Towards Open Respiratory Acoustic Foundation Models: Pretraining and Benchmarking

Yuwei Zhang[†1], **Tong Xia**[†1], **Jing Han**[1], **Yu Yvonne Wu**[1], **Georgios Rizos**[1], **Yang Liu**[1], **Mohammed Mosuily**[2], **Jagmohan Chauhan**[2], **Cecilia Mascolo**[1]

[1] **University of Cambridge,** [2] **University of Southampton, UK**
† joint first authors, equal contribution
{yz798, tx229}@cam.ac.uk

Abstract

Respiratory audio, such as coughing and breathing sounds, has predictive power for a wide range of healthcare applications, yet is currently under-explored. The main problem for those applications arises from the difficulty in collecting large labeled task-specific data for model development. Generalizable respiratory acoustic foundation models pretrained with unlabeled data would offer appealing advantages and possibly unlock this impasse. However, given the safety-critical nature of healthcare applications, it is pivotal to also ensure openness and replicability for any proposed foundation model solution. To this end, we introduce OPERA, an OPEn Respiratory Acoustic foundation model pretraining and benchmarking system, as the first approach answering this need. We curate large-scale respiratory audio datasets (∼136K samples, 440 hours), pretrain three pioneering foundation models, and build a benchmark consisting of 19 downstream respiratory health tasks for evaluation. Our pretrained models demonstrate superior performance (against existing acoustic models pretrained with general audio on 16 out of 19 tasks) and generalizability (to unseen datasets and new respiratory audio modalities). This highlights the great promise of respiratory acoustic foundation models and encourages more studies using OPERA as an open resource to accelerate research on respiratory audio for health. The system is accessible from `https://github.com/evelyn0414/OPERA`.

1 Introduction

Respiratory audio, such as coughing and breathing sounds generated by the respiratory system's airflow, contains multiple physiological characteristics of individuals and therefore its modeling could be instrumental in health monitoring and disease detection applications [49, 59]. For instance, audio recordings can be used to estimate respiratory rate and lung function [14, 53, 71], detect snoring and apnea events during sleep [37, 27, 52], assess the effect of smoking on health [43, 42] and diagnose diseases like flu and asthma [39, 36, 28, 50].

To enable the widespread adoption of these applications, high-performing algorithms are needed. Related studies rely on traditional signal processing methods [14, 53, 37, 27, 43, 39, 36], which require domain knowledge and often exhibit limited performance. Supervised deep acoustic models

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

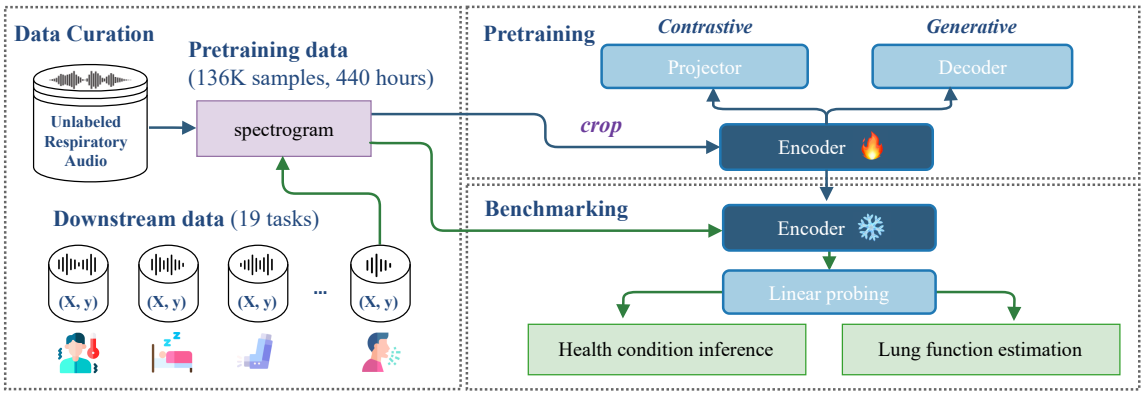


Figure 1: System overview of OPERA. After data curation, respiratory acoustic foundation models (Encoder) are pretrained and then evaluated on various downstream health tasks.

have been proposed [71, 28, 60] but their performance heavily depends on the volume and quality of available labels, which might be difficult and expensive to collect. Hence, foundation models pretrained with large unlabeled respiratory audio data have a high potential to improve performance through transfer learning and supervised fine-tuning [13, 68]. However, in contrast with other health data modalities like clinical imaging [48], electronic health records (EHRs) [66], and medical time series [74, 1, 15], foundation models for respiratory audio are largely under-explored.

Respiratory audio datasets are available but no comprehensive collection has been curated. Recent years have seen an ever-increasing accumulation of respiratory audio [69, 47, 12], exhibiting heterogeneous properties such as varying acquisition modalities and sampling rates. These datasets exhibit significant potential for acoustic model development and evaluation. However, no existing effort has curated such data systematically.

There is no open respiratory acoustic foundation model, impeding the field's growth and understanding. Existing open-source acoustic models like AudioMAE [35] and CLAP [17] are pretrained on general audio event datasets such as YouTube audio, containing very few (around 0.3%) respiratory sounds [38, 24]. These models may not be able to effectively capture the subtle nuances of respiratory sounds, which can vary in abrupt bursts, aperiodic components, and frequency distributions, particularly across different health conditions [49]. Although a model pretrained on respiratory sounds has been recently presented [6], it is not open-source, making it hard to analyze, replicate, or compare its workings. The insights on how to effectively train respiratory acoustic foundation models also remain limited.

There is no ready-to-use benchmark for respiratory audio research. Current task-specific studies evaluate their models on purposely collected datasets, leaving the models' generalizability to other tasks unclear [6]. A benchmark that combines multiple public datasets across diverse applications to enable fair and comprehensive evaluations of the developed foundation models is essential but currently lacking. This is crucial for safety-critical health applications, where models must be rigorously evaluated before use [67, 64].

To mitigate these gaps, in this paper, we put forward *OPERA*, an **OPE**n **R**espiratory **A**coustic foundation model pretraining and benchmarking system (Figure 1). It curates unlabeled respiratory audio datasets, pretrain three pioneering foundation models, and evaluates them against existing pretrained acoustic models across various applications. Specifically, our contributions are:

- We curate a unique large-scale (∼136K samples, 440 hours), multi-source (5 datasets), multi-modal (breathing, coughing, and lung sounds) and publicly available (or available on request) respiratory audio dataset for foundation model pretraining, orders of magnitude larger than the number of respiratory audio samples in datasets used for training existing open acoustic models.

- We pretrain 3 foundation models with the curated unlabeled data using the most common self-supervised approaches (a contrastive learning-based transformer, a contrastive learning-based CNN model, and a generatively pretrained transformer) to study the effect of the training designs.
- We employ 10 labeled datasets (6 not covered by pretraining) to formulate 19 respiratory health tasks (12 in health condition inference and 7 in lung function estimation), ensuring fair, comprehensive and reproducible downstream evaluation.
- We benchmark the performance of our 3 foundation models, one commonly used acoustic feature set, and 3 open pretrained acoustic models on these tasks as a starting point for future exploration.

Extensive experiments demonstrate that our foundation models outperform the models pretrained with general audio on 16 out of 19 benchmark tasks, confirming the power and promise of dedicated respiratory acoustic foundation models. Results also show that our models are generalizable across multiple downstream tasks, including new datasets and unseen respiratory audio modalities. This is a critical advancement towards realizing the potential of respiratory sounds as a mainstream technique for health monitoring.

Within our three models, we find that the contrastive pretraining model is better for classification-based downstream tasks, while the generative pretrained model performs better in regression tasks, possibly due to the nature of their training objectives: contrastive learning can capture the nuances of the local patterns to make features distinguishable while generative learning focuses more on global features which are vital for regression. Our transformer models generally outperform the CNN model because they have stronger modeling capability, though requiring more intensive computation. These findings provide insightful guidance to the development and application of such types of models.

In summary, this paper introduces *the first open-source respiratory acoustic foundation model pretraining and benchmarking system*. This represents a critical first step towards comprehensive and reproducible audio foundation models for health: future foundation model research can leverage our system as an experimental resource, and application studies can take advantage of our foundation models as feature extractors. This can facilitate progress in both machine learning and healthcare. These efforts will extend current machine learning capabilities, now able to *see* (via vision) and *read* (via natural languages), to also *listen to* (via audio) our health.

2 Related Work

2.1 Pretraining in Acoustic Modeling

Models pretrained on large-scale datasets have demonstrated great generalizability in diverse downstream tasks, especially when labeled data are limited [8, 16, 25, 35]. For audio-driven health applications, several general audio pretrained models can be used as feature extractors. One widely used model is *VGGish* [30], trained on 5.24 million hours of audio from YouTube videos to predict 30,871 categories of video labels. Other models have been developed for audio event classification tasks [40, 10, 35]. Among them, *AudioMAE* [35] is an open model trained via an auto-encoding objective without requiring any audio labels. Inspired by recent advances in large language models, language-supervised pretraining has also been explored. *CLAP* [17] is an open model pretrained in this manner. We have included these open models in our benchmark.

It is also worth noting that these open models are pretrained on general audio event datasets such as *AudioSet* [24], *FSD50K* [21], and *FreeSound* [22], which contains few samples of respiratory-related audio. For instance, AudioSet's 2 million clips include only 2334 snoring, 871 cough, 834 breathing, and 1200 sneeze clips, making up only 0.3% of the total. In face of this issue, we curate large-scale respiratory audio datasets to pretrain our foundation models for comparison.

In terms of pretraining methods, given the difficulty in collecting large-scale labeled health-related datasets, we consider self-supervised learning (SSL) to leverage unlabelled data for learning meaningful representations [62, 1, 6]. Main SSL methods fall into two categories: contrastive [11, 5, 54] and generative [29, 35, 46]. Contrastive learning trains models to distinguish between similar and dissimilar samples, while generative models are trained to reconstruct original audio data or features

from masked or corrupted versions. Since they have been demonstrated to be effective in general audio, We implement both methods in our system.

A recent work, *HeAR* [6], curated millions of respiratory audio clips from YouTube videos to pretrain a foundation model using a generative SSL approach. However, neither the data nor the model are publicly available, resulting in a lack of transparency and reproducibility. Limited exploration has been conducted on the reasoning behind the chosen SSL method for various downstream tasks. Our work investigates, for the first time, open pretraining respiratory acoustic foundation models to provide a better understanding of their limits and their potential.

2.2 Benchmarks in Respiratory Audio-based Applications

Current respiratory audio-based health studies typically evaluate their developed models using their self-formulated protocols[6, 70, 72], instead of following a uniform evaluation pipeline. This leads to weak reproducibility due to several challenges [28]: lack of implementation details or released code, absence of reliable training and testing division, and varying implementation frameworks (e.g., some in TensorFlow [28] while other in PyTorch [4]) making them difficult to compare.

High-quality benchmarks are essential in machine learning to ensure advancements are reliable and applicable to real-world problems. While several benchmarks exist for pretrained representation models on general audio event detection and speech recognition [63, 56, 26, 73], similar benchmarks are missing in respiratory audio for health, despite their equal importance. The only related benchmark [32] in this area compares supervised models for breath phase and adventitious sound detection using a single dataset, and is thus not applicable for evaluating foundation models. A comprehensive benchmarking effort of respiratory acoustic foundation models is lacking but has the potential to really shed light on the power of these techniques in the context of respiratory health tasks.

3 System Overview

As shown in Figure 1, OPERA comprises three main components: data curation (including unlabeled data for pretraining and labeled data for evaluation), general-purpose pretraining to develop acoustic foundation models (Encoder), and a benchmark comparing the pretrained models on various downstream tasks.

In OPERA, we employ five datasets for pretraining and ten datasets for benchmarking. Four of the downstream datasets overlap with the pretraining resources, but we ensure the testing data is held out before pretraining and thus is never seen by the models. During the pretraining step, we build two SSL strategies enabling the use of different encoder architectures. We then use the pretrained models to extract features and apply linear probing to report the performance for downstream tasks. Detailed information about data curation and pretraining methods is elaborated on in Section 4, and the benchmark data curation and evaluation results are summarized in Section 5.

