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 [70] . 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 [73, 51].

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 StatisticianâĂŹs 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 [48]. 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 [52]. 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.

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

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[70]	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[48]	EPFL	https://zenodo.org/records/4048312	CC BY 4.0
ICBHI[52]	*	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[71]	WHU	https://github.com/xiaoli1996/SSBPR	CC BY 4.0
MMlung[45]	UoS	https://github.com/MohammedMosuily/mmlung	Custom license
NoseMic[9]	UoC	https://github.com/evelyn0414/OPERA/tree/main/datasets/nosemic	Custom license

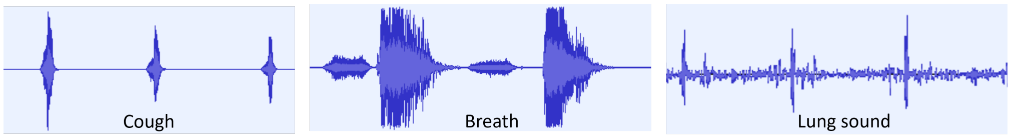


Figure 4: Examples of different respiratory audio modalities used.

(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.

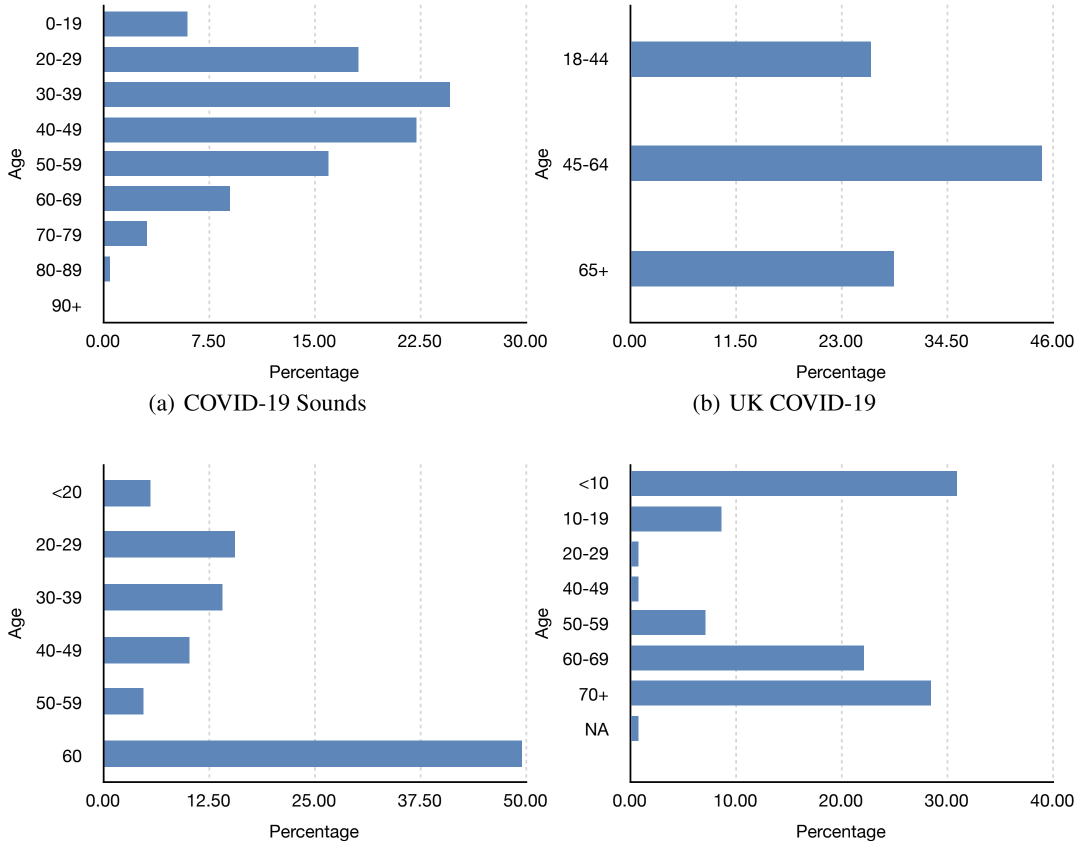
SSBPR [71] . SSBPR is a snore-based sleep body position recognition dataset consisting of 7570 snoring recordings, which comprises six distinct labels for sleep body position: supine, supine but left lateral head, supine but right lateral head, left-side lying, right-side lying and prone. One of the labels is only present in a few subjects and thus is excluded from the task following the 5-class setup in [71].

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 [45] . 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 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.



