A Dual-Purpose Deep Learning Model for Auscultated Lung and Tracheal Sound Analysis Based on Mixed Set Training

Fu-Shun Hsu[a,b,c], Shang-Ran Huang[c], Chang-Fu Su[a,d,e], Chien-Wen Huang[f], Yuan-Ren Cheng[c,g,h], Chun-Chieh Chen[f], Chun-Yu Wu[i], Chung-Wei Chen[b], Yen-Chun Lai[c,j], Tang-Wei Cheng[c], Nian-Jhen Lin[c,k], Wan-Ling Tsai[c], Ching-Shiang Lu[c], Chuan Chen[c], and Feipei Lai[a,*]

[a] Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan

[b] Department of Critical Care Medicine, Far Eastern Memorial Hospital, New Taipei, Taiwan

[c] Heroic Faith Medical Science Co., Ltd., Taipei, Taiwan

[d] Department of Anesthesia, Division of Medical Quality, En-Chu-Kong Hospital, New Taipei, Taiwan

[e] Department of Electronic Engineering, Oriental Institute of Technology, New Taipei, Taiwan

[f] Avalanche Computing Inc., Taipei, Taiwan

[g] Department of Life Science, College of Life Science, National Taiwan University, Taipei, Taiwan

[h] Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan

[i] Department of Anesthesiology, National Taiwan University Hospital, Taipei, Taiwan

[j] Department of Anesthesiology, Taipei Medical University Hospital, Taipei, Taiwan

[k] Division of Pulmonary Medicine, Far Eastern Memorial Hospital, New Taipei, Taiwan

[*]**Corresponding Author**: Feipei Lai, Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan. (phone: 886-2-3366-4961; Fax: 886-2-3366-3754; e-mail: flai@csie.ntu.edu.tw)

Abstract

Computerized respiratory sound analysis can bring evolutionary change in gradually forgotten auscultation technique. An expert is no longer required to interpret the auscultated sound at the point of care, which makes an automated tele-auscultation service and continuous respiratory sound monitoring possible. In this study, we constructed a tracheal sound database, HF_Tracheal_V1, containing 10958 15-s tracheal sound recordings, 23087 inhalation labels, 16728 exhalation labels, and 6874 continuous adventitious sound (CAS) labels. Tracheal sounds in HF_Tracheal_V1 and lung sounds in our previously built lung sound database, HF_Lung_V2, were either combined or used alone to train convolutional neural network bidirectional gate recurrent unit models for inhalation, exhalation and CAS detection in lung and tracheal sounds. Different training strategies were investigated and compared: (1) training on a single database (either lung or tracheal), (2) training on a mixed set of tracheal and lung sounds, and (3) using domain adaptation to fine-tune pretrained lung sound models with the tracheal sound data and vice versa. The results revealed that the models trained on only one database performed poorly when tested on the other database. Mixed set training and domain adaptation improved the performance for 1) inhalation, exhalation and CAS detection in lung sounds and 2) CAS detection in tracheal sounds. In particular, the model trained on the mixed set had great flexibility to be used in both lung and tracheal sounds.

Keywords: Auscultation, convolutional neural network, deep learning, domain adaptation, gated-recurrent unit, respiratory sound

1. Introduction

Respiratory auscultation [1] with a stethoscope is a longstanding diagnostic technique used to examine the respiratory system. Respiratory sounds can be classified into subtypes such as mouth sounds, tracheal sounds, bronchial sounds, bronchovesicular sounds, and vesicular (lung) sounds, based on the location at which the sound is auscultated [2]. Lung and tracheal sounds are the most frequently auscultated in clinical applications.

Lung auscultation is commonly used as a first line physical examination tool to diagnose pulmonary disease because it is noninvasive and inexpensive [3]. Healthy lungs generate normal lung sounds during breathing; unhealthy lungs may manifest continuous adventitious sounds (CAS) such as wheezes, stridor, or rhonchi or manifest discontinuous adventitious sounds (DAS) such as crackles or pleural friction rubs [1, 2]. Healthcare professionals can recognize abnormal pulmonary conditions by the presence, type, characteristics, and location of adventitious lung sounds [1-3].