4 Self-supervised Pretraining

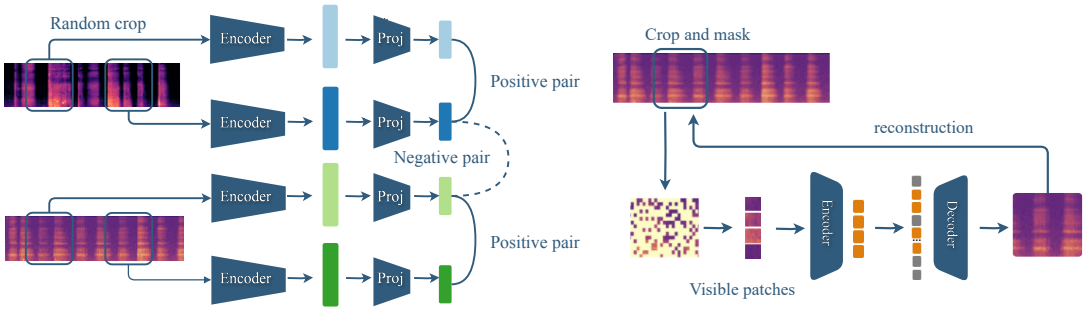
4.1 Pretraining Datasets

Five open data resources are curated in OPERA to enable the training of respiratory acoustic foundation models (Table 1). They were collected by different research institutions using various protocols, and are all publicly available or accessible upon request. Some recordings were made with a microphone near the mouth [69, 12, 47], while others used a digital stethoscope attached to the chest [51, 31]. This allows the pretrained models to see heterogeneous data for better generalizability.

We only include qualified samples (those identified as respiratory audio, not noise) in the pretraining step. Some labeled audio samples from these datasets, which can be used for downstream evaluations, are held out. We then trim the remaining audio recordings by removing the beginning and ending silence to further ensure the quality of the data. The statistics of the data after quality check are summarized in Table 1 (extended description can be found in Appendix A.1). As a result, the entire pretraining dataset consists of 135,944 samples, with a total duration of about 404.1 hours.

Table 1: Statistics of the data used for model pretraining (SR: sampling rate; Duration: mean [95% quantile range]; Crop: cropped length for pretraining).

Data name	Collected by	SR	Modality	#Sample	Duration (s)	Crop (s)
COVID-19 Sounds [69]	Microphone	16∼44.1kHz	Induced cough (3 times)	40866	6.1[2.6∼11.2]	2
			Deep breath (5 times)	36605	20.5[9.7∼31.6]	8
UK COVID-19 [12]	Microphone	48kHz	Induced cough (3 times)	19533	4.1[2.1∼9.2]	2
			Exhalation (5 times)	20719	7.7[4.2∼15.6]	4
COUGHVID [47]	Microphone	48kHz	Induced cough (up to 10s)	7179	6.9[2.4∼9.9]	2
ICBHI [51]	Stethoscope	4∼44.1kHz	lung sound (several breath cycles)	538	22.2[20.0∼65.9]	8
HF LUNG [31]	Stethoscope	4kHz	lung sound (several breath cycles)	10554	15.0[15.0∼15.0]	8



(a) Contrastive (OPERA-CT, OPERA-CE) (b) Generative (OPERA-GT)

Figure 2: Self-supervised learning methods used in our system.

Before pretraining, all recordings are resampled to 16 kHz and merged into a mono channel. They are then transformed into spectrograms using 64 Mel filter banks with a 64 ms Hann window that shifts every 32 ms [57, 75]. For example, a 4s recording will be converted into a spectrogram of $1 \times 126 \times 64$ dimension. Finally, these spectrograms are used to pretrain our respiratory acoustic foundation models.

4.2 Pretraining Models and Methods

We pre-train our models using a combination of the aforementioned data resources, dividing each dataset into equally-sized batches for consistent processing. We randomly shuffle the batches and reserve 10% for validation. Due to inherent variations in audio length within individual batches, we employ random cropping of spectrograms, with crop lengths specified in Table 1. Considering the unlabeled nature of the pretraining data, we adopt the most representative SSL methods: contrastive learning-based and generative pretraining-based objectives to pretrain our models. The rationale behind this choice is that if an encoder can distinguish the source of audio segments (contrastive) or reconstruct masked spectrograms (generative), it is expected to have encoded useful and generalizable acoustic features. The three foundation models we pretrained are:

- **OPERA-CT**: OPERA-CT is a contrastive learning based [54] transformer model. Two segments from the same spectrogram are regard as a positive pair, otherwise negative pairs. As shown in Figure 2(a), an encoder network (a transformer [10]) extracts features from these segments, and a projector maps them into a low-dimensional representation space, where bilinear similarity is calculated. The optimization objective aims to maximize the similarity between positive pairs and minimize it for negative pairs. The encoder has 31M trainable parameters.
- **OPERA-CE**: Similar to OPERA-CT, CE leverages a contrastive pre-training approach. However, it utilizes a more lightweight and efficient CNN encoder (EfficientNet-B0) [61], which has approximately 4M trainable parameters.
- **OPERA-GT**: OPERA-GT is a generatively pretrained transformer model [3]. As shown in Figure 2(b), the encoder (a vision transformer with 21M trainable parameters) is utilized to extract useful features from masked spectrograms, from which the decoder (a lightweight swin-transformer with 12M trainable parameters) can reconstruct the original spectrograms. To train the encoder and the decoder, spectrograms are cropped to equal lengths and then split into small patches. We randomly mask 70% of patches per spectrogram for reconstruction.

Table 2: Downstream task characteristics grouped by task category. Datasets in grey are entirely new (not used in pretraining), while others have test sets held out unseen. For T13-T19, FVC denotes forced vital capacity (L), FEV1 is the forced expiratory volume in 1 second, and FEV1/FVC refers to the ratio of the two.

Dataset	ID	Task	Modality	#Sam. (#Sub.)	Data Distribution
UK COVID-19 [12]	T1	Covid / Non-covid	Exhalation	2500 (2500)	840 / 1660
	T2	Covid / Non-covid	Cough	2500 (2500)	840 / 1660
COVID-19 Sounds [69]	T3	Symptomatic / Healthy	Breath	4138 (3294)	2029 / 2109
	T4	Symptomatic / Healthy	Cough	4138 (3294)	2029 / 2109
CoughVID [47]	T5	Covid / Non-covid	Cough	6175 (n/a)	547 / 5628
	T6	Female / Male	Cough	7263 (n/a)	2468 / 4795
ICBHI [51]	T7	COPD / Healthy	Lung sounds	828 (90)	793 / 35
Coswara [7]	T8	Smoker / Non-smoker	Cough	948 (n/a)	201 / 747
	T9	Female / Male	Cough	2496 (n/a)	759 / 1737
KAUH [23]	T10	Obstructive / Healthy	Lung sounds	234 (79)	129 / 105
Respiratory@TR [2]	T11	COPD severity	Lung sounds	504 (42)	72 / 60 / 84 / 84 / 204
SSBPR [70]	T12	Body position recognition	Snoring	7468 (20)	1638 / 1454 / 1269 / 1668 / 1439
MMlung [44]	T13	FVC	Deep breath	40 (40)	3.402 ± 1.032 L
	T14	FEV1	Deep breath	40 (40)	2.657 ± 0.976 L
	T15	FEV1/FVC	Deep breath	40 (40)	0.808 ± 0.190 L
	T16	FVC	O Vowels	40 (40)	3.402 ± 1.032 L
	T17	FEV1	O Vowels	40 (40)	2.657 ± 0.976 L
	T18	FEV1/FVC	O Vowels	40 (40)	0.808 ± 0.190 L
NoseMic [9]	T19	Respiratory rate	Breath	1297 (16)	13.915 ± 3.386 bpm

Detailed introduction to these three models can be found in Appendix A.2. We train them for up to 200 epochs and save the best model based on the held-out validation set (i.e., its performance on the pretraining objective). Model checkpoints are also released. More pretraining results and analysis are available in Appendix A.3.

5 Benchmarking

5.1 Benchmark Datasets and Tasks Setup

Tasks. To facilitate the evaluation of our pretrained models, existing acoustic models, and future emerging respiratory acoustic foundation models, we introduce a new benchmark. A total of 10 labeled respiratory audio datasets, encompassing 6 respiratory audio modalities, are curated for this benchmark. Among these 10 datasets, 6 are new and unseen during the pretraining stage.

Using these 10 datasets, we formulate 19 downstream tasks: 12 for health condition inference and 7 for lung function estimation. The first group covers disease detection such as COVID-19 and COPD (Chronic Obstructive Pulmonary Disease), participant attribute inference like smoker and gender, disease severity classification, and body position in sleep monitoring. Tasks 1-10 are binary classification, while Tasks 11-12 involve 5 classes. The second group includes spirometry test performance and respiratory rate estimation, which are regression tasks aimed at predicting continuous values. Data and task statistics are summarized in Table 2, with detailed descriptions and licenses provided in Appendix A.1.

All data in this benchmark are publicly available or under controlled access procedures. When available, we follow the official train-test split (Tasks 1-4 and 12-18); otherwise, we implement a random participant-independent split to ensure realistic evaluation (Tasks 5-11 and 19). Due to the limited number of participants in Tasks 13-19, we employ leave-one-subject-out evaluation. For all other tasks, we adopt a fixed random train-validation-test split.

Baselines. In addition to our pretrained models, we also include a commonly used acoustic feature set and three open pretrained acoustic models in this benchmark. They are **Opensmile** [18] (*Emobase* acoutic feature set), **VGGish** [30] (supervised pretrained), **AudioMAE** [35] (self-supervised pretrained) and **CLAP** [17] (language-supervised pretrained). We consider these four methods as baselines to be distinguished from our pretrained models.

Table 3: Mean reciprocal ranks on task groups (higher is better). The best model within each group is highlighted in pink and the second-best is highlighted in blue.

Task	#	Opensmile	VGGish	AudioMAE	CLAP	**OPERA-CT**	**OPERA-CE**	**OPERA-GT**
All	19	0.2920	0.1882	0.2861	0.3435	0.5576	0.4307	0.4947
Health condition inference	12	0.2073	0.1853	0.2044	0.4014	0.7361	0.4083	0.4500
Lung function estimation	7	0.4371	0.1932	0.4262	0.2442	0.2517	0.4690	0.5714

Evaluation protocol. All tasks are evaluated using the standard linear probe protocol [11, 54, 45]: training a single fully connected layer on top of the representations extracted from the frozen encoder. Linear evaluation focuses on the quality of learned representations and is applicable to some very small datasets. **AUROC** (area under the receiver operating characteristic) is reported for classification (Task 1-12) and **MAE** (mean absolute error) is reported for regression (Task 13-19). For a comprehensive overall evaluation, we report **MRR** (mean reciprocal rank) [58] across tasks.

For baselines, both the data pre-processing and feature extraction strictly follow their official implementation. For our pretrained models, the same audio preprocessing is used as in pretraining. We then segment our audio into short frames to feed into our foundation models to extract features, and use the averaged representation over these frames as the input for the linear layer [35]. An extended description of the implementing details can be found in Appendix A.2. *Note that the baselines and our pretrained models are implemented within the same pipeline, making our results easy to reproduce and our benchmark ready to use.*

5.2 Experimental Results

We report the MRR of different task groups in Table 3, with the detailed reciprocal ranks of all evaluated methods on each task provided in Appendix A.3. The performance metrics for each task are summarized in Table 4 and Table 5. Our benchmark demonstrates reliability, as our implementation of baselines achieves comparable performance to those reported in the literature (e.g., existing cough-based COVID-19 detection studies report an AUROC of about 0.65 [12, 69], aligning with our baseline results in Task 2) Through these extensive experimental results, we now answer the following two main research questions (RQs):

RQ1. Can pretraining a foundation model with diverse unlabeled respiratory audio data lead to better performance than baselines designed for general audio?

From results highlighted in Table 3, it is evident that our pretrained respiratory acoustic foundation models outperform both the acoustic feature set and existing general audio pretrained models. Among them, OPERA-CT and OPERA-GT achieve the highest MMR scores of 0.5576 and 0.4947, respectively. Looking at ✓ and * in Table 4 and 5, the best OPERA model outperforms the acoustic feature set on 18 tasks and the baseline pretrained models on 16 tasks out of the 19 evaluated tasks. This provides a clear positive answer to RQ1. This advantage likely stems from their exposure to *large-scale* and *heterogeneous* respiratory audio data, showing the power and promise of respiratory audio foundation models for health applications.

