npj Digital Medicine

https://doi.org/10.1038/s41746-026-02445-4

Article in Press

A device-invariant multi-modal learning framework for respiratory disease classification

Received: 5 September 2025

Accepted: 7 February 2026

Cite this article as: Yang, M., Liu, X., Du, W. *et al.* A device-invariant multi-modal learning framework for respiratory disease classification. *npj Digit. Med.* (2026). https://doi.org/10.1038/s41746-026-02445-4

Mo Yang, Xuefei Liu, Wei Du, Yang Liu, Wenyu Zhu, Zhaoyang Bu, Jiaxuan Mao, Qian Wang, Si Chen, Min Zhou & Jie-ming Qu

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

A Device-Invariant Multi-Modal Learning Framework for Respiratory Disease Classification

Mo Yang[1], Xuefei Liu[2,3†], Wei Du[2,3†], Yang Liu[1], Wenyu Zhu[1], Zhaoyang Bu[1], Jiaxuan Mao[1], Qian Wang[1*], Si Chen[1*], Min Zhou[2,3*], Jie-ming Qu[2,3*]

[1]Research&Development Department, Luca Healthcare, No.317 Xianxia Road, Shanghai, 200051, Shanghai, China.
[2]Department of Pulmonary and Critical Care Medicine, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, No. 197 Ruijin Second Road, Shanghai, 200025, Shanghai, China.
[3]Institute of Respiratory Diseases, Shanghai Jiao Tong University School of Medicine, No. 197 Ruijin Second Road, Shanghai, 200025, Shanghai, China.

*Corresponding author(s). E-mail(s): adam.wang@lucahealthcare.com; echo.chen@lucahealthcare.com; doctor_zhou_99@163.com; jmqu0906@163.com;
Contributing authors: myang5061@outlook.com; shjtdxlxf@163.com; duweiwilson@qq.com; gavin.liu@lucahealthcare.com; wendy.zhu@lucahealthcare.com; buzhaoyang@lucahealthcare.com; mary.mao@lucahealthcare.com;
†These authors contributed equally to this work.

Abstract

Recent advances in cough sound analysis using deep learning techniques enable smartphone-based respiratory disease screening suitable for self-management care in a home setting, yet their utility is limited by device heterogeneity, population diversity, and challenges in multimodal integration. We propose a device-invariant, multimodal deep learning framework that jointly models cough acoustics, demographic data, and symptom descriptions for multi-label classification of adult respiratory diseases. To address the issues of device effect, an adversarial branch is embedded in the audio encoder to enforce device-invariant feature learning, while an invariant risk minimization-augmented loss enhances robustness to non-structural shifts. To evaluate the effectiveness of our proposed method, a real-world, multi-center dataset containing over 10,000 cases spanning seven major respiratory conditions was curated. On the tasks of individual respiratory disease identification for chronic obstructive pulmonary disease (COPD), lower respiratory tract infection (LRTI) and pulmonary shadows (PS), our method achieves superior performance with the area under the receiver operating characteristic curve (AUROC) of 0.9698, 0.8483 and 0.8720, respectively. It also shows promising results in identifying the presence of comorbidities for 7 respiratory diseases with an overall AUROC of 0.8907. More importantly, extensive experimental results demonstrate our method mitigates the issues of device effect and facilitates the cross-device generalization for cough-based respiratory disease diagnoses. This work demonstrates a scalable and transferable AI-based approach for cough-driven respiratory screening, emphasizing the importance of multimodal fusion and robust representation learning in advancing clinical applicability.

Keywords: Device-Invariant, Multi-Modal, Feature Fusion, Joint Loss Function, Respiratory Disease Classification

Introduction

Respiratory diseases represent a major global public health concern, contributing significantly to morbidity and mortality, particularly among adults and the elderly[1]. Conditions such as chronic obstructive pulmonary disease (COPD), interstitial lung disease (ILD), chronic bronchitis (CB), and various infectious diseases (e.g., lower respiratory tract infection, LRTI) often present with non-specific early symptoms—such as cough, sputum production, and shortness of breath—with coughs being one of the most common and earliest clinical indicators. Traditional methods for respiratory disease screening typically rely on chest imaging, pulmonary function tests[2], or auscultation by trained clinicians. These approaches, however, often require specialized equipment and personnel[3][4], making them difficult to scale in community settings or resource-limited environments[5]. In recent years, advances in artificial intelligence (AI) and acoustic modeling have facilitated the development of automated cough-based diagnostic tools. These systems offer a non-invasive, low-cost, and accessible alternative with the potential for remote deployment, positioning them as a promising solution for early respiratory disease screening and intelligent diagnostic support[6][7].

Despite encouraging progress in cough sound analysis[8][9], several limitations still hinder the real-world deployment of such AI models: first, distributional shifts in audio data caused by device variability significantly compromise model stability and generalization[10]. In the home setting, cough sounds may be recorded using a wide range of devices that differ in brand, model, and microphone placement[11]. Without explicit mechanisms for device-invariant learning, models often suffer performance degradation or even complete failure when deployed on unseen hardware. Second, most existing approaches rely solely on unimodal audio features, neglecting the rich diagnostic information embedded in patient demographics (e.g., age, sex, body composition, etc.) and symptom descriptions (e.g., sputum production, fever, dyspnea, etc.). This limited scope restricts the model's ability to develop a holistic understanding of disease presentations. Furthermore, in real-world clinical settings, patients frequently present with multiple co-occurring respiratory conditions. However, most prior models are designed for single-label classification, making them ill-suited to capture the complex, multi-label pathologies seen in clinical practice[12].

To address the aforementioned challenges, we propose a device-invariant, multimodal deep learning framework for the identification and classification of adult respiratory diseases. Our approach integrates cough audio, demographic data, and symptom descriptions to systematically exploit the complementary information across modalities. We introduce an adversarial training mechanism into the audio encoder[13], employing a gradient reversal strategy to counteract a device classifier and learn representations invariant to recording device differences. To further enhance robustness to distributional shifts between training and deployment environments, we incorporate invariant risk minimization (IRM)[14] into a joint loss function[15]. Moreover, we design a unified multi-label learning framework capable of recognizing multiple respiratory diseases simultaneously, reflecting the real-world prevalence of comorbid conditions. Finally, we conduct a comprehensive evaluation on a multi-center real-world dataset comprising over 10,000 adult patient samples, demonstrating the model's superiority in terms of accuracy, robustness, and generalization.

The proposed framework bridges the gap between experimental research and real-world clinical deployment of cough-based AI systems. It shows strong potential for application in frontline healthcare, remote health monitoring, and chronic disease screening among elderly populations.

Results

We conducted a comprehensive evaluation of the proposed multimodal deep learning framework for respiratory disease classification, with a particular focus on its effectiveness in multimodal integration, robustness to device-related biases, and adaptability to real-world deployment scenarios.

Data Collection

We established a large-scale, multicenter cohort comprising 12,378 adult outpatients ($\geq$18 years) recruited from respiratory clinics at four independent clinical centers. Written consent was obtained from all participants. The study was conducted in accordance with relevant ethical guidelines and approved by the ethics committees of Ruijin Hospital Shanghai Jiaotong University School of Medicine Ethics Committee (No. 2023199), Ruijin Hospital Luwan Branch Ethics Committee (2023HXK-V1), Shanghai Jing'an District Central Hospital Ethics Committee (No. 2023-33), Shanghai Zhabei Central Hospital Ethics Committee (ZBLL2024030401001). The trial is registered at ClinicalTrials.gov (NCT06082791).

For each participant, at least 10 seconds of voluntary cough audio were recorded in a relatively quiet environment. All recordings underwent an internally validated quality control (QC) pipeline comprising event segmentation, cough detection, and validity assessment (as detailed in the Cough Sound Quality Control Process section). From each recording, multiple standardized 3-second cough segments were extracted, anchored at the cough onset and spanning 1 second before and 2 seconds after the burst, with recording device metadata documented for each segment.

To capture comprehensive clinical context, participants completed structured questionnaires encompassing three domains: demographic information (e.g., sex, age, height, weight), smoking history (current and former status), and respiratory symptom profiles, including cough frequency and duration, sputum characteristics, dyspnea, and other relevant signs. Detailed information on subject questionnaires are provided in Supplementary Figure 1. Disease labels were assigned based on final diagnoses by attending physicians and further verified through dual review by two independent senior pulmonologists, with multiple rounds of QC to ensure annotation accuracy. Detailed definitions and diagnostic criteria for the seven disease categories are provided in Supplementary Table 1, and the pulmonary shadows (PS) label is further described in Supplementary Table 2.

Partitioning the collected data in different ways resulted in several datasets designed to support distinct evaluation tasks. For the key binary classification analyses, three representative conditions were selected to enable independent positive–negative recognition: COPD (representing chronic airway disease, Figure 1a), LRTI (representing infectious disease, Figure 1b), and PS (representing potential tumors or structural abnormalities, Figure 1c). For the COPD and LRTI tasks, we conducted five-fold cross-validation. Recordings originating from the same device and clinical center were confined to a single fold to strictly prevent information leakage across folds. All three binary classification tasks exhibited considerable class imbalance. In the COPD task, the dataset comprised 813 positive cases (COPD) and 4,933 negative cases (Other), with positives accounting for 14.1% of the total. In the LRTI task, 720 subjects were labeled as positive (LRTI) and 4,894 as negative (Other), corresponding to a positive proportion of 12.8%. The PS task exhibited the greatest imbalance, with only 432 positive cases (PS) compared to 6,379 negatives (Other), corresponding to a positive rate of 6.3%. These distributions indicate that the PS classification problem poses the most challenging imbalance, while COPD and LRTI tasks remain moderately skewed toward negative samples.

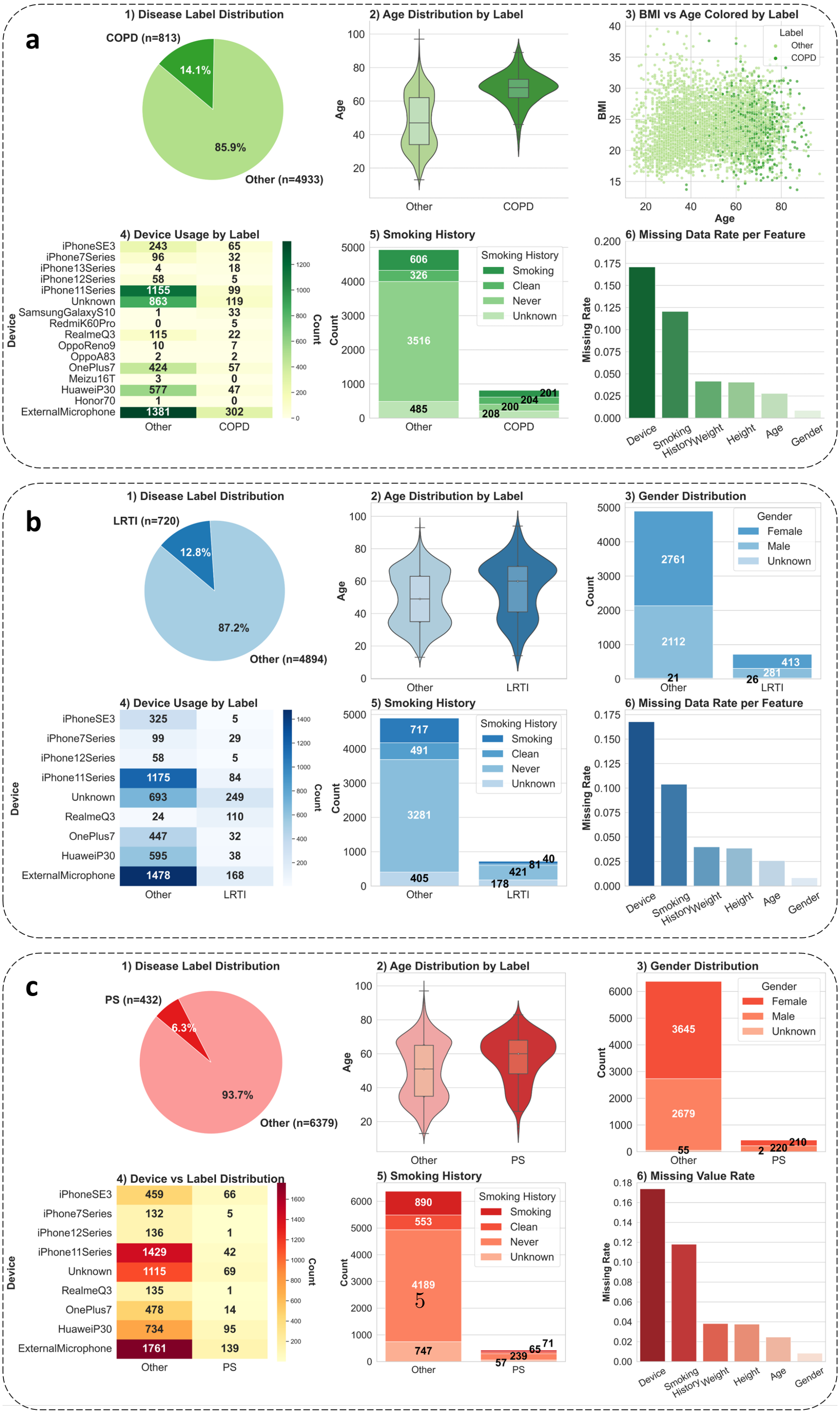

Fig. 1 | Data distribution across three binary classification tasks.

