OPERA

Towards Open Respiratory Acoustic Foundation Models: Pretraining and Benchmarking

Yuwei Zhang, Tong Xia, Jing Han, Yu Wu, Georgios Rizos, Yang Liu,
Mohammed Mosuily, Jagmohan Chauhan, Cecilia Mascolo

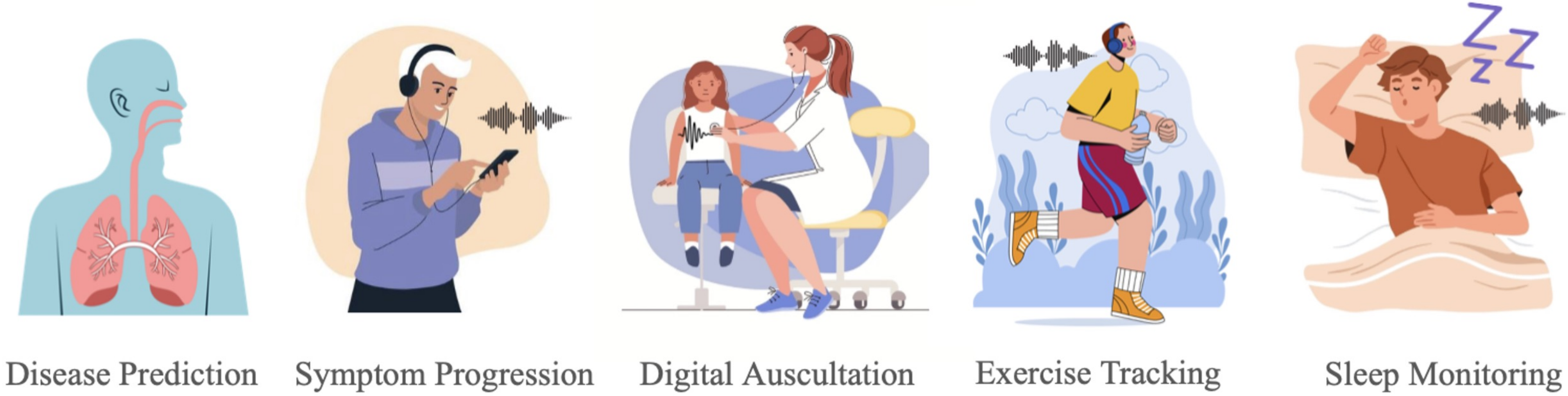
UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology

University of Southampton

Yuwei (Evelyn) Zhang, NeurIPS 2024

Motivation

- Potential of respiratory audio in healthcare
  - disease detection
  - health monitoring