Now let us dive into the task performance at a finer granularity. For classification, an AUROC exceeding 0.7 is typically desirable to demonstrate the utility of the extracted features [20]. When examining the AUROC in Table 4, OPERA models achieve an AUROC exceeding 0.7 on 6 of the 12 health condition inference tasks (Task 2, 6-7, 9-10, and 12), whereas the best baseline, CLAP, only surpasses this threshold on 4 tasks (Task 7, 9-10, and 12). This indicates that our models better encode health condition-related information from respiratory audio. Regarding lung function estimations (regression tasks), the model needs to capture the global dynamics from the entire audio sample and lower MAE indicates better performance. In Table 5, our pretrained models reduce the error in FEV1 estimation using breathing sounds (Task 14), FVC estimation using vowel sounds (Task 16), FEV1 estimation using vowel sounds (Task 17), and respiratory rate estimation (Task 19), with performance close to baselines on other tasks. Furthermore, OPERA-GT also achieves a lower standard deviation across subjects, suggesting better generalizability and robustness to different subjects, which are of great importance for healthcare applications.

Table 4: AUROC on health condition inference tasks (higher is better). The best model for each task is highlighted. We report mean and standard deviation from five independent runs. ✓ and * indicates superiority over the opensmile feature set and the other pretrained baselines respectively.

ID	Task Abbr.	Opensmile	VGGish	AudioMAE	CLAP	**OPERA-CT**	**OPERA-CE**	**OPERA-GT**	
T1	Covid (Exhale)	0.550 ± 0.015	0.580 ± 0.001	0.549 ± 0.001	0.565 ± 0.001	0.586 ± 0.008	0.551 ± 0.010	0.605 ± 0.001	✓*
T2	Covid (Cough)	0.649 ± 0.006	0.557 ± 0.005	0.616 ± 0.001	0.648 ± 0.003	0.701 ± 0.002	0.629 ± 0.006	0.677 ± 0.001	✓*
T3	Symptom (Breath)	0.571 ± 0.006	0.571 ± 0.003	0.583 ± 0.003	0.611 ± 0.006	0.603 ± 0.005	0.610 ± 0.004	0.613 ± 0.002	✓*
T4	Symptom (Cough)	0.633 ± 0.012	0.605 ± 0.004	0.659 ± 0.001	0.669 ± 0.002	0.680 ± 0.006	0.665 ± 0.001	0.673 ± 0.001	✓*
T5	Covid (Cough)	0.546 ± 0.008	0.602 ± 0.001	0.549 ± 0.005	0.603 ± 0.013	0.609 ± 0.004	0.584 ± 0.003	0.575 ± 0.006	✓*
T6	Gender (Cough)	0.639 ± 0.010	0.608 ± 0.000	0.666 ± 0.002	0.684 ± 0.002	0.801 ± 0.000	0.722 ± 0.004	0.762 ± 0.001	✓*
T7	COPD (Lung)	0.579 ± 0.043	0.605 ± 0.077	0.886 ± 0.017	0.933 ± 0.005	0.855 ± 0.012	0.872 ± 0.011	0.741 ± 0.011	✓
T8	Smoker (Cough)	0.534 ± 0.060	0.507 ± 0.027	0.549 ± 0.022	0.680 ± 0.009	0.685 ± 0.012	0.674 ± 0.013	0.650 ± 0.005	✓*
T9	Gender (Cough)	0.753 ± 0.008	0.606 ± 0.003	0.724 ± 0.001	0.742 ± 0.001	0.874 ± 0.000	0.801 ± 0.002	0.825 ± 0.001	✓*
T10	Obstructive (Lung)	0.502 ± 0.080	0.505 ± 0.110	0.614 ± 0.040	0.703 ± 0.036	0.719 ± 0.018	0.742 ± 0.014	0.700 ± 0.013	✓*
T11	COPD severity (Lung)	0.494 ± 0.054	0.590 ± 0.034	0.510 ± 0.021	0.635 ± 0.040	0.625 ± 0.038	0.683 ± 0.007	0.615 ± 0.019	✓*
T12	Position (Snoring)	0.772 ± 0.005	0.657 ± 0.002	0.649 ± 0.001	0.702 ± 0.001	0.781 ± 0.000	0.769 ± 0.000	0.742 ± 0.001	✓*

Table 5: MAE on lung function estimation tasks (lower is better). Best model per task is highlighted. We report mean and standard deviation across subjects.

ID	Task Abbr.	Opensmile	VGGish	AudioMAE	CLAP	**OPERA-CT**	**OPERA-CE**	**OPERA-GT**	
T13	FVC (Breath)	0.965 ± 0.589	1.545 ± 2.084	1.345 ± 0.792	1.138 ± 0.962	1.606 ± 1.312	1.023 ± 0.854	1.191 ± 0.721	*
T14	FEV1 (Breath)	0.859 ± 0.815	1.738 ± 2.967	1.081 ± 0.720	1.130 ± 0.845	1.459 ± 1.074	0.771 ± 0.752	0.996 ± 0.732	✓*
T15	FEV1/FVC (Breath)	0.194 ± 0.397	0.279 ± 0.629	0.143 ± 0.153	0.178 ± 0.151	0.155 ± 0.155	0.148 ± 0.165	0.155 ± 0.172	✓
T16	FVC (Vowel)	0.724 ± 0.532	0.900 ± 1.377	0.983 ± 0.721	0.710 ± 0.585	1.737 ± 1.041	0.672 ± 0.535	0.593 ± 0.414	✓*
T17	FEV1 (Vowel)	0.605 ± 0.541	1.103 ± 1.466	0.960 ± 0.741	0.838 ± 0.694	1.488 ± 1.005	0.736 ± 0.566	0.561 ± 0.348	✓*
T18	FEV1/FVC (Vowel)	0.179 ± 0.204	0.227 ± 0.301	0.150 ± 0.184	0.276 ± 0.300	0.179 ± 0.127	0.220 ± 0.217	0.245 ± 0.185	✓
T19	Breathing Rate	3.852 ± 1.060	2.611 ± 0.786	2.630 ± 0.832	2.615 ± 0.804	2.567 ± 0.785	2.623 ± 0.831	2.537 ± 0.782	✓*

RQ2. Are the pretrained respiratory acoustic foundation models generalizable to new data?

It is crucial that foundation models can generalize to new and unseen data once developed. In our benchmark, we have 12 tasks formulated from unseen datasets (Task 8-19) and unseen respiratory audio modalities (Task 12, 16-18) not used for pretraining. Notably, our respiratory acoustic foundation models demonstrate good generalization capabilities, achieving the best performance on 5 out of 5 classification tasks and 4 out of 7 regression tasks. They are able to outperform the acoustic feature set and general audio pretrained models which are supposed to exhibit generalizability. Specifically, in Table 4, Task 8-12 all have an AUROC higher than 0.68. Comparing Task 6 and Task 9 with the same prediction target, the performance on unseen data (Task 9) is comparable. Therefore, our foundation models are generalizable, likely due to the minimal assumptions made during SSL pretraining.

RQ3. How to design SSL methods and model architectures of respiratory acoustic foundation models with different applications in mind?

Within the OPERA system, we train foundation models using two different SSL strategies: contrastive and generative. From Table 3, 4, and 5, it can be observed that the models pretrained with a contrastive objective (OPERA-CT, OPERA-C) generally achieve superior performance on classification tasks (i.e., health condition inference), while the generative pretrained models (OPERA-GT and baseline AudioMAE) perform better on regression tasks (i.e., lung function estimation). This finding aligns with the inherent nature of the methods, as contrastive learning's discriminative training goal naturally aligns with the classification objective, and it discards the decoder in the architecture compared to generative models. It is also consistent with prior observations on various vision benchmarks [41].

We also compare CNN and transformer encoder architectures using the same SSL strategy. Overall, our results suggest a strong representation ability of the transformer architecture for audio. Specifically, OPERA-CT performs the best in 7 out of the 12 health condition inference tasks (Figure 15(a)), with a mean reciprocal rank as high as 0.7361 (Table 3). For lung function estimation tasks, OPERA-GT performs the best in 3 out of the 7 tasks (Figure 15(b), with the highest mean reciprocal rank of 0.5714 (Table 3) and achieves the second on health condition inference tasks. As a lightweight CNN model, OPERA-CE also demonstrates satisfactory results, with a mean reciprocal of 0.4690, and performs third and second best in the two groups of tasks respectively(Table 3). This shows the promise of training a lightweight foundation model for efficient computing and on-device learning for resource-constrained scenarios.

Table 6: AUROC (higher is better) for linear probing and finetuning on T4. Best model highlighted.

Method	# Train	AudioMAE	CLAP	OPERA-CT	OPERA-CE	OPERA-GT
Linear	1749	0.659 ± 0.001	0.669 ± 0.002	0.680 ± 0.006	0.665 ± 0.001	0.673 ± 0.001
Fine-tune	1749	0.672 ± 0.039	0.691 ± 0.008	0.710 ± 0.003	0.703 ± 0.003	0.715 ± 0.006
Fine-tune	6648	0.723 ± 0.010	0.723 ± 0.009	0.739 ± 0.008	0.733 ± 0.002	0.735 ± 0.005

6 Conclusion and Future Research Directions

In this paper, we present *OPERA*, the first open-source respiratory acoustic foundation model pretraining and benchmarking system. OPERA offers a unique curated dataset pool, a ready-to-use evaluation portal as well as a thorough analysis of performance across architectures and tasks. We discuss the limitations of our work and how it can serve as a foundation for future explorations:

(1) Studying data-efficient fine-tuning. Section 5 uses linear evaluation with frozen encoders following standard protocols and accommodating limited downstream data (see Table 2). We select some tasks with relatively abundant labeled data to examine fine-tuning performance (details in Appendix A.4). Results for Task 4 are presented in Table 6 . Using the same number of labeled data as in linear probing (1749 samples), all models show improved performance and the three OPERA models achieve an AUROC above 0.7. With more labeled data for fine-tuning (6648 samples), the best OPERA-GT model achieves an AUROC of 0.739. Similarly, OPERA-CT's performance on Task 12 (7468 samples) could be enhanced to 0.994 compared to 0.781 in linear evaluation.
However, most other tasks have a much smaller training set, and thus data efficient large model fine-tuning approaches are desirable. Methods have been proposed in the machine learning literature such as adapter tuning [34], prefix tuning [65], prompt tuning [19], and low-rank adaptation [33]. Yet, they are not designed for audio (spectrograms) or acoustic foundation models. Considering the properties of downstream health-related tasks which often exhibit limited and imbalanced data, novel audio-specific data efficient fine tuning methods need to be explored.

(2) Investigating scaling law in respiratory acoustic foundation models. Recent research on foundation models has uncovered their emergent abilities, largely arising from scaling up pretraining data and model size [55]. It is also interesting to study the scaling laws in respiratory acoustic foundation models. Our benchmark can help to quantify how increasing a model's scale and its training data can significantly enhance performance on downstream tasks. Based on the currently 404 hours of respiratory audio, our OPERA-CT (31M parameters) and OPERA-GT (21M) models surpass the lightweight OPERA-CE model (4M). With the rapid accumulation of respiratory audio datasets [68, 13], more evaluation of scaling laws should be conducted in future.

(3) Exploring novel pretraining strategies for unlabeled health audio. We have pretrained three models (OPERA-CT, OPERA-GT, OPERA-CE) and compared their performance. More configurations in terms of model size, architecture, and pretraining methods could be compared in the future. Among the two representative SSL approaches we adapted for pretraining, there exist limitations: For contrastive learning, defining positive and negative pairs is challenging due to downstream task diversity, and our definitions might not be optimal. In generative pretraining, using alternative objectives to reconstruction might improve performance on discriminative tasks. Combining these methods could be beneficial but presents challenges in balancing objectives, and previous studies suggest simple combinations do not improve performance [3]. Audio data also pose unique challenges like heterogeneous sound types, varying sampling rates and durations, and complex temporal-frequency correlations, requiring tailored solutions to better pretrain and apply the foundation models. OPERA provides a framework for exploring these technical challenges.