Tracheal auscultation can be used to detect pulmonary ventilation abnormalities such as abnormal respiratory rates, upper airway obstructions [4], and apnea. Respiratory rates can be estimated by identifying breath phases (inhalation and exhalation) in the tracheal sound [5, 6]. Partial upper airway obstruction is indicated by the presence of CAS-like patterns such as stridor [7, 8] or snoring [9]. Total upper airway obstruction is indicated by the pattern of a very short and high-intensity sound which represent a prematurely stopped inhalation resulted from high airway resistance (Fig. 1). Apnea can be inferred from the prolonged absence of inhalation and exhalation during tracheal auscultation [6, 8, 10-12]. Therefore, tracheal sound monitoring is recommended by some clinical guidelines for use if a patient's pulmonary ventilatory functions are at risk of being compromised, such as during a diagnostic or surgical procedure with the use of sedation [13, 14].

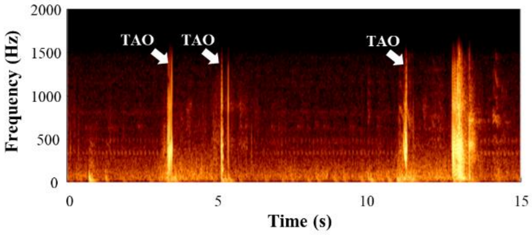

Fig 1. The patterns of total airway obstruction displayed on a spectrogram. TAO: total airway obstruction.

Although respiratory auscultation can be used in many clinical applications, it is gradually forgotten in daily practice because of the limitations of using a conventional stethoscope [15] and the advent of advanced diagnostic devices. Also, sudden loud sounds, such as a patient's unconscious mumbling, and drilling and spurting sounds in a dental procedure, may hurt the practitioner's ears during tracheal auscultation. Computerized respiratory sound analysis [16, 17] is key to overcoming the limitations of the old auscultation technique. Previous studies have comprehensively reviewed proposed methods [2, 18]. However, few research groups [19-21] have investigated breath phase and adventitious sound detection in lung sound at a recording level [2] using deep learning, and the research of tracheal sound analysis using deep learning [22] is also uncommon. In our previous studies, we established the lung sound databases HF_Lung_V1 [15] and HF_Lung_V2 [23]. Deep learning–based convolutional neural network (CNN)-bidirectional gated recurrent unit (BiGRU) models were proposed, and these networks were demonstrated to adequately detect inhalation, exhalation, CAS, and DAS events in lung sounds [15, 23]. However, we had not yet investigated computerized tracheal sound analysis. Thus, we aimed to create a tracheal sound database and train tracheal sound analysis models for the identification of breath phase and detection of CAS through deep learning. DAS detection was not included because crackles and pleural friction rubs were not labeled in the collected tracheal sounds.

Moreover, a large data set is key to training an accurate deep learning model [24, 25]. However, collecting and labeling data are inevitably laborious and expensive. To take full advantage of the established lung and tracheal sound databases, we aimed to investigate the feasibility and benefit of using both databases together in training lung and tracheal sound analysis models. Two feasible approaches were identified. First, the lung and tracheal sound files could be combined to form a mixed set to train a single model for both lung and tracheal sound analysis. Second, transfer learning [26], specifically domain adaptation [27], could be used to fine-tune a pretrained lung sound model for tracheal sound analysis (or vice versa) to further improve the model performance. In addition, we also investigated how well a single model can serve dual purposes, lung and tracheal sound analysis, at the same time.

2. Materials and Methods

2.1 Establishment of tracheal sound database

The protocol for the tracheal sound study was approved by the Joint Institutional Review Board of the Medical Research Ethical Foundation, Taipei, Taiwan (case number: 19-006-A-2). The protocol was further reviewed and approved by En Chu Kong Hospital (case number: ECKIRB1090303). This study was conducted in accordance with the 1964 Helsinki Declaration and its later amendments.

A total of 299 Taiwanese individuals aged ≥20 years who had undergone a diagnostic or surgical procedure with non-intubated intravenous sedation (procedural sedation) were enrolled in this study. We did not enroll individuals belonging to vulnerable groups (e.g., incarcerated individuals, indigenous people, persons with disabilities, or persons with mental illness), having a history of allergies that prevented contact with medical patches or artificial skin, or having a diagnosis of atrial fibrillation or arrhythmia. Tracheal sounds were collected between November 2019 and June 2020.

