SputumLocator: Enhancing Airway Clearance with Auscultation-based Sputum Localization

YANBIN GONG, The Hong Kong University of Science and Technology, Hong Kong SAR
WENTAO XIE, The Hong Kong University of Science and Technology, Hong Kong SAR
CHI XU, The Hong Kong University of Science and Technology, Hong Kong SAR
QIAN ZHANG*, The Hong Kong University of Science and Technology, Hong Kong SAR
SHIFANG YANG*, Guangdong Provincial People's Hospital and Southern Medical University, China

Airway clearance is essential for managing Muco-Obstructive Lung Diseases (MOLDs). Percussion, a widely used airway clearance technique (ACT) in community and home care settings, is favored for its ease of implementation compared to other complex techniques. However, percussion is time-consuming and physically demanding for both caregivers and patients, as caregivers typically perform percussion on the entire back to avoid missing accumulated sputum when its exact location is unknown. Therefore, accurate sputum localization can significantly enhance the percussion experience. Current clinical methods for sputum localization typically rely on imaging techniques, which are costly, expose patients to radiation, and are usually performed only once during diagnosis, thereby limiting their application to inpatient settings. Alternatively, some medical professionals combine auscultation with other clinical assessments, but this approach requires substantial clinical experience and is impractical for community or home care settings where medical experts are unavailable. To address these limitations, we introduce SputumLocator, an innovative sputum localization system based on digital stethoscopes. SputumLocator leverages standard auscultation procedures to detect accumulated sputum in the four quadrants of the back, which is straightforward and highly practical. SputumLocator comprises two components: SputumEmbedder, which extracts key abnormal sounds and their spatial features using a Transformer-based feature extractor, and SputumClassifier, which maps these features to determine sputum presence in each region via a Convolutional Block Attention Module (CBAM). Given the limited availability of annotated sputum data, we developed a pretraining method based on Embedding on Masked Data (EOM) and enhanced model robustness through a Teacher-Student Architecture (TSA) that integrates noisy data. In collaboration with a medical institution, we evaluate SputumLocator on 43 patients with diverse physiological characteristics and under varying recording conditions. Experimental results demonstrate that SputumLocator achieves high accuracy with an overall sensitivity of 0.97, specificity of 0.82, and F1-Score of 0.83, maintaining robustness across different thoracic regions, genders, and disease types.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Applied computing → Health informatics**.

*Corresponding authors.

Authors' addresses: Yanbin Gong, ygongae@connect.ust.hk, CSE, The Hong Kong University of Science and Technology, Hong Kong SAR; Wentao Xie, wxieaj@cse.ust.hk, CSE, The Hong Kong University of Science and Technology, Hong Kong SAR; Chi Xu, cxubs@cse.ust.hk, CSE, The Hong Kong University of Science and Technology, Hong Kong SAR; Qian Zhang, qianzh@cse.ust.hk, CSE, The Hong Kong University of Science and Technology, Hong Kong SAR; Shifang Yang, yangshifang@gdph.org.cn, Department of Pulmonary and Critical Care Medicine, Guangdong Provincial People's Hospital and Southern Medical University, China.

ACM 2474-9567/2025/6-ART30
https://doi.org/10.1145/3729472

ACM Reference Format:
Yanbin Gong, Wentao Xie, Chi Xu, Qian Zhang, and Shifang Yang. 2025. SputumLocator: Enhancing Airway Clearance with Auscultation-based Sputum Localization. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 2, Article 30 (June 2025), 28 pages. https://doi.org/10.1145/3729472

1 INTRODUCTION

Muco-Obstructive Lung Diseases (MOLDs), including Chronic Obstructive Pulmonary Disease (COPD) affecting 480 million people [13] and bronchiectasis affecting 55 million people [83], pose significant health risks due to their characteristic symptom of mucus hypersecretion [15, 35, 74]. Excess sputum accumulation can trigger severe complications, including airway obstruction, impaired oxygen absorption, and increased vulnerability to infections such as pneumonia, thereby accelerating disease progression and impeding recovery [33, 45]. Therefore, effective sputum management is essential to maintain respiratory health and prevent complications [11]. Airway clearance techniques (ACTs) serve as a cornerstone in sputum management by facilitating mucus removal, alleviating airway obstruction, improving lung function recovery, and preventing secondary infections, ultimately improving physiological health and slowing disease progression [51, 52, 82]. Recent evidence supports the implementation of community-based ACTs as an effective approach to ongoing respiratory care [18].

While there exists a wide range of ACTs, each offering distinct advantages yet presenting significant limitations, percussion - the method of rhythmically tapping areas of sputum accumulation, remains the most widely adopted technique, particularly in developing regions with limited medical training and equipment resources [11]. This preference stems from percussion's simplicity, effectiveness, and its seamless integration with other ACTs, such as Active Cycle of Breathing Technique (ACBT) or postural drainage (PD), allowing for customized treatment approaches based on individual patient conditions [20, 78]. The technique, typically performed by community caregivers and family members (hereafter referred to as caregivers), involves rhythmically tapping specific lung areas with a cupped palm to generate deep-penetrating vibrations. These vibrations break up and mobilize thick sputum, reducing its viscosity and facilitating its movement from small airways to central airways for expulsion through coughing or suctioning.

Given that percussion remains indispensable, particularly in resource-limited settings, enhancing its efficiency could significantly benefit respiratory disease treatment. Specifically, optimizing percussion precision would minimize ineffective applications, thereby reducing patient discomfort and alleviating caregiver burden [58]. While percussion should ideally target specific regions of sputum accumulation, current sputum localization mainly relies on medical imaging (X-rays or CT scans) [77]. Although these imaging methods can guide percussion for hospitalized patients, they are impractical for broader implementation in community or home settings (hereafter referred to as community) due to radiation exposure, cost constraints, and inability to provide real-time feedback. Therefore, developing a method that is suitable for community level, low-cost, non-invasive and can guide percussion in real time has important clinical significance and social value.

Recently, medical professionals assess sputum distribution by combining auscultation (listening to internal sounds of the chest using a stethoscope) with health records, as sputum accumulation produces distinct respiratory sounds [14]. While this clinical practice demonstrates the potential of acoustic-based sputum localization, its reliance on extensive expertise and clinical information limits its application in community settings. To address this limitation, researchers have explored machine learning approaches for identifying abnormal respiratory sounds [25, 43]. However, these studies only analyze single-point recordings without spatial information. Although sound source localization techniques using multi-point collection could potentially solve this problem [34], they require complex acoustic modeling and controlled environments. Despite advances in digital stethoscopes offering improved sound quality [3, 5], developing a practical system for community-based sputum localization remains challenging.

In this research, our objective is to develop an auscultation-based sputum localization system that enhances percussion efficiency using only a commercially available digital stethoscope. This system aims to provide accurate, low-cost, non-invasive, and easy-to-follow guidance for individuals requiring percussion at the community level. To achieve this objective, the designed system should: (i) be both robust and user-friendly, (ii) detect the presence of sputum, and (iii) localize the areas of sputum accumulation. Following standard auscultation protocols, our system collects data from 12 specific points on each patient's back, ensuring systematic and reproducible data acquisition [14]. To facilitate practical treatment, we simplified the thorax mapping into four primary regions: upper left, upper right, lower left, and lower right. This simplified segmentation enables caregivers to perform percussion more easily by following the system's step-by-step guidance for identified sputum-accumulated regions.

However, developing such a system faces several key challenges. First, the significant individual differences in lung structure and various diseases create heterogeneous audio features [36], with diseases affecting sound characteristics beyond simple changes in crackles or wheezes [75]. Second, respiratory sounds collected by digital stethoscopes typically exhibit a low signal-to-noise ratio (SNR) [60], making them susceptible to environmental noise, skin friction, and heart sounds. Third, high-quality datasets for lung auscultation are extremely limited, particularly those with sputum location labels. Finally, using a single stethoscope for standard auscultation means collecting asynchronous data from different points, requiring the model to construct spatial representations from temporally separated signals.

To address these challenges, we present SputumLocator, as shown in Fig. 3, the first sputum localization system designed to guide percussion in community settings through standard auscultation with digital stethoscopes. SputumLocator comprises two primary components: SputumEmbedder and SputumClassifier. SputumEmbedder, a Transformer-based feature extractor, undergoes two-stage pre-training using both labeled and unlabeled datasets to handle heterogeneous, low-SNR signals. It employs Embedding on Masked Data (EOM) for feature extraction and a Teacher-Student Architecture (TSA)-based contrastive learning approach [73] for noise reduction. SputumClassifier integrates these embeddings using a lightweight Convolutional Block Attention Module (CBAM) [87] to determine sputum accumulation across different thoracic regions, treating the task as a multi-label classification problem to capture inter-region dependencies.

We evaluated the performance of SputumLocator at a large medical institution. We used CT imaging and bronchoscopy results as the gold standard. To minimize additional radiation exposure or trauma and ensure efficient use of medical resources, we integrated data collection into patients' routine diagnostic procedures. For patients undergoing CT scans, we performed auscultation immediately after scanning to confirm that sputum distribution matched the imaging findings. For those requiring bronchoscopy, we conducted auscultation before the procedure or sputum suction to maintain the temporal relevance of the data. The study recruited 43 patients with various respiratory diseases to participate in the clinical trial. To assess the system's performance, we employed a five-fold cross-validation method. Results demonstrated that SputumLocator can accurately detect the presence of sputum in each thorax region, achieving 0.97 sensitivity, 0.82 specificity, and 0.83 F1-Score.

To summarize, the contributions of this work are as follows:

- We introduce SputumLocator, the first digital stethoscope-based solution for localizing accumulated sputum using the standard auscultation process. Requiring no specialized training or additional procedures, it can seamlessly integrate with conventional lung auscultation workflows.
- We develop a set of techniques to overcome the limitations of lung auscultation, including a two-stage pre-training framework that learns robust embeddings from low SNR and small datasets, and a lightweight CBAM-based classifier that integrates spatial information from different auscultation points and chest regions.

- We implement our system on a commercially available digital stethoscope, ensuring widespread accessibility. Experiments conducted on 43 patients in a large medical institution demonstrate SputumLocator's excellent performance and robustness, which achieves 0.97 sensitivity, 0.82 specificity, and 0.83 F1-Score on 5-fold cross validation.

The remainder of this paper is organized as follows: Section 2 provides preliminary knowledge about ACTs and the role of percussion. It also addresses the challenges in using auscultation for sputum localization, as well as the rationale behind our model design. Section 3 elaborates our detailed system architecture, followed by comprehensive evaluation results in Section 4. In Section 5, we discuss clinical implications, practical integration considerations, and explore current limitations and future research directions. Section 6 reviews related work in respiratory sound analysis and pulmonary disease management. Finally, we conclude the research in Section 7.

2 PRELIMINARY AND RATIONALE

Before delving into system design, we first establish the foundations and motivations of this research. Despite the availability of various airway clearance techniques (ACTs), percussion remains widely adopted due to its effectiveness and accessibility, particularly in community settings. However, the efficiency of percussion highly depend on accurate sputum localization, which relies on either expensive medical imaging or experienced clinicians' expertise that are unavailable in community settings. This section begins by examining different ACTs and highlighting why percussion is the predominant choice in community healthcare. We then explore the potential of auscultation as a cost-effective approach for sputum localization to guide percussion therapy. Although auscultation is a routine diagnostic tool widely available at the community level, utilizing it for accurate sputum localization faces several technical challenges. We systematically analyze these challenges, including the heterogeneity of audio features, low signal-to-noise ratio (SNR), and limited dataset availability, while presenting our corresponding design rationales to address each challenge.

2.1 Airway Clearance Techniques

2.1.1 Overview of Different Approachs.

There are wide range of ACTs, each offering distinct therapeutic approaches [11]. Equipment-free methods, such as Active Cycle of Breathing Technique (ACBT) and Autogenic Drainage (AD), utilize controlled breathing patterns and body positioning to mobilize secretions. ACBT combines breathing control with thoracic expansion and forced expiration, while AD employs different lung volumes to optimize airflow. While these techniques promote patient autonomy, they often face adherence challenges due to the complex execution requirements and extensive training needed [59]. Device-based methods include high-frequency chest wall oscillation (HFCWO), which uses an inflatable vest to generate airway vibrations, and positive expiratory pressure (PEP) devices that create resistance during exhalation. Although these mechanical approaches demonstrate clinical effectiveness, they present significant barriers including high equipment costs, potential patient discomfort, and limited accessibility, particularly in resource-constrained settings [11].Postural drainage (PD), which relies on gravity-assisted mucus clearance through specific body positioning, offers simplicity and ease of implementation. However, its effectiveness is often limited compared to other ACTs [11]. Percussion, a manual technique, characterized by rhythmic tapping with cupped hands on specific chest areas, is straightforward to execute and requires minimal equipment, though it typically needs caregiver assistance and may cause physical fatigue during prolonged sessions [20].