By introducing this open-source system, we hope to lay the groundwork for responsible, reliable, and sustainable development of foundation models in respiratory healthcare, paving the way for a healthier future for generations to come.

Acknowledgments and Disclosure of Funding

This work was supported by ERC Project 833296 (EAR), EPSRC Project RELOAD, Nokia Bell Labs through a donation, and the Institute for Life Sciences (IfLS) HEIF Research Stimulus Fund. Y. Z. is additionally supported by the Cambridge Trust Scholarship.

References

[1] S. Abbaspourazad, O. Elachqar, A. Miller, S. Emrani, U. Nallasamy, and I. Shapiro. Large-scale training of foundation models for wearable biosignals. In *The Twelfth International Conference on Learning Representations*, 2023.

[2] G. Altan, Y. Kutlu, Y. Garbi, A. Ö. Pekmezci, and S. Nural. Multimedia respiratory database (respiratorydatabase@ tr): Auscultation sounds and chest x-rays. *Natural and Engineering Sciences*, 2(3):59–72, 2017.

[3] A. Baade, P. Peng, and D. Harwath. Mae-ast: Masked autoencoding audio spectrogram transformer. *arXiv preprint arXiv:2203.16691*, 2022.

[4] S. Bae, J.-W. Kim, W.-Y. Cho, H. Baek, S. Son, B. Lee, C. Ha, K. Tae, S. Kim, and S.-Y. Yun. Patch-mix contrastive learning with audio spectrogram transformer on respiratory sound classification. *arXiv preprint arXiv:2305.14032*, 2023.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[6] S. Baur, Z. Nabulsi, W.-H. Weng, J. Garrison, L. Blankemeier, S. Fishman, C. Chen, S. Kakarmath, M. Maimbolwa, N. Sanjase, et al. Hear–health acoustic representations. *arXiv preprint arXiv:2403.02522*, 2024.

[7] D. Bhattacharya, N. K. Sharma, D. Dutta, S. R. Chetupalli, P. Mote, S. Ganapathy, C. Chandrakiran, S. Nori, K. Suhail, S. Gonuguntla, et al. Coswara: A respiratory sounds and symptoms dataset for remote screening of sars-cov-2 infection. *Scientific Data*, 10(1):397, 2023.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[9] K.-J. Butkow, T. Dang, A. Ferlini, D. Ma, Y. Liu, and C. Mascolo. An evaluation of heart rate monitoring with in-ear microphones under motion. *Pervasive and Mobile Computing*, 100:101913, 2024.

[10] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2022.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[12] H. Coppock, G. Nicholson, I. Kiskin, V. Koutra, K. Baker, J. Budd, R. Payne, E. Karoune, D. Hurley, A. Titcomb, et al. Audio-based ai classifiers show no evidence of improved covid-19 screening over simple symptoms checkers. *Nature Machine Intelligence*, pages 1–14, 2024.

[13] T. Dang, D. Spathis, A. Ghosh, and C. Mascolo. Human-centred artificial intelligence for mobile health sensing: challenges and opportunities. *Royal Society Open Science*, 10(11):230806, 2023.

[14] E. P. Doheny, B. P. O'Callaghan, V. S. Fahed, J. Liegey, C. Goulding, S. Ryan, and M. M. Lowery. Estimation of respiratory rate and exhale duration using audio signals recorded by smartphone microphones. *Biomedical Signal Processing and Control*, 80:104318, 2023.

[15] J. Dong, H. Wu, H. Zhang, L. Zhang, J. Wang, and M. Long. Simmtm: A simple pre-training framework for masked time-series modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[17] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[18] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[19] Y. Fathullah, C. Wu, E. Lakomkin, J. Jia, Y. Shangguan, K. Li, J. Guo, W. Xiong, J. Mahadeokar, O. Kalinli, et al. Prompting large language models with speech recognition abilities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13351–13355. IEEE, 2024.

[20] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[21] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

[22] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra. Freesound datasets: a platform for the creation of open audio datasets. In *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.* International Society for Music Information Retrieval (ISMIR), 2017.

[23] M. Fraiwan, L. Fraiwan, B. Khassawneh, and A. Ibnian. A dataset of lung sounds recorded from the chest wall using an electronic stethoscope. *Data in Brief*, 35:106913, 2021.

[24] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[25] Y. Gong, Y.-A. Chung, and J. Glass. Ast: Audio spectrogram transformer. *Proc. INTERSPEECH*, 2021.

[26] S. Grollmisch, E. Cano, C. Kehling, and M. Taenzer. Analyzing the potential of pre-trained embeddings for audio classification tasks. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 790–794. IEEE, 2021.

[27] M. Halevi, E. Dafna, A. Tarasiuk, and Y. Zigel. Can we discriminate between apnea and hypopnea using audio signals? In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3211–3214. IEEE, 2016.

[28] J. Han, T. Xia, D. Spathis, E. Bondareva, C. Brown, J. Chauhan, T. Dang, A. Grammenos, A. Hasthanasombat, A. Floto, et al. Sounds of covid-19: exploring realistic performance of audio-based digital testing. *NPJ digital medicine*, 5(1):16, 2022.

[29] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[30] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (ICASSP)*, pages 131–135. IEEE, 2017.

[31] F.-S. Hsu, S.-R. Huang, C.-W. Huang, Y.-R. Cheng, C.-C. Chen, J. Hsiao, C.-W. Chen, and F. Lai. A progressively expanded database for automated lung sound analysis: an update. *Applied Sciences*, 12(15):7623, 2022.

[32] F.-S. Hsu, S.-R. Huang, C.-W. Huang, C.-J. Huang, Y.-R. Cheng, C.-C. Chen, J. Hsiao, C.-W. Chen, L.-C. Chen, Y.-C. Lai, et al. Benchmarking of eight recurrent neural network variants for breath phase and adventitious sound detection on a self-developed open-access lung sound database—hf_lung_v1. *PLoS One*, 16(7):e0254134, 2021.

[33] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

[34] Z. Hu, L. Wang, Y. Lan, W. Xu, E.-P. Lim, L. Bing, X. Xu, S. Poria, and R. K.-W. Lee. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[35] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022.

[36] M. A. Islam, I. Bandyopadhyaya, P. Bhattacharyya, and G. Saha. Multichannel lung sound analysis for asthma detection. *Computer methods and programs in biomedicine*, 159:111–123, 2018.

[37] R. Jané, J. Solà-Soler, J. A. Fiz, and J. Morera. Automatic detection of snoring signals: validation with simple snorers and osas patients. In *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No. 00CH37143)*, volume 4, pages 3129–3131. IEEE, 2000.

[38] A. Jansen, J. F. Gemmeke, D. P. Ellis, X. Liu, W. Lawrence, and D. Freedman. Large-scale audio event discovery in one million youtube videos. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 786–790. IEEE, 2017.

[39] A. Jayadi, B. H. Prasetio, S. R. Akbar, E. R. Widasari, and D. Syauqy. Embedded flu detection system based cough sound using mfcc and knn algorithm. In *2022 International Conference of Science and Information Technology in Smart Administration (ICSINTESA)*, pages 1–5. IEEE, 2022.

[40] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

[41] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[42] Z. Ma, C. Bullen, J. T. W. Chu, R. Wang, Y. Wang, and S. Singh. Towards the objective speech assessment of smoking status based on voice features: a review of the literature. *Journal of Voice*, 37(2):300–e11, 2023.

[43] Z. Ma, Y. Qiu, F. Hou, R. Wang, J. T. W. Chu, and C. Bullen. Determining the best acoustic features for smoker identification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8177–8181. IEEE, 2022.

[44] M. Mosuily, L. Welch, and J. Chauhan. MMLung: Moving Closer to Practical Lung Health Estimation using Smartphones. In *Proc. INTERSPEECH 2023*, pages 2333–2337, 2023.

[45] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino. Byol for audio: Self-supervised learning for general-purpose audio representation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[46] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino. Masked modeling duo: Learning representations by encouraging both networks to model the input. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[47] L. Orlandic, T. Teijeiro, and D. Atienza. The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1):156, 2021.

[48] S. Pai, D. Bontempi, I. Hadzic, V. Prudente, M. Sokač, T. L. Chaunzwa, S. Bernatz, A. Hosny, R. H. Mak, N. J. Birkbak, et al. Foundation model for cancer imaging biomarkers. *Nature machine intelligence*, pages 1–14, 2024.

[49] S. Reichert, R. Gass, C. Brandt, and E. Andrès. Analysis of respiratory sounds: state of the art. *Clinical medicine. Circulatory, respiratory and pulmonary medicine*, 2:CCRPM–S530, 2008.

[50] G. Rizos, R. A. Calvo, and B. W. Schuller. Positive-pair redundancy reduction regularisation for speech-based asthma diagnosis prediction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[51] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni, et al. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological measurement*, 40(3):035001, 2019.

[52] H. E. Romero, N. Ma, G. J. Brown, and E. A. Hill. Acoustic screening for obstructive sleep apnea in home environments based on deep neural networks. *IEEE Journal of Biomedical and Health Informatics*, 26(7):2941–2950, 2022.

[53] G. Rudraraju, S. Palreddy, B. Mamidgi, N. R. Sripada, Y. P. Sai, N. K. Vodnala, and S. P. Haranath. Cough sound analysis and objective correlation with spirometry and clinical diagnosis. *Informatics in Medicine Unlocked*, 19:100319, 2020.

[54] A. Saeed, D. Grangier, and N. Zeghidour. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879. IEEE, 2021.

[55] R. Schaeffer, B. Miranda, and S. Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2024.

[56] R. V. Sharan, H. Xiong, and S. Berkovsky. Benchmarking audio signal representation techniques for classification with convolutional neural networks. *Sensors*, 21(10):3434, 2021.

[57] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

[58] Y. Shi, A. Karatzoglou, L. Baltrunas, M. Larson, N. Oliver, and A. Hanjalic. Climf: learning to maximize reciprocal rank with collaborative less-is-more filtering. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 139–146, 2012.

[59] A. Sovijarvi, F. Dalmasso, J. Vanderschoot, L. Malmberg, G. Righini, and S. Stoneman. Definition of terms for applications of respiratory sounds. *European Respiratory Review*, 10(77):597–610, 2000.

[60] A. Srivastava, S. Jain, R. Miranda, S. Patil, S. Pandya, and K. Kotecha. Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease. *PeerJ Computer Science*, 7:e369, 2021.

[61] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[62] C. I. Tang, I. Perez-Pozuelo, D. Spathis, S. Brage, N. Wareham, and C. Mascolo. Selfhar: Improving human activity recognition through self-training with unlabeled data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 5(1):1–30, 2021.

[63] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally, et al. Hear: Holistic evaluation of audio representations. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 125–145. PMLR, 2022.

[64] S. Vollmer, B. A. Mateen, G. Bohner, F. J. Király, R. Ghani, P. Jonsson, S. Cumbers, A. Jonas, K. S. McAllister, P. Myles, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *bmj*, 368, 2020.

[65] D. Vos, T. Döhmen, and S. Schelter. Towards parameter-efficient automation of data wrangling tasks with prefix-tuning. In *NeurIPS 2022 First Table Representation Workshop*, 2022.

[66] M. Wornow, Y. Xu, R. Thapa, B. Patel, E. Steinberg, S. Fleming, M. A. Pfeffer, J. Fries, and N. H. Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135, 2023.

[67] Q. Wu, M. A. Khan, S. Das, V. Nanda, B. Ghosh, C. Kolling, T. Speicher, L. Bindschaedler, K. P. Gummadi, and E. Terzi. Towards reliable latent knowledge estimation in llms: In-context learning vs. prompting based factual knowledge extraction. *arXiv preprint arXiv:2404.12957*, 2024.

[68] T. Xia, J. Han, and C. Mascolo. Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues. *Experimental Biology and Medicine*, 247(22):2053–2061, 2022.

[69] T. Xia, D. Spathis, J. Ch, A. Grammenos, J. Han, A. Hasthanasombat, E. Bondareva, T. Dang, A. Floto, P. Cicuta, et al. Covid-19 sounds: a large-scale audio dataset for digital respiratory screening. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*, 2021.