In parallel, a multi-label respiratory disease recognition task was constructed to capture disease co-occurrence, allowing each subject to be annotated with the presence (1) or absence (0) of 7 conditions (i.e. COPD, Asthma, ILD, upper respiratory tract infection (URTI), LRTI, CB and Bronchiectasis). Models were required to estimate the probability of presence for each disease, with task-specific data distributions presented in Figure 2, with asthma (10.3%) and COPD (9.4%) being most prevalent, followed by LRTI (7.8%) and multi-disease cases (4.1%). Recordings were obtained across heterogeneous devices, with iPhone series and external microphones dominating. The cohort spanned a wide age range (Figure 2 pannel 4) with expected gender differences in height, weight, and smoking history, the latter showing markedly higher prevalence among males.

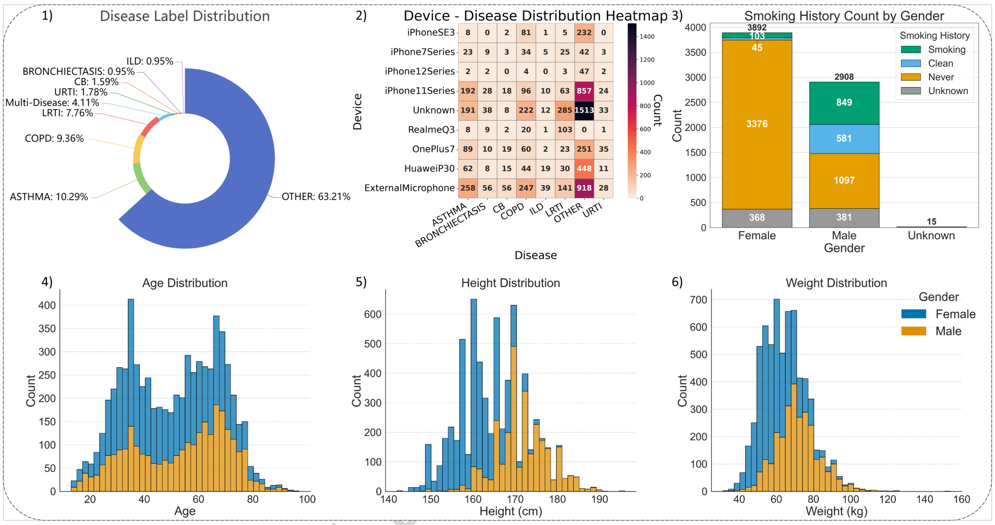

Fig. 2 | Data distribution of the multi-label respiratory disease classification task.

Binary Classification Results for Chronic Obstructive Pulmonary Disease

For the binary classification task of COPD, we employed a five-fold cross-validation strategy for performance evaluation. Table 1 summarizes the mean and standard deviation of area under the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPRC), and reports the mean and standard deviation of the F1-score, precision, sensitivity, specificity, and overall accuracy calculated under two thresholding strategies: a manually defined threshold and the Youden Index method. The AUROC, AUPRC, and their corresponding confidence intervals (AUROC_CI) for each fold, as well as detailed metric reports under different thresholds, are provided in Supplementary Tables 3-4. The results indicate that our proposed model achieved mean AUROC values exceeding 0.96 on both the validation and test sets, with standard deviations below 0.007, demonstrating high stability and superior discriminative capability for this task.

Table 1 | Indicators of the COPD binary classification task

	Validation		Test	
	Mean	Std	Mean	Std
AUROC	0.9674	0.0068	0.9698	0.0066
AUPRC	0.8785	0.0266	0.8851	0.0130
F1[1]	0.8146	0.0100	0.8228	0.0244
Precision[1]	0.7779	0.0389	0.7882	0.0426
Sensitivity[1]	0.8572	0.0278	0.8623	0.0290
Specificity[1]	0.961	0.0095	0.9629	0.0104
Accuracy[1]	0.9468	0.0051	0.9491	0.0092
F1[2]	0.8015	0.0204	0.7918	0.0472
Precision[2]	0.7269	0.0394	0.7021	0.0773
Sensitivity[2]	0.8951	0.0194	0.9143	0.0265
Specificity[2]	0.9463	0.0127	0.936	0.0266
Accuracy[2]	0.9393	0.0094	0.9331	0.0216

Note: For detailed results of each fold, please refer to Supplementary Tables 3-4.

[1] Cut off = 0.15. Manual cutoff thresholds were empirically determined based on validation set distributions to balance sensitivity and specificity for clinically relevant detection.

[2] Cut off value selected by Youden Index.

Binary Classification Results for Lower Respiratory Tract Infection

For the LRTI binary classification task, we employed the same five-fold cross-validation protocol used in the COPD experiments. Table 2 presents the primary performance metrics, while the complete set of results is provided in Supplementary Tables 5–6. Despite the etiological complexity and substantial phenotypic heterogeneity of LRTI in clinical practice, our model consistently achieved mean AUROC values above 0.84 on both the validation and test sets, with standard deviations below 0.03, highlighting its robustness in handling clinically complex and heterogeneous disease presentations.

Binary Classification Results for Pulmonary Shadows

To evaluate the potential of our model for early screening of pulmonary tumors, we defined the inclusion criteria for PS in consultation with clinical experts (see Supplementary Table 2) and excluded cases attributable to infectious etiologies. Given the relatively limited number of such cases, we randomly partitioned the dataset into

Table 2 | Indicators of the LRTI binary classification task

	Validation		Test	
	Mean	Std	Mean	Std
AUROC	0.8491	0.0293	0.8483	0.0198
AUPRC	0.6084	0.0484	0.5858	0.0339
F1[1]	0.6245	0.0177	0.5883	0.0352
Precision[1]	0.6630	0.0593	0.5757	0.0979
Sensitivity[1]	0.5969	0.0553	0.6218	0.0736
Specificity[1]	0.9538	0.0144	0.9259	0.0367
Accuracy[1]	0.9078	0.0056	0.8868	0.0252
F1[2]	0.5826	0.0464	0.5465	0.0512
Precision[2]	0.5013	0.0899	0.443	0.0744
Sensitivity[2]	0.7181	0.0713	0.7293	0.0499
Specificity[2]	0.8872	0.0441	0.8575	0.048
Accuracy[2]	0.8654	0.0309	0.8411	0.0379

Note: For detailed results of each fold, please refer to Supplementary Tables 5–6.

[1] Cut off = 0.10. Manual cutoff thresholds were empirically determined based on validation set distributions to balance sensitivity and specificity for clinically relevant detection.

[2] Cut off value selected by Youden Index.

training, validation, and test sets at a ratio of 3:1:1, ensuring comparable demographic distributions across splits. The evaluation metrics used were consistent with those applied in the COPD and LRTI experiments (Table 3).

The model achieved AUROC values above 0.87 across validation and test sets, indicating robust ranking capability despite the extreme class imbalance (positive rate 6.3%). However, the F1-score (val: 0.3556; test: 0.3418), precision (val: 0.2353; test: 0.2207), and an AUPRC of 0.4785—only moderately above the prevalence-based baseline—highlight the limited precision attainable for this rare-event outcome.

This discrepancy between AUROC and AUPRC reflects the heterogeneous and frequently clinically silent nature of PS, which ranges from incidental ground-glass nodules to more advanced tumors. Early-stage or peripheral lesions typically produce minimal or non-specific symptoms, whereas the model relies primarily on demographics, symptom questionnaires, and voluntary cough acoustics—features that are more indicative of obstructive airway disease or diffuse parenchymal involvement than of small or asymptomatic nodules. In view of these characteristics and the low prevalence of PS, we do not position the model as a stand-alone screening tool for pulmonary nodules or primary lung cancer. Instead, PS scores should be interpreted as a triage or risk-stratification aid to help prioritize individuals for confirmatory imaging, while a low predicted probability should not be used to defer further diagnostic evaluation.

Table 3 | Indicators of the PS binary classification task

Type	Val	Test
AUROC	0.8704	0.8720
AUROC 95% CI	[0.8110, 0.9178]	[0.8066, 0.9241]
AUPRC	0.4427	0.4785
F1[1]	0.3556	0.3418
Precision[1]	0.2353	0.2207
Sensitivity[1]	0.7273	0.7581
Specificity[1]	0.8549	0.8207
Accuracy[1]	0.8475	0.8168
F1[2]	0.3605	0.3125
Precision[2]	0.2360	0.1897
Sensitivity[2]	0.7636	0.8871
Specificity[2]	0.8482	0.7462
Accuracy[2]	0.8433	0.7551

Note:

[1] Cut off = 1×10^{-5}. Manual cutoff thresholds were empirically determined based on validation set distributions to balance sensitivity and specificity for clinically relevant detection.

[2] Cut off value selected by Youden Index.

Performance on Multilabel Respiratory Disease Classification

In the multilabel classification task involving seven distinct respiratory diseases, our proposed framework was benchmarked against a range of established models (Figure 3 a). Leveraging a large-scale, multi-center, multi-device dataset, the model achieved the highest performance in both AUROC (Figure 3 b) and AUPRC (Figure 3 c) metrics. These results significantly outperform baseline approaches, underscoring the framework's strong generalization ability and its potential for deployment in complex clinical environments involving heterogeneous data modalities (Table 4).

Table 4 | AUC results of different models for each label in multi-label tasks

Disease	EAT[16]	BTS[17]	FrameATST[18] + Roberta[19][1]	BEATs[20] + Roberta[19][1]	Ours
URTI	0.7222	0.7507	0.8047	0.8392	0.8432
LRTI	0.7877	0.7903	0.8532	0.8140	0.8316
ILD	0.7485	0.7054	0.6916	0.8011	0.7732
COPD	0.9195	0.9288	0.9382	0.9299	0.9430
CB	0.7264	0.7633	0.8123	0.7878	0.8641
BRONCH[2]	0.7778	0.8090	0.7043	0.7757	0.8005
ASTHMA	0.7570	0.7687	0.7824	0.7726	0.7993
ALL[3]	0.7770	0.7880	0.7981	0.8172	0.8364
ALL[4]	0.8346	0.8463	0.8667	0.8712	0.8907

[1] The same architecture as Figure 1 a (without adding the device adversarial module) is used.

[2] BRONCH is an abbreviation for BRONCHIECTASIS.

[3] ROC AUC Macro of all labels.

[4] ROC AUC Micro of all labels.

[5] An additional stratified multi-label evaluation by introducing background subgroups can be seen in Supplementary Table 7.

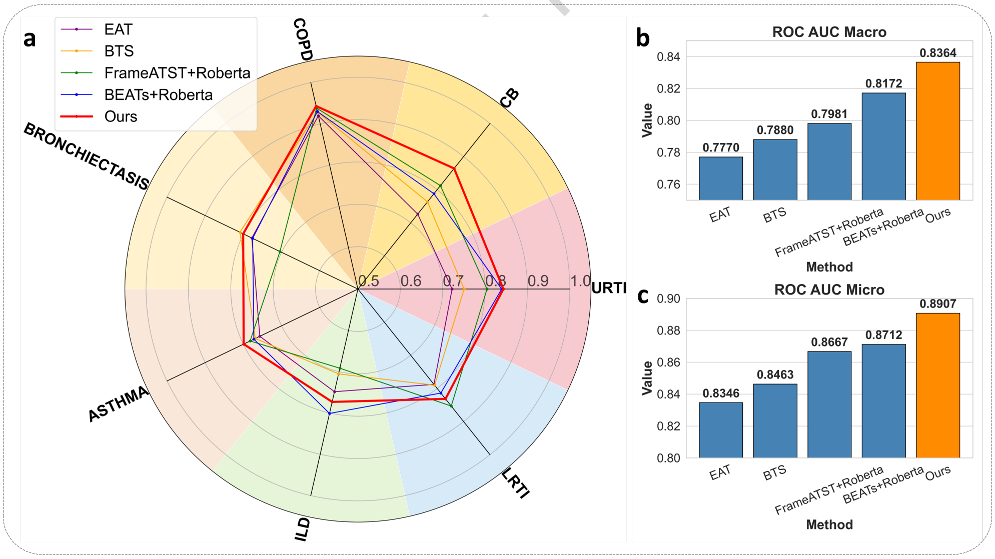

Fig. 3 | Comparison of different models in multi-label tasks.