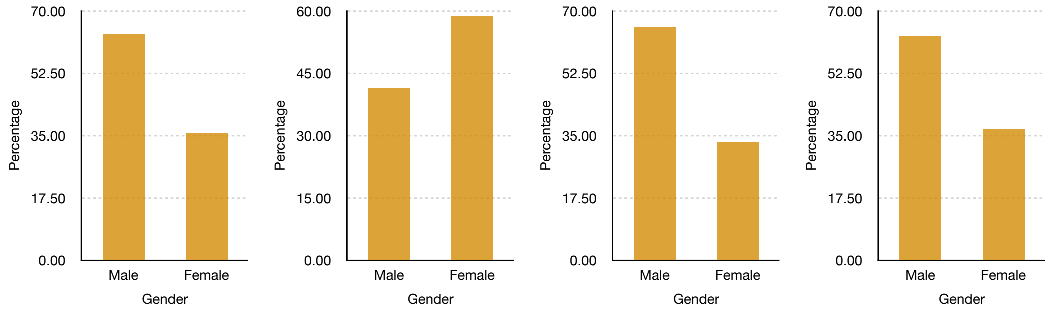
(a) COVID-19 Sounds

(b) UK COVID-19

(c) COUGHVID

(d) ICBHI

Figure 5: Age distribution of the pretraining datasets.



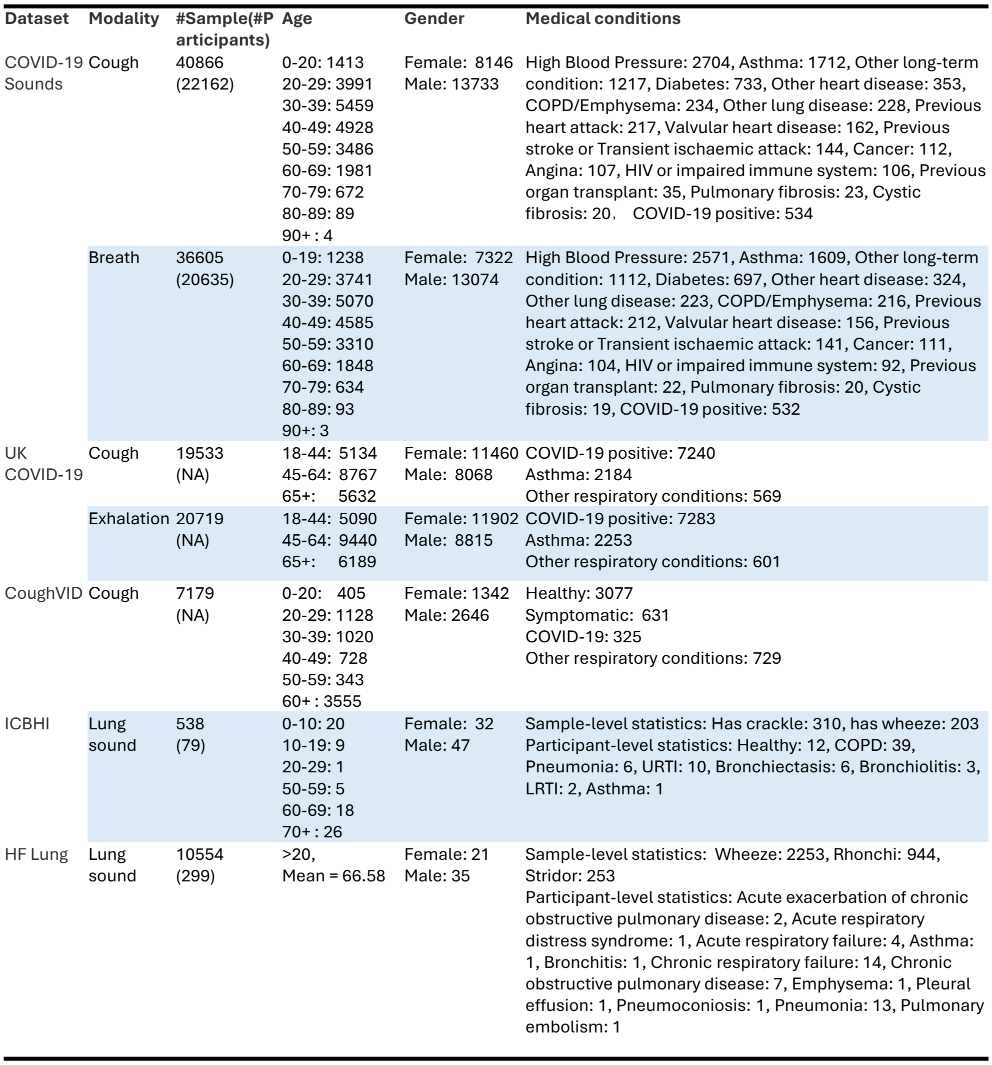
(a) COVID-19 Sounds (b) COVID-19 UK (c) COUGHVID (d) ICBHI

Figure 6: Gender distribution of the pretraining datasets.

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 5 and Figure 6 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]).