Disease Prediction

Symptom Progression

Digital Auscultation

Exercise Tracking

Sleep Monitoring

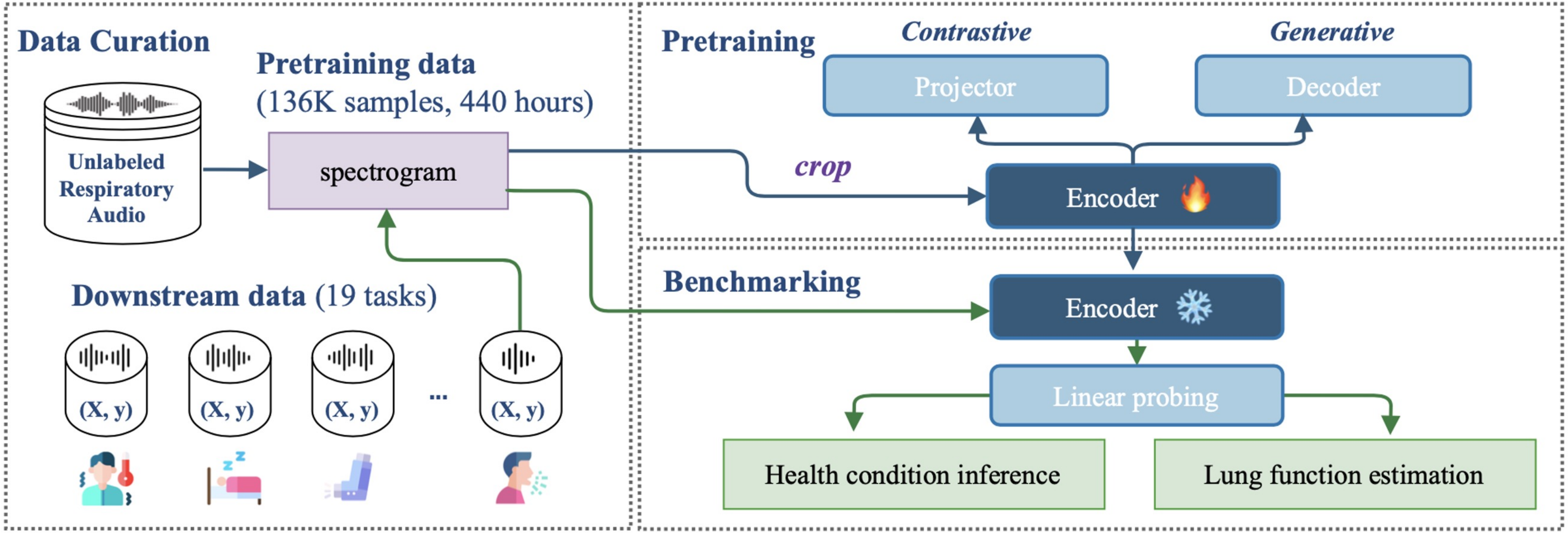
- Challenges in collecting large labeled datasets for specific tasks
- Need for ***generalizable*** and ***open*** foundation models

Current Literature and Challenges

- **Data**: Large amounts of respiratory audio data exist, but a comprehensive, curated collection is missing.
- **Model**: There is a lack of open-source foundation models specifically designed for respiratory audio analysis.
- **Benchmark**: No ready-to-use benchmark exist for evaluating the performance of respiratory audio foundation models.

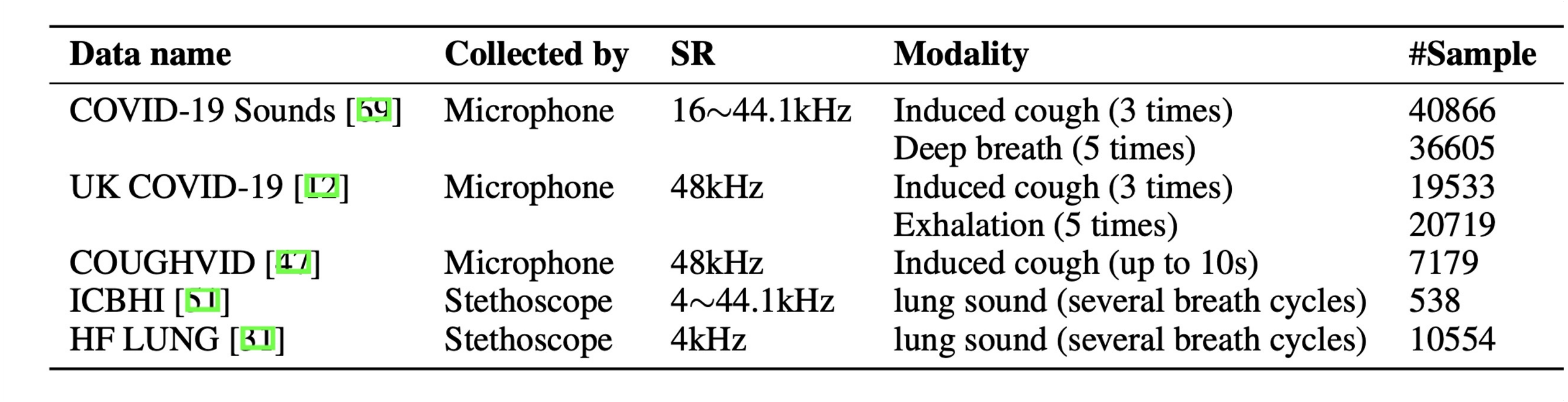
Introduction to OPERA

Goals: Curate large ***datasets***, pretrain acoustic ***models***, ***benchmark*** on various tasks