[70] L. Xiao, X. Yang, X. Li, W. Tu, X. Chen, W. Yi, J. Lin, Y. Yang, and Y. Ren. A snoring sound dataset for body position recognition: Collection, annotation, and analysis. *Proc. INTERSPEECH*, 2023.

[71] W. Xie, Q. Hu, J. Zhang, and Q. Zhang. Earspiro: Earphone-based spirometry for lung function assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–27, 2023.

[72] H. Xue and F. D. Salim. Exploring self-supervised representation ensembles for covid-19 cough classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1944–1952, 2021.

[73] Q. Yang, J. Xu, W. Liu, Y. Chu, Z. Jiang, X. Zhou, Y. Leng, Y. Lv, Z. Zhao, C. Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*, 2024.

[74] C.-C. M. Yeh, X. Dai, H. Chen, Y. Zheng, Y. Fan, A. Der, V. Lai, Z. Zhuang, J. Wang, L. Wang, et al. Toward a foundation model for time series data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4400–4404, 2023.

[75] Q. Zhou, J. Shan, W. Ding, C. Wang, S. Yuan, F. Sun, H. Li, and B. Fang. Cough recognition based on mel-spectrogram and convolutional neural network. *Frontiers in Robotics and AI*, 8:580080, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We did not identify any.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix A.2.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See Appendix A.1.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix A.1.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix A.1.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix for OPERA

Contents

A.1 Datasets Overview

We have used 11 datasets in our benchmark. Their statistics are summarized in Table 1 and Table 2 in the main paper. Here, we supplement their access methods and licenses in Table 7 with a more detailed description below. It can be noted that all datasets contain an audio set and a metadata part. Audio data used are anonymous and the metadata do not contain personally identifiable information or offensive content.

COVID-19 Sounds [69] . The COVID-19 Sounds dataset consists of 53,449 audio samples (over 552 hours in total) crowd-sourced from 36,116 participants through the COVID-19 Sounds app. This dataset is comprehensive in terms of demographics and spectrum of health conditions. It also provides participants' self-reported COVID-19 testing status with 2,106 samples tested positive. It consists of three modalities including breathing, cough, and voice recordings. Only breathing and cough modalities are used in this paper.

This dataset is crowdsourced through the COVID-19 Sounds project, approved by the Ethics Committee of the Department of Computer Science and Technology at the University of Cambridge. Informed consent was obtained from all the participants. The dataset is accessible under controlled access through a Data Transfer Agreement and has been widely shared and used [72, 50].

UK COVID-19 [12]. The UK COVID-19 Vocal Audio Dataset is designed for the training and evaluation of machine learning models that classify SARS-CoV-2 infection status or associated respiratory symptoms using vocal audio. The UK Health Security Agency recruited voluntary participants through the national Test and Trace programme and the REACT-1 survey in England from March 2021 to March 2022, during dominant transmission of the Alpha and Delta SARS-CoV-2 variants and some Omicron variant sublineages. Audio recordings of volitional coughs, exhalations, and speech (speech not included in open access version, nor used in this paper) were collected in the 'Speak up to help beat coronavirus' digital survey alongside demographic, self-reported symptom and respiratory condition data, and linked to SARS-CoV-2 test results.

The study has been approved by The National Statistician's Data Ethics Advisory Committee (reference NSDEC(21)01) and the Cambridge South NHS Research Ethics Committee (reference 21/EE/0036) and Nottingham NHS Research Ethics Committee (reference 21/EM/0067). Participants reviewed the participant information and confirmed their informed consent to take part.

COUGHVID [47]. The COUGHVID dataset provides over 25,000 crowdsourced cough recordings representing a wide range of participant ages, genders, geographic locations, and COVID-19 statuses.

All of the data collection and annotation was done in compliance with relevant ethical regulations. Informed consent was obtained by all participants who uploaded their cough sounds and metadata.

ICBHI [51]. The ICBHI Respiratory Sound Database contains audio samples, collected independently by two research teams in two different countries, over several years. Ethical approval was obtained from the ethics committees of the appropriate institutions.

Table 7: Dataset availability. *ICBHI and HF Lung datasets coming from multiple sources, please refer to the text description below. COVID-19 Sounds, SSBPR, MMLung and NoseMic are available upon request. The custom license is detailed in the DTA (data transfer agreement).

Dataset	Source	Access	license
COVID-19 Sounds[69]	UoC	https://covid-19-sounds.org/blog/neurips_dataset	Custom license
UK COVID-19 [12]	IC	https://zenodo.org/records/10043978	OGL 3.0
CoughVID[47]	EPFL	https://zenodo.org/records/4048312	CC BY 4.0
ICBHI[51]	*	https://bhichallenge.med.auth.gr	CC0
HF Lung [31]	*	https://gitlab.com/techsupportHF/HF_Lung_V1	CC BY 4.0
		https://gitlab.com/techsupportHF/HF_Lung_V1_IP	CC BY-NC 4.0
Coswara[7]	IISc	https://github.com/iiscleap/Coswara-Data	CC BY 4.0
KAUH[23]	KAUH	https://data.mendeley.com/datasets/jwyy9np4gv/3	CC BY 4.0
Respiratory@TR[2]	ITU	https://data.mendeley.com/datasets/p9z4h98s6j/1	CC BY 4.0
SSBPR[70]	WHU	https://github.com/xiaoli1996/SSBPR	CC BY 4.0
MMlung[44]	UoS	https://github.com/MohammedMosuily/mmlung	Custom license
NoseMic[9]	UoC	https://github.com/evelyn0414/OPERA/tree/main/datasets/nosemic	Custom license

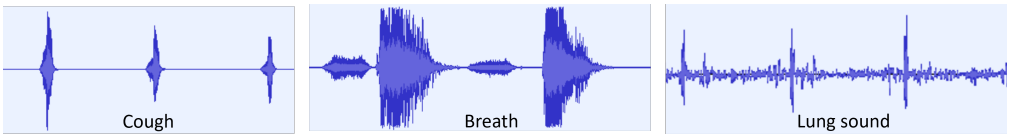


Figure 3: Examples of different respiratory audio modalities used.

Most of the database consists of audio samples recorded by the School of Health Sciences, University of Aveiro (ESSUA) research team at the Respiratory Research and Rehabilitation Laboratory (Lab3R), ESSUA and at Hospital Infante D. Pedro, Aveiro, Portugal. The second research team, from the Aristotle University of Thessaloniki (AUTH) and the University of Coimbra (UC), acquired respiratory sounds at the Papanikolaou General Hospital, Thessaloniki and at the General Hospital of Imathia (Health Unit of Naousa), Greece. The database consists of a total of 5.5 hours of recordings in 920 annotated audio samples from 126 subjects.

HF Lung [31] . HF Lung V2 dataset comprises of HF Lung V1 and HF Lung V1 IP: The lung sound recordings of HF Lung V1 come from two sources. The first source was a database used in a datathon in Taiwan Smart Emergency and Critical Care (TSECC), 2020, under the license of Creative Commons Attribution 4.0 (CC BY 4.0), provided by the Taiwan Society of Emergency and Critical Care Medicine (TSECCM). Lung sound recordings in the TSECC database were acquired from 261 patients. The second source was sound recordings acquired from 18 residents of a respiratory care ward (RCW) or a respiratory care center (RCC) in Northern Taiwan between August 2018 and October 2019. The recordings were approved by the Research Ethics Review Committee of Far Eastern Memorial Hospital (case number: 107052-F). Written informed consent was obtained from the 18 patients.

The lung sound recordings of HF Lung V1 IP come from two sources. The Lung sound recordings from the first source are provided by Taiwan Society of Emergency and Critical Care Medicine (TSECCM) acquired from 32 patients by using a commercial digital stethoscope Littmann 3200 (3M). The lung sound recordings of the second source are acquired by from 7 residents of a respiratory care ward (RCW) or a respiratory care center (RCC) in Northern Taiwan between August 2019 and December 2019. The recordings were approved by the Research Ethics Review Committee of Far Eastern Memorial Hospital (case number: 107052-F). Written informed consent was obtained from the 7 patients or their statutory agents.

Coswara [7]. The Coswara dataset contains respiratory sounds recorded between April 2020 and February 2022 from 2635 individuals (1819 SARS- CoV-2 negative, 674 positive, and 142 recovered subjects). The respiratory sounds contained nine sound categories associated with variants of breathing, cough and speech. The metadata contains demographic information associated with age, gender and geographic location, as well as the health information relating to the symptoms, pre-existing respiratory ailments, comorbidity and SaRS-CoV-2 test status.

The data collection procedure was approved by the Institutional Human Ethics Committee, at the Indian Institute of Science, Bangalore. The informed consent was obtained from all participants who uploaded their data records. All the data collected was anonymized and excluded any participant identity information.

KAUH [23]. The KAUH dataset includes sounds from seven ailments (i.e., asthma, heart failure, pneumonia, bronchitis, pleural effusion, lung fibrosis, and chronic obstructive pulmonary disease (COPD) as well as normal breathing sounds. The dataset contains the audio recordings from the examination of the chest wall at various vantage points using an electronic stethoscope. The stethoscope placement on the subject was determined by the specialist physician performing the diagnosis. Each recording was replicated three times corresponding to various frequency filters that emphasize certain bodily sounds. The dataset can be used for the development of automated methods that detect pulmonary diseases from lung sounds or identify the correct type of lung sound.

All study participants (or their parents in the case of underage subjects) provided written informed consent to be included in the study and allowed their data to be shared. This study was approved by the institutional review board at King Abdullah University Hospital and Jordan University of Science and Technology, Jordan (Ref. 91/136/2020). The data collection was carried out under the relevant guidelines and regulations. The authors have the right to share the data publicly.

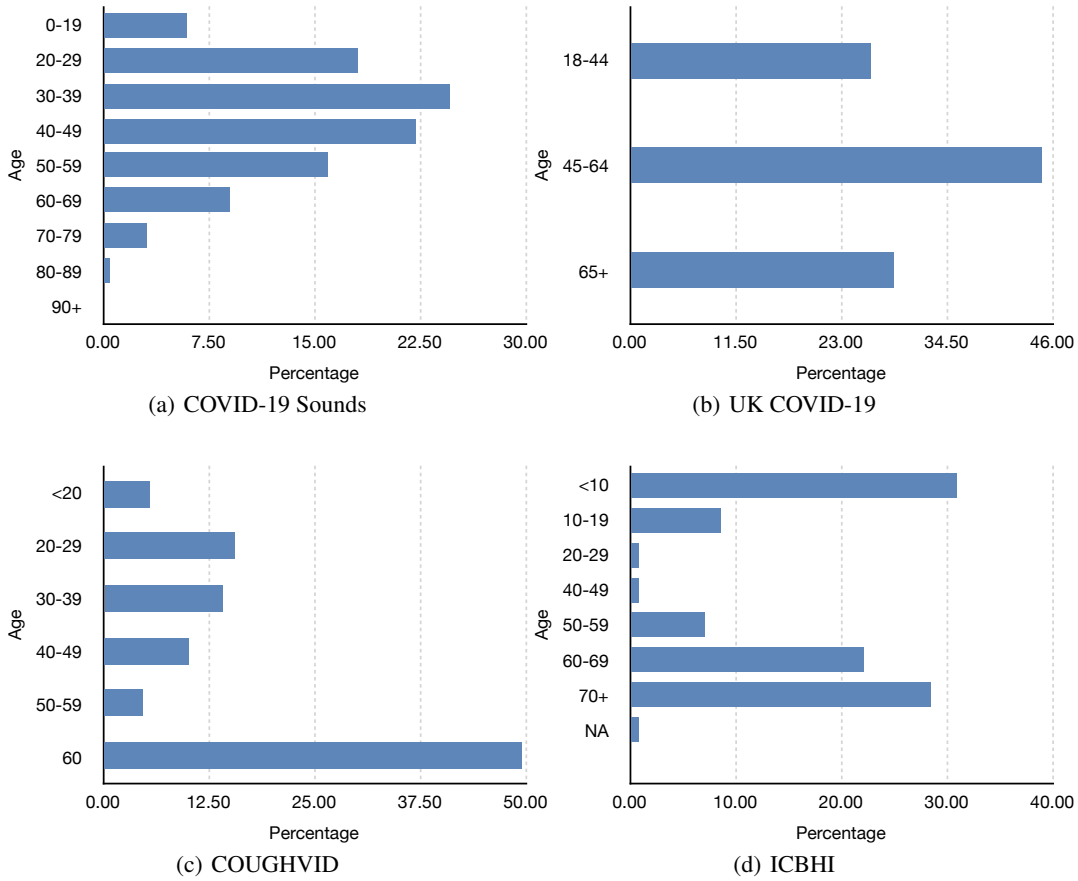
Respiratory@TR [2]. Respiratory@TR contains lung sounds recorded from left and right sides of posterior and anterior chest wall and back using two digital stethoscopes in Antakya State Hospital. The chest X-rays and the pulmonary function test variables and spirometric curves, the St. George respiratory questionnaire (SGRQ-C) are collected as multimedia and clinical functional analysis variables of the patients. The 12 channels of lung sounds are focused on upper lung, middle lung, lower lung and costophrenic angle areas of posterior and anterior sides of the chest. The recordings are validated and labeled by two pulmonologists evaluating the collected chest X-ray, PFT and auscultation sounds of the subjects. Labels fall into 5 COPD severities (COPD0, COPD1, COPD2, COPD3, COPD4). The dataset was released by Iskenderun Technical University, Turkey. Voluntary admittance was evaluated on a voluntary basis form with minimal information. The patients aged 38 to 68 are selected from different occupational groups, socio-economic status and genders for an accomplished analysis of the disorders.

