 'B' (time domain, spectral) and 'C' (cepstral, tonal). However, given all considered features in a single feature vector, the Coswara-shallow dataset still shows its highest accuracy on 'BC' samples since cepstral and tonal features are very influential overall.

MFCC (cepstral), S-CONT (spectral), and C-ENS/C-STFT (tonal) are the highest-scoring features in their categories, whereas the time domain is more variable. Although S-CONT is the only spectral composite feature (7-D), Figure 5 clearly shows individual sub-features outperforming most other spectral features. We conclude that S-CONT's high COVID-19 classification accuracy is based on informative sub-features rather than just its increased dimensionality.

Lastly, we note a surprising trend for MFCC. A prevalent rule of thumb suggests 12–13 coefficients for audio classification [3,6,18,21]. However, Figure 6 shows that higher-order features provide discriminative information for the identification of COVID-19 on par with (Coswara-deep) or significantly outperforming (Cambridge) lower orders. This phenomenon is most noticeable in the 'BC' and 'B' features and MFCC's derivatives. Since higher-order features contain information about fine details such as pitch and tone quality [11], we extrapolate that timbral information is very relevant to COVID-19 audio classification.

Discussion. Our extensive analysis, comparison, and ranking of 15 features has found recurring patterns of predictive efficiency for COVID-19 audio

Table 5: *5-fold CV ROC-AUC mean μ and standard deviation σ.* 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: time domain, tonal, spectral, and cepstral. The same pattern can be found across datasets and models.

(a) SVM results on the Cambridge dataset.

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) SVM results on the Coswara-deep dataset.

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

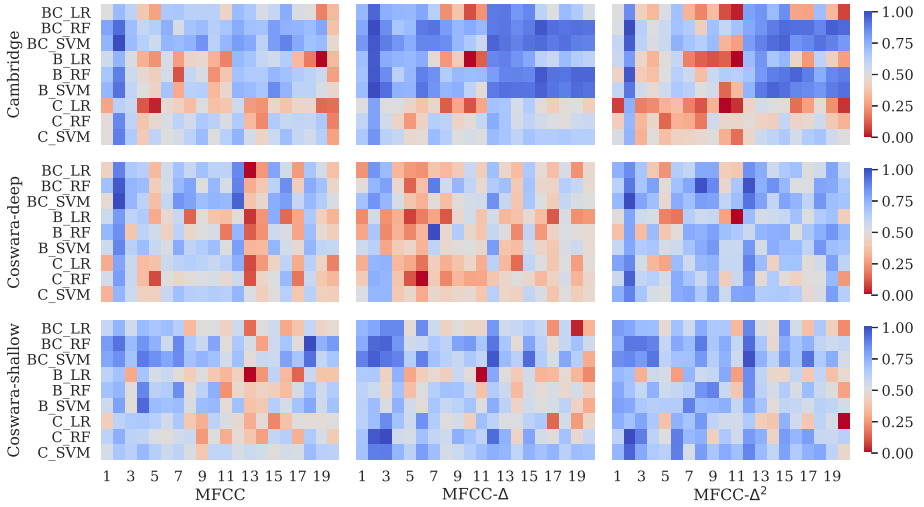


Fig. 6: *Normalised ROC-AUC of MFCC and derivatives for 'breathcough' (BC), 'breath' (B), and 'cough' (C).* Contrary to a common rule of thumb [3,6,18,21], 13+ features provide significant discriminatory data, and shows that timbral information is especially relevant to COVID-19 classification.

classification across independent datasets. There is a distinct category ranking consistent across models, sample types, and datasets (increasing): time domain, tonal, spectral, and cepstral. Contrary to the intuitive expectation, some 'complex' categories provide less discriminative information than 'simpler' ones (e.g. tonal/spectral features). However, this is justified when considering that tonal features describe pitch and so are more suited to tasks with melodic content.

The ranking underlines the significance of frequency-based features by elevating the spectral and cepstral categories describing timbral aspects and tone quality/colour. We have also shown that the common guideline to use only the first 13 MFCC features [3,6,18,21] is not applicable to COVID-19. Indeed, the higher-order (timbre) features' predictive efficiency provides significantly more discriminatory information, especially for the 'BC' and 'B' 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 'BC' feature vector outperforms the individual 'B' and 'C' vectors in most cases.