Pretraining datasets

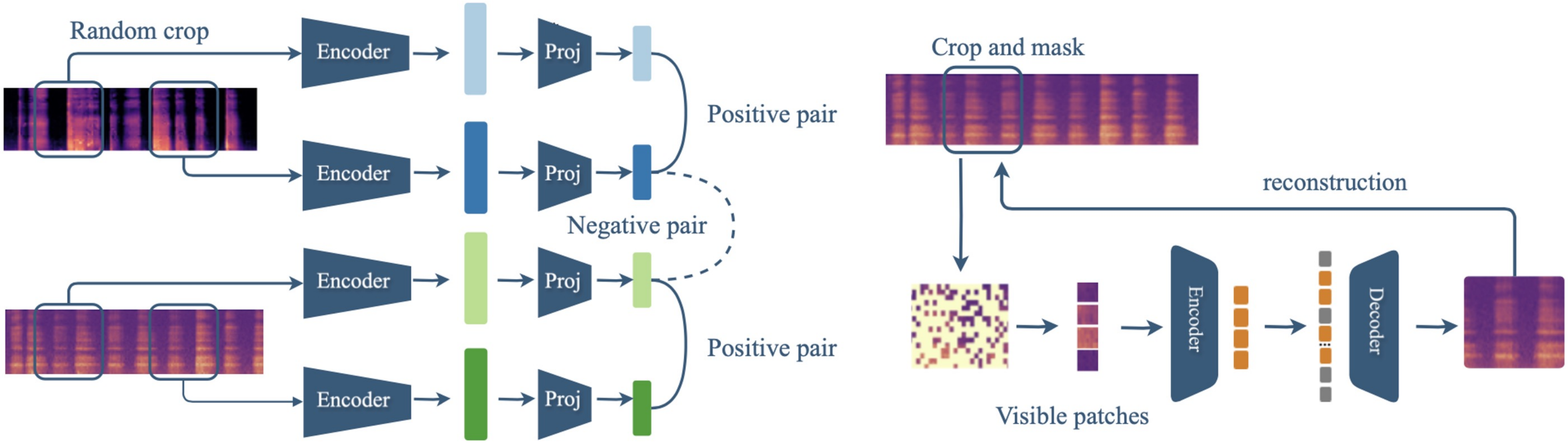
- Curated datasets (136K samples, 440 hours)
  - Sources: COVID-19 Sounds, UK COVID-19, COUGHVID, ICBHI, HF LUNG
  - Types: Breathing, coughing, lung sounds



Data name	Collected by	SR	Modality	#Sample
COVID-19 Sounds [69]	Microphone	16∼44.1kHz	Induced cough (3 times)	40866
			Deep breath (5 times)	36605
UK COVID-19 [12]	Microphone	48kHz	Induced cough (3 times)	19533
			Exhalation (5 times)	20719
COUGHVID [47]	Microphone	48kHz	Induced cough (up to 10s)	7179
ICBHI [51]	Stethoscope	4∼44.1kHz	lung sound (several breath cycles)	538
HF LUNG [31]	Stethoscope	4kHz	lung sound (several breath cycles)	10554