2.1.2 Percussion: The Predominant Choice.

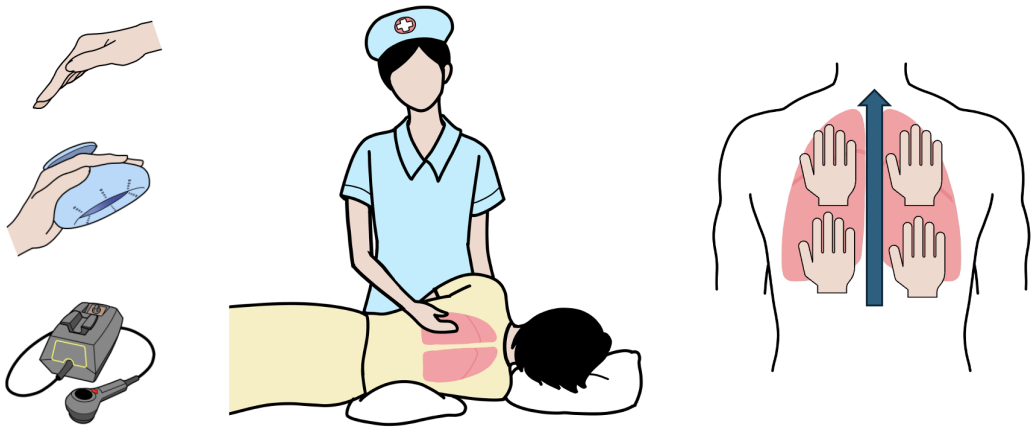

Fig. 1. Percussion Setup

Percussion involves rhythmically tapping specific thorax areas with a cupped palm, generating vibrations that penetrate deep into lung tissue. These controlled vibrations effectively break up viscous sputum, facilitate its detachment from airways, and enhance its fluidity, enabling easier expulsion through coughing or suctioning.

Despite the availability of various ACTs, percussion remains widely adopted, particularly in resource-limited settings. Its prominence stems from several key advantages: simple execution requiring minimal training, proven effectiveness in sputum clearance, seamless integration with other ACTs, i.e., ACBT or PD, and adaptability to various patient conditions [20, 78].

To assess real-world demand in our region, we conducted a preliminary survey involving 22 family members of patients with respiratory problems from community healthcare centers or district hospitals. The survey results revealed the following prescribed interventions:

- 15 were advised by physicians to employ percussion for expectoration
- 7 were instructed in alternative methods including ACBT, PEP or medication interventions only

The result demonstrates the popularity of percussion. Among the 15 individuals choosing percussion:

- 9 invested in simple or electric percussors
- 6 performed percussion manually
- 2 engaged home care services (personal caregivers) for percussion

The result shows the caregivers wish to provide high quality percussion with the help of simple devices. Notably, all patients who employed percussion reported its effectiveness in expelling sputum. Furthermore, there was agreement among these participants on initiatives aimed at improving the efficiency of percussion therapy.

These findings strongly suggest that percussion is not only widely utilized, but is also perceived as an effective method for airway clearance in our region. The high adoption rate and positive feedback underscore the potential value of developing more efficient percussion techniques, which could significantly impact patient care and caregiver support in respiratory disease management.

2.1.3 Potential Improvements.

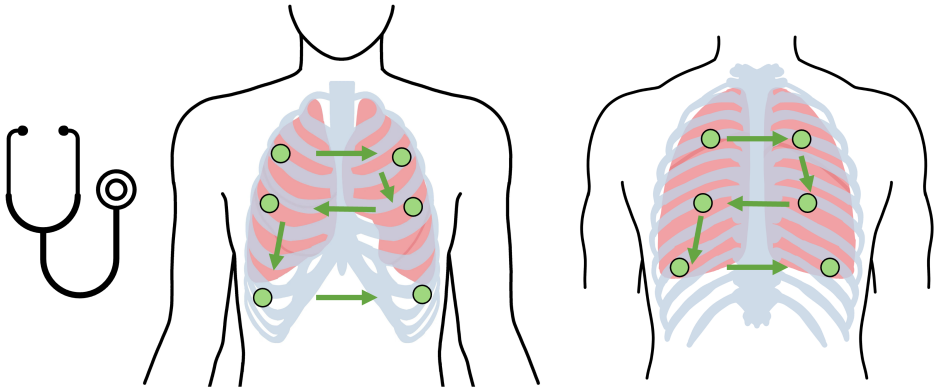

Fig. 2. Auscultation Procedures

As illustrated in Fig. 1, the percussion process is ideally targeted at specific thorax regions where sputum accumulates to improve efficiency and minimize patient discomfort [58].

In hospital settings, medical imaging (CT scans or X-rays) can precisely identify areas of sputum accumulation [77], thus guides the percussion. However, this approach is impractical for broader implementation due to cost, radiation exposure concerns, and limited accessibility. Alternatively, experienced clinicians can combine auscultation with comprehensive assessments to guide percussion. Clinical evidence suggests that skilled practitioners can interpret distinct respiratory sounds associated with sputum accumulation [14], but this approach relies heavily on individual expertise and detailed patient history.

However, in community healthcare settings, where advanced imaging or other clinical assessments are often unavailable, caregivers typically perform whole-chest percussion on both sides of the lungs. This approach is time-consuming and inefficient. For example, a comprehensive percussion session for an elderly patient can take 20 to 30 minutes, whereas accurately locating sputum may reduce this time to 5 to 10 minutes [46]. Given that patients often require three to four sputum extractions daily, the cumulative time significantly burdens both caregivers and patients. Additionally, prolonged tapping can cause discomfort and reduce treatment compliance.

These challenges underscore the urgent need for a low-cost, easy-to-operate method that enables accurate localization of the sputum in community settings. Such a solution could improve the efficiency of percussion techniques, reduce patient discomfort, lower treatment costs, and promote airway clearance more effectively. One promising approach is the use of the stethoscope, a routine diagnostic tool for respiratory diseases that is widely available at the community level. This strategy leverages existing, familiar tools in a novel way, potentially offering a practical solution to the challenges of percussion therapy in diverse healthcare environments.

2.2 Auscultation

2.2.1 Auscultation Procedure and Respiratory Sound Characteristics.

During auscultation, clinicians identify two main categories of lung sounds: normal sounds from healthy respiratory airflow, and abnormal sounds indicating potential lung diseases. Abnormal sounds may manifest as additional noises that overlay normal sounds, diminished or absent normal sounds, or left-right lung asymmetry. In clinical respiratory auscultation, the digital stethoscope should be placed at specific points on the thorax and maintained for at least one complete respiratory cycle to thoroughly assess lung conditions. As illustrated in Fig. 2, to evaluate the upper lobe, the stethoscope should be positioned at the second intercostal space on both the left

and right anterior chest sides, as well as the suprascapular area at the corresponding level. The fourth intercostal space and the interscapular area correspond to the left upper lobe (lingular segment) and the right middle lobe, respectively. When assessing the lower lobe, auscultation should be conducted at the eighth intercostal space on both sides and the subscapular area. This systematic approach ensures any trained caregiver can collect lung sound data comprehensively.

Table 1. Characteristics of Abnormal Lung Sounds

Sound Type	Acoustic Features	Timing	Clinical Significance
Fine Crackles [21]	• High-pitched, ~ 650 Hz • Short duration: ~ 5 ms • Discontinuous	More common during mid-to-late inspiration	Often indicate fluid or secretions in alveoli or small airways; commonly seen in **pulmonary fibrosis** and **pneumonia**
Coarse Crackles [21, 64]	• Lower-pitched, ~ 350 Hz • Short duration: ~ 15 ms • Discontinuous	More common during early inspiration	Usually caused by secretions in **larger airways**; commonly observed in **COPD**, **bronchiectasis**, **asthma**
Wheezes [14]	• High-pitched, 100 - 5000 Hz • Duration: > 80 ms • Continuous	More common during expiration	Associated with airway narrowing, often with sputum; commonly found in **asthma** and **COPD**
Rhonchi [61]	• Lower-pitched, ~ 150 Hz • Duration: > 80 ms • Continuous	More clear during expiration	Usually due to **thick or excessive bronchial secretions**

Different auscultation points yield distinct normal lung sounds, including tracheal, bronchial, vesicular, and bronchovesicular sounds [30]. These sounds originate from airflow through various anatomical structures, occur in different phases of the respiratory cycle, and possess unique acoustic characteristics. Abnormal lung sounds, as summarized in Tab. 1, can be categorized into discontinuous (crackles) and continuous (wheezes and rhonchi) types based on their acoustic features. These sounds provide crucial information about the presence, distribution, and characteristics of sputum in the respiratory tract [14]. The acoustic characteristics of lung sounds vary across different auscultation points due to individual variations in lung structure and disease progression. By analyzing and integrating information from multiple auscultation locations, the system can more accurately determine specific areas of sputum accumulation, thereby improving diagnostic accuracy and treatment targeting.

2.2.2 Challenges to Overcome.

Although auscultation remains a commonly employed routine clinical diagnostic method, it provides comparatively limited information compared to imaging techniques, which offer rich and intuitive diagnostic insights. The effectiveness of auscultation heavily relies on the doctor's experience and often needs to be supplemented with other evaluation methods to achieve a complete assessment. In view of this, we systematically identified the

main challenges of localizing sputum by auscultation and proposed corresponding design rationales for each challenge.

- *Heterogeneity of audio features.* Lung auscultation audio features exhibit high heterogeneity due to multiple factors. Different respiratory diseases uniquely affect lung sounds. For example, COPD can cause prolonged exhalation and wheezing, while pneumonia may produce localized moist rhonchi. Additionally, disease states significantly alter sound conduction characteristics, for instance, lung consolidation enhances sound transmission, whereas pneumothorax diminishes it [14]. Individual variations in lung structure and function, such as lung capacity and airway diameter, further contribute to differences. Factors such as age, gender, and body shape also impact sound conduction, i.e., thinner chest walls allow clearer breath sounds, whereas obesity can diminish sound transmission. Moreover, variations in breathing patterns, including respiratory rate and depth, affect auscultation results [71]. This complex heterogeneity makes it extremely challenging to extract universal features from auscultation audio. Even when sputum accumulates in the same area, the resulting sound characteristics may differ, causing models to perform poorly with new, unseen data. Consequently, developing robust models remains a significant challenge.
  Corresponding Design: *To address the challenge of audio feature heterogeneity, our design philosophy is to learn common patterns from diverse backgrounds. Specifically, we used self-supervised learning (SSL) technology for further pre-training to obtain a more robust feature extractor. We designed a model architecture that can capture different types of breathing-related features, and introduced diverse data augmentation techniques during training to simulate different changes in lung sounds. We used masked embedder to learn basic features from a large-scale general audio dataset, and then fine-tuned it for lung sound data to improve the model's adaptability to different features.*
- *Low SNR.* This issue arises from the interplay of multiple factors. First, breath sounds under normal conditions have low intensity and are easily masked by ambient noise. Second, background noises in medical settings, such as hospital conversations, medical equipment operations, and patient activities, further reduce the SNR by overlaying on the breath sounds. Additionally, internal noise sources like skin friction sounds and continuous heartbeats mix with breath signals, complicating signal separation. These overlapping noise factors make it extremely challenging to accurately extract meaningful breath sound information from the raw audio. If not effectively addressed, this can lead to errors in feature extraction, reduce classification accuracy, and ultimately undermine the reliability of clinical diagnoses.
  Corresponding Design: *To address the problem of low SNR, we first apply a series data pre-processing techniques based on the sputum-related audio features. We further introduce a contrastive learning method to enhance the denoising ability of the model by mixing noisy data. During the training process, we simulated different degrees of noise interference to improve the robustness of the model.Specifically, we adopted a teacher-student framework, using the teacher model to guide the student model to learn the denoised feature representation, further improving the performance of the model in low SNR environments.*
- *Limited Dataset Size.* High-quality lung sound datasets are extremely scarce, mainly because collecting and labeling professional lung sound data requires a lot of time and expertise. In addition, patient privacy and ethical issues limit the acquisition of large-scale data. Different disease states, severity, and individual characteristics lead to uneven data distribution, and there is a lack of unified data collection and labeling standards. In particular, datasets with ground truth annotations of sputum locations are more difficult, because the location of sputum can only be obtained through imaging methods, and the location of sputum may change due to behaviors such as coughing. Therefore, auscultation data must be collected immediately after the patient completes the imaging examination to ensure data consistency. The limited size of the dataset may cause the model to overfit, reduce its generalization ability, and thus affect the performance of the model in practical applications.