Among the five datasets, COVID-19 Sounds and CoughVID were collected globally, while UK COVID-19 and ICBHI were primarily collected in European countries, and HF Lung was collected



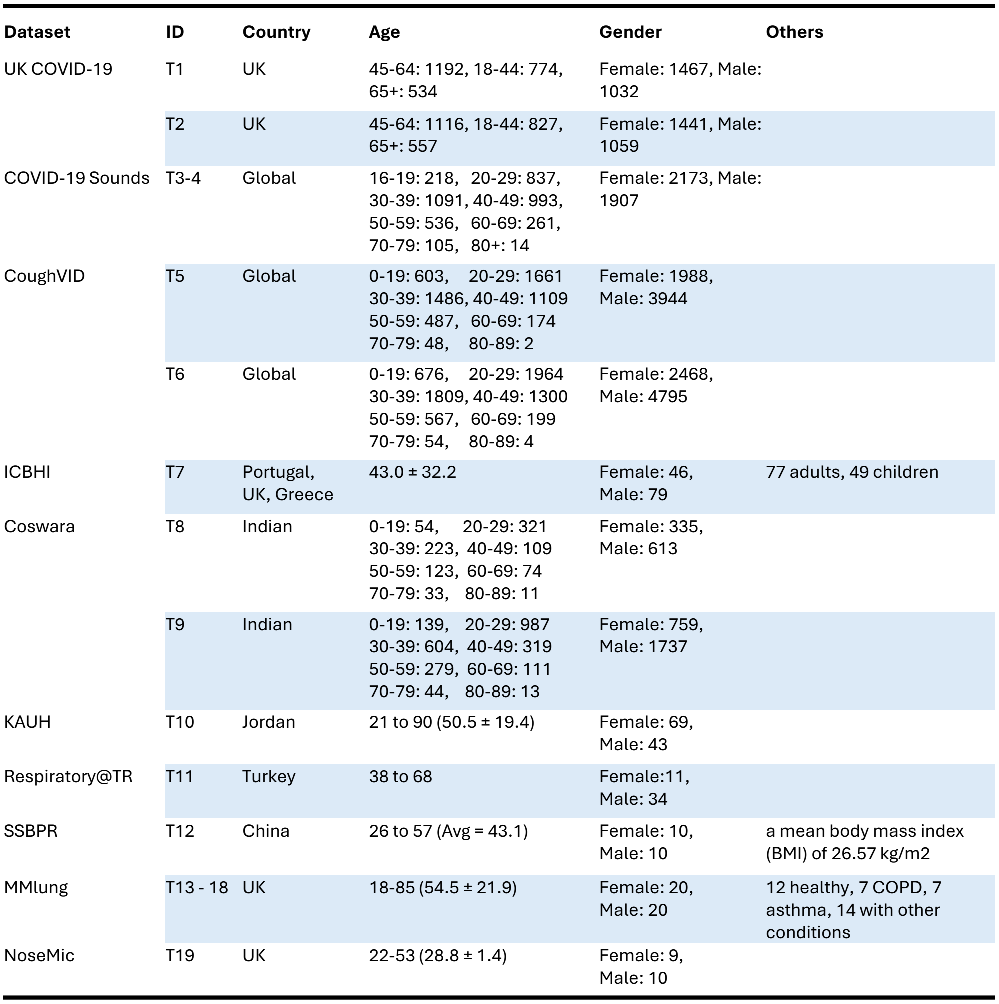
Dataset	Modality	#Sample(#Participants)	Age	Gender	Medical conditions
COVID-19 Sounds	Cough	40866 (22162)	0-20: 1413 20-29: 3991 30-39: 5459 40-49: 4928 50-59: 3486 60-69: 1981 70-79: 672 80-89: 89 90+ : 4	Female: 8146 Male: 13733	High Blood Pressure: 2704, Asthma: 1712, Other long-term condition: 1217, Diabetes: 733, Other heart disease: 353, COPD/Emphysema: 234, Other lung disease: 228, Previous heart attack: 217, Valvular heart disease: 162, Previous stroke or Transient ischaemic attack: 144, Cancer: 112, Angina: 107, HIV or impaired immune system: 106, Previous organ transplant: 35, Pulmonary fibrosis: 23, Cystic fibrosis: 20， COVID-19 positive: 534
	Breath	36605 (20635)	0-19: 1238 20-29: 3741 30-39: 5070 40-49: 4585 50-59: 3310 60-69: 1848 70-79: 634 80-89: 93 90+: 3	Female: 7322 Male: 13074	High Blood Pressure: 2571, Asthma: 1609, Other long-term condition: 1112, Diabetes: 697, Other heart disease: 324, Other lung disease: 223, COPD/Emphysema: 216, Previous heart attack: 212, Valvular heart disease: 156, Previous stroke or Transient ischaemic attack: 141, Cancer: 111, Angina: 104, HIV or impaired immune system: 92, Previous organ transplant: 22, Pulmonary fibrosis: 20, Cystic fibrosis: 19, COVID-19 positive: 532
UK COVID-19	Cough	19533 (NA)	18-44: 5134 45-64: 8767 65+: 5632	Female: 11460 Male: 8068	COVID-19 positive: 7240 Asthma: 2184 Other respiratory conditions: 569
	Exhalation	20719 (NA)	18-44: 5090 45-64: 9440 65+: 6189	Female: 11902 Male: 8815	COVID-19 positive: 7283 Asthma: 2253 Other respiratory conditions: 601
CoughVID	Cough	7179 (NA)	0-20: 405 20-29: 1128 30-39: 1020 40-49: 728 50-59: 343 60+ : 3555	Female: 1342 Male: 2646	Healthy: 3077 Symptomatic: 631 COVID-19: 325 Other respiratory conditions: 729
ICBHI	Lung sound	538 (79)	0-10: 20 10-19: 9 20-29: 1 50-59: 5 60-69: 18 70+ : 26	Female: 32 Male: 47	Sample-level statistics: Has crackle: 310, has wheeze: 203 Participant-level statistics: Healthy: 12, COPD: 39, Pneumonia: 6, URTI: 10, Bronchiectasis: 6, Bronchiolitis: 3, LRTI: 2, Asthma: 1
HF Lung	Lung sound	10554 (299)	>20, Mean = 66.58	Female: 21 Male: 35	Sample-level statistics: Wheeze: 2253, Rhonchi: 944, Stridor: 253 Participant-level statistics: Acute exacerbation of chronic obstructive pulmonary disease: 2, Acute respiratory distress syndrome: 1, Acute respiratory failure: 4, Asthma: 1, Bronchitis: 1, Chronic respiratory failure: 14, Chronic obstructive pulmonary disease: 7, Emphysema: 1, Pleural effusion: 1, Pneumoconiosis: 1, Pneumonia: 13, Pulmonary embolism: 1