Pretraining approaches

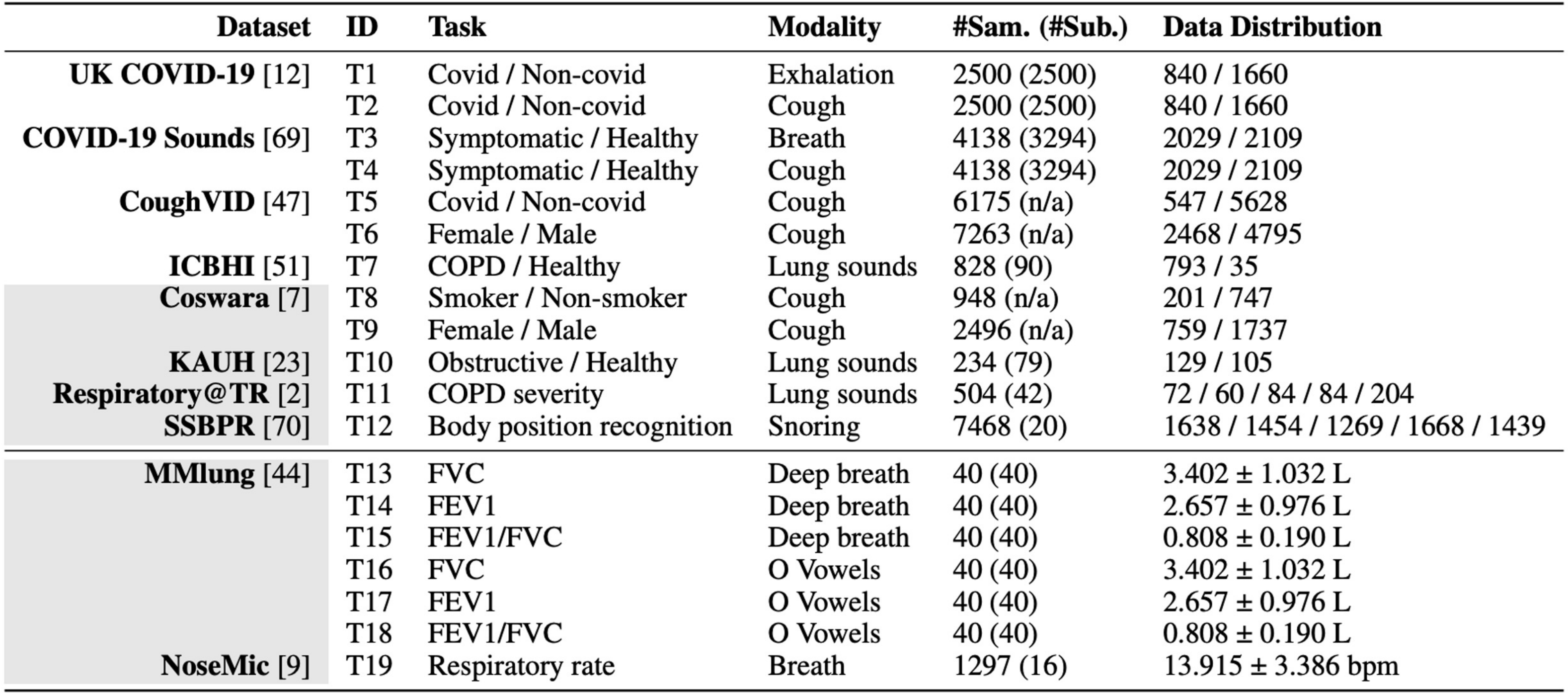
- Contrastive Learning
  - Transformer-based (OPERA-CT)
  - CNN-based (OPERA-CE)
- Generative Pretraining (OPERA-GT)
  - Vision Transformer with masked spectrograms



(a) Contrastive (OPERA-CT, OPERA-CE)

(b) Generative (OPERA-GT)