Ablation Studies and the Impact of Adversarial Training

We conducted an ablation study on the LRTI binary classification task to systematically evaluate the contribution of different modalities and their combinations to disease prediction (Figure 4). We assessed the independent and combined predictive utility of audio recordings, demographic information, and symptom descriptions. The results revealed substantial performance discrepancies across single modalities: models relying solely on demographic and symptom information (text modality) achieved the lowest AUROC (0.6392), whereas the use of audio alone markedly improved performance (AUROC=0.7914). Dual-modality integration further enhanced classification accuracy, with audio combined with demographics (AUROC=0.7966) or with symptoms (AUROC=0.8061) both outperforming single-modality models, highlighting the complementary value of cross-modal information. Incorporating all three modalities—audio, demographics, and symptoms—yielded an additional performance gain (AUROC=0.8393). This shows that audio-only models demonstrated strong baseline performance, while demographic and symptom features provided complementary benefits, particularly in multi-label recognition tasks. Notably, the inclusion of adversarial training in this multi-modal setting achieved the best performance (AUROC=0.8571) and exhibiting superior generalizability and clinical utility, underscoring its effectiveness in mitigating device-related bias and improving model generalizability. Collectively, these findings emphasize the critical role of multimodal integration and adversarial optimization in advancing disease classification, and the synergistic value of multimodal medical data in respiratory disease diagnosis.

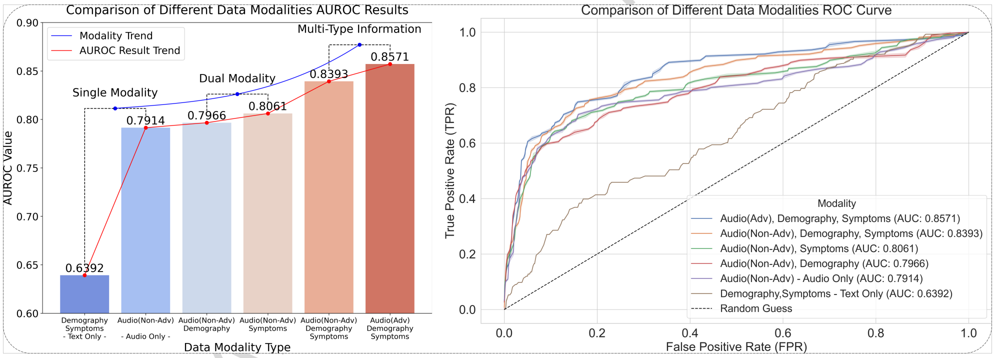

Fig. 4 | Modality ablation study.

Additionally, to assess the effectiveness of adversarial training in mitigating device-related biases, we systematically compared model variants trained with and without adversarial mechanisms across all classification tasks (Figure 5). Consistent improvements in AUC were observed with adversarial training, indicating its efficacy in suppressing device-induced variability and enhancing model robustness in heterogeneous deployment settings.

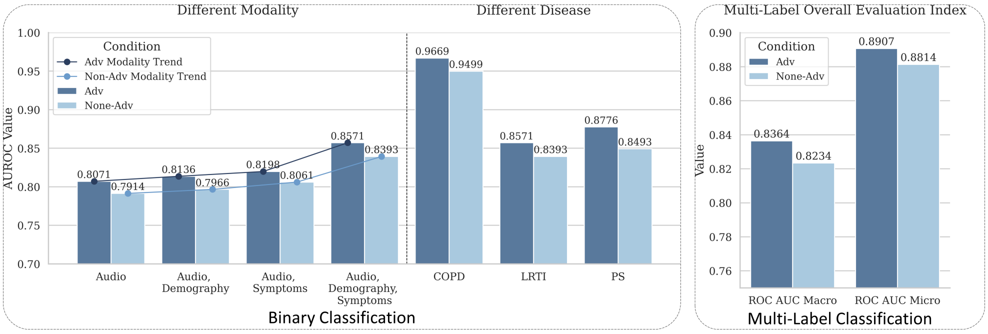

Fig. 5 | Comparison of adversarial and non-adversarial modules in different tasks.

Adversarial Training Enables Device-Invariant Feature Learning

To visualize the impact of adversarial training on learned audio representations, we extracted CLS token features from the audio encoder[21] and projected them into a 2D space using t-SNE[22] (Figure 6). Without adversarial training, samples from different recording devices exhibited distinct clustering patterns, revealing clear device bias. In contrast, models trained with the adversarial mechanism produced tighter clusters by disease category and exhibited markedly reduced device-based separation. These findings suggest that our approach enables the learning of device-invariant, yet discriminative, audio embeddings.

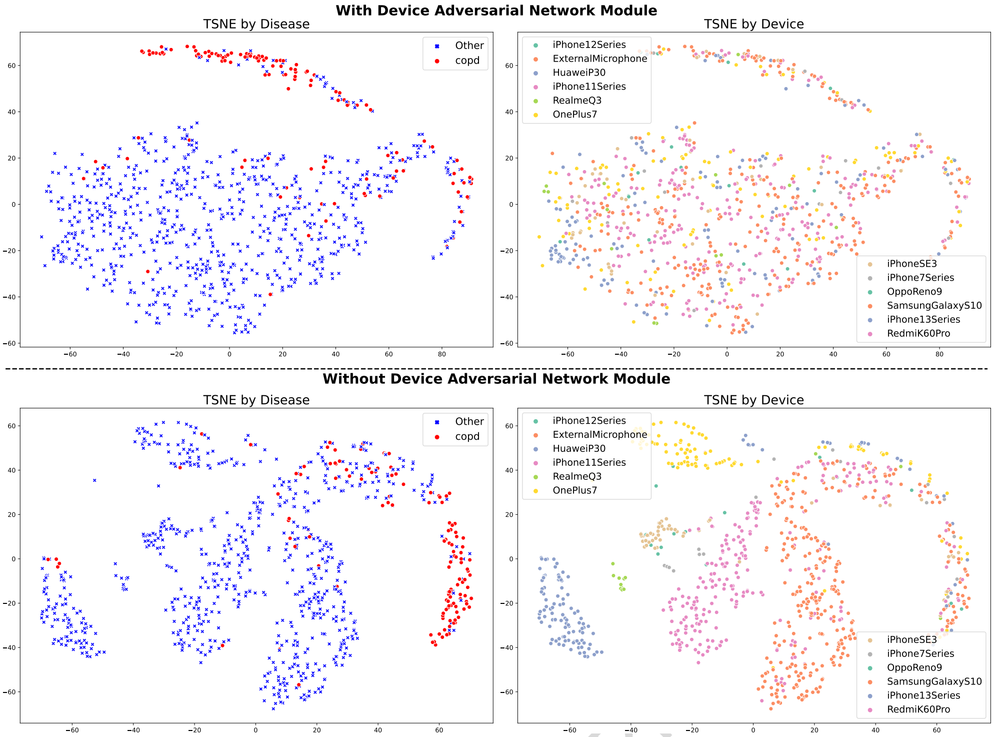

Fig. 6 | Feature space distribution of disease labels and device labels in adversarial/non-adversarial training tasks.

The Adversarial Model Demonstrates Superior Cross-Device Generalizability

To further assess the adaptability of our framework to deployment across heterogeneous devices in real-world scenarios, we retrained both the adversarial and non-adversarial models on the multi-label dataset after excluding all samples collected from “Unknown” devices. We then conducted evaluation on an independent external test cohort consisting of cough recordings acquired from multiple smartphone models that were never used during training or validation (Figure 7).

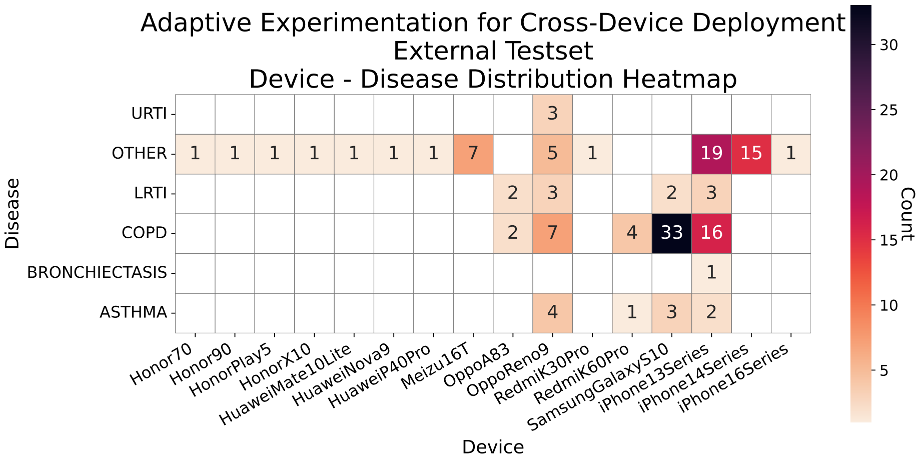

Fig. 7 | Disease and label statistics in adaptive experiments for cross-device deployment.

The results revealed that the non-adversarial model exhibited a marked decline in classification performance on unseen devices, with AUROC notably lower than that observed on the training devices, suggesting that its learned representations were device-dependent (Table 5). In contrast, the adversarially trained model maintained stable performance under the same testing conditions, with substantially attenuated AUROC degradation.

Table 5 | Comparison of results between adversarial and non-adversarial models on external testsets

	ROC AUC Micro			ROC AUC Macro		
	Internal Validation	External Test	∆	Internal Validation	External Test	∆
Adv	0.9189	0.9056	0.0133	0.8917	0.8903	0.0014
None-Adv	0.8862	0.8684	0.0178	0.8322	0.7649	0.0673

Note: The training and validation datasets used for model comparison were derived from the multi-label task dataset after excluding samples labeled as ”Unknown” devices. The external test set consisted of a fully independent collection of samples acquired from multiple smartphone models that were not present in either the training or validation sets.

These findings highlight that the device-adversarial module effectively mitigates distributional shifts induced by variations in audio acquisition hardware, thereby preserving discriminative power and robustness in cross-device deployment. This property is critical for ensuring reliable application of the proposed framework in multi-center, cross-regional, and resource-limited healthcare settings.

Invariant Risk Minimization Enhances Device Robustness

To further address distributional shifts introduced by diverse recording hardware, we incorporated both device-adversarial learning and IRM strategies into our framework. In cross-device transfer tests, models without adversarial components experienced substantial performance degradation. However, models employing the combined adversarial + IRM strategy demonstrated significantly improved classification accuracy on unseen target devices, evidencing enhanced invariance and stability across deployment conditions (Table 6). Although the absolute performance gain in terms of AUC under identical experimental settings is numerically modest, repeated experiments with different random seeds on a relatively small COPD dataset revealed a consistent improvement trend (8 out of 10 cases). Specifically, the adversarial model augmented with IRM outperformed its non-IRM counterpart in the majority of runs, while exhibiting reduced performance variance, indicating improved training stability rather than stochastic fluctuation. Due to the limited dataset size, we did not perform formal statistical significance testing; nevertheless, the observed consistency across repeated trials suggests that the performance gain is systematic rather than incidental.

From an optimization perspective, incorporating IRM into the device-adversarial framework increases the computational cost per iteration. However, this is offset by substantially faster convergence, as the adversarial + IRM model reaches its optimal performance with significantly fewer training epochs compared to the adversarial-only baseline. This improved convergence behavior further supports the role of IRM in stabilizing training dynamics under distributional shifts. Collectively, these findings indicate that the combined adversarial + IRM strategy offers a favorable trade-off between computational overhead and robustness, and holds strong potential for enabling reliable and scalable deployment of intelligent respiratory disease recognition models in multi-center clinical environments.