Corresponding Design: *We collaborate with large medical institutions to perform auscultation and collect relevant data when patients undergo normal diagnostic processes, especially after imaging examinations. It is fully integrated into the existing diagnosis and treatment process, does not bring additional physical burden or psychological pressure to patients, and ensures the ethical compliance of data collection. Considering the differences in the amount of data available for different tasks, we carefully designed a dual-component architecture: a powerful feature extractor (SputumEmbedder) and a lightweight classifier (SputumClassifier). This design allows us to train a general feature extractor on large-scale data while fine-tuning a lightweight classifier on small-scale data for specific tasks, thereby achieving a balance between model performance and computational efficiency. In addition, to further enhance the generalization and robustness of the model, we implemented a series of advanced data enhancement techniques.*

3 SYSTEM DESIGN

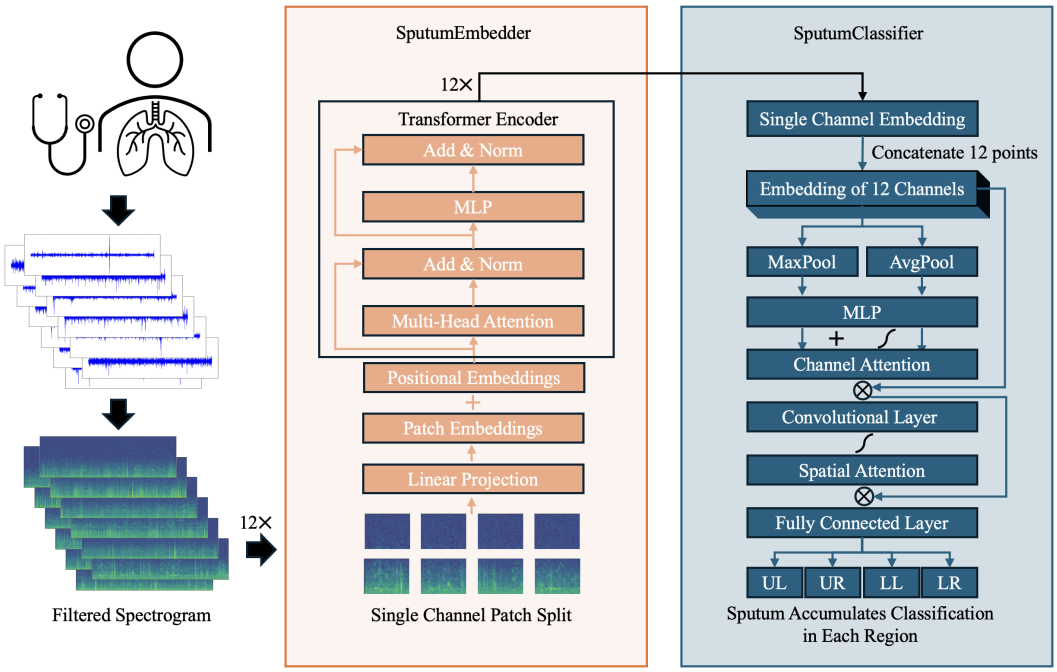

Fig. 3. SputumLocator overview

This section details the architectural design of the SputumLocator system. As illustrated in Fig. 3, the system comprises two primary modules: SputumEmbedder and SputumClassifier. Although these modules operate sequentially, their training processes are largely independent, ensuring that each can perform its specific function optimally. The workflow begins with pre-processing the input audio, which consists of standard 15-second recordings sampled at 4000 Hz. The raw audio signal is first passed through a 100-1900 Hz bandpass filter to target the frequency range of abnormal sounds based on clinical knowledge [14], effectively eliminating consistent noises like heart sounds. Subsequently, the audio is upsampled to 16 kHz to align with the sampling rates of common pretraining datasets and converted into a Mel-spectrogram with 80-dimensional log Mel filterbank features

($n_{mel} = 80$), computed with window size of 25 ms ($win_{length=400}$) and hop size of 10 ms ($hop_{length} = 160$), which is empirically used by related works [9, 27, 29]. Then, the processed spectrogram is fed to the SputumEmbedder, which extracts relevant features of breath sounds and sputum-induced sounds from each individual channel. This step condenses complex audio data into an implicit feature space, capturing the critical characteristics of sputum sounds. Subsequently, SputumClassifier integrates the embeddings from all 12 channels. By merging these multi-dimensional feature representations, SputumClassifier accurately identifies the presence of sputum accumulation in each chest region. This dual-module design allows independent optimization of feature extraction and classification tasks. In this way, SputumLocator can more effectively adapt to diverse auscultation scenarios and varying data characteristics, providing robust and accurate diagnostic support.

3.1 SputumEmbedder

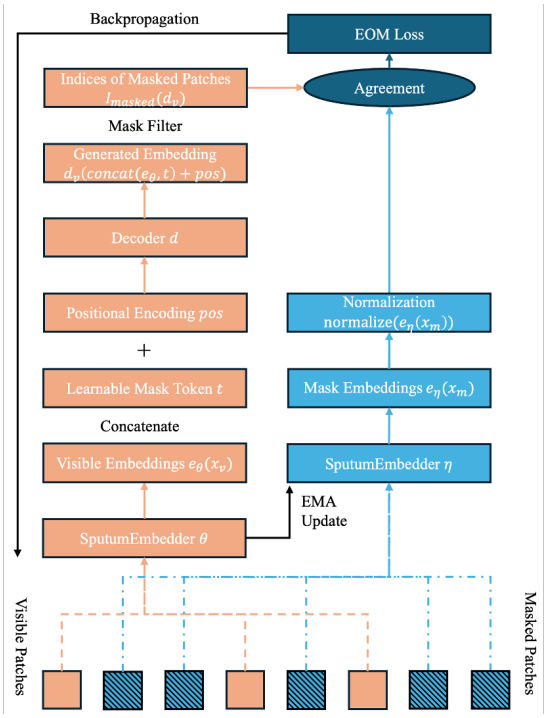

Fig. 4. Embedding on Masked Data

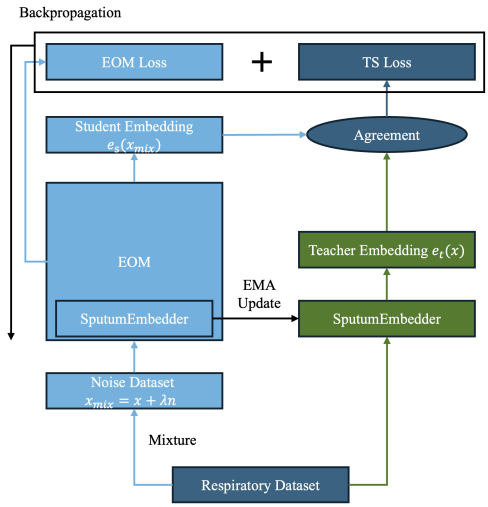

Fig. 5. Teacher Student Architecture

In the preliminary stages of our module design, we incorporate data augmentation using SpecAugment [70] in our preprocessing pipeline. This augmentation strategy has proven particularly effective for respiratory audio analysis tasks [40]. We normalize both our self-collected data and the ICBHI dataset [80] to zero mean and 0.5 standard deviation. This normalization ensures consistency across different recording conditions and equipment, facilitating more effective model training. Following normalization, we apply frequency masking and time masking techniques, which randomly obscure portions of the spectrogram in both frequency and time domains, respectively. These augmentations enhance the model's robustness to variations in acoustic conditions and partial occlusions of respiratory sounds, commonly encountered in real-world clinical settings. Notably, we

deliberately omit time warping from our augmentation pipeline, recognizing the critical importance of preserving short-term temporal correlations in respiratory acoustics, as emphasized by Bohadana et al. [14].

To effectively capture and leverage these temporal dynamics, we adopt a transformer-based solution. Bae et al. [10] demonstrate the superior performance of the Audio Spectrogram Transformer (AST) [28] on small-scale respiratory datasets when pretrained on large general datasets, outperforming networks trained from scratch. Inspired by this, and given our need for a highly robust model capable of excellent performance on a small, specialized dataset, we choose to leverage a pretrained transformer model. Our model employs the Transformer structure as its backbone, incorporating several key modifications to optimize it for respiratory sound analysis. The process begins by dividing the two-dimensional spectrogram into fixed-size patches of 16×16 pixels. These patches are then mapped to an embedding dimension of 768 using a `Conv2d` layer, which allows for efficient processing of the spatial information within each patch. To preserve crucial time-frequency relationships, we apply 2D positional encoding to each embedding. This step ensures that the model retains information about the relative positions of features within the spectrogram. We then prepend a special `CLS` token to the sequence, which serves as a global feature representation, aggregating information from all patches throughout the network. The resulting sequence of latent features is then processed through a stack of 12 Transformer blocks. Each block consists of multi-head self-attention mechanisms and position-wise feed-forward networks (FFNs). The multi-head attention allows the model to focus on different aspects of the input simultaneously, while the FFNs introduce non-linearity and increase the model's capacity to learn complex patterns, which greatly matches our need to learn sputum-related audio features while maintaining the inherent spatial audio information. The final step in *SputumEmbedder* involves concatenating the encoded features along the time axis to obtain the ultimate embedding that preserves the temporal structure of the respiratory sounds.

3.1.1 *General Respiratory Features Embedding.*

To address the challenge of limited clinical annotations, we adopt a self-supervised learning (SSL) approach for respiratory feature extraction. Inspired by Masked Modeling Duo (M2D) [65], we integrate the concepts of Masked Autoencoder (MAE) [38] and Bootstrap Your Own Latent (BYOL) [32] to develop a two-stream architecture that generates embedding on masked data (EOM). As illustrated in Fig. 4, we divide the input data x into masked x_m and visible x_v components with a masking ratio of 0.6. This ratio was empirically determined through ablation studies to balance information preservation and learning efficiency. These components are processed through two Transformer encoders with identical structures but separate parameters, utilizing the standard Vision Transformer (ViT) Base model considering our dataset size [19].

In the visible stream, the input x_v is processed through an encoder e_θ to generate an intermediate representation $e_\theta(x_v)$. This representation is augmented by concatenating learnable mask tokens and adding positional encodings to preserve spatial information. The augmented sequence is subsequently processed by the MAE decoder:

$$d\left(\text{concat}(e_\theta(x_v), \text{mask\_token}) + pos\right)$$

The final visible stream embedding is derived by selecting the masked positions from the decoder output:

$$\text{embedding}_v = I_{masked}(d_v)$$

The masked stream processes x_m through a separate encoder e_η. The masked stream embedding is obtained as:

$$\text{embedding}_m = \text{normalize}(e_\eta(x_m)) = \frac{e_\eta(x_m) - \text{mean}(e_\eta(x_m))}{\sqrt{\text{var}(e_\eta(x_m))}}$$

Unlike traditional MAE frameworks, our approach extends beyond input reconstruction by introducing a contrastive learning objective in the embedding space. The training objective is to minimize the L2 distance between the embeddings of masked regions obtained from both streams. For the visible stream, we first obtain

embeddings of the masked regions through the decoder's reconstruction, while the masked stream directly processes these regions:

$$\mathcal{L} = \|I_{masked}(d(e_\theta(x_v))) - \text{normalize}(e_\eta(x_m))\|_2^2$$

This approach encourages alignment between two different views of the same masked regions: one reconstructed from visible context through the decoder, and another directly encoded from the masked input. Such a design helps the model learn contextually-aware representations that capture essential acoustic features and their relationships. Unlike traditional MAE approaches that focus on pixel-space reconstruction, our embedding-space alignment better captures semantic features of respiratory sounds.