Benchmarking



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

⬆️ **Unseen data sources**

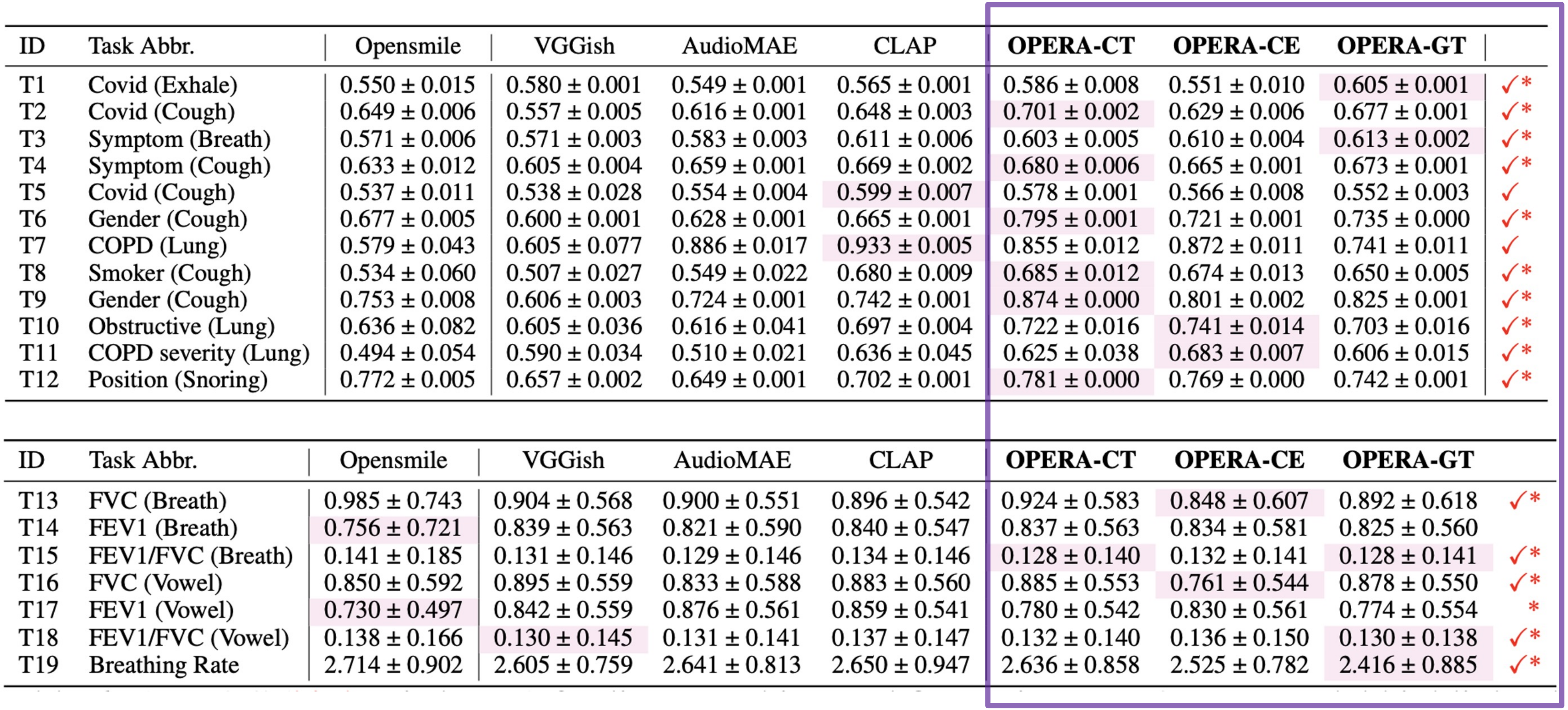
Comparing with baselines (linear evaluation)

- OPERA pretrained models
  - **OPERA-CT**: Contrastive learning with transformers
  - **OPERA-CE**: Contrastive learning with CNN
  - **OPERA-GT**: Generative learning with transformers
- **OpenSMILE** feature set
- General Audio Pretrained Models
  - **VGGish** [1]: supervised pretraining
  - **AudioMAE** [2]: unsupervised pretraining
  - **CLAP** [3]: language supervised pretraining

[1] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Wilson, K. "CNN architectures for large-scale audio classification." ICASSP 2017.
[2] Huang, P. Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., ... & Feichtenhofer, C. "Masked autoencoders that listen." NeurIPS 2022.
[3] Elizalde, B., Deshmukh, S., Al Ismail, M., & Wang, H. "Clap learning audio concepts from natural language supervision." ICASSP 2023.