SSBPR [70] . SSBPR is a snore-based sleep body position recognition dataset consisting of 7570 snoring recordings, which comprises six distinct labels for sleep body position: supine, supine but left lateral head, supine but right lateral head, left-side lying, right-side lying and prone. One of the labels is only present in a few subjects and thus is excluded from the task following the 5-class setup in [70].

The data were collected from 20 adult patients who underwent overnight PSG at a local Sleep Medicine Research Center within the hospital. The study was conducted with the approval of the local medical ethics committee, and patients provided signed consent for their participation, including audio and video recordings during sleep. The personal information of the study subjects was collected and stored anonymously to ensure privacy protection.

MMLung [44] . This data was collected from 40 participants (20 male, 20 female) with an age range of 18-85 years old. All participants are English speakers from the UK. Among them, 12 were healthy participants, while the others consisted of seven self-reported COPD patients, seven self-reported asthma patients, and 14 people with other long-term conditions. Ethics approval for this study was obtained from the University of Southampton.

Three devices were used to collect the data: Google Pixel 6 Smartphone with an app installed for the data collection, and an Easy on-PC ultrasonic spirometer by ndd Medical Technologies. The audio data collection from smartphones was conducted in stereo mode at a sampling rate of 44100 Hz. The data was saved in the *WAV* format. The collection took place in a silent room conditions. The process consisted of collecting data for four audio modalities i.e. cough, vowels, mobile spirometry, and speech via a series of tasks from each participant in a single session. In this paper, we only include the deep breath and the vowel sound of ‘o’. Ground truth data were collected using a medical-grade



(a) COVID-19 Sounds (b) UK COVID-19

(c) COUGHVID (d) ICBHI

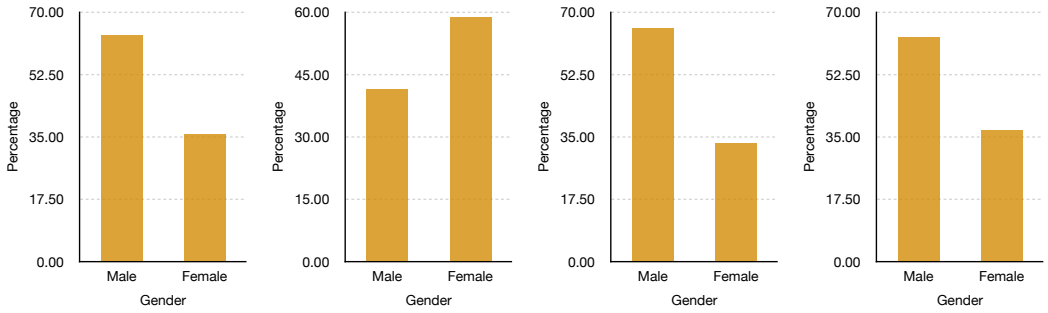
Figure 4: Age distribution of the pretraining datasets.

spirometer by a healthcare professional as per European Respiratory Society (ATS/ERS) clinical standards. However, it should be noted that with any objective measure that is reliant on individual effort, there may always be unforeseen errors (effort dependent blows). This data is available upon request.

NoseMic [9] . NoseMic is a subset of the data collected for a respiratory rate estimation project. The audio data was collected using microphones attached close to the nose, and the respiratory dynamics were measured with a Zephyr pressure sensor on the chest. The data was collected in stationary settings, both before and after the participants exercised. A total number of 21 participants were involved, while data from some participants were excluded because of the poor sensing quality. Audio recordings before and after running were included in our benchmark. Each recording was segmented into 30-second windows with a 15-second overlap. The average respiratory rate of each window was used as the ground truth.

A.1.1 Pretraining Data Demographics

Diversity and representativeness of the training data are important for a generalizable model. We examine the demographic distribution of the five datasets used for model pretraining. The bar plots in Figure 4 and Figure 5 illustrate the age and gender distributions across four of these datasets. While the demographic details of HF Lung are not publicly available, the data includes 35 male and 21 female subjects, with an average age of 66.58 (according to the paper [31]). By integrating these diverse datasets in OPERA, we achieve a more representative and unbiased demographic distribution compared to any single data source. This highlights the importance of uniting varied sources for



(a) COVID-19 Sounds (b) COVID-19 UK (c) COUGHVID (d) ICBHI

Figure 5: Gender distribution of the pretraining datasets.

pretraining a foundational model: not only increasing the number of data samples but also ensuring a more comprehensive distribution.

A.1.2 Downstream Task Description

Here we give a detailed description of all 19 tasks formulated in the OPERA benchmark. The tasks are categorized into three types:

- **Binary Classification (Tasks 1-10)**: Tasks requiring prediction of a binary outcome (positive/negative, smoker/non-smoker, etc.) based on respiratory audio recordings.
- **Multi-Class Classification (Tasks 11, 12)**: Tasks involving classification of respiratory audio recordings into one of several predefined categories (5 classes of COPD severity, sleeping position)
- **Regression (Tasks 13-19)**: Tasks aiming to predict continuous values (lung function metrics, respiratory rate) from respiratory audio data.

Task 1. Each of the audio in UK COVID-19 [12] has a binary label indicating the COVID-19 test result of the participant. This task is to predict whether the test result is positive based on the exhalation recording, consisting of three successive "ha" exhalation sounds.

Task 2. The data source and prediction target is the same as Task 1, while Task 2 is based on the cough recording consisting of three successive volitional coughs.

Task 3. The audio samples in COVID-19 Sounds [69] have the reported symptoms at the moment of participation. This task aims at predicting respiratory abnormalities, where the symptomatic group consists of participants who reported any respiratory symptoms, including dry cough, wet cough, fever, sore throat, shortness of breath, runny nose, headache, dizziness, and chest tightness, while asymptomatic controls are those who reported no symptoms. The audio data consists of 3 to 5 deep breathing sounds. This task follows the subset and split from [69], with the training set downsampled.

Task 4. The dataset and prediction target is the same as Task 3, but the audio includes three coughs.

Task 5. Each of the audio in CoughVID[47] contains a cough and is associated with labels of self-reported demographics and COVID-19 status. This task involves predicting the COVID-19 status based on the cough recording.

Task 6. The dataset and audio modality are the same as Task 5, while the prediction target is gender as reported in demographics.

Task 7. The ICBHI [51] dataset contains labels of the diagnosis of the subjects. We use the subset of COPD patients and healthy controls to formulate a binary classification of COPD detection.

Task 8. Each audio in the Coswara [7] dataset contains a binary label of smoker in the metadata. This task aims to predict the smoker from non-smokers from the cough-shallow audio modality in the dataset, aligning with the implementation in [6].

Task 9. Each audio in the Coswara [7] dataset contains a label of sex in the metadata. This task aims to predict this label from the cough-shallow audio modality in the dataset, aligning with the implementation in [6].

Task 10. The KAUH [23] dataset contains the disease diagnosis labels of the participants. This task aims to use lung sound audio to distinguish patients with COPD and asthma (obstructive lung diseases) from healthy controls.

Task 11. The Respiratory@TR [2] dataset associates each audio with a COPD severity label from 0 to 4. This task aims to predict this severity level from lung sounds.

Task 12. The SSBPR [70] dataset associates each snoring audio with a label of the body position: supine, supine but left lateral head, supine but right lateral head, left-side lying, right-side lying and prone. The last class is excluded here as it is only present in some of the male participants. Thus this task aims to predict one of the five body positions from the snoring sounds.

Task 13. Spirometry is a gold standard for diagnosing Long-term respiratory illnesses like COPD and Asthma. It is a lung health test that requires specialized equipment and trained healthcare experts, making it expensive and difficult to scale. Moreover, blowing into a spirometer can be quite hard for people suffering from pulmonary illnesses. To address this problem, researchers aim to develop audio-based testing methods without requiring the best efforts from patients. MMLung [44] was collected for this purpose. Task 13 evaluates how accurate the forced vital capacity (FCV) can be estimated from a deep breath sound.

Task 14. Similar with Task 13 , Task 14 evaluates how accurate the forced expiratory volume in 1 second (FEV1) can be estimated from a deep breath sound.

Task 15. While FEV1 and FVC are very personal, the ratio between them is the proportion of lung capacity that can be exhaled in the first second. It is expressed as a percentage and is used to diagnose and determine the severity of obstructive and restrictive lung diseases. Task 15 uses breathing sounds to estimate this ratio.

Task 16. Task 16 again aims to evaluate an individual's FVC, similar to Task 13. However, a vowel sound is used, i.e., the participant speaks out the 'o' sound for as long as possible.

Task 17. Task 17 involves the use of 'o' vowel sound for FEV1 estimation.

Task 18. This task predicts the ratio between FEV1 and FVC from the collected 'o' vowel sounds.

Task 19. Continuous respiratory rate (RR) monitoring is integral to mobile healthcare and fitness tracking, offering valuable insights into longitudinal health and wellness due to its strong correlations with both physical and mental health. This task involves the estimation of RR from 30 seconds of breathing sounds.

A.2 Implementation Details

All of the experiments are implemented in Python 3.10.4, with main supporting libraries: PyTorch, Librosa, PyTorch Lightning, numpy, with the exact environment detailed in 'environment.yml' in the code repository. All our experiments are conducted using a NVIDIA A100 GPU with 80GB memory. Our code is accessible from `https://github.com/evelyn0414/OPERA`.

A.2.1 Pretraining Models and Methods

We pre-train our models on a combination of seven sets of data derived from the first five data sources in Table 7 (including separate modalities from COVID-19 Sounds and UK COVID-19). Each set of data is split into batches of equal length to ensure consistent data processing. These batches maintain both modality and source homogeneity. We then randomly shuffle the batches and reserve 10% for validation. Due to inherent variations in audio length within individual batches, we employ random cropping of spectrograms. Crop lengths for each of the seven datasets are detailed in Table 1, and the crop methods depend on the pretraining methods, which will be elaborated on

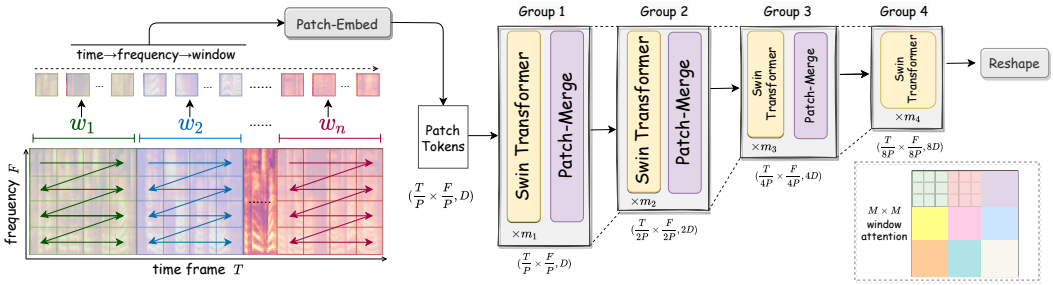


Figure 6: The hierarchical token-semantic audio transformer architecture, from [10].