A key aspect of our methodology is the asymmetric parameter update strategy: the parameters θ of the visible encoder are updated through conventional backpropagation, allowing rapid adaptation to the current batch of data, while the parameters η of the masked encoder are updated more gradually via Exponential Moving Average (EMA) of θ:

$$\eta_t = \beta\eta_{t-1} + (1-\beta)\theta_t$$

where β is the EMA decay rate. This temporal smoothing serves multiple purposes: it provides a more stable learning target, creates a pseudo-ensemble effect potentially improving generalization, and mitigates the risk of representation collapse or rapid oscillations during training.

This dual-stream architecture, with its asymmetric parameter update mechanism, enables the model to learn robust features from partially obscured inputs, leveraging principles of masked modeling and contrastive learning. The contrast between embedding_v and embedding_m forms the basis of the self-supervised learning objective, encouraging the model to capture meaningful audio features even in the absence of explicit labels. Additionally, EMA-based parameter updates further stabilize the training process.

For downstream applications in *SputumLocator*, we utilize only the encoder-generated embeddings, discarding the decoder after pre-training. This design choice leverages the rich acoustic features learned during self-supervised training, which is more effective for sputum localization than reconstructed representations. The entire framework effectively transforms unlabeled respiratory audio data into meaningful feature representations, addressing the fundamental challenge of limited clinical annotations while maintaining robustness to variations in respiratory sound patterns.

3.1.2 Robust Embedding Enhancement.

As previously analyzed, respiratory auscultation presents a significant challenge due to extremely low SNR. Unlike the periodic and predictable nature of heart sounds, respiratory audio signals are frequently disrupted by randomly occurring, intensity-unpredictable environmental noises and artifacts such as skin or clothing friction sounds. These disturbances not only substantially increase the difficulty of signal processing but also can significantly degrade the model's performance and accuracy. To effectively address this challenge and enhance the model's robustness in low SNR environments, we adopt an innovative Teacher-Student architecture-based method [42]. Instead of directly training the ViT model with mixed noise, our approach performs signal enhancement at the EOM layer. This strategy allows us to address noise issues at a higher abstraction level, potentially capturing more meaningful representations of the underlying respiratory sounds.

As illustrated in Fig. 5, we initialize both Teacher and Student networks with the pre-trained EOM architecture. In the Teacher network, we freeze the model parameters and generate embeddings of the raw data through the encoder , expressed as: $\text{embedding}_t = e_t(x)$ This process serves to extract relatively pure respiratory sound feature representations. Concurrently, the Student network processes noisy inputs constructed by mixing clean signals with environmental sounds:

$$x_{mix} = x + \lambda n$$

where $n \sim \mathcal{N}(FSD50K)$ denotes noise samples drawn from FSD50K [22], a comprehensive dataset containing over 500 sound classes, and λ is a mixing coefficient that controls the signal-to-noise ratio. This mixing strategy simulates real-world acoustic interference patterns.

While the Student network shares the same EOM architecture and initialization as the Teacher, its training incorporates two complementary objective functions. The first maintains the original EOM objective for noisy inputs:

$$l_{EOM} = \|e_s(x_{mix,v}) - e_s(x_{mix,m})\|_2^2$$

where $x_{mix,v}$ and $x_{mix,m}$ represent the visible and masked components of the noisy input x_{mix}, respectively. This loss ensures the preservation of the masked modeling capability under noisy conditions. And the Teacher-Student Loss, defined as

$$l_{TS} = \|e_t(x) - e_s(x_{mix})\|_2^2$$

which quantifies the discrepancy between the embeddings generated by the Student network encoder and those produced by the Teacher network, aiming to guide the Student network to learn noise-robust feature representations. It is important to note that l_{EOM} operates on the split streams of x_{mix} following the original EOM framework, while l_{TS} utilizes the complete x_{mix} to generate student embeddings. The total loss is formulated as their weighted combination:

$$L_{total} = \alpha l_{EOM} + (1 - \alpha) l_{TS}$$

where α is a tunable hyperparameter that balances the contribution of EOM learning and noise robustness. Following the momentum-based update strategy established in EOM, the Student network parameters θ_s are updated through standard backpropagation, allowing rapid adaptation to the current batch of data, while the Teacher network parameters θ_t evolve through EMA of the Student network. Through preliminary research, we set $\alpha = 0.5$ to give equal weight to l_{EOM} and l_{TS}. This approach enhances the model's robustness to background noise while preserving sensitivity to respiratory sounds, captures more abstract features by addressing noise in the embedding space, and utilizes a Teacher-Student architecture for effective knowledge distillation, enabling the Student network to learn key features from the Teacher and perform better in noisy environments.

3.2 SputumClassifier

The SputumClassifier model proposed in this study is a specially designed multi-channel lightweight classifier for processing audio data from 12 different stethoscope positions. The audio from each position is first processed through our SputumEmbedder, generating highly abstract embeddings. Let $embedding_i \in \mathbb{R}^d$ be the embedding of the i-th stethoscope position, where d is the embedding dimension. The input to SputumClassifier can then be represented as:

$$X = [embedding_1, embedding_2, ..., embedding_{12}] \in \mathbb{R}^{12 \times d}$$

Given the challenge of extremely small data volume, we adopted a "heavy front, light back" strategy, concentrating complexity in the SputumEmbedder while choosing a relatively simple structure for the classification task to avoid overfitting.

As shown in 3, The core of SputumClassifier is an attention mechanism based on CBAM (Convolutional Block Attention Module), which includes channel attention M_c and spatial attention M_s. The channel attention is computed as follows:

$$M_c(X) = \sigma(MLP(AvgPool(X)) + MLP(MaxPool(X)))$$

where σ is the sigmoid function and MLP is a multi-layer perceptron. The spatial attention is calculated as:

$$M_s(X) = \sigma(Conv([AvgPool(X); MaxPool(X)]))$$

where we use a 1D convolution with kernel size 7. The output of CBAM, X', can be expressed as:

$$X' = M_s(M_c(X) \odot X) \odot (M_c(X) \odot X)$$

Next, we apply batch normalization (BN) and dropout, and the classification result is obtained through a fully connected layer:

$$y = softmax(W \cdot flatten(Dropout(BN(X'))) + b)$$

where $y \in \mathbb{R}^4$ is the probability distribution over four categories. We selected BCEWithLogitsLoss, which combines a sigmoid activation layer with binary cross-entropy loss, as the loss criterion.

This design fully utilizes the high-quality features extracted by SputumEmbedder from 12 stethoscope positions while maintaining good generalization ability under limited data conditions. The CBAM mechanism allows the model to adaptively allocate importance among the 12 stethoscope positions. Through this carefully balanced multi-channel architecture, SputumClassifier can effectively integrate sputum sound information from different parts of the body. The mathematical expressions of the model clearly demonstrate how it achieves comprehensive analysis of multi-position auscultation data while maintaining simplicity. This design not only adapts to the constraints of small datasets but also fully leverages the advantages of multi-position auscultation, promising to provide more comprehensive and reliable analysis results for precise localization and classification of sputum sounds.

4 EVALUATION

4.1 Evaluation Setup

4.1.1 Device and Platform Selection.

We utilized the 3M™ Littmann® CORE Digital Stethoscope for data collection [1]. This FDA-certified, mainstream digital stethoscope is widely employed in medical institutions. Data acquisition was performed using the Eko 5.6.0 application on iPhone 13 running iOS 17. Each audio recording segment lasted 15 seconds and was sampled at a rate of 4000 Hz. To maintain data integrity, we used raw sampling data that had not been processed by the Eko filtering algorithm. Model development was conducted on the Google Colab platform, leveraging Python version 3.10.12 and PyTorch version 2.4.1+cu121, and executed on an NVIDIA A100 GPU. This configuration ensured efficient and effective model training and inference processes.

4.1.2 Data Collection.

To ensure precise localization of sputum, we employed CT scans or bronchoscopy to determine the presence of sputum in various regions of the lungs. Acknowledging that invasive procedures may cause patient discomfort or even harm, we seamlessly integrated data collection into the patients' routine examination schedules. Due to environmental constraints, it was not feasible to simultaneously collect audio signals and establish ground truth. Therefore, data was collected immediately after CT scans or before bronchoscopy suctioning. We minimize the time interval between assessment and data acquisition, ensuring that sputum distribution remains consistent without significant deviation.In this research, we adopted a method for dividing lung regions that is more closely aligned with clinical practice. Instead of using the traditional anatomical lobe divisions, we segmented the patient's back into four primary areas, left upper, left lower, right upper, and right lower, which is based on the practical needs of sputum percussion procedures. This approach simplifies the sputum extraction process, enhancing its applicability and operability within community healthcare settings.

Data collection was conducted in real clinical settings across 11 hospital wards (W1-W8 and W10-W12) in Department of Pulmonary and Critical Care Medicine at a tertiary hospital, each presenting distinctive acoustic environments with varying levels of background noise. Ward 9 was not included in our study as no eligible patients with the required clinical labels were present in this ward during our data collection period. The wards differed significantly in physical configuration and occupancy, ranging from smaller 2-bed rooms to larger 8-bed

shared spaces (specifically containing 4, 5, 2, 2, 4, 8, 4, 2, 2, 2, and 6 beds respectively), with no single-patient private rooms available in our study setting. The acoustic characteristics varied substantially between ward types, where larger multi-bed wards (particularly W6 with 8 beds and W12 with 6 beds) exhibited higher ambient noise levels due to increased verbal communication among patients, family members, and medical staff, creating a more complex soundscape with frequent conversational overlap. Medical equipment generated distinctive sounds that varied by ward (W3, W6, and W10) with prominent ventilator sounds with characteristic cycling patterns, while general medical wards contained more intermittent monitoring device alerts and infusion pump signals. Medical professionals performed the auscultation process following standard protocols, which involved listening to 12 symmetrical locations: the second, fourth, and eighth intercostal spaces on both sides of the anterior chest, as well as the suprascapular, interscapular, and subscapular areas on both sides of the back. Our study included 43 patients (32 males and 11 females), yielding 63 data points [1]. The average age was 60 years (SD = 14). The primary reasons for hospitalization were cancer (20 patients), pneumonia (11 patients), and COPD or its acute exacerbation (6 patients). 26 of 43 patients had sputum accumulation confirmed by CT scans or bronchoscopy. When divided by region, sputum was identified in the upper left, upper right, lower left, and lower right areas in 8, 6, 16, and 16 cases, respectively, indicating a higher tendency for sputum buildup in the lower lung regions. Among patients with sputum accumulation, 14 patients had sputum in 1 region, 8 patients in 2 regions, and only 4 patients in all four regions. This distribution pattern highlights the potential clinical value of targeted percussion therapy, as it could focus treatment efforts on specific affected regions rather than applying percussion to the entire back.

4.1.3 Performance Metrics.

We assess each region by classifying it as either containing sputum or not, allowing us to measure the accuracy of our detection method. Since most regions do not contain sputum (negative samples outnumber positive samples as most patients have sputum in only 1-2 regions), we need metrics that can properly evaluate performance on imbalanced datasets. We utilize three evaluation metrics commonly used in medical diagnosis: sensitivity, specificity, and F1 score, defined as follows:

(1) **Sensitivity (True Positive Rate)**: The proportion of actual sputum-containing regions correctly identified by the model.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

(2) **Specificity (True Negative Rate)**: The proportion of regions without sputum that are correctly identified as negative.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

(3) **F1 Score**: The harmonic mean of precision and recall, which helps evaluate model robustness on imbalanced datasets.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

[1]Experiments were conducted following the ethical policies of our institutions.