Results

outperforming in ***16 out of 19*** tasks



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)	0.537 ± 0.011	0.538 ± 0.028	0.554 ± 0.004	0.599 ± 0.007	0.578 ± 0.001	0.566 ± 0.008	0.552 ± 0.003	✓
T6	Gender (Cough)	0.677 ± 0.005	0.600 ± 0.001	0.628 ± 0.001	0.665 ± 0.001	0.795 ± 0.001	0.721 ± 0.001	0.735 ± 0.000	✓*
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)	0.636 ± 0.082	0.605 ± 0.036	0.616 ± 0.041	0.697 ± 0.004	0.722 ± 0.016	0.741 ± 0.014	0.703 ± 0.016	✓*
T11	COPD severity (Lung)	0.494 ± 0.054	0.590 ± 0.034	0.510 ± 0.021	0.636 ± 0.045	0.625 ± 0.038	0.683 ± 0.007	0.606 ± 0.015	✓*
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	✓*

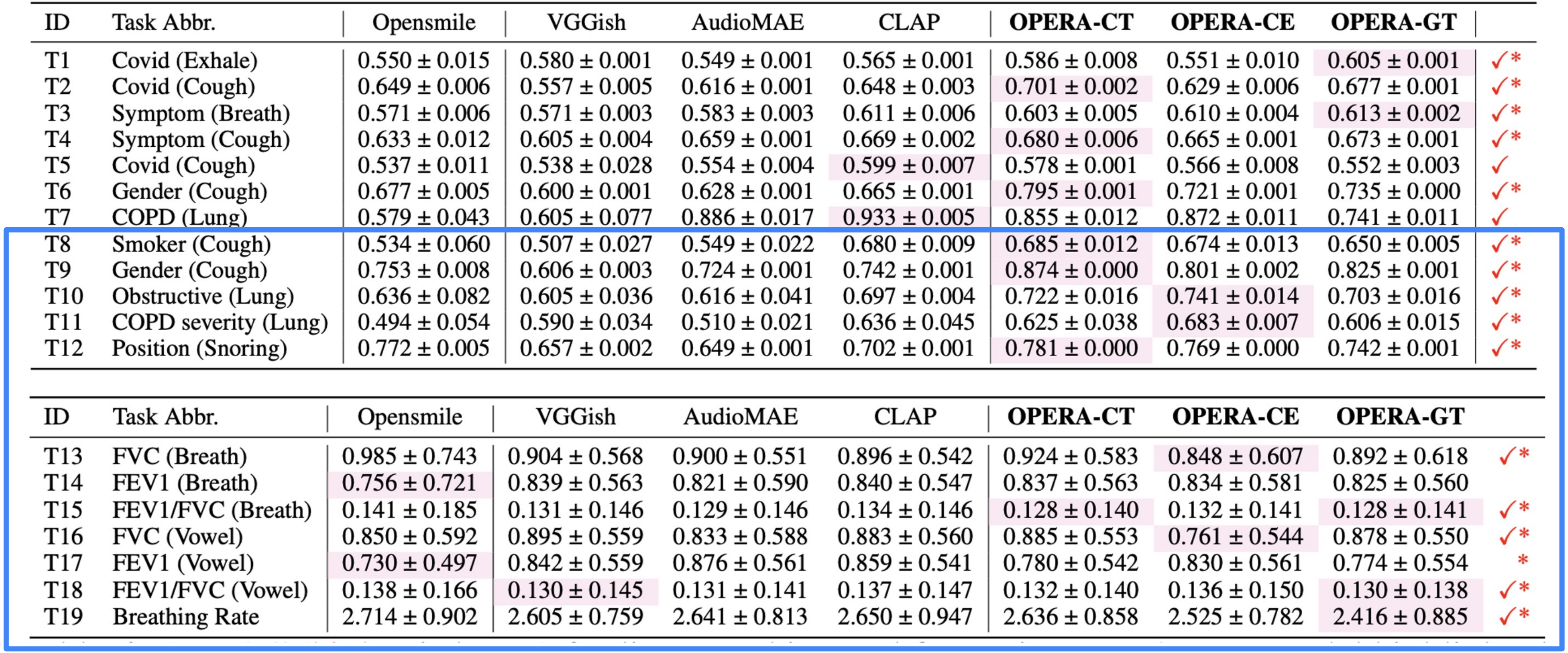
ID	Task Abbr.	Opensmile	VGGish	AudioMAE	CLAP	**OPERA-CT**	**OPERA-CE**	**OPERA-GT**	