Figure 7: Statistics of demographics and medical conditions for datasets used for pretraining.

in Asian regions. Therefore, our curated data presents a comprehensive geo-distribution, covering participants from different ethnic backgrounds and speaking various languages.

Figure 7 summarizes in detail all demographics and medical conditions for the five datasets used for model pre-training. The five datasets used cover a wide range of respiratory medical conditions. COVID-19 Sounds, UK COVID-19, and CoughVID were collected during the pandemic and include some participants who tested positive or negative for COVID-19. Some of the participants had other conditions such as asthma, COPD, pulmonary fibrosis, cancer, etc. The ICBHI and HF Lung datasets include participants who were either healthy or had various respiratory diseases including asthma, COPD, URTI, Pneumonia, etc. Recordings feature both healthy individuals and those with symptoms such as wheeze, crackles, or rhonchi.

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



Dataset	ID	Country	Age	Gender	Others
UK COVID-19	T1	UK	45-64: 1192, 18-44: 774, 65+: 534	Female: 1467, Male: 1032	
	T2	UK	45-64: 1116, 18-44: 827, 65+: 557	Female: 1441, Male: 1059	
COVID-19 Sounds	T3-4	Global	16-19: 218, 20-29: 837, 30-39: 1091, 40-49: 993, 50-59: 536, 60-69: 261, 70-79: 105, 80+: 14	Female: 2173, Male: 1907	
CoughVID	T5	Global	0-19: 603, 20-29: 1661 30-39: 1486, 40-49: 1109 50-59: 487, 60-69: 174 70-79: 48, 80-89: 2	Female: 1988, Male: 3944	
	T6	Global	0-19: 676, 20-29: 1964 30-39: 1809, 40-49: 1300 50-59: 567, 60-69: 199 70-79: 54, 80-89: 4	Female: 2468, Male: 4795	
ICBHI	T7	Portugal, UK, Greece	43.0 ± 32.2	Female: 46, Male: 79	77 adults, 49 children
Coswara	T8	Indian	0-19: 54, 20-29: 321 30-39: 223, 40-49: 109 50-59: 123, 60-69: 74 70-79: 33, 80-89: 11	Female: 335, Male: 613	
	T9	Indian	0-19: 139, 20-29: 987 30-39: 604, 40-49: 319 50-59: 279, 60-69: 111 70-79: 44, 80-89: 13	Female: 759, Male: 1737	
KAUH	T10	Jordan	21 to 90 (50.5 ± 19.4)	Female: 69, Male: 43	
Respiratory@TR	T11	Turkey	38 to 68	Female:11, Male: 34	
SSBPR	T12	China	26 to 57 (Avg = 43.1)	Female: 10, Male: 10	a mean body mass index (BMI) of 26.57 kg/m2
MMlung	T13 - 18	UK	18-85 (54.5 ± 21.9)	Female: 20, Male: 20	12 healthy, 7 COPD, 7 asthma, 14 with other conditions
NoseMic	T19	UK	22-53 (28.8 ± 1.4)	Female: 9, Male: 10	

Figure 8: Statistics of demographics for downstream tasks.

A.1.2 Downstream Task Description

Here we give a detailed description of all 19 tasks formulated in the OPERA benchmark. The demographic statistics are summarized in Figure 8. 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 âĂIJhaâĂİ 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 [70] 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 [70], 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[48] 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 [52] 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 [71] 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 [45] 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.