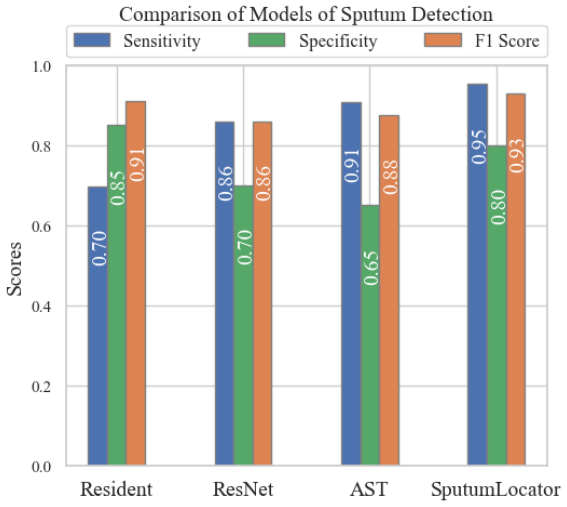

Fig. 6. Performance of Sputum Detection

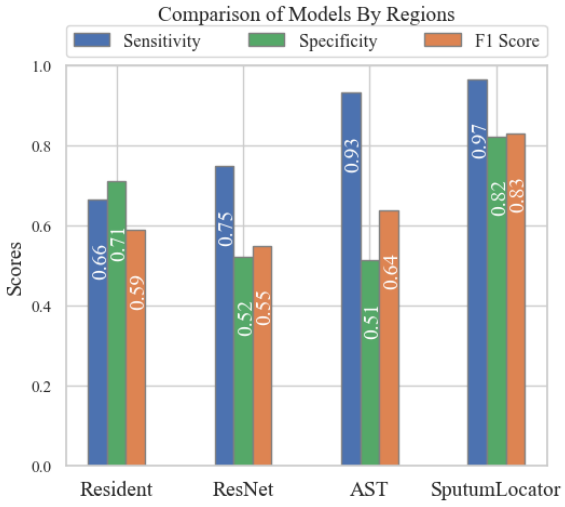

Fig. 7. Performance of Sputum Localization

4.2 Overall Performance

We conducted five-fold cross-validation on the dataset, grouping data by patient to balance computational efficiency and ensure reliable performance while preventing data leakage. The SputumEmbedder was trained for 1000 epochs at each stage, and the SputumClassifier was trained for 100 epochs. We selected ResNet-50 as the benchmark, which is well-suited for audio spectrogram data [39, 41]. We also compare our approach with AST, as it represents the state-of-the-art (SOTA) in general audio tasks [28]. As a clinical baseline, we recruited a medical resident with two years of residency training experience, including one year in respiratory medicine. The resident performed manual classification of the audio samples without access to electronic health records (EHR). In addition to evaluating the models' ability to identify sputum presence in individual regions (sputum localization), we also assessed their sputum detection capability at the patient level (sputum detection). Rather than training additional model heads, we adopted a straightforward aggregation strategy: a sample was classified as negative only if all four regions were predicted as sputum-free, and positive if sputum was detected in any region.

As shown in Fig. 6, all models demonstrated superior performance in the sputum detection task due to its relative simplicity. Typically, the absence of sputum corresponds to normal breath sounds, while the presence of sputum may not always produce distinct abnormal sounds or may generate very subtle acoustic signals. In the detection task, where most patients were positive cases, all models successfully identified sputum presence from the 12-channel audio recordings. However, human assessment showed relatively low sensitivity due to auditory limitations, which aligns with clinical observations reported in [50]. The sputum localization task presented a contrasting data distribution, with negative regions being the majority and positive regions the minority. The task was further complicated by potential acoustic interference, where sputum-induced sounds from one region could affect the assessment of adjacent regions. As illustrated in Fig. 7, ResNet fails to effectively extract sputum sound features and their spatial characteristics from the complex data. The AST model pre-trained on general audio datasets, demonstrates slightly better performance. However, during the training process, we observed that it struggles to balance precision and recall, and the model appears more adept at capturing the acoustic features associated with sputum but fails to retain their inherent acoustic properties. Residency assessment, benefiting

from the ability to distinguish environmental and skin friction noises by human experience, showed better performance in identifying negative samples compared to these models. In contrast, SputumLocator accurately extracts both sputum-induced features and spatial information, effectively reducing false positives in the task of determining whether sputum has accumulated in each region, and achieves sensitivity of 0.97, specificty of 0.82 and F1 score of 0.83.

4.3 System Performance

Table 2. System Performance Comparison

Model	Parameters (M)	Inference Time (ms)
ResNet-50	25	47
AST	87	158
SputumLocator	87	170

While our application doesn't mandate real-time processing or deployment in hardware-constrained environments, as cloud computing resources can be leveraged similar to commercial solutions, we still conducted comprehensive system performance analysis. We evaluated two key metrics: average inference time and model size (parameters).

As illustrated in Tab. 2, our experiments revealed that ResNet-50, serving as our baseline, demonstrates the most efficient performance with 25M parameters and an average inference time of 47ms. The AST model, while more sophisticated, requires 87M parameters and takes 158ms per inference. Our proposed SputumLocator, building upon the AST architecture, maintains the same parameter count (87M) but requires slightly more computation time (170ms) due to the additional localization components. Though SputumLocator system is slightly more time-consuming, but it remains practical for real-world applications, as inference is performed only once per use.

4.4 Demographic Analysis

We also compare the system's performance in different demographics, the result is shown in Fig. 8

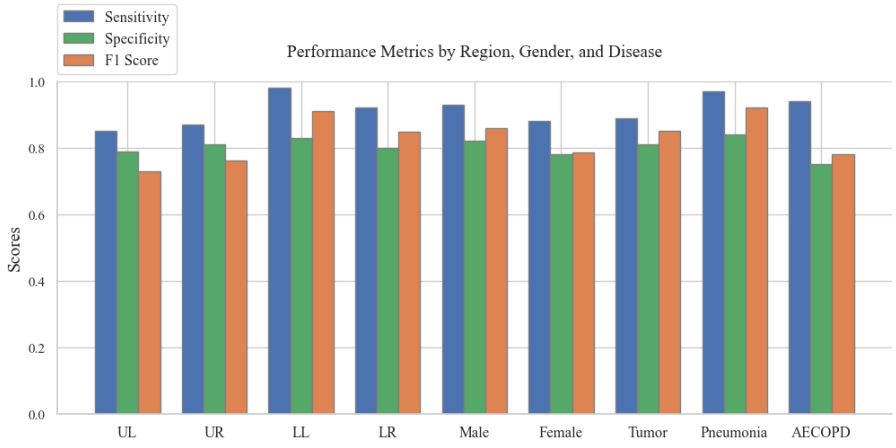

Fig. 8. Demographic Analysis

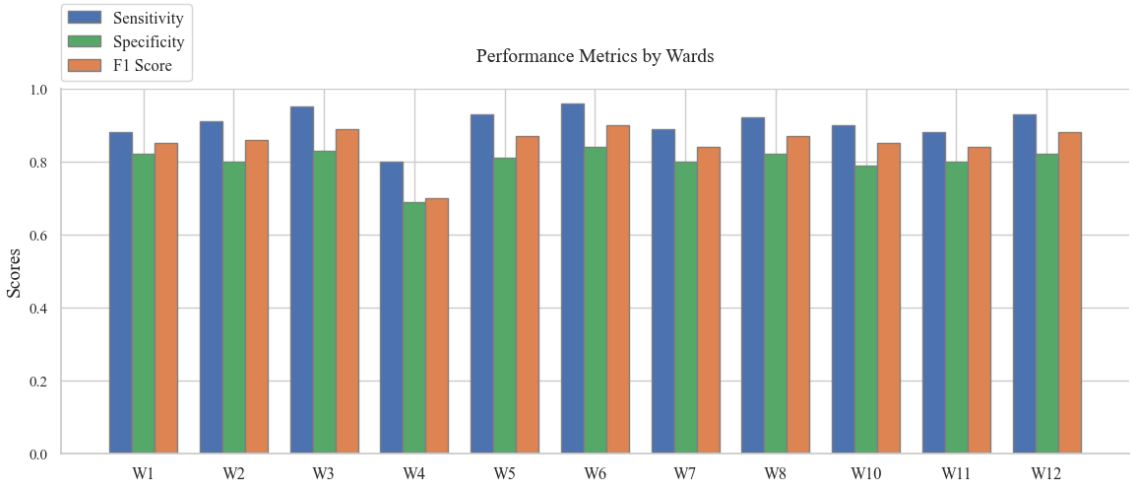

Fig. 9. Performance in Difference Wards

4.4.1 *Performance on Thorax Regions.*

The model performs similarly across all four regions, excelling in the lower lungs likely due to a larger dataset. It shows slightly reduced performance in the upper left lung, possibly caused by stronger heart murmurs and the use of cardiac pacemakers in some patients, which may affect auscultation accuracy.

4.4.2 *Performance on Genders.*

We observed that the model performs better on male than on female due to several factors. Firstly, males generally have higher respiratory intensity, resulting in stronger signals with abnormal respitory audio features, especially in patients with weaker respiratory strength. Secondly, our dataset contains fewer female samples, limiting training diversity and potentially reducing the model's ability to generalize to female test samples. Nonetheless, the model remains robust across both genders, delivering acceptable results.

4.4.3 *Performance on Diseases.*

SputumLocator exhibits slightly differentiated performance between groups of patients with various diseases. For patients undergoing maintenance chemotherapy for cancer, the model's performance is comparable to the overall average, since their respiratory function is relatively less affected by the disease and the sample size is the largest. In hospitalized patients with acute pneumonia, the model achieves the highest precision, which is attributable to the pronounced sputum accumulation features caused by inflammation and the higher respiratory intensity in these patients, which facilitates the recognition of sputum sounds. However, in the group of patients with acute exacerbation of chronic obstructive pulmonary disease (AECOPD), it has very high sensitivity but relatively low specificity. This characteristic performance pattern likely stems from the complex pathophysiological features of AECOPD. These patients typically present with diffuse airway inflammation, increased sputum production, and altered breath sounds throughout the respiratory tract. Given these widespread abnormalities, the model shows strong capability in detecting genuine sputum sounds, but may struggle to distinguish between actual sputum sounds and other similar respiratory artifacts. These characteristics present substantial challenges for the accurate identification and localization of sputum sounds.

4.4.4 *Performance in Different Wards.*

As shown in Fig. 9, our model demonstrated consistent performance across different wards, with no substantial

variations in effectiveness. Notably, the eight-bed ward (W6) even exhibited slightly superior performance metrics, though it's the most noisy environment. The only notable exception was W4, a two-bed ward, where performance metrics showed a relative decline. Upon detailed case review, we identified that one patient in W4 presented with severely diminished breath sounds with significant parenchymal lung changes, which adversely affected the model's performance in this ward. It should be noted that W9 was not included in our data collection process during the study period, hence no performance metrics were recorded for this ward.

4.4.5 Statistical Parity Analysis.

We conducted a analysis of demographic parity and equality of opportunity across different grouping variables: Chest Region, Gender, Disease Type, and Wards. The assessment was based on the widely-adopted Four-Fifths Rule [23]. For demographic parity, we compared the proportion of positive predictions across different groups, while for equality of opportunity, we examined the true positive rates among positive samples. The rule stipulates that the maximum disparity between groups should not exceed 20%.

As shown in Tab. 3, our analysis revealed that demographic parity could not be achieved for Chest Region, Disease and Gender categories. For Chest Region, this disparity primarily stems from the physiological tendency of sputum accumulation in lower lung regions due to gravitational effects and bronchial tree anatomy, resulting in a naturally higher prevalence of sputum sounds in these areas. The disparities in Disease and Gender categories can be attributed to two main factors: the inherent variation in positive sample distribution across different disease types, and the strong correlation between Gender and Disease in our limited sample size. For instance, all AECOPD patients in our cohort were male. However, equality of opportunity was successfully maintained across all grouping variables.