T13	FVC (Breath)	0.985 ± 0.743	0.904 ± 0.568	0.900 ± 0.551	0.896 ± 0.542	0.924 ± 0.583	0.848 ± 0.607	0.892 ± 0.618	✓*
T14	FEV1 (Breath)	0.756 ± 0.721	0.839 ± 0.563	0.821 ± 0.590	0.840 ± 0.547	0.837 ± 0.563	0.834 ± 0.581	0.825 ± 0.560	
T15	FEV1/FVC (Breath)	0.141 ± 0.185	0.131 ± 0.146	0.129 ± 0.146	0.134 ± 0.146	0.128 ± 0.140	0.132 ± 0.141	0.128 ± 0.141	✓*
T16	FVC (Vowel)	0.850 ± 0.592	0.895 ± 0.559	0.833 ± 0.588	0.883 ± 0.560	0.885 ± 0.553	0.761 ± 0.544	0.878 ± 0.550	✓*
T17	FEV1 (Vowel)	0.730 ± 0.497	0.842 ± 0.559	0.876 ± 0.561	0.859 ± 0.541	0.780 ± 0.542	0.830 ± 0.561	0.774 ± 0.554	*
T18	FEV1/FVC (Vowel)	0.138 ± 0.166	0.130 ± 0.145	0.131 ± 0.141	0.137 ± 0.147	0.132 ± 0.140	0.136 ± 0.150	0.130 ± 0.138	✓*
T19	Breathing Rate	2.714 ± 0.902	2.605 ± 0.759	2.641 ± 0.813	2.650 ± 0.947	2.636 ± 0.858	2.525 ± 0.782	2.416 ± 0.885	✓*