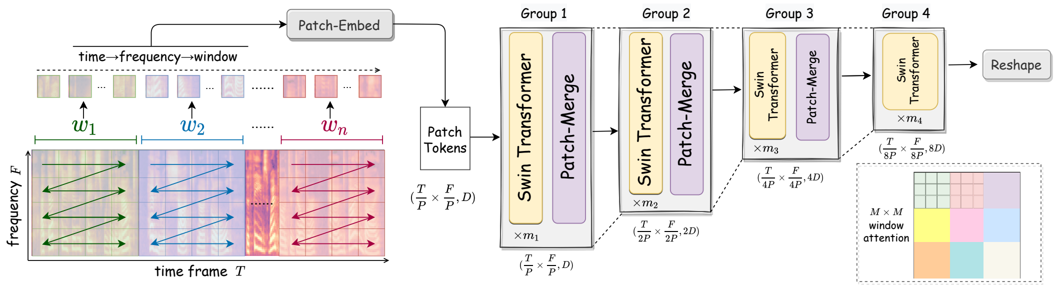


Figure 9: The hierarchical token-semantic audio transformer architecture, from [10].

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 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 [55], 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) [62]. The architecture is detailed in Table 8. This encoder outputs a feature dimension of 1280 and has approximately 4M trainable parameters.

Table 8: The EfficientNet-B0 architecture.

Layer	Kernel Size	#channels	#layers
Input	-	32	1
MBConv1	3x3	16	1
MBConv6	3x3	24	2
MBConv6	5ÃŮ5	40	2
MBConv6	3x3	80	3
MBConv6	5x5	112	3
MBConv6	5x5	192	4
MBConv6	3x3	320	1
Conv head & Avg Pooling		1280	1

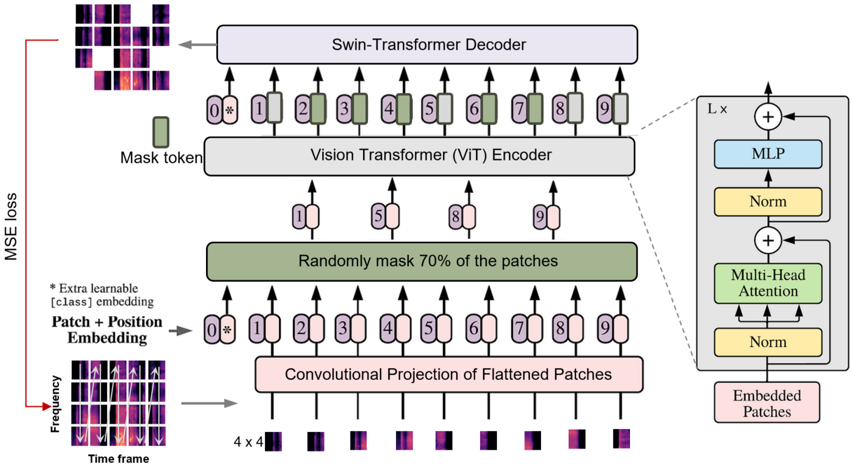

Swin-Transformer Decoder
0 * 1 2 3 4 5 6 7 8 9
Mask token
Vision Transformer (ViT) Encoder
1 5 8 9
Randomly mask 70% of the patches
MSE loss
* Extra learnable [class] embedding
Patch + Position Embedding
0 * 1 2 3 4 5 6 7 8 9
Convolutional Projection of Flattened Patches
Frequency
Time frame
4 x 4
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

Figure 10: OPERA-GT architecture.

OPERA-GT: OPERA-GT is a generative pretrained transformer model. It uses a masked autoencoder to extract useful features from masked spectrograms, which a decoder then uses to reconstruct 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 10.

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 10. 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.

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

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.

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.

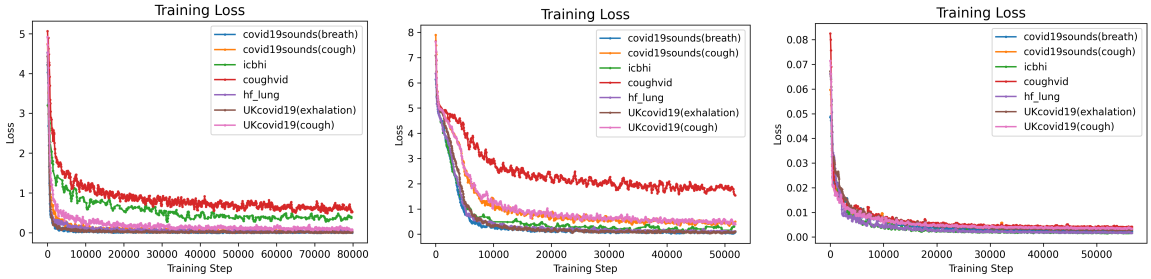
For baselines, both the data pre-processing and feature extraction strictly follow their official implementation. For our pretrained models, the same audio preprocessing is used as in pretraining. The required audio input length is also summarized in Table 9.