Table 3. Statistical Parity Analysis Results

Group Variable	**Demographic Parity**		**Equality of Opportunity**	
	Max Disparity	Satisfied	Max Disparity	Satisfied
Chest Region	0.35	✗	0.13	✓
Gender	0.25	✗	0.05	✓
Disease	0.39	✗	0.08	✓
Wards	0.17	✓	0.17	✓

4.5 Ablation Study

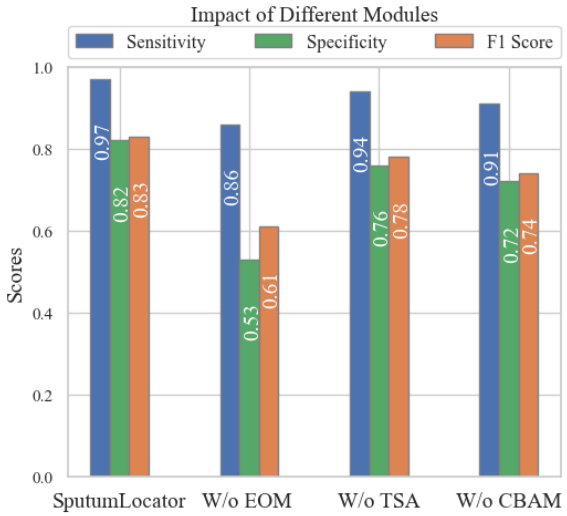

Fig. 10. Ablation Study

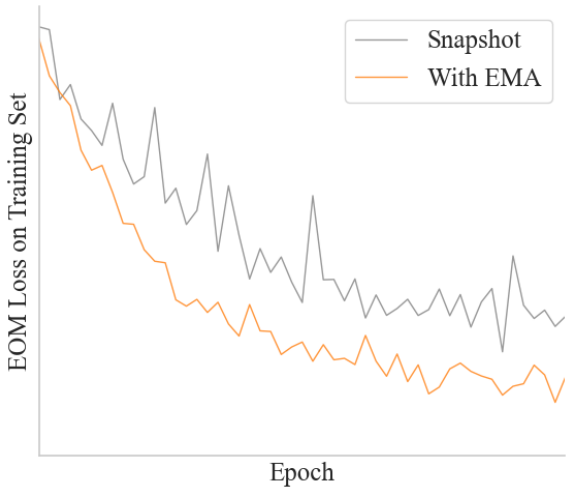

Fig. 11. Impact of EMA

4.5.1 *Benefits of EOM.*

Theoretically, we could train directly on mixed data. However, we found that the model was difficult to converge. This difficulty arises because the random masks introduce additional randomness, and excessive data augmentation may further complicate model training [38]. To assess the impact of pretraining on breathing data using EOM, we directly trained TSA with a noise mixture for 1,000 epochs for comparison. The results demonstrate that without training EOM on breathing data, the model struggles to extract features from complex, noisy data, performing even worse than the AST model pre-trained on a general dataset. Additionally, the increased complexity of our model contributes to this performance gap.

4.5.2 *Benefits of TSA with Noise Data.*

The primary purpose of TSA is to enhance the model's robustness. To compare the results, we directly trained EOM for 1,000 epochs and then input the generated embeddings into the SputumClassifier for evaluation. We observed that without TSA to augment the model on noise data, the model's performance declined 3% sensitivity, 6% specificity, and 5% F1-score.

4.5.3 *Benefits of CBAM.*

We directly input the SputumEmbedder generated embeddings into the 12-channel data through a fully connected layer. This practice is very common in the Transformer architecture and helps improve the performance of the model. We see that each auscultation point intrincically have some spation information of sputum, and naive FC can also extract some basic information, but the performance if much weaker than CBAM. The result indicates it's crucial to understand all channel information to determine if one channel is without sputum while there exists sputum in other regions.

4.5.4 *Impact of SpecAugment.*

Since our framework relies on MAE structure, augmenting the raw data becomes quite challenging. During the

training of SputumEmbedder, some patches are already randomly masked. When we further mask portions of the frequency and time dimensions, the model's training task would become excessively difficult. On the other hand, the masked data also generate partial embeddings, meaning that this information is effectively being reused. SpecAugment, however, simply removes parts of the information to enhance the model's robustness. In our study, the selection of both the SpecAugment mask ratio and the patch mask ratio is crucial. Through experiments, we chose a 5% SpecAugment mask ratio and a 60% patch mask ratio, thereby improving the model's generalization ability and robustness without significantly increasing the training difficulty.

4.5.5 Impact of EMA.

The selection of the teacher model is crucial in self-supervised learning. While directly using the snapshot of the student model from the previous iteration as the teacher is straightforward, it can lead to training instability due to rapid parameter changes. Following [57, 81], we adopt Exponential Moving Average (EMA) to maintain a stable teacher model. EMA creates a temporal ensemble of parameters by assigning higher weights to recent iterations while gradually discounting historical information, thus smoothing parameter fluctuations and ensuring consistent predictions. By averaging the weights, EMA reduces the impact of sudden changes, stabilizes the training process, smooths out the loss landscape, and prevents overfitting by making the averaged weights less susceptible to noise in the training data, ultimately leading to improved model convergence and generalization. Our experimental results on EOM Loss convergence over 50 epochs, Fig. 11 demonstrate that EMA achieves notably smoother convergence compared to the snapshot approach, corroborating the findings in [81] and confirming its effectiveness in stabilizing the training process.

5 DISCUSSION

While SputumLocator demonstrates promising performance in automated sputum localization through digital auscultation, several aspects warrant discussion regarding its clinical implications and future development.

5.1 Clinical Implications and Practice Integration

5.1.1 Enhanced Treatment in Resource-Limited Settings.

For long-term management of patients with MOLDs, SputumLocator serves as an auscultation-guided percussion tool in community and home care settings. By providing standardized auscultation and automated sputum localization, it helps caregivers deliver more precise and effective percussion therapy. This support is particularly valuable in areas where specialized respiratory expertise is limited, as it enables caregivers to perform quality percussion despite minimal training. The guidance of the system reduces uncertainty in percussion locations and timing, allowing caregivers to focus on proper technique and patient comfort during treatment.

5.1.2 Supporting Remote Care and Patient Engagement.

SputumLocator's digital capabilities integrate sputum localization into remote care systems. While its primary function is to assist caregivers in locating sputum through auscultation for more effective percussion, the system also captures and analyzes sputum-related data. When integrated with electronic health records (EHR), this quantified sputum information enables healthcare providers to track changes in sputum distribution patterns and respiratory conditions over time. Through standardized home-based monitoring, caregivers can perform more targeted percussion with real-time guidance, while the collected data supports clinical decision-making and treatment adjustments. This systematic approach to percussion therapy not only improves treatment quality but also enhances patient engagement and caregiver confidence, leading to better adherence to prescribed respiratory care regimens.

5.2 Limitations and Future Research Directions

5.2.1 Demographic Representation and Clinical Validation.

Our current data collection was limited to inpatients, primarily those with lung cancer and pneumonia, resulting in sampling bias towards severe conditions. This may not adequately represent key beneficiary populations like stable chronic lung disease patients and the elderly in community settings. Additionally, factors such as subcutaneous fat thickness affecting auscultation quality could not be properly controlled. Furthermore, our analysis revealed room for improvement in demographic parity, which is essential for ensuring that our model is fair and equitable across different populations. To address these limitations, we plan to partner with primary care institutions to integrate data collection into routine health screenings, enabling more diverse and representative sampling. By expanding our dataset to include a broader range of patients, we aim to improve the generalizability of our model, reduce potential biases, and revisit our approach to demographic parity, ultimately leading to a more robust and equitable model that can effectively serve diverse populations and improve health outcomes.

5.2.2 Device Compatibility.

Current validation is limited to the Eko 3M™ Littmann® CORE Digital Stethoscope. Given that different digital stethoscopes have varying hardware specifications and acoustic characteristics [8], future work will expand testing to other mainstream devices to evaluate SputumLocator's performance across different hardware configurations.

5.2.3 Integration with Clinical Systems.

Future development will focus on integrating SputumLocator with electronic health records (EHR) to create a comprehensive respiratory care platform. This integration aims to combine auscultation data with clinical parameters, treatment records, and patient-reported outcomes for enhanced disease monitoring and management. A multi-center randomized controlled trial comparing standard versus system-guided percussion therapy is proposed to validate clinical effectiveness and guide implementation strategies.

5.2.4 Ausculataion Guidance.

While clinical specialists and trained caregivers who know anatomical structure can easily position the stethoscope for auscultation, expanding the system's impact requires consideration of users without such expertise, such as the family members of MOLDs patients. To address this, designated guidance is necessary to facilitate accurate stethoscope placement. Furthermore, future improvements should focus on enhancing the model's robustness to inaccurate stethoscope positioning and evaluating its performance in real-world scenarios.

6 RELATED WORKS

In this section, we review related works from both technical and contextual perspectives. Specifically, we categorize existing studies into two main areas: (1) respiratory sound analysis and (2) community-level management of pulmonary diseases.

6.1 Respiratory Sound Analysis

Respiratory sounds contain rich clinical information. Extracting abnormal sound features from auscultation is not only fundamental for localizing sputum but also crucial for diagnosing respiratory diseases. Numerous researchers have employed machine learning techniques to conduct objective and quantitative assessments of lung health [69]. Sen et al. successfully classified lung sounds as normal or abnormal by feeding mathematical features into SVM and GMM frameworks. Similarly, Mondal et al. distinguished respiratory patients from healthy subjects using a MLP, achieving an accuracy of 92.8% [63]. In [7], researchers classified patients as having COPD or being healthy using Boltzmann machines, reaching an accuracy of 93.7%. García et al. [26] utilized CNN to classify subjects as healthy, COPD patients, or non-COPD patients. Fraiwan et al. further extended this approach by classifying subjects as healthy or having one of five specific diseases [24]. *mWheeze* [17] leverages

pervasive smartphone sensors (IMU and microphone) placed on a patient's chest to detect breathing patterns, wheeze sounds, and assess airway obstruction severity. While *mWheeze* achieves reliable wheeze detection and severity classification, it focuses on overall obstruction assessment rather than providing spatial information of sputum that could guide percussion. However, a significant limitation in identifying patients with different types of respiratory diseases lies in the datasets, which typically include only a limited range of disease types. Consequently, models struggle to classify unseen diseases. Therefore, detecting specific lung sounds, such as crackles, as an auxiliary task to diagnose specific diseases is more clinically valuable. Following the release of ICBHI 2017 Challenge respiratory sound database and the introduction of evaluation criteria like the ICBHI Score (average of specificity and sensitivity) [72], many studies have focused on classifying normal sounds, crackles, wheezes, and combinations of crackles and wheezes. Li et al. [53] proposed augmenting attention convolution within ResNet blocks for classification, achieving a score of 53.90. Wang et al. [86] improved their score to 55.30 by augmenting data through domain transfer and utilizing ResNeXt. With the popularity of transformer architectures, Bae et al. [10] employed the AST to achieve a score of 59.55, marking a significant improvement over CNN-based models. They observed that transformer models require more data and thus proposed Patch-Mix contrastive learning for data augmentation. Kim et al. [48] noted that the device type can impact performance and introduced stethoscope-guided supervised contrastive learning for cross-domain adaptation, achieving a score of 61.71. Masked Modeling Duo, which utilizes contrastive learning and pre-training in large general data sets, reached a score of 62.73 [66]. Recently, the BTS model, the first multimodal text-audio model that incorporates respiratory sound metadata, achieved the highest score of 63.54 [49]. Although these studies provide valuable insights and evidence on the extraction of information from respiratory data, they do not specifically target sputum localization and consequently lack designs capable of extracting intrinsic spatial information from different auscultation points. There are also researchs target sputum detection, but they merely determine if the abnormal sound induced by sputum is detected, but not the location of the sputum [47, 67].

6.2 Community-Level Management of Pulmonary Diseases

For people with respiratory diseases, particularly those with chronic conditions, routine assessment and monitoring at the community level are crucial. Among vital signs, respiratory rate (RR) is a significant indicator [56]. Numerous studies have successfully estimated RR from Photoplethysmography (PPG) signals [55, 62], while others have attempted to capture subtle movements associated with breathing using Inertial Measurement Units (IMUs) [37, 54, 79]. Currently, sensors integrated into smartwatches have effectively monitored breathing at rest and have received FDA approval for the detection of sleep apnea [2, 4]. In addition, some research has focused on identifying RR using the microphone on mobile phones. Radio-Frequency (RF) based techniques, such as active acoustic solutions, WiFi, and millimeter waves, have also been used to provide non-invasive respiratory monitoring [6, 31, 84, 85, 88, 91]. Beyond RR, the analysis of breathing patterns is also important. Passive alterations in breathing patterns can indicate respiratory compensation and indicates the deterioration of the respiratory system [76]. Conversely, patients with chronic respiratory diseases often engage in active diaphragmatic breathing practice to improve their quality of life (QoL) [16]. BreathMentor uses a microphone array and active sonar to distinguish between chest and abdominal breathing patterns[68]. However, this study was conducted on healthy subjects. Another approach by [68] employs a motion capture system to determine the breathing pattern, which is not cost-efficient. Recently, DeepBreath [89] has been developed to measure chest and abdomen movements using a depth camera, simultaneously determining RR, breathing patterns, and breathing volume.