Findings

- ***Superiority*** over existing acoustic models
    - outperforming in ***16 out of 19*** tasks

- ***Generalizability*** to unseen data sources and respiratory audio types
    - 12 tasks from unseen datasets and respiratory audio types
    - OPERA models achieving the best performance on ***10 out of 12***

Results



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)	0.537 ± 0.011	0.538 ± 0.028	0.554 ± 0.004	0.599 ± 0.007	0.578 ± 0.001	0.566 ± 0.008	0.552 ± 0.003	✓
T6	Gender (Cough)	0.677 ± 0.005	0.600 ± 0.001	0.628 ± 0.001	0.665 ± 0.001	0.795 ± 0.001	0.721 ± 0.001	0.735 ± 0.000	✓*
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)	0.636 ± 0.082	0.605 ± 0.036	0.616 ± 0.041	0.697 ± 0.004	0.722 ± 0.016	0.741 ± 0.014	0.703 ± 0.016	✓*
T11	COPD severity (Lung)	0.494 ± 0.054	0.590 ± 0.034	0.510 ± 0.021	0.636 ± 0.045	0.625 ± 0.038	0.683 ± 0.007	0.606 ± 0.015	✓*
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	✓*

ID	Task Abbr.	Opensmile	VGGish	AudioMAE	CLAP	**OPERA-CT**	**OPERA-CE**	**OPERA-GT**	
T13	FVC (Breath)	0.985 ± 0.743	0.904 ± 0.568	0.900 ± 0.551	0.896 ± 0.542	0.924 ± 0.583	0.848 ± 0.607	0.892 ± 0.618	✓*
T14	FEV1 (Breath)	0.756 ± 0.721	0.839 ± 0.563	0.821 ± 0.590	0.840 ± 0.547	0.837 ± 0.563	0.834 ± 0.581	0.825 ± 0.560	
T15	FEV1/FVC (Breath)	0.141 ± 0.185	0.131 ± 0.146	0.129 ± 0.146	0.134 ± 0.146	0.128 ± 0.140	0.132 ± 0.141	0.128 ± 0.141	✓*
T16	FVC (Vowel)	0.850 ± 0.592	0.895 ± 0.559	0.833 ± 0.588	0.883 ± 0.560	0.885 ± 0.553	0.761 ± 0.544	0.878 ± 0.550	✓*
T17	FEV1 (Vowel)	0.730 ± 0.497	0.842 ± 0.559	0.876 ± 0.561	0.859 ± 0.541	0.780 ± 0.542	0.830 ± 0.561	0.774 ± 0.554	*
T18	FEV1/FVC (Vowel)	0.138 ± 0.166	0.130 ± 0.145	0.131 ± 0.141	0.137 ± 0.147	0.132 ± 0.140	0.136 ± 0.150	0.130 ± 0.138	✓*
T19	Breathing Rate	2.714 ± 0.902	2.605 ± 0.759	2.641 ± 0.813	2.650 ± 0.947	2.636 ± 0.858	2.525 ± 0.782	2.416 ± 0.885	✓*

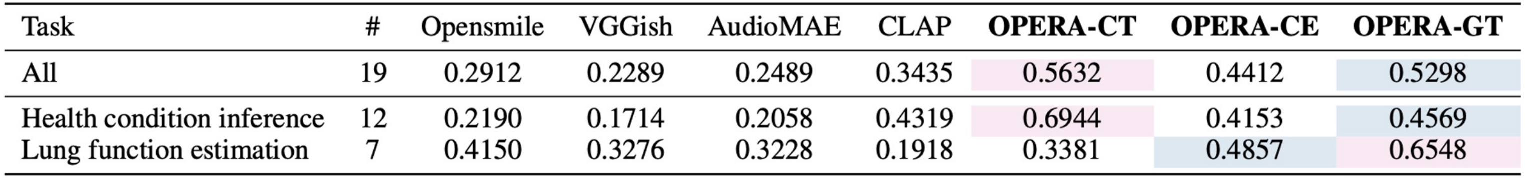
Generalizability to unseen data sources and respiratory audio types

Findings

- ***Superiority*** over existing acoustic models
  - outperforming in ***16 out of 19*** tasks

- ***Generalizability*** to unseen data sources and respiratory audio types
  - 12 tasks from unseen datasets and respiratory audio types
  - OPERA models achieving the best performance on ***10 out of 12***

- ***Training design:***
  - Contrastive models excel in classification tasks
  - Generative models excel in regression tasks



Task	#	Opensmile	VGGish	AudioMAE	CLAP	**OPERA-CT**	**OPERA-CE**	**OPERA-GT**
All	19	0.2912	0.2289	0.2489	0.3435	0.5632	0.4412	0.5298
Health condition inference	12	0.2190	0.1714	0.2058	0.4319	0.6944	0.4153	0.4569
Lung function estimation	7	0.4150	0.3276	0.3228	0.1918	0.3381	0.4857	0.6548

Conclusion and Future Directions

- Importance of open-source models and datasets for research growth
  - Availability of OPERA resources on GitHub: https://github.com/evelyn0414/OPERA
  - Model checkpoints on HuggingFace: https://huggingface.co/evelyn0414/OPERA/tree/main.
- Future Directions
  - data efficient fine-tuning
  - the scaling law
  - novel pretraining strategies for unlabeled health audio

THANK YOU!

Web Page
https://opera-benchmark.github.io

OPERA

GitHub
evelyn0414/OPERA

UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology

Yuwei (Evelyn) Zhang
yz798@cl.cam.ac.uk