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 10) 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.

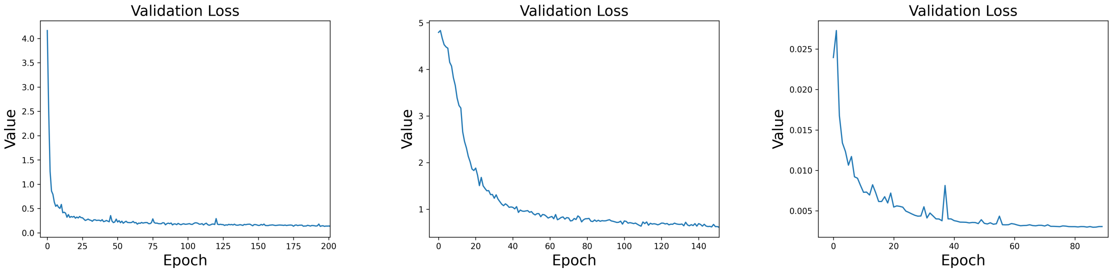
A.3 Pretraining Results

Pretraining loss. We showcase the training process of our three OPERA models here. Specifically, Figure 11 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 12 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

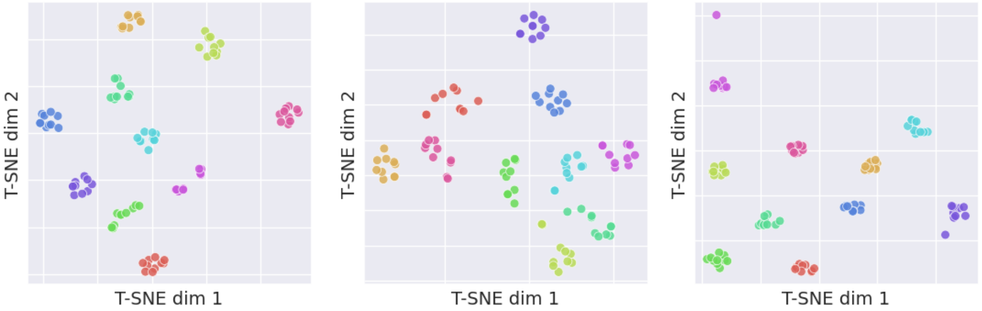
Figure 11: Training loss of the three OPERA models. The OPERA-GT and OPERA-CE use contrastive instance discrimination loss, while OPERA-GT uses generative mean square error loss.



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

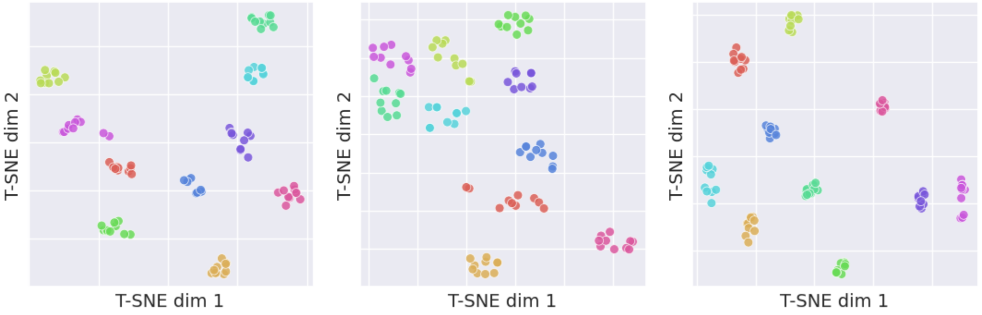
Figure 12: Validation loss of the three OPERA models. The OPERA-GT and OPERA-CE use contrastive instance discrimination loss, while OPERA-GT uses generative mean square error loss.

Embedding distribution analysis for constructive pretraining. Figure 13 and Figure 14 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)

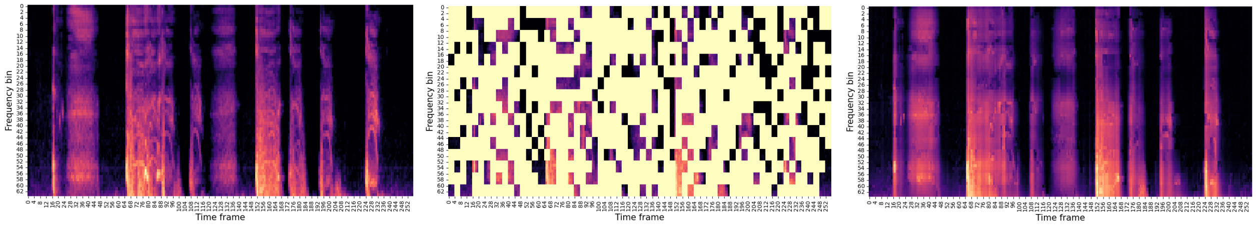
Figure 13: T-SNE visualization result of features from OPERA-CT on the held-out validation of pretraining data. Each dot is an audio segment and the same color represents the same audio recording. It can be seen that audio segments from the same recording are close to each other while far away from other recordings in the embedding space.