BreathTrack [44] leverages smartphone IMU sensors to guide microphone to detect breathing phases and assess respiratory conditions. While *BreathTrack* also employs TSA, where the IMU sensor acts as a teacher

to guide the acoustic model, our work proposes a fundamentally different TSA. Unlike *BreathTrack*'s cross-modal knowledge transfer, our TSA operates within the same modality and model architecture, where harder augmented samples serve as students to learn from easier samples (teachers), thus improves the model robustness. Assessing lung condition remotely is also crucial. Recent innovations in this field include Listen2Cough , which implements a passive, end-to-end cough detection system on smartphones for lung health evaluation [90]. Similarly, PulmoListener leverages smartwatch technology to collect audio data and assess symptom severity in COPD patients [12]. The technological advances serve as valuable complements to airway clearance management protocols.

These studies contribute to respiratory health management from various perspectives, thereby aiding patients in achieving a better QoL at the community level.

7 CONCLUSION

In this research, we present SputumLocator, a novel digital stethoscope-based sputum localization system, which leverages standard auscultation procedures without relying on complex operations, making it suitable for widespread use in community settings. SputumLocator employs a data-driven approach, with an innovative design with a powerful SputumEmbedder and a lightweight SputumClassifier, which effectively utilize multi-level data features. By applying two-stage distinct pretraining methods, we developed a resilient feature extractor with limited data. In collaboration with a large medical institution, we collected comprehensive and standardized auscultation data from 43 patients immediately after CT scans or before bronchial aspiration, ensuring consistency in sputum distribution and the labels. The experimental results demonstrate that SputumLocator achieves an overall sensitivity of 0.97, specificity of 0.82, and F1-Score of 0.83 and exhibits excellent robustness across thoracic regions, genders, and disease types. This system has the potential to benefit p opulations r equiring airway clearance at the community level.

In this research, we present SputumLocator, a novel digital stethoscope-based sputum localization system, which leverages standard auscultation procedures without relying on complex operations, making it suitable for widespread use in community settings. SputumLocator employs a data-driven approach, with an innovative design with a powerful SputumEmbedder and a lightweight SputumClassifier, which effectively utilize multi-level data features. By applying two-stage distinct pretraining methods, we developed a resilient feature extractor with limited data. In collaboration with a large medical institution, we collected comprehensive and standardized auscultation data from 43 patients immediately after CT scans or before bronchial aspiration, ensuring consistency in sputum distribution and the labels. The experimental results demonstrate that SputumLocator achieves an overall Sensitivity of 0.97, Specificity of 0.82, and F1-Score of 0.83 and exhibits excellent robustness across thoracic regions, genders, and disease types. This system has the potential to benefit populations requiring airway clearance at the community level.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous editors and reviewers for the valuable comments and helpful suggestions. This work is partially supported by the RGC under Contract CERG 16204418, 16203719, 16204820, R8015 and MOST18FYT08.

REFERENCES

[1] 2024. https://www.ekohealth.com/products/3m-littmann-core-digital-stethoscope?variant=39307014209632 Accessed Oct 21, 2024.

[2] 2024. Apple introduces groundbreaking health features to support conditions impacting billions of people. https://www.apple.com/hk/en/newsroom/2024/09/apple-introduces-groundbreaking-health-features/#:~:text=Apple%20introduces%20groundbreaking%20new%20features,experiences%20that%20enrich%20usersâĂŹ%20lives. Accessed Oct 19, 2024.

[3] 2024. Eko Healthv Digital Stethoscope. https://www.ekohealth.com/?srsltid=AfmBOoq0wSFs8__V2MZBwaMatTtfXHOgcWq63hsySNS_1mMy2bOCCpy6 Accessed Oct 16, 2024.
[4] 2024. Samsung's Sleep Apnea Feature on Galaxy Watch First of Its Kind Authorized by US FDA. https://news.samsung.com/global/samsungs-sleep-apnea-feature-on-galaxy-watch-first-of-its-kind-cleared-by-us-fda Accessed Oct 19, 2024.
[5] 2024. Thinklabs One Digital Stethoscope. https://www.thinklabs.com/ Accessed Oct 16, 2024.
[6] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C. Miller. 2015. Smart Homes That Monitor Breathing and Heart Rate. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 837–846. https://doi.org/10.1145/2702123.2702200
[7] Gokhan Altan, Yakup Kutlu, and Novruz Allahverdi. 2019. Deep learning on computerized analysis of chronic obstructive pulmonary disease. *IEEE journal of biomedical and health informatics* 24, 5 (2019), 1344–1350.
[8] Yi Yang Ang, Li Ren Aw, Vivian Koh, and Rex X Tan. 2023. Characterization and cross-comparison of digital stethoscopes for telehealth remote patient auscultation. *Medicine in Novel Technology and Devices* 19 (2023), 100256.
[9] Sangmin Bae, June-Woo Kim, Won-Yang Cho, Hyerim Baek, Soyoun Son, Byungjo Lee, Changwan Ha, Kyongpil Tae, Sungnyun Kim, and Se-Young Yun. 2023. Patch-Mix Contrastive Learning with Audio Spectrogram Transformer on Respiratory Sound Classification. In *INTERSPEECH 2023*. 5436–5440. https://doi.org/10.21437/Interspeech.2023-1426
[10] Sangmin Bae, June-Woo Kim, Won-Yang Cho, Hyerim Baek, Soyoun Son, Byungjo Lee, Changwan Ha, Kyongpil Tae, Sungnyun Kim, and Se-Young Yun. 2023. Patch-mix contrastive learning with audio spectrogram transformer on respiratory sound classification. *arXiv preprint arXiv:2305.14032* (2023).
[11] Stefano Belli, Ilaria Prince, Gloria Savio, Elena Paracchini, Davide Cattaneo, Manuela Bianchi, Francesca Masocco, Maria Teresa Bellanti, and Bruno Balbi. 2021. Airway clearance techniques: the right choice for the right patient. *Frontiers in medicine* 8 (2021), 544826.
[12] Sejal Bhalla, Salaar Liaqat, Robert Wu, Andrea S Gershon, Eyal de Lara, and Alex Mariakakis. 2023. PulmoListener: Continuous Acoustic Monitoring of Chronic Obstructive Pulmonary Disease in the Wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–24.
[13] Elroy Boers, Meredith Barrett, Jason G Su, Adam V Benjafield, Sanjeev Sinha, Leanne Kaye, Heather J Zar, Vy Vuong, Daniela Tellez, Rahul Gondalia, et al. 2023. Global burden of chronic obstructive pulmonary disease through 2050. *JAMA Network Open* 6, 12 (2023), e2346598–e2346598.
[14] Abraham Bohadana, Gabriel Izbicki, and Steve S Kraman. 2014. Fundamentals of lung auscultation. *New England Journal of Medicine* 370, 8 (2014), 744–751.
[15] Richard C Boucher. 2019. Muco-obstructive lung diseases. *New England Journal of Medicine* 380, 20 (2019), 1941–1953.
[16] Lawrence P Cahalin, Malinda Braga, Yoshimi Matsuo, and Edgar D Hernandez. 2002. Efficacy of diaphragmatic breathing in persons with chronic obstructive pulmonary disease: a review of the literature. *Journal of Cardiopulmonary Rehabilitation and Prevention* 22, 1 (2002), 7–21.
[17] Soujanya Chatterjee, Md Mahbubur Rahman, Tousif Ahmed, Nazir Saleheen, Ebrahim Nemati, Viswam Nathan, Korosh Vatanparvar, and Jilong Kuang. 2020. Assessing severity of pulmonary obstruction from respiration phase-based wheeze-sensing using mobile sensors. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[18] Laura Cooper, Kylie Johnston, and Marie Williams. 2024. Physiotherapy-led, community-based airway clearance services for people with chronic lung conditions: a retrospective descriptive evaluation of an existing model of care. *BMC Health Services Research* 24, 1 (2024), 98.
[19] Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
[20] Patricia A Downie, DM Innocenti, and SE Jackson. 1987. Cash's textbook of chest, heart and vascular disorders for physiotherapists. (1987).
[21] B Flietstra, Natasha Markuzon, Andrey Vyshedskiy, and R Murphy. 2011. Automated analysis of crackles in patients with interstitial pulmonary fibrosis. *Pulmonary medicine* 2011, 1 (2011), 590506.
[22] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2021. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2021), 829–852.
[23] James R Foulds and Shimei Pan. 2020. Are Parity-Based Notions of {AI} Fairness Desirable? *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering* 43, 4) (2020).
[24] Mohammad Fraiwan, Luay Fraiwan, Mohanad Alkhodari, and Omnia Hassanin. 2022. Recognition of pulmonary diseases from lung sounds using convolutional neural networks and long short-term memory. *Journal of Ambient Intelligence and Humanized Computing* (2022), 1–13.
[25] Juan P Garcia-Mendez, Amos Lal, Svetlana Herasevich, Aysun Tekin, Yuliya Pinevich, Kirill Lipatov, Hsin-Yi Wang, Shahraz Qamar, Ivan N Ayala, Ivan Khapov, et al. 2023. Machine learning for automated classification of abnormal lung sounds obtained from public databases: a systematic review. *Bioengineering* 10, 10 (2023), 1155.

[26] María Teresa García-Ordás, José Alberto Benítez-Andrades, Isaías García-Rodríguez, Carmen Benavides, and Héctor Alaiz-Moretón. 2020. Detecting respiratory pathologies using convolutional neural networks and variational autoencoders for unbalancing data. *Sensors* 20, 4 (2020), 1214.
[27] Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*. 571–575. https://doi.org/10.21437/Interspeech.2021-698
[28] Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021).
[29] Yuan Gong, Yu-An Chung, and James Glass. 2021. PSLA: Improving Audio Tagging with Pretraining, Sampling, Labeling, and Aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2021). https://doi.org/10.1109/TASLP.2021.3120633
[30] Yanbin Gong, Wentao Xie, Qian Zhang, and Shifang Yang. 2024. Hypergradient Descent Based Multi-Task Learning on Auscultation Point Guided Respiratory Sound Classification. In *2024 IEEE 20th International Conference on Body Sensor Networks (BSN)*. IEEE, 1–4.
[31] Yanbin Gong, Qian Zhang, Bobby H.P. NG, and Wei Li. 2022. BreathMentor: Acoustic-Based Diaphragmatic Breathing Monitor System. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 53 (jul 2022), 28 pages. https://doi.org/10.1145/3534595
[32] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* 33 (2020), 21271–21284.
[33] Seema Grover. 2022. Challenges in physiotherapy of managing respiratory diseases in elderly population. *Indian Journal of Tuberculosis* 69 (2022), S280–S286.
[34] Pierre-Amaury Grumiaux, Srđan Kitić, Laurent Girin, and Alexandre Guérin. 2022. A survey of sound source localization with deep learning methods. *The Journal of the Acoustical Society of America* 152, 1 (2022), 107–151.
[35] Wei-jie Guan, Xiao-rong Han, David de la Rosa-Carrillo, and Miguel Angel Martinez-Garcia. 2019. The significant global economic burden of bronchiectasis: a pending matter. *European Respiratory Journal* 53, 2 (2019). https://doi.org/10.1183/13993003.02392-2018 arXiv:https://publications.ersnet.org//content/erj/53/2/1802392.full.pdf
[36] Honorata Hafke-Dys, Anna Bręborowicz, Paweł Kleka, Jędrzej Kociński, and Adam Biniakowski. 2019. The accuracy of lung auscultation in the practice of physicians and medical students. *PLoS One* 14, 8 (2019), e0220606.
[37] Tian Hao, Chongguang Bi, Guoliang Xing, Roxane Chan, and Linlin Tu. 2017. MindfulWatch: A Smartwatch-Based System For Real-Time Respiration Monitoring During Meditation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3, Article 57 (Sept. 2017), 19 pages. https://doi.org/10.1145/3130922
[38] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.
[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[40] Wentao He, Yuchen Yan, Jianfeng Ren, Ruibin Bai, and Xudong Jiang. 2024. Multi-View Spectrogram Transformer for Respiratory Sound Classification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8626–8630.
[41] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.
[42] Guanzhe Hong, Zhiyuan Mao, Xiaojun Lin, and Stanley H Chan. 2021. Student-teacher learning from clean inputs to noisy inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12075–12084.
[43] Dong-Min Huang, Jia Huang, Kun Qiao, Nan-Shan Zhong, Hong-Zhou Lu, and Wen-Jin Wang. 2023. Deep learning-based lung sound analysis for intelligent stethoscope. *Military Medical Research* 10, 1 (2023), 44.
[44] Bashima Islam, Md Mahbubur Rahman, Tousif Ahmed, Mohsin Yusuf Ahmed, Md Mehedi Hasan, Viswam Nathan, Korosh Vatanparvar, Ebrahim Nemati, Jilong Kuang, and Jun Alex Gao. 2021. BreathTrack: detecting regular breathing phases from unannotated acoustic data captured by a smartphone. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–22.
[45] Jean-Paul Janssens and Karl-Heinz Krause. 2004. Pneumonia in the very old. *The Lancet infectious diseases* 4, 2 (2004), 112–124.
[46] Sherri Lynne Katz. 2023. Airway clearance. In *Pulmonary Assessment and Management of Patients with Pediatric Neuromuscular Disease*. Elsevier, 91–110.
[47] Hyunbum Kim, Daeyeon Koh, Yohan Jung, Hyunjun Han, Jongbaeg Kim, and Younghoon Joo. 2023. Breathing sounds analysis system for early detection of airway problems in patients with a tracheostomy tube. *Scientific reports* 13, 1 (2023), 21029.
[48] June-Woo Kim, Sangmin Bae, Won-Yang Cho, Byungjo Lee, and Ho-Young Jung. 2024. Stethoscope-Guided Supervised Contrastive Learning for Cross-Domain Adaptation on Respiratory Sound Classification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1431–1435.
[49] June-Woo Kim, Miika Toikkanen, Yera Choi, Seoung-Eun Moon, and Ho-Young Jung. 2024. BTS: Bridging Text and Sound Modalities for Metadata-Aided Respiratory Sound Classification. *arXiv preprint arXiv:2406.06786* (2024).
[50] Yoonjoo Kim, YunKyong Hyon, Sung Soo Jung, Sunju Lee, Geon Yoo, Chaeuk Chung, and Taeyoung Ha. 2021. Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning. *Scientific reports* 11, 1 (2021), 1–11.