Two devices, HF-Type-2 and HF-Type-3 devices, were used to record tracheal sounds. HF-Type-2 (Fig. 2a) is an electronic stethoscope (AS-101, Heroic Faith Medical Science, Taipei, Taiwan) connected to a smartphone (Mi 9T Pro, Xiaomi, Beijing, China). HF-Type-3 (Fig. 2b) is constructed from a chestpiece (603P, Spirit Medical, New Taipei, Taiwan), a stethoscope tubing, a microphone (ECM-PC60, Sony, Tokyo, Japan), and a smartphone (Mi 9T pro, Xiaomi, Beijing, China). A customized app was installed in the smartphone to record the received tracheal sounds. Tracheal sounds from each participant were recorded at the flat area of the left or right thyroid cartilage as displayed in Fig. 3 by using one of the devices. Although HF-Type-2 supported multichannel recording, only one channel was used for tracheal sound recording. Tracheal sounds were collected at a sampling rate of 4000 Hz and 16-bit depth. Tracheal sounds were recorded while the participants were undergoing a procedure under non-intubated intravenous sedation. The recording began before the first administration of the anesthetic and ended when the procedure was completed. The recording time varied depending on the necessity of tracheal sound monitoring; most recordings ranged from a few minutes to less than 20 minutes. The majority of the recording was collected when the participants were under moderate sedation. However, some recording covers the periods of mild and deep sedation. The audio recordings were subsequently truncated to 15-s files using a sliding window with a step size of 15 s; thus, the truncated files did not overlap. Any tracheal sound file shorter than 15 s was deleted.

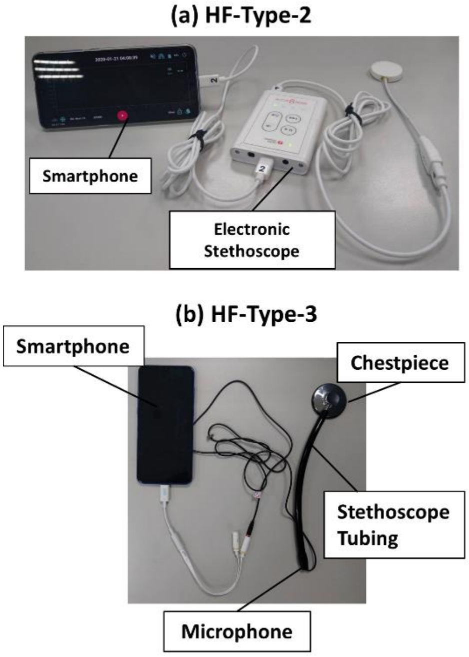

Fig. 2. (a) HF-Type-2 and (b) HF-Type-3 devices used for tracheal sound recording.

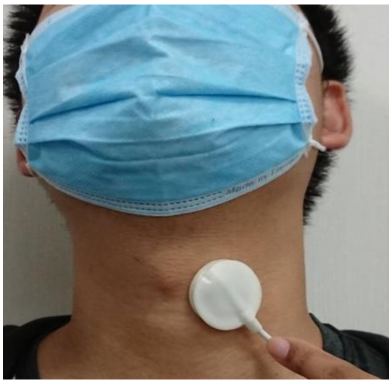

Fig. 3. Tracheal sound recording location.

Each of the 15-s audio files was subsequently labeled by one labeler and inspected by an inspector. The labeling was done either by a board-certified respiratory therapist (NJL) with 8 years of clinical experience or a board-certified nurse (WLT) with 13 years of clinical experience. The quality of the labeling was verified by another board-certified respiratory therapist (CC) with 6 years of clinical experience or another board-certified nurse (CSL) with 4 years of clinical experience. If the inspector and the labeler did not agree, the labels were further reviewed and amended until mutual agreement was reached. A self-developed labeling software was used for labeling [28]. The consensus on labeling criteria were maintained throughout the labeling process by holding regular meetings. Labelers were asked to label the start and end times of inhalation (I), exhalation (E), and CAS (C) events. Unlike labels in HF_Lung_V1 and HF_Lung_V2, we did not specifically differentiate tracheal sound CAS events as a wheeze, stridor, or rhonchus. However, CAS labels in this study included snoring.

The tracheal sound recordings and the corresponding labels were divided into a training set and a testing set at a ratio of approximately 4:1. Files from the same participant were all assigned to the same set (either training or testing). The training set and testing set formed the HF_Tracheal_V1 (Tracheal_V1) database.