Given our insights, we compare our ML results to the published baselines, summarised in Table 6. The evaluated models are of similar type and complexity; The major difference is our introduction of new training features. We can see that our improved feature vectors significantly outperform both the Cambridge and Coswara baseline accuracies by 10–17%, validating our feature selection.

Table 6: *Comparison to dataset papers' 5-fold CV baseline results.* We select the most comparable configuration (feature pre-processing and classification model).

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
[12]	Cos-Unknown	C	RF	67.45	—	—	—	—	—

4 Related work

During in- and exhalation, air travelling through the respiratory tract undergoes turbulence and produces sounds. Consequently, any physical changes to the airways or lungs (e.g. caused by diseases such as COVID-19) also alter the produced respiratory sounds [19]. Even though listening and evaluating lung sounds manually is inherently subjective, medical professionals have long used this technique to diagnose a wide variety of respiratory diseases non-invasively [2].

The popularisation of digital signal processing techniques and Machine Learning (ML) have made the automatic classification of respiratory sounds possible as a less subjective, low-cost, and patient-friendly (pre-)screening method. A literature review of existing implementations shows that ML can reliably pick up on subtle cues in audio signals for a variety of diseases.

Smartwatches and wearable devices have made audio monitoring for healthcare purposes feasible. Nguyen et al. apply a dynamically activated respiratory event detection mechanism to detect cough and sneeze events non-intrusively [14]. [1] presents classifiers distinguishing between asthma and pneumonia in pediatric patients. Lastly, an image classification solution with comparable results is developed in [2], using spectrograms as the input.

One of the first COVID-19 audio datasets containing breath and cough samples was presented in [3]. Using standard ML and audio processing techniques, the authors report 71% ROC accuracy for COVID classification. [12] and [21] consider further recording types such as vowel intonation and sequence counting, achieving 67% and 66% accuracy with ML models respectively.

5 Conclusion and future work

Our extensive comparative analysis of 15 audio features has provided significant insight into ML feature selection for COVID-19 respiratory audio classification and addressed the research questions laid out in Section 3.1. Primarily, we identify the most informative feature characteristics and verify their ranking across two independent datasets. Since the two feature rankings show considerable overlap, we conclude that the features' relative salience is likely inherent to the respiratory signals rather than the evaluated datasets.

Throughout our analysis, a number of informative audio features are newly incorporated in the context of COVID-19 classification. In combination with our feature ranking, we achieve 88% and 77% accuracy on the Cambridge and Coswara datasets. Since the complexity of the signal processing and ML models is comparable to the baselines, the increase of up to 17% and 10% respectively is a consequence of our feature selection. Our established feature ranking can benefit future sound-based COVID-19 classification applications.

This paper provides a starting point for the holistic evaluation of respiratory audio features for COVID-19 classification. Considerations that could be addressed in future work are a comprehensive strategy to regularise different sample lengths and to identify the most informative audio features for complex architectures such as Deep Learning neural networks.

Although sound-based COVID-19 detection is the primary purpose of this research, many other respiratory diseases and disorders could benefit from the development and improvement of automatic audio detection systems for diagnosis, treatment, and management. 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 University of Brighton's Connected Futures, Radical Futures' initiatives, and Santander's Global Challenges Research grant.

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. 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)
5. 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)

6. 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)
7. Korzeniowski, F., Widmer, G.: Feature learning for chord recognition: The deep chroma extractor. In: Proceedings of the 17th ISMIR Conference. pp. 37–43. International Society for Music Information Retrieval (ISMIR), New York, USA (2016)
8. 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)
9. Madhu, N.: Note on measures for spectral flatness. Electronics letters **45**(23), 1195–1196 (2009)
10. 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)
11. 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)
12. 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)
13. Müller, M., Kurth, F., Clausen, M.: Audio matching via chroma-based statistical features. In: ISMIR. vol. 2005, p. 6 (2005)
14. 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)
15. Panagiotakis, C., Tziritas, G.: A speech/ music discriminator based on RMS and zero-crossings. IEEE Transactions on multimedia **7**(1), 155–166 (2005)
16. 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)
17. 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)
18. 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)
19. 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)
20. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one **10**(3), e0118432 (2015)
21. 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)
22. 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)