[51] Annemarie L Lee, Angela T Burge, and Anne E Holland. 2015. Airway clearance techniques for bronchiectasis. *Cochrane Database of Systematic Reviews* 11 (2015).

[52] Mary K Lester and Patrick A Flume. 2009. Airway-clearance therapy guidelines and implementation. *Respiratory care* 54, 6 (2009), 733–753.

[53] Jizuo Li, Jiajun Yuan, Hansong Wang, Shijian Liu, Qianyu Guo, Yi Ma, Yongfu Li, Liebin Zhao, and Guoxing Wang. 2021. LungAttn: advanced lung sound classification using attention mechanism with dual TQWT and triple STFT spectrogram. *Physiological Measurement* 42, 10 (2021), 105006.

[54] Daniyal Liaqat, Mohamed Abdalla, Pegah Abed-Esfahani, Moshe Gabel, Tatiana Son, Robert Wu, Andrea Gershon, Frank Rudzicz, and Eyal De Lara. 2019. WearBreathing: Real World Respiratory Rate Monitoring Using Smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 56 (jun 2019), 22 pages. https://doi.org/10.1145/3328927

[55] Yue-Der Lin, Ya-Hsueh Chien, and Yi-Sheng Chen. 2017. Wavelet-based embedded algorithm for respiratory rate estimation from PPG signal. *Biomedical Signal Processing and Control* 36 (2017), 138–145.

[56] Haipeng Liu, John Allen, Dingchang Zheng, and Fei Chen. 2019. Recent development of respiratory rate measurement technologies. *Physiological measurement* 40, 7 (2019), 07TR01.

[57] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. 2021. Unbiased Teacher for Semi-Supervised Object Detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[58] Alyssa Yue Hui Low, Just Sen Tan, Rethinam Ganesan, Jaclyn Tan, and Albert Yick Hou Lim. 2020. Systematic review on adherence, barriers to treatment and impact of airway clearance in bronchiectasis. *Archives of Clinical and Biomedical Research* 4, 5 (2020), 481–497.

[59] Maggie Patricia McIlwaine, Nicole Marie Lee Son, and Melissa Lynn Richmond. 2014. Physiotherapy and cystic fibrosis: what is the evidence base? *Current opinion in pulmonary medicine* 20, 6 (2014), 613–617.

[60] Ian McLane, Dimitra Emmanouilidou, James E West, and Mounya Elhilali. 2021. Design and comparative performance of a robust lung auscultation system for noisy clinical settings. *IEEE Journal of Biomedical and Health Informatics* 25, 7 (2021), 2583–2594.

[61] Hasse Melbye, Luis Garcia-Marcos, Paul Brand, Mark Everard, Kostas Priftis, and Hans Pasterkamp. 2016. Wheezes, crackles and rhonchi: simplifying description of lung sounds increases the agreement on their classification: a study of 12 physicians' classification of lung sounds from video recordings. *BMJ open respiratory research* 3, 1 (2016), e000136.

[62] D. J. Meredith, D. Clifton, P. Charlton, J. Brooks, C. W. Pugh, and L. Tarassenko. 2012. Photoplethysmographic derivation of respiratory rate: a review of relevant physiology. *J Med Eng Technol* 36, 1 (Jan 2012), 1–7.

[63] Ashok Mondal, Parthasarathi Bhattacharya, and Goutam Saha. 2014. Detection of lungs status using morphological complexities of respiratory sounds. *The scientific world journal* 2014, 1 (2014), 182938.

[64] Raymond Murphy and Andrey Vyshedskiy. 2010. Acoustic findings in a patient with radiation pneumonitis. *New England Journal of Medicine* 363, 20 (2010), e31.

[65] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. 2023. Masked modeling duo: Learning representations by encouraging both networks to model the input. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[66] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. 2024. Masked Modeling Duo: Towards a Universal Audio Pre-Training Framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).

[67] Jinglong Niu, Yan Shi, Maolin Cai, Zhixin Cao, Dandan Wang, Zhaozhi Zhang, and Xiaohua Douglas Zhang. 2018. Detection of sputum by interpreting the time-frequency distribution of respiratory sound signal using image processing techniques. *Bioinformatics* 34, 5 (2018), 820–827.

[68] Yulia Orlova, Alexander Gorobtsov, Oleg Sychev, Vladimir Rozaliev, Alexander Zubkov, and Anastasia Donsckaia. 2023. Method for determining the Dominant type of human breathing using motion capture and Machine Learning. *Algorithms* 16, 5 (2023), 249.

[69] Rajkumar Palaniappan, Kenneth Sundaraj, and Nizam Uddin Ahamed. 2013. Machine learning in lung sound analysis: a systematic review. *Biocybernetics and Biomedical Engineering* 33, 3 (2013), 129–135.

[70] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779* (2019).

[71] Sandra Reichert, Raymond Gass, Christian Brandt, and Emmanuel Andrès. 2008. Analysis of respiratory sounds: state of the art. *Clinical medicine. Circulatory, respiratory and pulmonary medicine* 2 (2008), CCRPM–S530.

[72] Bruno M Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P Kahya, Nikša Jakovljevic, Tatjana L Turukalo, Ioannis M Vogiatzis, Eleni Perantoni, et al. 2019. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological measurement* 40, 3 (2019), 035001.

[73] Aaqib Saeed, David Grangier, and Neil Zeghidour. 2021. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3875–3879.

[74] Saeid Safiri, Kristin Carson-Chahhoud, Maryam Noori, Seyed Aria Nejadghaderi, Mark JM Sullman, Javad Ahmadian Heris, Khalil Ansarin, Mohammad Ali Mansournia, Gary S Collins, Ali-Asghar Kolahi, et al. 2022. Burden of chronic obstructive pulmonary disease and its attributable risk factors in 204 countries and territories, 1990-2019: results from the Global Burden of Disease Study 2019. Bmj 378 (2022).

[75] Malay Sarkar, Irappa Madabhavi, Narasimhalu Niranjan, and Megha Dogra. 2015. Auscultation of the respiratory system. *Annals of thoracic medicine* 10, 3 (2015), 158–168.
[76] John T Sharp, Joseph Danon, Walter S Druz, Norma B Goldberg, Howard Fishman, and Wanda Machnach. 1974. Respiratory muscle function in patients with chronic obstructive pulmonary disease: its relationship to disability and to respiratory therapy. *American Review of Respiratory Disease* 110, 6P2 (1974), 154–167.
[77] Ian E Smith, Erik Jurriaans, Stefan Diederich, Nabeel Ali, John M Shneerson, and CD Flower. 1996. Chronic sputum production: correlations between clinical features and findings on high resolution computed tomographic scanning of the chest. *Thorax* 51, 9 (1996), 914–918.
[78] Maria Luísa Soares, Margarida Torres Redondo, and Miguel R Gonçalves. 2016. Implications of Manual Chest Physiotherapy and Technology in Preventing Respiratory Failure after Extubation. *Noninvasive Mechanical Ventilation and Difficult Weaning in Critical Care: Key Topics and Practical Approaches* (2016), 57–62.
[79] Xiao Sun, Li Qiu, Yibo Wu, Yeming Tang, and Guohong Cao. 2017. SleepMonitor: Monitoring Respiratory Rate and Body Position During Sleep Using Smartwatch. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 104 (sep 2017), 22 pages. https://doi.org/10.1145/3130969
[80] Zhiqiang Sun. 2023. ICBHI 2017 challenge. https://doi.org/10.7910/DVN/HT6PKI
[81] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems* 30 (2017).
[82] Teresa A Volsko. 2013. Airway clearance therapy: finding the evidence. *Respiratory care* 58, 10 (2013), 1669–1678.
[83] Lu Wang, Jiajia Wang, Guixiang Zhao, and Jiansheng Li. 2024. Prevalence of bronchiectasis in adults: a meta-analysis. *BMC Public Health* 24, 1 (2024), 2675.
[84] Tianben Wang, Zhishen Wang, Xiantao Liu, Wenbo Liu, Leye Wang, Yuanqing Zheng, Jin Hu, Tao Gu, and Daqing Zhang. 2023. OmniResMonitor: Omnimonitoring of Human Respiration using Acoustic Multipath Reflection. *IEEE Transactions on Mobile Computing* (2023), 1–14. https://doi.org/10.1109/TMC.2023.3281928
[85] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW Based Contactless Respiration Detection Using Acoustic Signal. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 170 (jan 2018), 20 pages. https://doi.org/10.1145/3161188
[86] Zijie Wang and Zhao Wang. 2022. A domain transfer based data augmentation method for automated respiratory classification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9017–9021.
[87] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.
[88] Wentao Xie, Runxin Tian, Jin Zhang, and Qian Zhang. 2021. Noncontact Respiration Detection Leveraging Music and Broadcast Signals. *IEEE Internet of Things Journal* 8, 4 (Feb. 2021), 2931–2942. https://doi.org/10.1109/JIOT.2020.3021915 Conference Name: IEEE Internet of Things Journal.
[89] Wentao Xie, Chi Xu, Yanbin Gong, Yu Wang, Yuxin Liu, Jin Zhang, Qian Zhang, Zeguang Zheng, and Shifang Yang. 2024. DeepBreath: Breathing Exercise Assessment with a Depth Camera. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–26.
[90] Xuhai Xu, Ebrahim Nemati, Korosh Vatanparvar, Viswam Nathan, Tousif Ahmed, Md Mahbubur Rahman, Daniel McCaffrey, Jilong Kuang, and Jun Alex Gao. 2021. Listen2cough: Leveraging end-to-end deep learning cough detection model to enhance lung health assessment using passively sensed audio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–22.
[91] Zhicheng Yang, Parth H. Pathak, Yunze Zeng, Xixi Liran, and Prasant Mohapatra. 2016. Monitoring Vital Signs Using Millimeter Wave. In *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing* (Paderborn, Germany) *(MobiHoc '16)*. Association for Computing Machinery, New York, NY, USA, 211–220. https://doi.org/10.1145/2942358.2942381