Table 6 | Comparisons of adversarial with IRM, adversarial without IRM and non-adversarial tasks

Task Name	Loss Function	Average Training Duration Per Epoch / H	Training Time to Achieve the Best AUC Result / H	AUC
Adv with IRM	Figure 10 a (1)	1.03	53.56 (52 epochs)	0.8571
Adv without IRM	Figure 10 a (2)	0.52	49.40 (95 epochs)	0.8529
Non-Adv	Figure 10 a (3)	0.48	44.64 (93 epochs)	0.8393

Note: Timeliness comparison based on the same batch size and computing resources.

Tradeoff Between Model Performance and Efficiency

To investigate the trade-off between model complexity and deployment efficiency, we compared two versions of our framework: a base-size model and a large-size model with approximately double the parameter count (Table 7). While the large-size variant achieved a higher mean AUROC of 0.8571 in the multilabel task—slightly outperforming the base-size model (AUROC=0.8405)—the lightweight base-size version maintained competitive accuracy while offering substantial advantages in inference latency and memory usage. These attributes make it particularly well-suited for deployment in resource-constrained environments such as mobile devices, edge hardware, or primary care settings.

Table 7 | Comparison of model parameters

Model Size Level	Parameters of Model Components				Total Parameters	AUC
	Encoder Module	Fusion + Disease Classifier Module	Device Adversarial Module	Uncertainty Weighting (Determined by the gradient)		
base	215M	10.6M	1.2M	3	226.2M	0.8405
large	440M	18.6M	1.2M	3	459.8M	0.8571

This comparison highlights the critical importance of model compression and efficient deployment strategies in real-world medical AI applications, offering valuable insights for the future design of edge-intelligent diagnostic systems.

Unless otherwise specified, all experiments were conducted using the large-sized model (460M parameters) in this study. The base variant (226M parameters) was evaluated only for ablation and scalability analyses.

Discussion

Clinical Significance and Deployment Feasibility

Cough is a hallmark symptom across a wide spectrum of respiratory diseases, and its acoustic characteristics carry rich pathological information. Compared with conventional diagnostic approaches relying on imaging, laboratory tests, or auscultation, cough-based acoustic analysis offers several compelling advantages: it is non-invasive, low-cost, easily scalable, and suitable for remote deployment. These properties make it particularly advantageous for primary care, telehealth platforms, and mobile health applications in resource-constrained settings[23, 24].

The proposed multimodal deep learning framework achieved high diagnostic accuracy across diverse respiratory conditions, even without relying on structured clinical inputs or specific device brands. The model demonstrated robust generalization across heterogeneous devices and multi-center data sources, confirming its potential for real-world clinical adaptation. These findings lay a strong foundation for its future deployment in community health centers, telemedicine platforms, and edge-intelligent medical devices.

Model Scalability and Cross-Scenario Adaptability

The architecture of the proposed framework is inherently modular and extensible, facilitating its adaptation to broader medical applications:

(1) Modal expansion: Beyond integrating audio, demographic features, and symptom questionnaire data, the model is readily extendable to incorporate additional clinical modalities such as chest X-rays, body temperature, heart rate, and peripheral oxygen saturation (SpO_2), thereby enhancing both disease detection sensitivity and clinical interpretability;

(2) Task transferability: The audio encoder and device-adversarial modules can potentially be transferred to other healthcare-related acoustic domains, including cardiac auscultation analysis, snore detection, asthma exacerbation prediction, or screening for sleep-related breathing disorders;

(3) Acquisition flexibility: The model's ability to handle device variability allows it to operate effectively across multiple audio input sources, including smartphones, remote microphones, and wearable sensors, thereby supporting deployment across a wide range of technical infrastructures and environmental conditions.

Looking ahead, we plan to leverage our ongoing longitudinal data collection efforts to integrate medical imaging and other data streams, further advancing the model's diagnostic accuracy, cross-domain generalizability, and clinical transparency.

Limitations and Future Directions

Despite the promising results achieved across multiple diagnostic tasks, several limitations remain and warrant further investigation:

(1) Objectivity and consistency of diagnostic labeling: The current ground-truth labels are primarily derived from initial clinical diagnoses, which may be affected by physician experience, documentation quality, and institutional practices[4]. Such subjectivity may introduce potential bias and variability in label assignment. To mitigate this, future work will incorporate multi-round diagnostic consensus, cross-validation with imaging and laboratory findings, and long-term follow-up data to establish a more objective and standardized labeling framework;

(2) Class imbalance and rare disease detection: The dataset exhibits considerable label imbalance, with certain respiratory diseases—especially rare conditions or early-stage pulmonary lesions—being severely underrepresented. This limited data availability compromises model performance on these classes, as reflected in lower AUC and recall scores compared to more prevalent conditions. This issue is particularly pronounced in the multilabel setting. To address this, future work will incorporate strategies such as class reweighting, oversampling, synthetic data augmentation, and transfer learning to mitigate imbalance-related performance bottlenecks and improve small-sample generalization;

(3) Limited model interpretability: Although attention mechanisms and adversarial components enhance model transparency to some extent, the decision-making process remains largely opaque. Incorporating explainable AI techniques[25] in future iterations will be critical for elucidating the model's rationale in a clinically meaningful way;

(4) Lack of temporal disease modeling: The current framework is designed for static snapshot-based prediction and does not yet incorporate temporal patterns in disease progression. Future work will focus on longitudinal modeling and patient trajectory analysis to support disease monitoring and individualized risk prediction;

(5) Environmental constraints and noise robustness: All audio recordings were collected in relatively quiet clinical environments to ensure data consistency. However, this controlled setting may not fully reflect the acoustic variability encountered in real-world scenarios such as home or community settings. Moreover, each audio event was pre-segmented into 3-second clips to facilitate model input standardization, which assumes clean segmentation—a condition that may not always hold in uncontrolled environments. Future studies will expand data collection to diverse acoustic contexts and develop segmentation-free or noise-robust learning paradigms to improve model generalization under realistic conditions.

(6) Handling weakly labeled and noisy data: Incomplete or noisy annotations are inherent in large-scale clinical datasets. To address this, future work will explore robust and data-efficient learning strategies such as causal inference[26], active learning[27], and semi-supervised learning[28], thereby improving adaptability under weak supervision and enhancing model reliability in real-world applications.

(7) Population homogeneity and ethnic generalizability: Although data were collected from multiple clinical sites, the enrolled population is relatively homogeneous in ethnicity and geographic distribution. Such population uniformity may limit the model's generalizability to other ethnic groups, healthcare systems, and socio-environmental contexts. Future work will emphasize dataset diversification through international collaborations and multi-regional data acquisition, as well as the integration of domain generalization and fairness-aware learning strategies to enhance model robustness, equity, and clinical transferability across diverse populations.

(8) Clinical applicability in complex comorbid conditions: While the present work demonstrates promising generalization across devices and recording settings, its performance in patients with complex comorbidities remains to be further validated. Conditions such as heart failure, endocrine disorders, and gastroesophageal reflux disease (GERD) may produce cough sounds or respiratory manifestations that overlap with pulmonary pathologies, thereby challenging model specificity in differential diagnosis. Future studies will focus on stratified validation across such comorbid subgroups and integrate multimodal clinical information—including cardiovascular and metabolic parameters—to improve diagnostic discrimination and real-world clinical applicability, particularly in primary care and community health scenarios. Although the current study demonstrates promising cross-device generalization in clinical datasets, prospective evaluations in real-world settings—including home monitoring and primary care environments—are currently ongoing. Results from these

studies will be reported in future work to further validate the model's clinical utility, generalizability, and practical deployment.

Methods

Pipeline Overview

The proposed framework comprises a complete end-to-end pipeline that integrates audio signal processing, text encoding, and multimodal learning to identify respiratory diseases from cough sounds and associated demographic information.

At the foundation of the pipeline lies a rigorous cough audio QC process that ensures the reliability of the raw recordings. Each audio sample undergoes segmentation, cough event detection, and cough validity verification to discriminate valid cough segments from background noise, whisper, throat clearance, and speech. The resulting high-quality cough clips form the basis for subsequent acoustic analysis.

Each validated cough segment is transformed into a two-dimensional mel-spectrogram representation, capturing both spectral and temporal variations relevant to disease-specific acoustic signatures. These representations are then processed by an audio encoder, based on the Efficient Audio Transformer (EAT) architecture, which generates compact and discriminative latent embeddings. In parallel, patient demographic data and textual attributes are encoded through a text encoder, built upon a pretrained Robustly Optimized BERT Pretraining Approach (RoBERTa) model fine-tuned for structured medical text.

The encoded audio and text embeddings are aligned and fused within the multimodal fusion classifier, which learns a shared latent space for cross-modal reasoning. This component integrates heterogeneous information to capture complex audio–text correlations that underpin clinical disease patterns. The final classifier makes a robust multimodal diagnostic decision.

To further enhance the model's robustness by reducing the impact of device variance, the training process incorporates adversarial and IRM strategies, as illustrated in Figures 9 and 10. The loss design enables the model to disentangle disease-related cues from device or environment-specific variations, thereby improving generalization across recording conditions and data sources.

Despite the relatively large total parameter count of the proposed multi-modal model, several deliberate design choices were adopted to ensure effective learning and to mitigate the risk of overfitting under the limited data regime. Extensive regularization strategies were employed throughout training. Specifically, dropout was applied across multiple layers, mixup-based data augmentation was used to improve robustness, and early stopping based on validation performance was adopted to prevent overfitting. And more importantly, the model heavily relies on pretrained initialization rather than learning from scratch. Both the audio encoder and the text encoder were initialized using publicly released pretrained weights provided by their original authors, which were trained on large-scale uni-modal datasets. This pretrained initialization substantially reduces the effective degrees of freedom during optimization and enables stable convergence even with limited task-specific data.

Crucially, the entire pretrained model was not fully fine-tuned. Instead, a partial fine-tuning strategy was adopted to further control model capacity. In particular, the text encoder was entirely frozen during training, and no layers within the text encoder were updated. This design choice ensures that the rich semantic representations learned from large-scale corpora are preserved, while completely eliminating the risk of overfitting within the text branch. In contrast, the remaining components of the model, including the audio encoder and the downstream fusion and classification modules, were fine-tuned to adapt modality-specific and cross-modal representations to the classification tasks. By freezing the text encoder and restricting gradient updates to only a subset of the overall architecture, the number of trainable parameters was substantially reduced relative to the full model size.

Overall, this unified pipeline seamlessly bridges signal-level processing and high-level multimodal inference, ensuring both clinical scalability across diverse real-world settings.

Cough Sound Quality Control Process

The Figure 8 illustrates a multi-level QC workflow for audio data applied during both training and validation/testing stages.

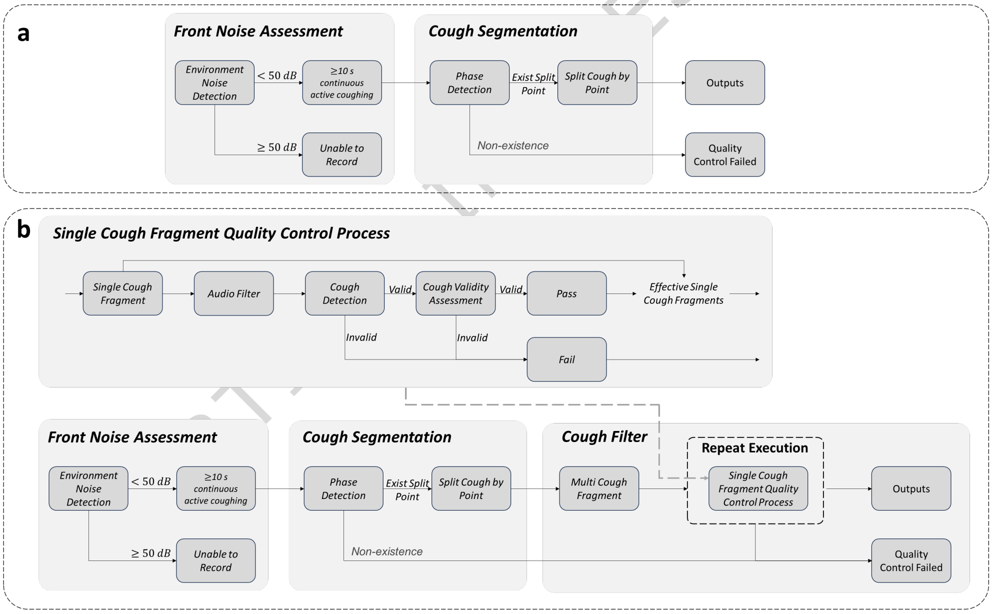

Fig. 8 | Audio data quality control workflow.