Table 8: The EfficientNet-B0 architecture.

Layer	Kernel Size	#channels	#layers
Input	-	32	1
MBConv1	3x3	16	1
MBConv6	3x3	24	2
MBConv6	5×5	40	2
MBConv6	3x3	80	3
MBConv6	5x5	112	3
MBConv6	5x5	192	4
MBConv6	3x3	320	1
Conv head & Avg Pooling		1280	1

below. Two representative SSL approaches are adopted: contrastive learning-based methods and generative pretraining-based methods, to pretrain three models. The high-level reasoning behind this is that if an encoder can distinguish the source of audio segments (contrastive) or reconstruct masked spectrograms (generative), it is expected to encode useful and generalizable acoustic features. Specifically:

OPERA-CT: OPERA-CT is a contrastive learning-based transformer model. Following [54], we randomly crop two segments from a spectrogram and regard them as a positive pair. Segments from different samples within one batch are regarded as negative pairs. As shown in Figure 2(a), an encoder network (a transformer here) extracts features from these segments, and a projector (a multi-layer perception) maps them into a low-dimensional representation space, where bilinear similarity is calculated as,

$$s(x, x') = g(f(x))^T W g(f(x')). \tag{1}$$

The optimization objective aims to maximize the similarity between positive pairs and minimize it for negative pairs. The loss function for this instance discrimination objective is a multi-class cross entropy applied to similarities,

$$\mathcal{L} = -\log \frac{\exp\left(s(x, x^+)\right)}{\sum_{x^- \in \mathcal{X}^-(x) \cup \{x^+\}} \exp\left(s(x, x^-)\right)}, \tag{2}$$

where x^+ is the positive anchor for x and $\mathcal{X}^-(x)$ refers to negative distractors.

Specifically, the transformer we employ is a hierarchical token-semantic audio transformer [10], which improves the computing and memory efficiency of the typical vision transformer for spectrograms. A patch size of 4×4 is used and the output feature dimension is 768. The encoder has 31M trainable parameters.

OPERA-CE: Similar to OPERA-CT, CE leverages a contrastive pre-training approach. However, it utilizes a more lightweight and efficient CNN encoder (EfficientNet-B0) [61]. The architecture is detailed in Table 8. This encoder outputs a feature dimension of 1280 and has approximately 4M trainable parameters.

OPERA-GT: OPERA-GT is a generative pretrained transformer model. It uses a masked auto-encoder to extract useful features from masked spectrograms, which a decoder then uses to reconstruct

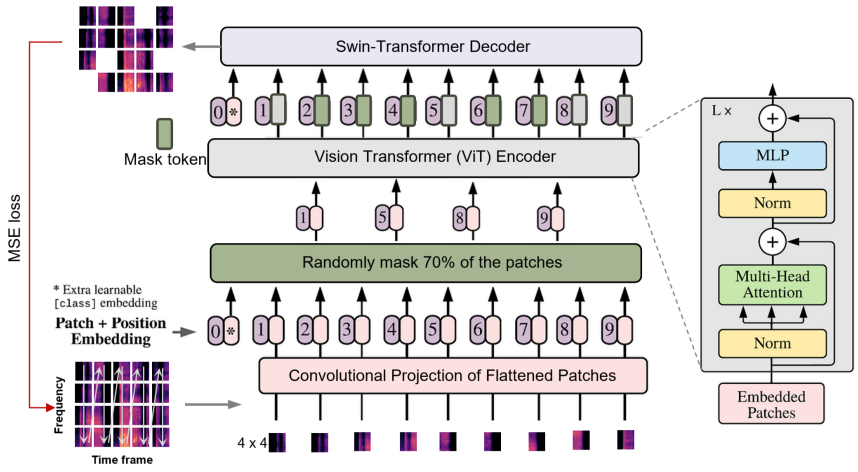


Figure 7: OPERA-GT architecture.

the original spectrograms, as illustrated in Figure 2(b). Following [3], we employ a vision transformer as the encoder (21M trainable parameters) and a lightweight swin-transformer (12M trainable parameters) as the decoder. The detailed architecture is shown in Figure 7.

To train this model, spectrograms from each dataset are cropped to equal lengths, as summarized in Table 1, and then split into patches of 4×4. Considering the varying lengths of different modalities, our model uses a unique patching order and accommodates any input length (no larger than the number of positional embeddings), as indicated by the arrows in Figure 7. Each patch is converted into a patch embedding via a 2-dimensional convolutional layer with a kernel size of 4×4 and a channel number of 384. We randomly mask 70% of patches per spectrogram and only feed the embeddings of the visible patches into the encoder. The encoder is a typical vision transformer with $l = 12$ blocks and 2 heads in each block. The output feature dimension is 384.

To reconstruct the spectrograms, both the embeddings of the masked patches and the new embeddings from the encoder are fed into the decoder. The decoder is a typical swin-transformer with both local and global attention. The output of the decoder is an array resembling a spectrogram. Mean square error loss is used for optimization, and only the masked pixels are considered in the loss,

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2, \tag{3}$$

where y is the vector only with the masked pixels in the i-th spectrogram.

A.2.2 Benchmark implementation details

Within our benchmark of downstream tasks, we have four baselines to compare with the OPERA models. Opensmile is chosen as a baseline representing the traditional feature extraction methods. VGGish, AudioMAE and CLAP are chosen as baselines for this study since they are open-source pretrained models representing the cutting edge of deep learning approaches.

Opensmile. OpenSMILE [18] is a powerful tool for extracting features from audio data. It offers pre-defined feature sets designed to capture various aspects of an audio signal. This established toolkit serves as a strong baseline for traditional feature extraction. It offers a diverse set of handcrafted features, providing a foundation for comparison.

VGGish. The VGGish model [30] is a modified VGG model using mel spectrograms as input, pretrained to classify the soundtracks of a dataset of 70M training videos (5.24 million hours) with 30,871 video-level labels.

Table 9: Number of parameters and feature dimension of all the models.

	Opensmile	VGGish	AudioMAE	CLAP	**OPERA-CT**	**OPERA-CE**	**OPERA-GT**
# Parameters (M)	-	62	86	80	31	4	21
Input length (s)	-	1	10	5	<32	>1.5	<8.18
Feature Dim.	988	128	768	1024	768	1280	384

AudioMAE. AudioMAE [35] leverages self-supervised learning for audio, inspired by image-based Masked Autoencoders (MAE) [29]. During training, AudioMAE masks a high proportion (70%) of the spectrogram patches and feeds the remaining unmasked tokens through a transformer encoder, which then attempts to reconstruct the original spectrogram. This process forces the model to learn robust features by relying on context and relationships within the spectrogram.

CLAP. The CLAP model is trained under natural language supervision, leveraging text descriptions to learn about audio concepts. It utilizes two encoders: one for processing audio spectrograms and another for handling text descriptions. Through a contrastive learning approach, CLAP brings these audio and text features into a shared space and encourages similarity within the same audio-text pair.

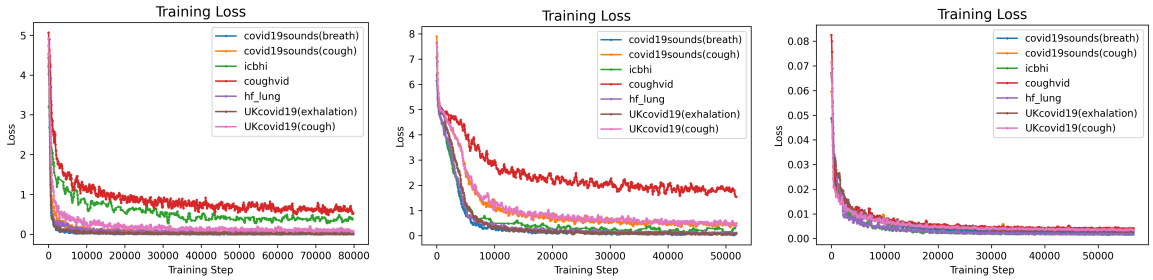
For baselines, both the data pre-processing and feature extraction strictly follow their official implementation. For our pretrained models, the same audio preprocessing is used as in pretraining. The required audio input length is also summarized in Table 9.

Our OPERA models can accept audio input of different lengths. Specifically, OPERA-CT has an interpolation step that transforms all spectrogram inputs to the same size, fitting the hierarchical structure of the model [10]. Audio longer than the maximum input length of about 32 seconds will need to be cropped, although this is not relevant to our downstream tasks. OPERA-CT is a CNN model with a pooling layer, allowing it to always output fixed-length features. However, it requires a minimum length of 1.5 seconds (the input size must be larger than the kernel size). OPERA-GT, a transformer model, incorporates a special patching method (see Figure 7) that allows it to accept varying lengths of audio shorter than its maximum input length of 8.18 seconds. For input audio exceeding 8 seconds, we segment the audio into short frames with overlaps, feed them into the model, and use the averaged representation of these frames as the final embedding [35].

Our evaluation employs linear evaluation for all downstream tasks. This technique leverages the pre-trained model's weights without modification, preserving their learned features. A new linear layer, sized according to the feature dimension (see Table 9) and the number of output classes (or 1 dimension for regression) in the specific downstream task, is added on top of the pre-trained model's output. This approach offers an efficient way to transfer the knowledge of the pre-trained models without extensive fine-tuning of the entire model and can be used for tasks with very limited data size. For classification tasks, a standard cross-entropy loss is used. For regression tasks, an MAE loss is used. A L2 regularization of 10^{-5} is employed.

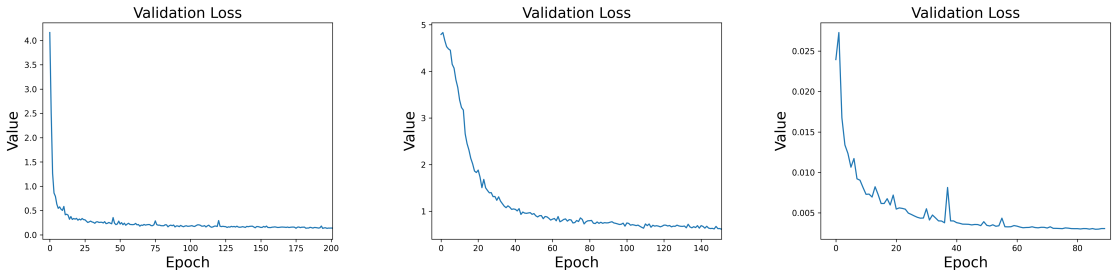
A.3 Pretraining Results

Pretraining loss. We showcase the training process of our three OPERA models here. Specifically, Figure 8 exhibits the training loss of different subsets of the data, converging at different speeds and levels, due to heterogeneity in data quality, data modality, etc. Figure 9 present the evolution of the loss on the validation set (a set combined a small proportion from all the data resource). It demonstrates a continued decay until convergence.



(a) OPERA-CT (b) OPERA-CE (c) OPERA-GT

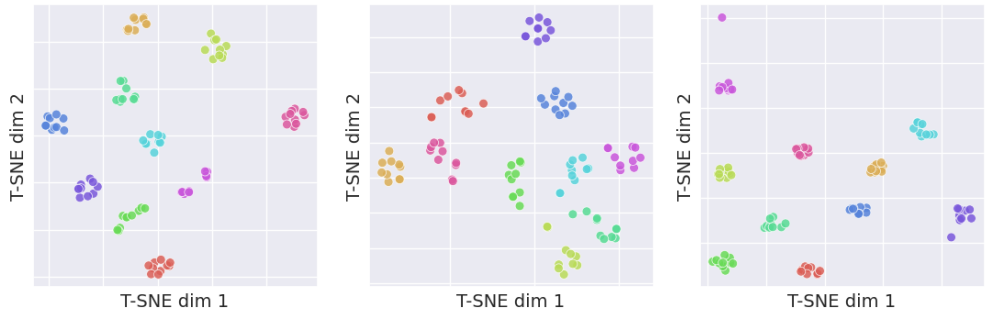
Figure 8: Training loss of the three OPERA models. The OPERA-GT and OPERA-CE use contrastive instance discrimination loss, while OPERA-GT uses generative mean square error loss.