2.2 Deep learning pipeline

The CNN-BiGRU model (Fig. 4) outperformed all the other benchmark models in the lung sound analysis of our previous study [15]. Therefore, the CNN-BiGRU model was used in this study.

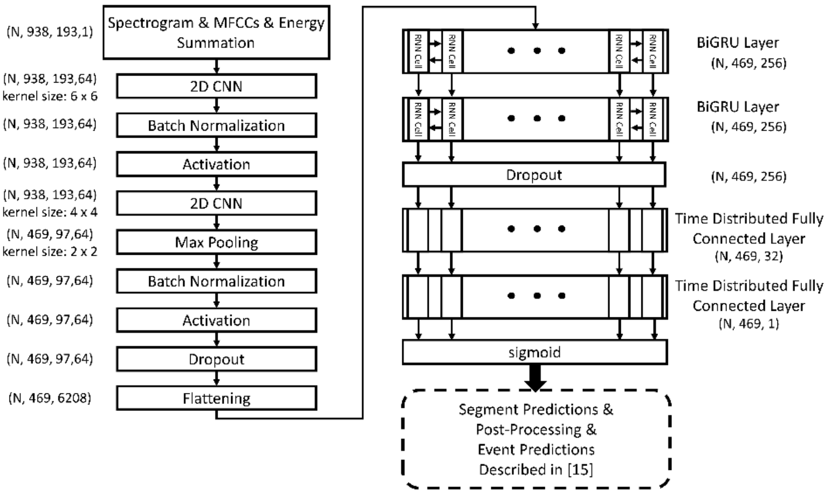

Fig. 4. Architecture of the CNN-BiGRU model.

The major tasks for the models were to do inhalation, exhalation and CAS event detection at the recording level, which was clearly defined in our previous studies [15, 23]. The deep learning pipeline is presented in Fig. 5, and it was the same as that in our previous studies [15, 23]. The 15-s signals were first filtered by a Butterworth high-pass filter with a cutoff frequency at 80 Hz. The spectrogram was then computed from the 15-s filtered signal using a short-time Fourier transform (STFT) [29] with a Hanning window of size 256, hop length of size 64, and no zero padding. A 938 × 129 matrix was generated; 938 was the number of time frames (segments) and 129 was the number of the frequency bins. The mel frequency cepstral coefficients (MFCCs) [19], including 20 static coefficients, 20 delta coefficients, and 20 acceleration coefficients were derived from every time segment of the spectrogram resulting in three 938 × 20 MFCC matrices. The energy in the four frequency bands of the spectrogram (0–250, 250–500, 500–1000, and 0–2000 Hz) were summed to produce four 938 × 1 energy summation vectors. The spectrogram, each of the three MFCC matrices, and each of the energy summation vectors were then normalized. The concatenation of the normalized spectrogram, MFCCs, and energy summation were fed into the CNN-BiGRU model. The output of the CNN-BiGRU model was a 469 × 1 probability vector. Thresholding was then applied to the probability vector to generate a binarized vector. Elements of the binary vector with a value of 1 indicated that sounds of inhalation, exhalation, or CAS were detected in the corresponding time segment. After the segment detection results were obtained, the vectors were sent for postprocessing to merge neighboring segments and remove burst events to, in turn, generate the event-detection results, as described in our previous studies [15, 23].

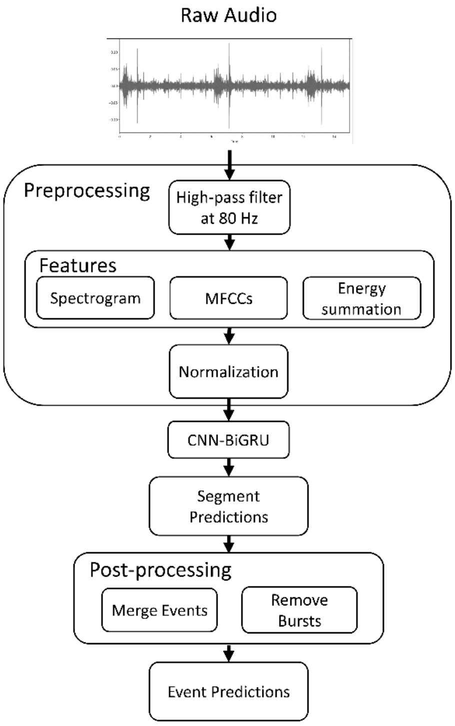

Fig. 5. Deep learning pipeline.