During training, audio data first undergoes an environmental noise assessment (Front Noise Assessment Module). Audio segments are considered usable if the background noise is below 50 dB and at least 10 seconds of continuous cough activity is detected. Segments passing the noise filter are then processed through cough segmentation (as detailed in the Cough Segmentation section), which uses phase detection to identify the onset, peak, and intermediate phases of each cough event and to determine the split points used to extract individual cough segments. If valid segments are obtained, they are retained for model training; otherwise, the data is marked as failed QC.

For validation, testing, and inference, a more refined single-cough segment QC process is applied. Each cough segment is subsequently evaluated for validity using both a cough detection model and a cough effectiveness assessment model (as detailed in the Cough Filter section). Valid segments pass QC and are included in the analysis, whereas invalid segments are discarded. For multi-segment cough recordings, candidate segments are initially selected using the same environmental noise assessment and cough segmentation procedures. Each candidate segment then undergoes the single-cough segment QC process, with only valid segments retained for inference or analysis. Segments deemed invalid at any stage are discarded.

Overall, this workflow ensures that cough audio used in both training and inference is rigorously filtered for environmental noise, accurately segmented by cough phase, and verified for validity. This process enhances model robustness and reliability while minimizing the impact of background noise or invalid cough data on model training and inference.

Cough Segmentation

We employed a multi-stage signal processing pipeline to segment cough phases from audio recordings, enabling the extraction of precise temporal markers for the onset, peak, and intermediate phases of each cough event (Algorithm 1).

First, the raw audio signal $x(t)$ was loaded and normalized by dividing the waveform by its maximum absolute amplitude:

$$\tilde{x}(t) = \frac{x(t)}{\max|x(t)|}, \tag{1}$$

which scales the signal to the range $[-1, 1]$. Candidate cough onset locations were detected using a frequency-domain onset detection procedure. The audio waveform was first transformed into the time–frequency domain via a short-time Fourier transform (STFT). For each frame, the spectral energy ratio between consecutive time steps was computed, and abrupt increases in high-frequency energy were identified. Frames exhibiting energy ratio changes exceeding an adaptive threshold were marked as potential onsets, and temporally adjacent detections were merged to produce the final set of candidate cough onset locations. For each candidate onset, a 0.7 s segment surrounding the onset (0.2 s before and 0.5 s after) was extracted. Within each segment, the precise *explosive phase* start was determined using root-mean-square (RMS) amplitude analysis and thresholding. The end of the explosive phase was detected via a spectral comparison method, where consecutive frames were analyzed for frequency-domain changes, and the first significant drop in the frequency energy ratio indicated the phase termination. Subsequently, the *intermediate phase* end was identified by detecting valleys in the low-pass filtered RMS signal.

All detected markers were corrected for segment offsets and filtered to remove closely spaced duplicates, ensuring that consecutive coughs were separated by at least 0.1 s. The resulting set of temporal markers for each cough event is represented as

$$\text{markers} = \{\text{explosive start}, \text{explosive end}, \text{intermediate end}\}.$$

Finally, after obtaining the temporal markers of each cough event, the corresponding audio signals were segmented for subsequent analysis. Specifically, each segment was centered on the detected explosive start marker, extending 1 s before and 2 s after it, thereby capturing the complete acoustic dynamics surrounding the cough phase transition. This segmentation ensured that both the onset and decay characteristics of the cough were preserved for downstream feature extraction and model training.

This segmentation process enables high-fidelity delineation of cough phases, which is critical for downstream feature extraction and classification.

Algorithm 1 Cough Phase Segmentation and Marker Detection

```
Require: Audio $x(t)$, sampling rate $F_s$, detection parameters $\theta$
Ensure: Cough phase markers markers = $[t_{\text{start}}, t_{\text{explosive\_end}}, t_{\text{intermediate\_end}}]$
Normalize audio: $x \leftarrow x / \max(|x|)$
Preliminary cough onsets: candidate_onsets $\leftarrow$ OnsetCough$(x, F_s, \theta)$
Initialize marker list: markers_list $\leftarrow \emptyset$
for each $t_{\text{cand}}$ in candidate_onsets do
    $segment \leftarrow x[t_{\text{cand}} - 0.2s : t_{\text{cand}} + 0.5s]$        ▷ Segment around candidate
    $t_{\text{start}} \leftarrow$ RMS_Threshold_Search$(segment)$
    $[ratio, t_{\text{explosive\_end}}] \leftarrow$ Phase1EndDetect$(segment, t_{\text{start}})$
    $t_{\text{intermediate\_end}} \leftarrow$ Phase2EndDetect$(segment, t_{\text{start}}, t_{\text{explosive\_end}})$
    Correct to original signal: $t_{\text{markers}} \leftarrow [t_{\text{start}}, t_{\text{explosive\_end}}, t_{\text{intermediate\_end}}] + t_{\text{cand}} - 0.2s$
    if $t_{\text{explosive\_end}} \neq$ None then
        markers_list $\leftarrow$ markers_list $\cup\ t_{\text{markers}}$
    end if
end for
Remove duplicates with minimum 0.1 s separation: markers_list $\leftarrow$ RemoveDuplicates(markers_list, $0.1s$)
return markers_list
```

Cough Filter

After segmentation, each cough-centered audio clip underwent a standardized preprocessing procedure to suppress device-related noise and harmonic interference. To mitigate baseline drift introduced by external microphones, a fourth-order Butterworth high-pass filter with a cutoff frequency of 50 Hz was applied to remove low-frequency noise components. Subsequently, harmonic-percussive source separation (HPSS) was performed to further attenuate stationary harmonic noise while preserving the transient cough components. The resulting signal was then used for downstream cough detection and validity assessment.

Cough detection was implemented to identify the presence of cough events within audio filter results. This stage was trained on a dataset of 3,000 manually annotated cough segments, capturing a wide variety of cough patterns and acoustic environments. A recurrent neural network[29] was employed to model the temporal dynamics of cough sounds, allowing robust detection across diverse recording conditions. The output of this module serves as the initial filter, selecting segments that likely contain cough events for further analysis.

Following detection, each candidate cough segment was evaluated for its validity and suitability for downstream analysis, such as disease classification. This step was trained on a subset of 1,000 carefully curated and high-quality annotated cough segments. A random forest[30] classifier was employed to assess the segment quality, incorporating features such as signal-to-noise ratio, temporal consistency, and spectral characteristics. Only segments passing this validity check were retained for subsequent model training and evaluation.

Data Preprocessing and Feature Extraction

To ensure high-quality input for downstream modeling, each recorded audio sample underwent a systematic preprocessing and feature extraction pipeline. This pipeline converts raw discrete-time waveforms into normalized log-mel filterbank representations, which capture both spectral and temporal characteristics of the signal while reducing sensitivity to amplitude variations and environmental noise. Specifically, the processing consists of waveform quantization, frame segmentation and windowing, short-time Fourier analysis, mel-scale filtering, logarithmic compression, and feature normalization. The following formalism describes each step in detail:

For each recorded audio sample, we denote the discrete-time waveform by

$$\mathbf{x} = \{x_n\}_{n=0}^{N-1}, \tag{2}$$

where x_n is the amplitude of the n-th sample, N is the total number of samples, and F_s is the sampling rate in Hz. The waveform is transformed into a normalized log-mel filterbank (FBank) representation via three steps: waveform quantization, short-time spectral analysis with mel-filtering, and feature normalization.

1. **Waveform quantization**
   The floating-point waveform is scaled to a 16-bit integer range:

$$x'_n = x_n \cdot 2^{15}, \tag{3}$$

   where x'_n is the quantized sample. For simplicity, we denote it as x_n in the subsequent steps.

2. **Frame segmentation and windowing**
   Frames of length $W_{\text{ms}} = 25$ ms with overlap $O_{\text{ms}} = 10$ ms are used. In samples:

$$W = \left\lfloor W_{\text{ms}} \cdot \frac{F_s}{1000} \right\rfloor, \quad O = \left\lfloor O_{\text{ms}} \cdot \frac{F_s}{1000} \right\rfloor, \quad H = W - O, \tag{4}$$

   where H is the hop size (frame shift).
   The k-th frame is extracted as

$$\mathbf{x}_f^{(k)} = \{x_{n+kH}\}_{n=0}^{W-1}, \quad k = 0, 1, \ldots, K-1, \tag{5}$$

   where $K = \left\lfloor \frac{N-W}{H} \right\rfloor + 1$ is the total number of frames.
   A Hamming window $\mathbf{w} = \{w_n\}_{n=0}^{W-1}$ is applied:

$$w_n = 0.54 - 0.46 \cos\left(\frac{2\pi n}{W-1}\right), \quad \mathbf{x}_w^{(k)} = \mathbf{x}_f^{(k)} \odot \mathbf{w}, \tag{6}$$

   where $\odot$ denotes element-wise multiplication.

3. **Short-time Fourier transform and power spectrum**
   For each windowed frame, the FFT of length N_{FFT} is

$$X_m^{(k)} = \sum_{n=0}^{N_{\text{FFT}}-1} x_{w,n}^{(k)}\, e^{-j2\pi mn/N_{\text{FFT}}}, \quad m = 0, 1, \ldots, N_{\text{FFT}} - 1, \tag{7}$$

   and the power spectrum is

$$P_m^{(k)} = \left| X_m^{(k)} \right|^2, \quad f_m = \frac{m}{N_{\text{FFT}}} F_s. \tag{8}$$

4. **Mel filter bank and energy computation**
   A triangular mel filter bank $\{H_p(f)\}_{p=1}^{M}$ with $M = 512$ filters is applied:

$$H_p(f) = \begin{cases} 0, & f < f_{p-1}, \\ \dfrac{f - f_{p-1}}{f_p - f_{p-1}}, & f_{p-1} \le f \le f_p, \\ \dfrac{f_{p+1} - f}{f_{p+1} - f_p}, & f_p \le f \le f_{p+1}, \\ 0, & f > f_{p+1}, \end{cases} \tag{9}$$

   where f_{p-1}, f_p, f_{p+1} are the corner frequencies of the p-th filter.
   The mel energy of frame k is

$$E_p^{(k)} = \sum_{m=0}^{\lfloor N_{\text{FFT}}/2 \rfloor} P_m^{(k)}\, H_p(f_m), \quad p = 1, 2, \ldots, M. \tag{10}$$

5. **Logarithmic compression**

$$\mathcal{F}_p^{(k)} = \log\left(E_p^{(k)} + \varepsilon\right), \quad p = 1, 2, \ldots, M, \tag{11}$$

where ε is a small constant (e.g., 10^{-8}) for numerical stability.
Stacking all frames yields the log-mel spectrogram

$$\mathcal{F} = \left[\mathcal{F}^{(0)}, \mathcal{F}^{(1)}, \ldots, \mathcal{F}^{(K-1)}\right] \in \mathbb{R}^{K \times M}. \tag{12}$$

6. **Feature normalization**

$$\mathcal{F}_{\text{norm}} = \frac{\mathcal{F} - \mu}{2\sigma}, \tag{13}$$

where μ and σ are the mean and standard deviation over the training set.
The factor of 2 in the denominator is empirically chosen for training stability.

7. **Summary**
The entire process from waveform to normalized log-mel features can be compactly written as

$$\mathcal{F}(\mathbf{x}) = \log\left(\sum_{m=0}^{\lfloor N_{\text{FFT}}/2 \rfloor} |X_m|^2 H_p(f_m)\right), \quad p = 1, \ldots, M. \tag{14}$$

Structured textual information—including patient demographics (e.g., age, gender, height, weight), smoking history, and symptom profiles—was first consolidated into standardized textual strings. For each patient, individual variables were concatenated with descriptive labels to form a coherent text representation. As an illustrative example, a subject's information might be formatted as:

"gender: male; age: 64 years; height: 161 cm; weight: 59 kg; smoking history: never."

In cases of missing values, the corresponding entry was consistently replaced with the placeholder *"unknown"*, ensuring that the resulting text strings remained complete and structurally uniform for all subjects. For instance, if the *weight* value were unavailable, the corresponding string would become:

"gender: male; age: 64 years; height: 161 cm; weight: unknown; smoking history: never."

These textual representations were subsequently tokenized using a pretrained tokenizer (RobertaTokenizerFast[19]), which converts each token into its corresponding vocabulary ID, and generates an attention mask identifying valid tokens. The resulting tensors—comprising both the input IDs and attention masks—served as the encoded input to the downstream model.