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

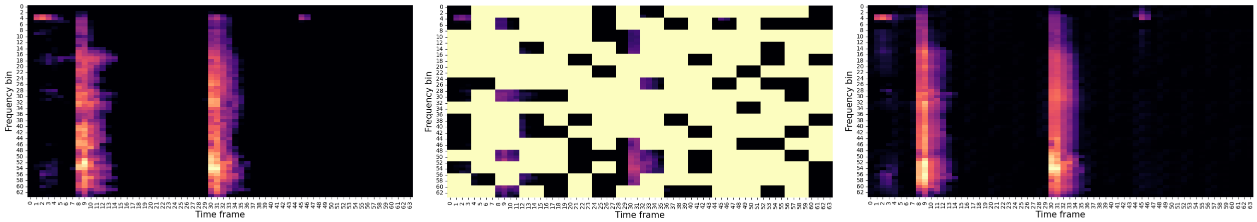
Figure 14: T-SNE visualization result of features from OPERA-CE on the validation data.

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 12(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 15, Figure 16, Figure 17. 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.



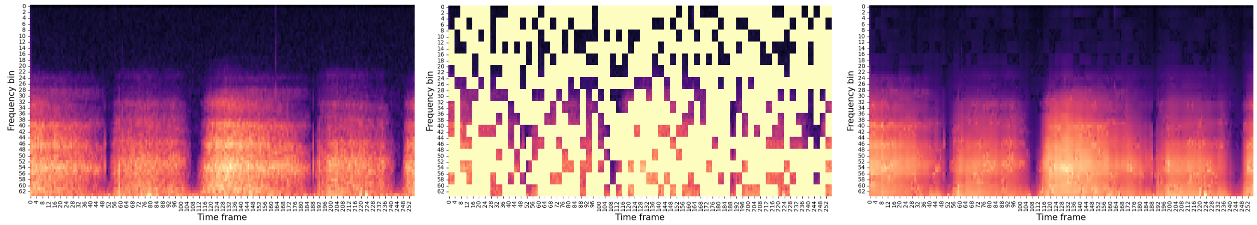
(a) Original spectrogram (b) Masked spectrogram (c) Reconstructed spectrogram

Figure 15: 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 16: Reconstruction result for a cough sound recording (cropped into 2s) from COUGHVID dataset.

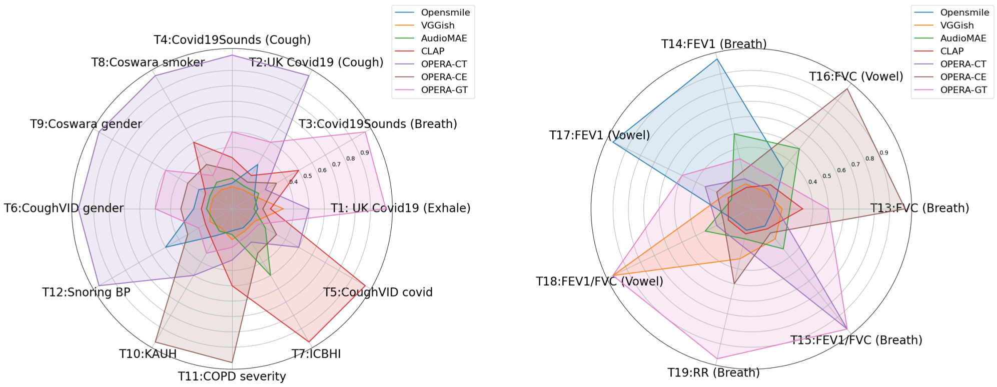


(a) Original spectrogram (b) Masked spectrogram (c) Reconstructed spectrogram

Figure 17: Reconstruction result for a lung sound recording (cropped into 8s) from ICBHI dataset.

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 18.



(a) Health Condition Inference

(b) Lung Function Estimation

Figure 18: 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)	0.329 ± 0.338	0.298 ± 0.252	0.299 ± 0.245	0.295 ± 0.222	0.304 ± 0.259	0.278 ± 0.261	0.291 ± 0.247	✓*
T14	FEV1 (Breath)	0.353 ± 0.469	0.394 ± 0.444	0.392 ± 0.480	0.396 ± 0.435	0.399 ± 0.449	0.381 ± 0.447	0.392 ± 0.466	
T15	FEV1/FVC (Breath)	0.178 ± 0.219	0.167 ± 0.165	0.164 ± 0.163	0.174 ± 0.177	0.161 ± 0.152	0.166 ± 0.149	0.162 ± 0.150	✓*
T16	FVC (Vowel)	0.277 ± 0.238	0.294 ± 0.246	0.280 ± 0.253	0.292 ± 0.247	0.292 ± 0.233	0.264 ± 0.260	0.293 ± 0.255	✓*
T17	FEV1 (Vowel)	0.342 ± 0.363	0.396 ± 0.446	0.417 ± 0.462	0.402 ± 0.409	0.359 ± 0.372	0.398 ± 0.455	0.368 ± 0.440	*
T18	FEV1/FVC (Vowel)	0.175 ± 0.183	0.167 ± 0.164	0.167 ± 0.157	0.176 ± 0.170	0.167 ± 0.153	0.171 ± 0.162	0.167 ± 0.158	✓*
T19	Breathing Rate	0.212 ± 0.080	0.205 ± 0.080	0.207 ± 0.086	0.207 ± 0.084	0.207 ± 0.099	0.193 ± 0.065	0.186 ± 0.071	✓*