2.3 Investigation of mixed set training and domain adaptation

To investigate whether using the lung and tracheal sound data together can train a better model for lung and tracheal sound analysis, different training strategies were applied. First, full training [30] (training from scratch) was used. Lung sound models were trained using the training set of HF_Lung_V2 (Lung_V2_Train) alone, and tracheal sound models were trained using the training set of HF_Tracheal_V1 (Tracheal_V1_Train) alone. Second, the recordings in Lung_V2_Train and Tracheal_V1_Train were combined, and models were trained on this mixed set. Third, we used domain adaptation [27] to fine-tune the pretrained lung sound models for tracheal sound analysis, and we fine-tuned the pretrained tracheal sound models for lung sound analysis. All trained models were tested separately on both testing set of HF_Lung_V2 (Lung_V2_Test) and testing set of HF_Tracheal_V1 (Tracheal_V1_Test). The positive controls (PCs) were 1) models trained on Lung_V2_Train alone and tested on Lung_V2_Test and 2) models trained on Tracheal_V1_Train alone and tested on Tracheal_V1_Test. The negative controls (NCs) were 1) models trained on Lung_V2_Train alone and tested on Tracheal_V1_Test and 2) models trained on Tracheal_V1_Train alone and tested with Lung_V2_Test. Only recordings containing at least one I, E or C label were used to train the corresponding detection model and evaluate the model performance.

2.4 Training environment and parameters

Models were trained on a server (OS: Ubuntu 18.04; CPU: Intel Xeon Gold 6154@3.00 GHz; RAM: 90 GB) provided by the National Center for High-performance Computing in Taiwan. We used TensorFlow 2.10 as the framework for the deep neural networks. GPU acceleration was provided by a NVIDIA Titan V100 GPU on the CUDA 10 and CuDNN 7 frameworks.

To train the model, we set the batch size to 64 and the number of epochs to 5000. Note that we also set the early stop policy with patience of 50 during the training procedure. We used Adam optimizer with an initial learning rate of 0.0001 for training. In this study, 5-fold cross validation was applied.

For domain adaptation, we used the model trained on either HF_Lung_V2 or HF_Tracheal_V1 as the pretrained model without freezing the weights. Then, we fine-tune the pre-trained model for 50 epochs with the training parameters mentioned above.

2.5 Performance evaluation

The event-detection performance of the models at the recording level were evaluated [15, 23]. We first had the ground-truth event labels in the 15-s recordings (red horizontal bars in Fig. 6a) after the labelers did the labeling. The results of segment prediction (blue vertical bars in Fig. 6b) were obtained after the model finished the inference process. The number of segments in a 15-s recording was determined by the parameters of STFT, which was 938 in this study. After the segment prediction results postprocessing, the event prediction results were obtained (Fig. 6c). Subsequently, the Jaccard index (JI) [20] was used to determine whether the models correctly detected an event. First, we used the ground truth labels as a reference and examined whether each ground truth label had a matching predicted event (JI $\geq$ 0.5). If the label had such an event, it was counted as a TP event (orange horizontal bar in Fig. 6d); otherwise, it was counted as an FN event (yellow horizontal bars in Fig. 6d). Then, conversely, we used the event prediction results as a reference; we checked whether we could find a matching ground truth label for each predicted event (JI $\geq$ 0.5). If we could do so, the predicted event was designated as a TP event (orange horizontal bar in Fig. 6e); if not, it was counted as an FP event (black horizontal bars in Fig. 6e). TN events were not counted because the background phase was not considered an event. Note that TP events were counted twice (orange horizontal bars in Fig. 6d and Fig. 6e) during the evaluation process. Therefore, a pair of TPs was considered a single TP event in computing the evaluation metrics. However, all the FP and FN events were used to compute the evaluation metrics despite the possibility of inducing undesirable bias.