Model Architecture

In this study, we present a device-invariant multimodal deep learning framework designed to jointly model cough audio signals and structured textual information, enabling robust detection of respiratory diseases in heterogeneous clinical settings (see

Figure 9 a, while the hyperparameter and configuration are provided in Supplementary Tables 8-9). The proposed architecture comprises five principal modules: the input module, the encoder module, the multimodal fusion module, the disease classifier, and the adversarial learning branch.

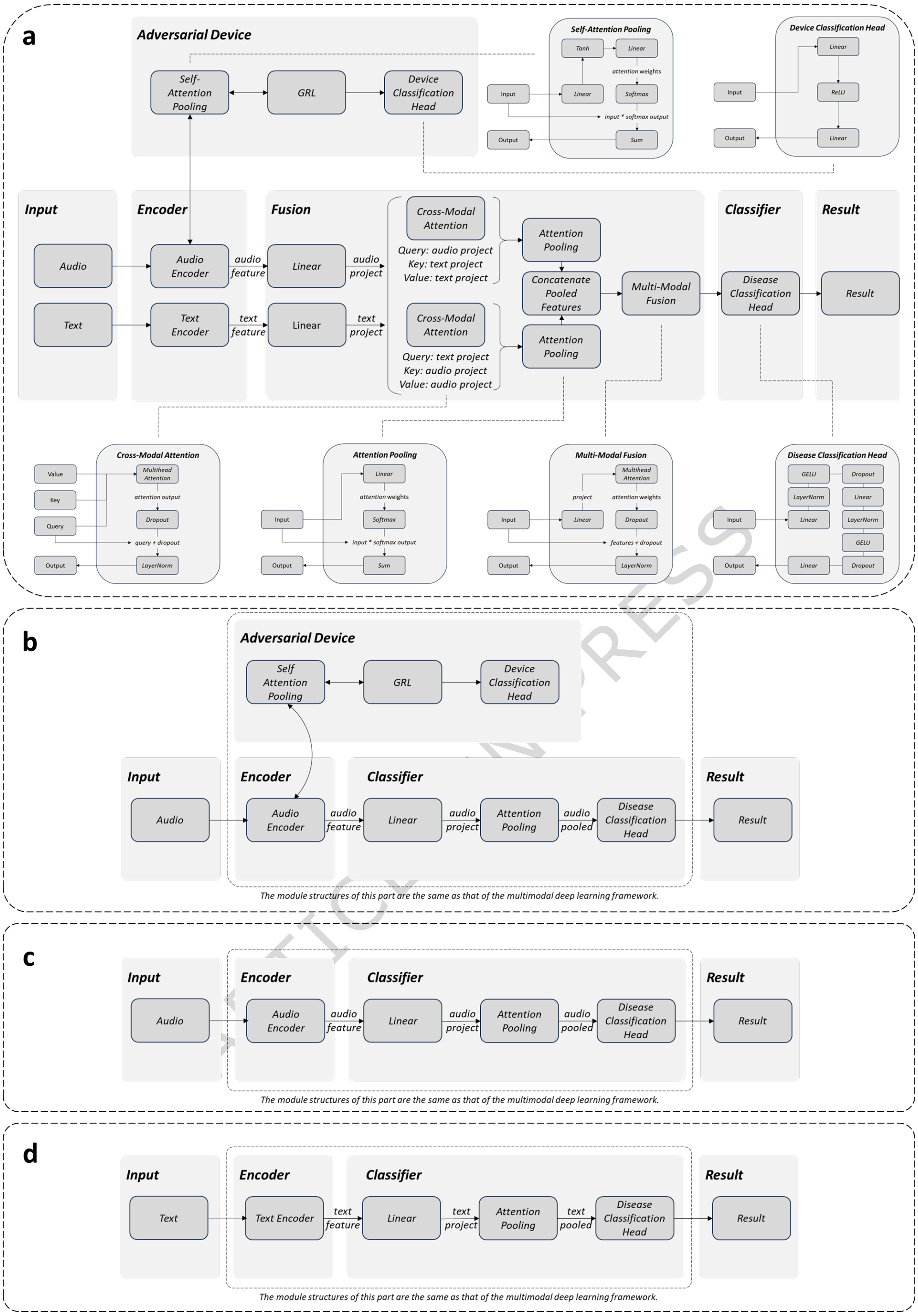

Fig. 9 | All model frameworks.

The input module processes two complementary data modalities: (i) preprocessed cough audio represented as mel-spectrogram tensors, and (ii) textual information including symptom descriptions and demographic attributes, each converted into unified vector representations.

The encoder module consists of both an audio encoder and a text encoder, enabling the model to capture complementary acoustic and semantic information.

The audio encoder is built upon the EAT, a transformer-based architecture specifically designed for compact and robust modeling of temporal–spectral patterns in acoustic signals. Following preprocessing, the cough spectrogram is segmented into non-overlapping 16 × 16 patches, which are linearly projected into embedding vectors and augmented with learnable positional encodings. These embeddings are then processed through a series of transformer encoder blocks, each comprising multi-head self-attention, feed-forward sublayers, residual connections, and layer normalization. For a batch of input spectrograms, the audio encoder produces frame-level acoustic features of shape $[\text{batch}, T, d]$, where T denotes the number of frames and d is the embedding dimension. In the base configuration, the audio encoder comprises 12 transformer layers with a hidden dimension of 768 and 8 attention heads, whereas the large configuration employs 24 layers with a hidden dimension of 1024 and 16 attention heads. To obtain a fixed-length representation, the frame-level features are temporally aggregated via mean pooling and subsequently projected into the shared multimodal fusion space for downstream integration with text embeddings.

The text encoder leverages a RoBERTa-based transformer to capture contextual linguistic and semantic information from symptom descriptions and clinical narratives. RoBERTa enhances language understanding by employing dynamic masking, larger training corpora, and longer sequences during pretraining. We adopt the RoBERTa model (12 layers, 12 attention heads, hidden dimension=768) for the configuration. Input sequences are tokenized with a maximum length of 256 tokens, and the [CLS] token from the final layer is used as the aggregated text embedding. The output textual features are projected into the shared latent space for multimodal integration.

To mitigate device-induced distribution shifts, an adversarial device branch is introduced during training. Audio features are first aggregated using self-attentive pooling[31], then passed through a gradient reversal layer (GRL)[32] before being fed into a device classifier. This adversarial optimization encourages the encoder to learn device-invariant representations. For sequence-level compression, attention pooling is applied: attention weights are obtained via linear projection, normalized through a softmax function, and used to compute a weighted sum of sequence elements, yielding fixed-length vectors for both intra- and inter-modal aggregation. The adversarial attention and hidden layers are aligned with the corresponding fusion dimensions in each configuration to ensure consistent representational capacity.

Multimodal fusion is implemented through a bidirectional cross-modal attention mechanism[33] that enables effective interaction between audio and text modalities. Initially, audio and text embeddings are linearly projected into a common latent space, forming modality-specific subspaces. Two symmetric cross-modal attention streams—audio-to-text and text-to-audio—allow each modality to selectively attend to salient cues from the other, with each stream composed of multi-head attention layers, followed by dropout and layer normalization. The resulting modality-specific outputs are then aggregated via attention pooling to derive compact embeddings. These embeddings are concatenated and further processed through a self-attentive multimodal fusion block, consisting of multi-head attention[34] layers with layer normalization, to enhance inter-modality dependency modeling.

Finally, the fused representation is fed into an enhanced classification head to generate the disease classification logits. The disease classifier is implemented as a multi-layer perceptron (MLP) integrating layer normalization[35], dropout[36], gaussian error linear unit (GELU) activations[37], and linear output layers. It produces a probability vector of shape $[\mathrm{batch}, \mathrm{num\_labels}]$, where num_labels corresponds to the number of diagnostic categories. In the base configuration, the classifier employs a 768-dimensional fusion space, 8 attention heads, and a dropout rate of 0.3. The large configuration expands these parameters to a 1024-dimensional fusion space with 16 attention heads and a reduced dropout rate of 0.2, thereby enhancing representational capacity.

Loss Function Design

To achieve both accurate recognition of respiratory diseases and robust modeling against device variability, we developed a dual-branch multi-objective loss framework that simultaneously optimizes two tasks: device-invariance modeling (device loss) and multi-label disease identification (disease joint loss). The overall architecture is illustrated in Figure 10 a (1).

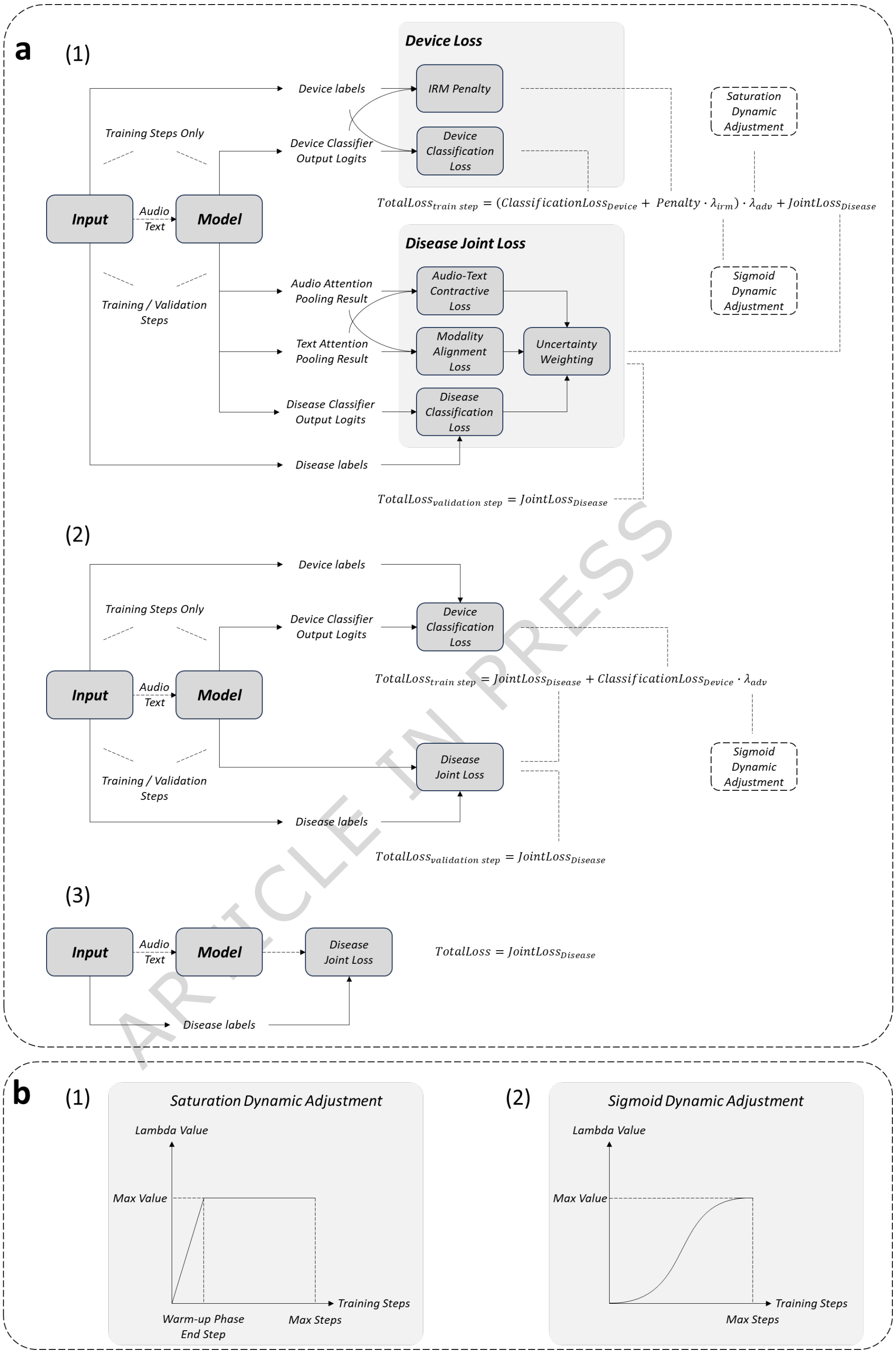
a
(1)
Device Loss
Device labels
IRM Penalty
Training Steps Only
Device Classifier Output Logits
Device Classification Loss
Saturation Dynamic Adjustment
Input
Audio
Text
Model
$TotalLoss_{train\ step} = (ClassificationLoss_{Device} + Penalty \cdot \lambda_{irm}) \cdot \lambda_{adv} + JointLoss_{Disease}$
Disease Joint Loss
Sigmoid Dynamic Adjustment
Audio Attention Pooling Result
Audio-Text Contractive Loss
Training / Validation Steps
Text Attention Pooling Result
Modality Alignment Loss
Uncertainty Weighting
Disease Classifier Output Logits
Disease Classification Loss
Disease labels
$TotalLoss_{validation\ step} = JointLoss_{Disease}$
(2)
Device labels
Training Steps Only
Device Classifier Output Logits
Device Classification Loss
Input
Audio
Text
Model
$TotalLoss_{train\ step} = JointLoss_{Disease} + ClassificationLoss_{Device} \cdot \lambda_{adv}$
Training / Validation Steps
Disease Joint Loss
Sigmoid Dynamic Adjustment
Disease labels
$TotalLoss_{validation\ step} = JointLoss_{Disease}$
(3)
Input
Audio
Text
Model
Disease Joint Loss
$TotalLoss = JointLoss_{Disease}$
Disease labels
b
(1)
Saturation Dynamic Adjustment
Lambda Value
Max Value
Training Steps
Warm-up Phase End Step
Max Steps
(2)
Sigmoid Dynamic Adjustment
Lambda Value
Max Value
Training Steps
Max Steps

Fig. 10 | Loss function calculation flow chart.