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

A.4.3 Cross-domain Zero-shot Performance

Zero-shot capacity is an particularly interesting trait for foundation models, especially LLM-based models. Though this is uncommon for models trained solely with unlabeled non-textual data, we also explore cross-domain zero-shot performance following [40]. We train a linear probe on source Task A and test it on target Task B, using T6 → T9 and T7 → T10 as examples, given their similarity (ref. Table 2). Table below shows that OPERA-CT outperforms the baselines.

Table 13: AUROC (higher is better) for cross domain zero-shot performance. Best model highlighted.

Method	Opensmile	VGGish	AudioMAE	CLAP	OPERA-CT
T6 → T9	0.534 ± 0.048	0.537 ± 0.025	0.472 ± 0.003	0.457 ± 0.005	0.600 ± 0.009
T7 → T10	0.682 ± 0.014	0.588 ± 0.002	0.692 ± 0.003	0.722 ± 0.002	0.823 ± 0.001

A.4.4 Performance for different model architectures

To investigate whether models trained using OPERA data consistently outperforms models trained with general audio data, comparison using consistent model architectures is also important. We used the same ViT from AudioMAE in OPERA-GT. Similarly, for CNNs, we pretrained VGG (same as VGGish) and CNN14 (as CLAP) using the contrastive objective. While in the main paper we chose to showcase OPERA-CE for its competitive performance and potential in constrained scenarios, we include the results here. The better performance of our models suggests the superiority of our curated respiratory audio data and pretrained models for respiratory health.

Table 14: Average AUROC (higher is better) for the health condition inference tasks.

Model	VGG	CNN14	ViT
General audio	0.584	0.676	0.627
OPERA data	0.653	0.692	0.674

A.4.5 Performance for a hybrid model

Given that contrastive and generative pretraining objectives bring different strengths and weaknesses, we also explored training a model that combines both. Using the ViT encoder, we employed a projection head for contrastive learning and a decoder to reconstruct the spectrogram. Preliminary results indicate that while this combined objective yields a model with more balanced performance, it does not consistently outperform the single-objective pretraining approach. We report the performance in Table 15 and Table 16, which can be compared with Table 4 and Table 5.

Table 15: AUROC (higher is better) of the hybrid model for the health condition inference tasks.

Task ID	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
Hybrid	0.575	0.692	0.622	0.711	0.558	0.730	0.886	0.671	0.759	0.652	0.655	0.737

Table 16: MAE (lower is better) of the hybrid model for the health condition inference tasks.

Task ID	T13	T14	T15	T16	T17	T18	T19
Hybrid	0.886	0.797	0.124	0.889	0.805	0.133	2.457

A.4.6 Significance tests

We conducted significance tests for all tasks and the p values indicating significance is shown in Table 17. Compared to the baselines, our models show a significant improvement in most cases. When compared to the best baseline, OPERA-CT performs better (a higher average of AUROC) on 8 tasks, with 5 of these improvements being statistically significant. Our github repo also provides an easy-to-use significance test function for benchmarking purposes and further use.

Table 17: P-values for significance tests (t-test) for Tasks 1-12. Significant values are highlighted in yellow (p<0.01). The cases where OPERA models outperform the best baseline are underlined.

Dataset	ID	Best baseline	OPERA-CT	OPERA-CE	OPERA-GT
UK COVID-19	T1	VGGish	0.0001	0.0002	0.4230
	T2	CLAP	0.0000	0.0022	0.0000
COVID-19 Sounds	T3	CLAP	0.0075	0.2155	0.9161
	T4	CLAP	0.0558	0.0000	0.0011
CoughVID	T5	CLAP	0.0003	0.0003	0.0000
	T6	CLAP	0.0000	0.0000	0.0000
ICBHI	T7	CLAP	0.0000	0.0000	0.0000
Coswara	T8	CLAP	0.1586	0.8547	0.0000
	T9	Opensmile	0.0000	0.0000	0.0000
KAUH	T10	CLAP	0.0183	0.0003	0.9875
Respiratory@TR	T11	CLAP	0.7182	0.0439	0.4200
SSBPR	T12	Opensmile	0.0027	0.9944	0.0000