(a) OPERA-CT (b) OPERA-CE (c) OPERA-GT

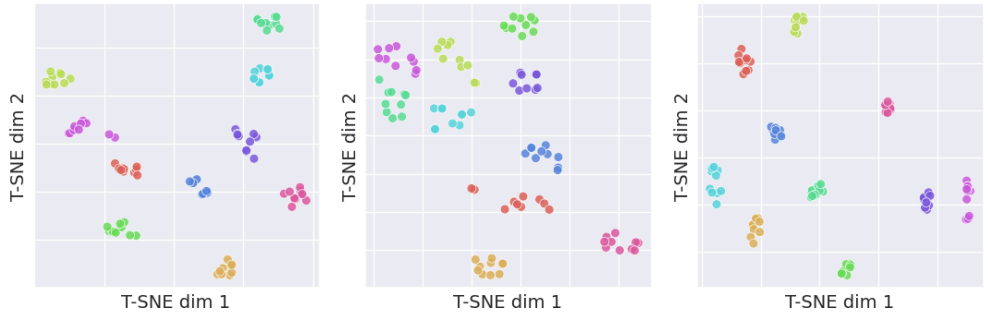
Figure 9: Validation loss of the three OPERA models. The OPERA-GT and OPERA-CE use contrastive instance discrimination loss, while OPERA-GT uses generative mean square error loss.

Embedding distribution analysis for constructive pretraining. Figure 10 and Figure 11 present the T-SNE visualization applied to features extracted from the contrastive pretraining models on the held-out test set of pretraining data. The visualization depicts four random crops of the same audio sample (the same color) close together in the embedding space. This suggests that the model can effectively capture the underlying characteristics of the audio data despite variations introduced by cropping.



(a) COVID-19 Sounds (breath) (b) UK COVID-19 (cough) (c) HF Lung (lung sounds)

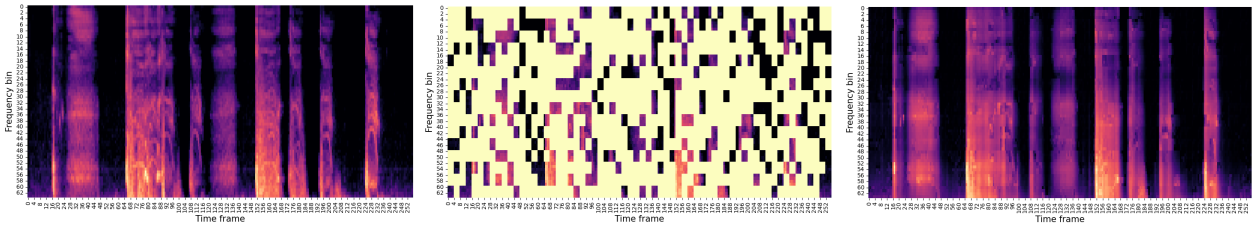
Figure 10: T-SNE visualization result of features from OPERA-CT on the held-out validation of pretraining data. Each dot is an audio segment and the same color represents the same audio recording. It can be seen that audio segments from the same recording are close to each other while far away from other recordings in the embedding space.



(a) COVID-19 Sounds (breath) (b) UK COVID-19 (cough) (c) HF Lung (lung sounds)

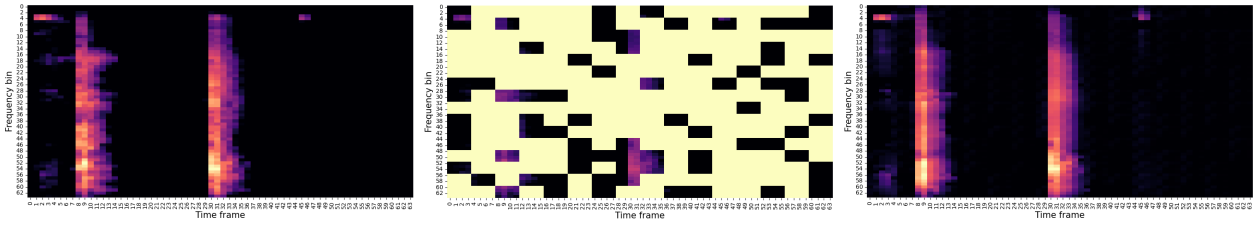
Figure 11: T-SNE visualization result of features from OPERA-CE on the validation data.

Spectrogram reconstruction result for generative pretraining. OPERA-GT aims to learn a useful encoder by extracting features that can be used to reconstruct the entire spectrogram. Figure 9(c) demonstrates a very small MSE loss on the validation set when the model converges, suggesting a good reconstruction ability. To show it more straightforward, some examples are visualized in Figure 12, Figure 13, Figure 14. From the visualization, it is clear that our pretrained encoder can capture both the local and global distribution of the spectrograms and the decoder can accurately recover the original information.



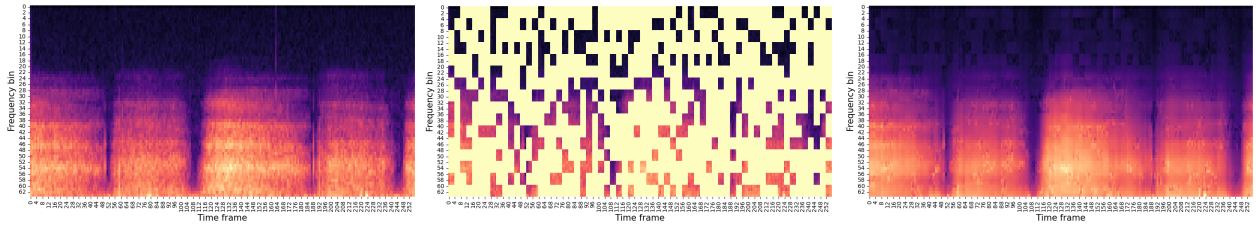
(a) Original spectrogram (b) Masked spectrogram (c) Reconstructed spectrogram

Figure 12: Reconstruction result for a breath sound recording (cropped into 8s) from COVID-19 Sounds dataset.



(a) Original spectrogram (b) Masked spectrogram (c) Reconstructed spectrogram

Figure 13: Reconstruction result for a cough sound recording (cropped into 2s) from COUGHVID dataset.



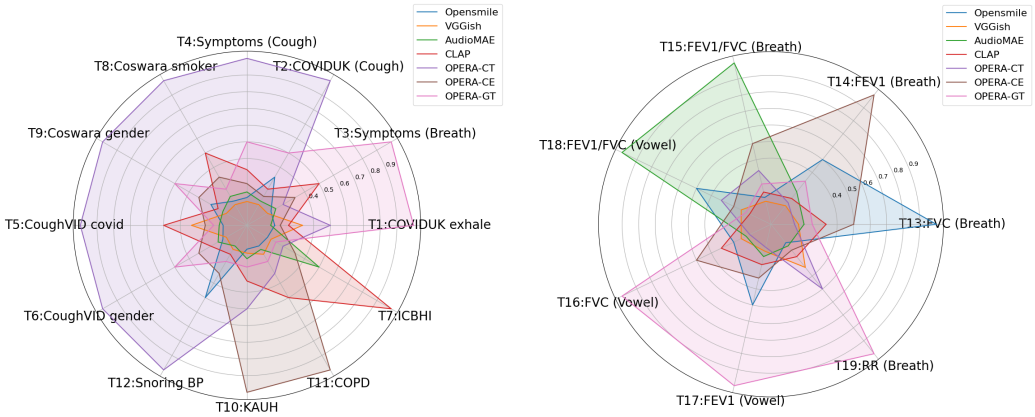
(a) Original spectrogram (b) Masked spectrogram (c) Reconstructed spectrogram

Figure 14: Reconstruction result for a lung sound recording (cropped into 8s) from ICBHI dataset.

[Correction] The reciprocal ranks of all the 19 tasks are detailed in Figure 15 in Appendix A.4.

A.4 Additional Evaluation Results

Table 3 summarized the over mean reciprocal ranks, with the reciprocal ranks of all the 19 tasks detailed in Figure 15.



(a) Health Condition Inference (b) Lung Function Estimation

Figure 15: Radar plot of reciprocal ranks on two groups of tasks.

A.4.1 Another Metric for Lung Function Estimation Tasks

While AUROC, used for classification, ranges from 0.5 to 1, MAE, used for regression, doesn't have a bounded range for comparison. Hence, here we additionally report the relative error for the estimation measured by MAPE (Mean Absolute Percentage Error) in Table 10. MAPE ranges from 0 to 1, with a lower value indicating better estimations.

Table 10: MAPE on lung function estimation tasks (lower is better). The best model per task is highlighted. We report mean and standard deviation across subjects.

ID	Task Abbr.	Opensmile	VGGish	AudioMAE	CLAP	**OPERA-CT**	**OPERA-CE**	**OPERA-GT**	
T13	FVC (Breath)	0.300 ± 0.189	0.535 ± 0.961	0.419 ± 0.272	0.375 ± 0.397	0.515 ± 0.509	0.363 ± 0.502	0.399 ± 0.370	*
T14	FEV1 (Breath)	0.386 ± 0.481	0.834 ± 1.646	0.474 ± 0.457	0.483 ± 0.437	0.592 ± 0.425	0.367 ± 0.493	0.460 ± 0.554	✓*
T15	FEV1/FVC (Breath)	0.264 ± 0.563	0.333 ± 0.687	0.177 ± 0.149	0.230 ± 0.194	0.199 ± 0.194	0.197 ± 0.221	0.193 ± 0.184	✓
T16	FVC (Vowel)	0.237 ± 0.206	0.356 ± 0.910	0.335 ± 0.320	0.240 ± 0.267	0.581 ± 0.449	0.217 ± 0.213	0.184 ± 0.142	✓*
T17	FEV1 (Vowel)	0.287 ± 0.383	0.571 ± 1.095	0.456 ± 0.516	0.350 ± 0.334	0.650 ± 0.508	0.319 ± 0.302	0.246 ± 0.208	✓*
T18	FEV1/FVC (Vowel)	0.228 ± 0.240	0.315 ± 0.578	0.188 ± 0.211	0.350 ± 0.393	0.234 ± 0.209	0.279 ± 0.257	0.304 ± 0.210	
T19	Breathing Rate	0.299 ± 0.113	0.202 ± 0.076	0.203 ± 0.078	0.204 ± 0.083	0.199 ± 0.074	0.201 ± 0.076	0.196 ± 0.074	✓*

A.4.2 Fine-tuning Performance

Apart from the standard linear evaluation, we also explore the effect of fine-tuning in improving the performance, using some of the tasks with a comparatively sufficient number of samples.

For OPERA-CE, due to the small number of parameters that could easily overfit and forget the pretraining, we freeze two-thirds of the blocks and only fine-tune the first 5 blocks dealing with the input data (along with the classification head). For all other models and baselines, we fine-tune the entire model together with the classifier.

In addition to the result for Task 4 detailed in Section 6, the performance of Task 7 and 12 after fine-tuning are presented in Table 11 and Table 12. It is obvious that the performance can be greatly improved after fine-tuning, and the two transformer-based OPERA models demonstrate superior performance.

Table 11: AUROC (higher is better) for linear probing and finetuning on T7 (COPD detection). Best model highlighted.

Method	# Train	AudioMAE	CLAP	OPERA-CT	OPERA-CE	OPERA-GT
Linear	828	0.886 ± 0.017	0.933 ± 0.005	0.855 ± 0.012	0.872 ± 0.011	0.741 ± 0.011
Fine-tune	828	0.984 ± 0.012	0.980 ± 0.007	0.957 ± 0.024	0.808 ± 0.032	0.986 ± 0.006

Table 12: AUROC (higher is better) for linear probing and finetuning on T12 (snoring based body position recognition). Best model highlighted.

Method	# Train	AudioMAE	CLAP	OPERA-CT	OPERA-CE	OPERA-GT
Linear	7468	0.649 ± 0.001	0.702 ± 0.001	0.781 ± 0.000	0.769 ± 0.000	0.742 ± 0.001
Fine-tune	7468	0.981 ± 0.002	0.935 ± 0.004	0.994 ± 0.001	0.981 ± 0.002	0.986 ± 0.003