1. **Device Loss**
   The device branch is designed to enhance the model's generalization capability across heterogeneous acquisition devices and comprises the following two components:

   - IRM Penalty: Leveraging the IRM framework, a gradient-based penalty is introduced after partitioning data according to device environments, encouraging the model to learn a shared feature space across devices. Specifically, the primary task loss gradients are computed independently within each device-specific subenvironment, and their variability is minimized to enforce invariance. The process can be formulated as follows:

$$\mathcal{L}_{\text{IRM}}(w, \Phi) = \sum_{e \in \mathcal{E}_{\text{train}}} \left\| \nabla_w \, \mathcal{L}^e \left(w \cdot \Phi(X^e), \, Y^e \right) \right\|^2 , \tag{15}$$

     where $\mathcal{E}_{\text{train}}$ is the set of training environments. (X^e, Y^e) denotes the input and label pairs from environment e. Φ is the feature extractor shared across all environments. w is a linear classifier applied on top of Φ. $\mathcal{L}^e(\cdot)$ is the empirical risk (e.g., cross-entropy loss) for environment e. $\nabla_w \mathcal{L}^e$ is the gradient of the loss with respect to w. The squared norm $\| \cdot \|^2$ penalizes deviations from invariance across environments.
   - Device Classification Loss: A GRL is inserted between the audio feature extractor and the device classifier to adversarially train the audio encoder, thereby confounding device-specific discriminative cues. This objective is implemented using the cross-entropy (CE) loss[38]:

$$\mathcal{L}_{\text{DCL}} = \text{CE} \left(f_{\text{device}} \left(z_{\text{audio}} \right) , \; y_{\text{device}} \right) . \tag{16}$$

Since the IRM penalty is a gradient-based regularization term, it is not suitable to apply it from the early stages of training. To address this, we employ a sigmoid-based dynamic gradient scaling (Figure 10 b (1)) for the penalty, which gradually increases its influence, thereby avoiding the performance degradation that can occur if IRM is introduced prematurely. The sigmoid-based dynamic gradient scaling is computed as:

$$\lambda_{\text{IRM}}(t) = \lambda_{\max} \cdot \frac{1}{\exp\left[-k \left(\frac{t}{T} - 0.5 \right) \right]} , \tag{17}$$

where t is the current training step. T is the total number of steps, computed as:

$$T = \left\lfloor \frac{\mathit{num\_samples}}{\mathit{batch\_size} \times \mathit{accumulate\_grad\_batches}} \right\rfloor \times \mathit{max\_epochs},$$

$\lambda_{\max}$ is the maximum weight assigned to the IRM loss. k is the *steepness* parameter, controlling the growth rate of λ_{IRM}. $\frac{t}{T}$ represents the normalized training progress, ranging from 0 to 1.

λ_{IRM} acts as a stability enhancement mechanism to gradually adjust the adversarial loss coefficient, and forms the device loss term through a weighted combination with $\mathcal{L}_{\text{IRM}}$ and $\mathcal{L}_{\text{DCL}}$:

$$\mathcal{L}_{\text{device}} = \lambda_{\text{IRM}} \cdot \mathcal{L}_{\text{IRM}} + \mathcal{L}_{\text{DCL}}. \tag{18}$$

2. **Disease Joint Loss**

For the disease joint loss, considering the complementarity between modalities and the inherent data uncertainty, we designed a unified framework comprising three sub-losses:

- Disease Classification Loss: Serving as the primary task loss, binary cross-entropy is used to independently classify each disease.
- Audio-Text Contrastive Loss: To enhance the consistency of multimodal representations, an InfoNCE-style contrastive loss[39] is employed, pulling closer the audio and text embeddings from the same patient while pushing apart those from different patients.
- Modality Alignment Loss: The fused audio/text features are aligned in the shared feature space using KL divergence[40] as a regularization constraint.

Given the differing optimization difficulty and contribution of each sub-task, the overall joint loss is dynamically fused via an uncertainty weighting mechanism[41], where each sub-task is associated with a learnable log-variance parameter:

$$\mathcal{L}_{\text{disease}} = \sum_{i=1}^{M} \frac{1}{2\sigma_i^2} \mathcal{L}_i + \log \sigma_i, \tag{19}$$

where M is the total number of tasks or loss components. $\mathcal{L}_i$ is the loss associated with the i-th task. σ_i^2 denotes the task-dependent variance (uncertainty), which is a learnable parameter. $\frac{1}{2\sigma_i^2}$ adaptively scales the i-th loss according to its uncertainty. $\log \sigma_i$ acts as a regularization term to avoid trivial solutions.

To prevent the device-adversarial task from strongly interfering with disease classification in the early stages of training, we adopt a linear warmup strategy[34] (Figure 10 b (2)), allowing the audio encoder to prioritize learning disease-relevant features first. Once disease classification stabilizes, device-adversarial constraints are gradually introduced:

$$\lambda_{\text{adv}}(t) = \begin{cases} \lambda_{\max 2} \cdot \frac{t}{T}, & t \le t_{\text{warm up}}, \\ \lambda_{\max 2}, & t > t_{\text{warm up}}, \end{cases} \tag{20}$$

The final optimization objective integrates device robustness and disease recognition into a single weighted total loss:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{adv}} \cdot \mathcal{L}_{\text{device}} + \mathcal{L}_{\text{disease}}, \tag{21}$$

Experimental Setup and Evaluation Metrics

To evaluate model performance in real-world clinical scenarios, we designed a comprehensive experimental protocol:

1. **Modality Ablation**
   We systematically evaluate the contribution of different modalities through ablation studies (Table 8):

Table 8 | Comparison of model parameters

Modalities	Model Structure
Audio only	Figure 9 c
Audio + Demographics	Figure 9 a (w/o Adversarial Device Module)
Audio + Symptoms	Figure 9 a (w/o Adversarial Device Module)
Demographics+Symptoms (no audio)	Figure 9 d
Full multimodal[1]	Figure 9 a (w/o Adversarial Device Module)
Full multimodal[1] + Adversarial training	Figure 9 a

Note:
[1] Full multimodal (Audio+Demographics+Symptoms)

2. **Adversarial Strategy Validation**
   We examine two key adversarial strategies:

   - Domain Adversarial Training (DAT): Device discriminator + GRL for device-invariant audio encoding.
   - IRM-based Optimization: Enhances consistency across heterogeneous device distributions.

   These are evaluated across multi-label, binary, and ablation tasks. To further assess adversarial effectiveness, we visualize [CLS] token embeddings from the audio encoder using t-SNE, comparing cluster separation (device-wise vs. disease-wise) before and after adversarial training. We also compare training time and performance metrics across three regimes: Non-adversarial, Adversarial (w/o IRM), Adversarial (with IRM)
3. **Model Scaling and Deployment Readiness**
   We benchmark two model sizes:

   - base-size: Moderate parameter count, optimized for low-resource deployment.
   - large-size: Approximately twice the parameters of base, to explore performance ceilings.

We report AUC, inference latency, and resource utilization to inform future deployment on mobile or edge devices.

4. **Evaluation Metrics**
   We adopt clinically meaningful metrics, including:
   - AUROC: Discrimination across thresholds; robust to class imbalance.
   - AUPRC: Emphasizes precision in low-prevalence classes.
   - F1 Score: Balances precision and recall.
   - Accuracy, Precision, Sensitivity (Recall): Standard binary classification measures.
   - Youden Index: Determines optimal threshold by maximizing (Sensitivity + Specificity - 1), aligning with diagnostic utility.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to the inclusion of sensitive clinical information collected under institutional and regulatory data-use agreements, as well as proprietary components that cannot be openly released, but are available from the corresponding author upon reasonable request.

Code availability

The code developed in this study is proprietary and has substantial commercial potential; accordingly, it cannot be made publicly available. Due to intellectual property protections and ongoing commercialization activities, the source code cannot be shared at this time. All model development, training, and analysis were conducted using Python 3.10 with PyTorch 2.1.0 (which can be accessed at https://pytorch.org/get-started/previous-versions/). Specific training configurations and parameters used to generate and analyze the datasets are detailed in the Methods section.

References

[1] Wang, Z., Lin, J., Liang, L., Huang, F., Yao, X., Peng, K., Gao, Y., Zheng, J.: Global, regional, and national burden of chronic obstructive pulmonary disease and its attributable risk factors from 1990 to 2021: an analysis for the global burden of disease study 2021. Respiratory Research **26**(1), 2 (2025) https://doi.org/10.1186/s12931-024-03051-2

[2] Bhakta, N.R., McGowan, A., Ramsey, K.A., *et al.*: European respiratory society/american thoracic society technical statement: standardisation of the measurement of lung volumes, 2023 update. European Respiratory Journal **62**(4), 2201519 (2023) https://doi.org/10.1183/13993003.01519-2022

[3] Thawanaphong, S., Nair, P.: Contemporary concise review 2024: Chronic obstructive pulmonary disease. Respirology **30**(7), 574–586 (2025) https://doi.org/10.1111/resp.70062

[4] Agusti, A., Vogelmeier, C.F.: Gold 2024: a brief overview of key changes. Jornal Brasileiro de Pneumologia : Publicacao Oficial da Sociedade Brasileira de Pneumologia e Tisilogia **49**(6), 20230369 (2023) https://doi.org/10.36416/1806-3756/e20230369

[5] Kim, S.H., Han, M.K.: Challenges and the future of pulmonary function testing in chronic obstructive pulmonary disease (copd): Toward earlier diagnosis of copd. Tuberculosis and Respiratory Diseases **88**(3), 413–418 (2025) https://doi.org/10.4046/trd.2025.0009

[6] Chu, Y., Wang, Q., Zhou, E., Fu, L., Liu, Q., Zheng, G.: Cycleguardian: a framework for automatic respiratory sound classification based on improved deep clustering and contrastive learning. Complex & Intelligent Systems **11**(4), 200 (2025) https://doi.org/10.1007/s40747-025-01800-4

[7] Isangula, K.G., Haule, R.J.: Leveraging ai and machine learning to develop and evaluate a contextualized user-friendly cough audio classifier for detecting respiratory diseases: Protocol for a diagnostic study in rural tanzania. JMIR Research Protocols **13**, 54388 (2024) https://doi.org/10.2196/54388

[8] Sharan, R.V., Xiong, H.: Wet and dry cough classification using cough sound characteristics and machine learning: A systematic review. International Journal of Medical Informatics **199**, 105912 (2025) https://doi.org/10.1016/j.ijmedinf.2025.105912

[9] Huddart, S., Yadav, V., Sieberts, S.K., Omberg, L., Raberahona, M., Rakotoarivelo, R., Lyimo, I.N., Lweno, O., Christopher, D.J., Nhung, N.V., Theron, G., Worodria, W., Yu, C.Y., Bachman, C.M., Burkot, S., Dewan, P., Kulhare, S., Small, P.M., Cattamanchi, A., Jaganath, D., Grandjean Lapierre, S.: A dataset of solicited cough sound for tuberculosis triage testing. Scientific Data **11**(1), 1149 (2024) https://doi.org/10.1038/s41597-024-03972-z

[10] Morocutti, T., Schmid, F., Koutini, K., Widmer, G.: Device-Robust Acoustic Scene Classification via Impulse Response Augmentation (2023). https://arxiv.org/abs/2305.07499

[11] Mezza, A.I., Habets, E.A.P., Müller, M., Sarti, A.: Unsupervised Domain Adaptation for Acoustic Scene Classification Using Band-Wise Statistics Matching (2020). https://arxiv.org/abs/2005.00145

[12] Ma, C., Wang, H., Hoi, S.C.H.: Multi-label Thoracic Disease Image Classification with Cross-Attention Networks (2020). https://arxiv.org/abs/2007.10859

[13] Lei, T., Hu, Q., Hou, Z., Lu, J.: Enhancing real-world far-field speech with supervised adversarial training. Applied Acoustics **229**, 110407 (2025) https://doi.org/10.1016/j.apacoust.2024.110407

[14] Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant Risk Minimization (2020). https://arxiv.org/abs/1907.02893

[15] Wang, J., Liu, X., Zhou, X., Hu, G., Zhai, D., Jiang, J., Ji, X.: Joint Asymmetric Loss for Learning with Noisy Labels (2025). https://arxiv.org/abs/2507.17692

[16] Chen, W., Liang, Y., Ma, Z., Zheng, Z., Chen, X.: EAT: Self-Supervised Pre-Training with Efficient Audio Transformer (2024). https://arxiv.org/abs/2401.03497

[17] Kim, J.-W., Toikkanen, M., Choi, Y., Moon, S.-E., Jung, H.-Y.: Bts: Bridging text and sound modalities for metadata-aided respiratory sound classification. In: Interspeech 2024, pp. 1690–1694 (2024). https://doi.org/10.21437/Interspeech.2024-492

[18] Shao, N., Li, X., Li, X.: Fine-tune the pretrained atst model for sound event detection. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 911–915 (2024). https://doi.org/10.1109/ICASSP48485.2024.10446159

[19] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019). https://arxiv.org/abs/1907.11692

[20] Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Wei, F.: BEATs: Audio Pre-Training with Acoustic Tokenizers (2022). https://arxiv.org/abs/2212.09058

[21] Gong, Y., Chung, Y.-A., Glass, J.: AST: Audio Spectrogram Transformer (2021). https://arxiv.org/abs/2104.01778

[22] Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)

[23] Santosh, K.C., Rasmussen, N., Mamun, M., Aryal, S.: A systematic review on cough sound analysis for covid-19 diagnosis and screening: is my cough sound covid-19? PeerJ Computer Science **8**, 958 (2022) https://doi.org/10.7717/peerj-cs.958 . © 2022 Santosh et al. KC Santosh is an Academic Editor for PeerJ.

[24] Santamaria, M., Christakis, Y., Demanuele, C., Zhang, Y., Tuttle, P.G., Mamashli, F., Bai, J., Landman, R., Chappie, K., Kell, S., Samuelsson, J.G., Talbert, K., Seoane, L., Roberts, W.M., Kabagambe, E.K., Capelouto, J., Wacnik, P., Selig, J., Adamowicz, L., Khan, S., Mather, R.J.: Longitudinal voice monitoring

in a decentralized bring your own device trial for respiratory illness detection. npj Digital Medicine **8**(1), 202 (2025) https://doi.org/10.1038/s41746-025-01584-4

[25] Mersha, M., Lam, K., Wood, J., AlShami, A.K., Kalita, J.: Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. Neurocomputing **599**, 128111 (2024) https://doi.org/10.1016/j.neucom.2024.128111

[26] Baiardi, A., Naghi, A.A.: The value added of machine learning to causal inference: evidence from revisited studies. The Econometrics Journal **27**(2), 213–234 (2024) https://doi.org/10.1093/ectj/utae004 https://academic.oup.com/ectj/article-pdf/27/2/213/58305113/utae004.pdf

[27] Poinsot, A., Panayiotou, P., Leite, A., Chesneau, N., Şimşek, Schoenauer, M.: Position: Causal Machine Learning Requires Rigorous Synthetic Experiments for Broader Adoption (2025). https://arxiv.org/abs/2508.08883

[28] Guo, L.-Z., Jia, L.-H., Shao, J.-J., Li, Y.-F.: Robust semi-supervised learning in open environments. Frontiers of Computer Science **19**(8), 198345 (2025) https://doi.org/10.1007/s11704-024-40646-w

[29] Schmidt, R.M.: Recurrent Neural Networks (RNNs): A gentle Introduction and Overview (2019). https://arxiv.org/abs/1912.05911

[30] Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001) https://doi.org/10.1023/A:1010933404324

[31] Chen, F., Datta, G., Kundu, S., Beerel, P.: Self-Attentive Pooling for Efficient Deep Learning (2022). https://arxiv.org/abs/2209.07659

[32] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-Adversarial Training of Neural Networks (2016). https://arxiv.org/abs/1505.07818

[33] Luo, J., Phan, H., Wang, L., Reiss, J.D.: Bimodal Connection Attention Fusion for Speech Emotion Recognition (2025). https://arxiv.org/abs/2503.05858

[34] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (2023). https://arxiv.org/abs/1706.03762

[35] Aly, H., Al-Ali, A.K., Suganthan, P.N.: Boosted multilayer feedforward neural network with multiple output layers. Pattern Recognition **156**, 110740 (2024) https://doi.org/10.1016/j.patcog.2024.110740

[36] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of

Machine Learning Research **15**(56), 1929–1958 (2014)

[37] Hendrycks, D., Gimpel, K.: Gaussian Error Linear Units (GELUs) (2023). https://arxiv.org/abs/1606.08415

[38] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986) https://doi.org/10.1038/323533a0

[39] Wang, Z., Xu, B., Yuan, Y., Shen, H., Cheng, X.: Infonce is a free lunch for semantically guided graph contrastive learning. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 719–728. ACM, ??? (2025). https://doi.org/10.1145/3726302.3730007 . http://dx.doi.org/10.1145/3726302.3730007

[40] He, K., Ding, Y., Wang, H., Li, F., Teng, C., Ji, D.: DALR: Dual-level Alignment Learning for Multimodal Sentence Representation Learning (2025). https://arxiv.org/abs/2506.21096

[41] Kendall, A., Gal, Y., Cipolla, R.: Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics (2018). https://arxiv.org/abs/1705.07115

Figure Legends

Figure 1 Data distribution across three binary classification tasks.
a COPD vs. other. 1) Disease label distribution shows that COPD cases account for 14.1% (n=813) of the cohort; 2) Age distribution by label demonstrates that COPD patients are generally older than the non-COPD group; 3) Scatter plot of BMI versus age, colored by label, indicates overlapping but distinguishable patterns between COPD and other cases; 4) Device usage by label illustrates heterogeneous recording sources, with iPhone and external microphones as the most frequently used devices; 5) Smoking history, stacked with counts, reveals a higher proportion of smokers among COPD patients; 6) Missing data rate per feature highlights device type and smoking history as the most incomplete variables. **b** LRTI vs. other. 1) Disease label distribution indicates that LRTI cases account for 12.8% (n=720); 2) Age distribution by label suggests that LRTI cases are more common among younger participants compared to non-LRTI individuals; 3) Gender distribution shows a slightly higher prevalence of LRTI among males; 4) Device usage by label demonstrates similar heterogeneity as in COPD, with substantial representation from iPhone devices and external microphones; 5) Smoking history, stacked with counts, shows a higher proportion of smokers in the LRTI group than in the other group; 6) Missing data rate per feature again emphasizes device type and smoking history as primary sources of missingness. **c** PS vs. other. 1) Disease label distribution indicates that PS cases account for 6.3% (n=432); 2) Age distribution by label shows PS cases to be concentrated in middle-to-older age ranges compared to non-PS individuals; 3) Gender distribution highlights a slight male predominance among PS cases; 4) Device usage by label shows broad representation across devices, with iPhone and external microphones being the most common; 5) Smoking history, stacked with counts, reveals relatively balanced proportions of smokers and non-smokers in the PS group; 6) Missing data rate per feature again identifies device type and smoking history as the most incomplete variables.

Figure 2 Data distribution of the multi-label respiratory disease classification task.
1) Disease label distribution: Pie chart showing the proportion of different respiratory diseases, with the majority labeled as OTHER (63.21%), followed by ASTHMA (10.29%), COPD (9.36%), LRTI (7.76%), Multi-disease (4.11%), URTI (1.78%), CB (1.55%), BRONCHIECTASIS (0.95%), and ILD (0.95%); 2) Device–disease heatmap: Heatmap illustrating the cross-distribution of recording devices (e.g., iPhone models, Huawei, OnePlus, external microphone, etc.) and annotated disease labels. Notably, recordings from “Unknown” devices and iPhone 11 Series dominate across multiple disease categories; 3) Smoking history by gender: Stacked bar plot comparing smoking history distributions between Female (n=3892), Male (n=2908), and Unknown (n=15) groups. Categories include Smoking, Clean, Never, and Unknown, with females showing a higher proportion of Clean history and males higher in Smoking and Never; 4) Age distribution: Histogram showing the age spread of participants stratified by gender. Both males and females are broadly distributed across ages

18–90, with peaks around 40 and 65 years; 5) Height distribution: Histogram of participants' height (cm), stratified by gender. Females cluster around 155–165 cm, while males center around 165–175 cm; 6) Weight distribution: Histogram of participants' weight (kg), stratified by gender. Females are concentrated in the 45–70 kg range, while males center around 60–85 kg.

Figure 3 Comparison of different models in multi-label tasks.
a Radar chart of AUC results for 7 diseases using different models. **b** Bar chart of AUROC results for different models. **c** Bar chart of AUPRC results for different models.

Figure 4 Modality ablation study.
Performance comparison of the multi-modal model when selectively removing individual modalities. Each bar chart corresponds to one removal configuration and a corresponding ROC curve. The full-modal model achieved the highest diagnostic performance, indicating that the complementarity of cough sounds, demographic information, and symptom information collectively enhances the learning of disease-related features. Removing the audio modality resulted in the most significant performance degradation, highlighting its dominant role in capturing respiratory phenotypes, while demographic and symptom information provided additional, non-redundant contributions.

Figure 5 Comparison of adversarial and non-adversarial modules in different tasks.
Performance comparison across multiple downstream tasks when integrating a device-invariant adversarial module. The adversarial version consistently enhances model generalization and reduces device-induced bias.

Figure 6 Feature space distribution of disease labels and device labels in adversarial/non-adversarial training tasks.
t-SNE projections of latent embeddings trained with (top) and without (bottom) the device adversarial module. Left panels: embeddings colored by disease label (COPD vs. Other). Right panels: embeddings colored by recording device (e.g., iPhone series, Huawei P30, Samsung Galaxy S10). Without adversarial training, device-specific clusters dominate the feature space, indicating strong hardware-driven confounding. Incorporating adversarial learning effectively collapses device clusters while preserving disease-relevant structure, demonstrating successful removal of device biases during representation learning.

Figure 7 Disease and label statistics in adaptive experiments for cross-device deployment.

Distribution of diseases and corresponding device used in external testing experiments.

Figure 8 Audio data quality control workflow.
a Quality control process for training. **b** Quality control process for validation, testing/inference.

Figure 9 All model frameworks.
a Multimodal device adversarial model framework. **b** Single audio modality device adversarial model framework. **c** Single audio modality non-device adversarial model framework. **d** Single text modality model framework.

Figure 10 Loss function calculation flow chart.
a (1) IRM-based adversarial strategies. (2) Adversarial strategies without IRM. (3) Non-adversarial strategies. **b** (1) Saturation dynamic adjustment. (2) Sigmoid dynamic adjustment.

Acknowledgements

This work was supported by the National Key R&D Program of China (2022YFC2010005).

Author information

These authors contributed equally: Xuefei Liu, Wei Du.

Authors and Affiliations

Research&Development Department, Luca Healthcare, Shanghai, China
Mo Yang, Yang Liu, Wenyu Zhu, Zhaoyang Bu, Jiaxuan Mao, Qian Wang, Si Chen.
Department of Pulmonary and Critical Care Medicine, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
Xuefei Liu, Wei Du, Min Zhou, Jie-ming Qu
Institute of Respiratory Diseases, Shanghai Jiao Tong University School of Medicine, Shanghai, China
Xuefei Liu, Wei Du, Min Zhou, Jie-ming Qu

Contributions

M.Y. conceived and designed the study. X.L., W.D., Y.L., W.Z., Z.B., and J.M. collected the data. M.Y., W.Z. and Z.B. analyzed the data. M.Y., Q.W. and Y.L. drafted the manuscript. Q.W., S.C., M.Z. and J.Q. revised the draft.

Corresponding authors

Correspondence to Qian Wang, Si Chen, Min Zhou and Jie-ming Qu

Ethics declarations

Competing interests
All authors declare no financial or non-financial competing interests.