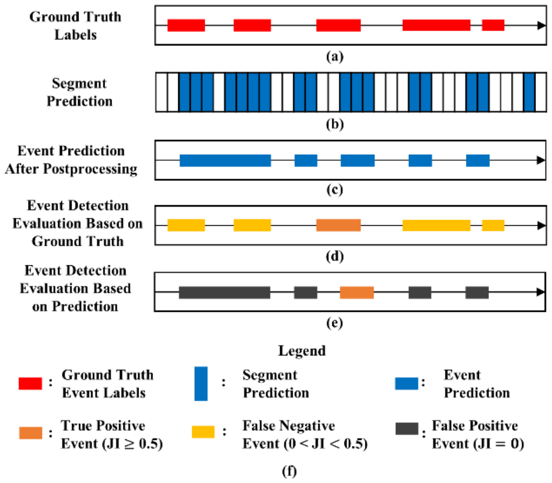
Ground Truth Labels
(a)
Segment Prediction
(b)
Event Prediction After Postprocessing
(c)
Event Detection Evaluation Based on Ground Truth
(d)
Event Detection Evaluation Based on Prediction
(e)
Legend
Ground Truth Event Labels
Segment Prediction
Event Prediction
True Positive Event (JI ≥ 0.5)
False Negative Event (0 < JI < 0.5)
False Positive Event (JI = 0)
(f)

Fig 6. Illustration of event detection evaluation. (a) Ground-truth event labels, (b) segment prediction, (c) event prediction after postprocessing, (d) event detection evaluation based on ground-truth event labels, (e) event detection evaluation based on event prediction, and (f) legend. JI: Jaccard index.

The threshold producing the best accuracy for segment prediction was used to generate event prediction results. We used F1 score [19] to evaluate the event-detection performance.

3. Results

3.1 Demographic data

Demographic data for participants whose tracheal sounds were acquired are summarized in Table 1. A total of 299 participants were enrolled in the study (137 men and 162 women). The average age was 45.7 ± 13.9 years. The average height and weight were 161.5 ± 8.2 cm and 63.9 ± 12.9 kg. The average body mass index was 24.3 ± 3.7 kg/m^2. The numbers of patients whose tracheal sound recorded by HF-Type-2 and HF-Type-3 were 176 and 123, respectively.

The information of participants enrolled to construct HF_Lung_V2 can be found in our previous study [23].

Table 1. Demographic characteristics of participants.

Items	Subjects (n=299)
Gender (M/F)	137/162
Age (year)	45.7 ± 13.9
Height (cm)	161.5 ± 8.2
Weight (kg)	63.9 ± 12.9
BMI (kg/m^2)	24.3 ± 3.7
Recording Devices	
HF-Type-2	176
HF-Type-3	123

Values in parenthesis represent the 95% confidence interval (CI).

3.2 Summary of HF_Lung_V2 and HF_Tracheal_V1 databases

The summary statistics of the HF_Lung_V2 and HF_Tracheal_V1 databases are listed in Table 2. A total of 14,138 15-s recordings, 49,379 I labels, 24,576 E labels, and 22,468 C labels were included in HF_Lung_V2. Tracheal_V1 contains 10,958 15-s recordings, 23,087 I labels, 16,728 E labels, and 6874 C labels. The mean duration of I labels significantly differed ($p < 0.001$) between HF_Lung_V2 and HF_Tracheal_V1 (0.95 ± 0.29 and 1.09 ± 0.39 s, respectively). The mean durations of E labels significantly differed ($p < 0.001$) between HF_Lung_V2 and HF_Tracheal_V1 (0.92 ± 0.49 and 0.99 ± 1.07 s, respectively). The mean durations of C labels significantly differed between HF_Lung_V2 and HF_Tracheal _V1 (0.82 ± 0.46 and 1.07 ± 0.58 s, respectively).

Table 2. Summary of HF_Lung_V2 and HF_Tracheal_V1 databases.

Recordings or Labels	Attributes	HF_Lung_V2	HF_Tracheal_V1
Recordings	No.	14138	10958
	Total duration (min)	3534.5	2739.5
I	No.	49379	23087
	Total duration (min)	784.06	420.56
	Mean duration (s)	0.95 ± 0.29	1.09 ± 0.39*
E	No.	24576	16728
	Total duration (min)	375.59	275.36
	Mean duration (s)	0.92 ± 0.49	0.99 ± 1.07*
C	No.	22468	6874
	Total duration (min)	307.27	122.15
	Mean duration (s)	0.82 ± 0.46	1.07 ± 0.58*

I: inhalation labels, E: exhalation labels, and C: continuous adventitious sound labels.

* $p < 0.001$ between HF_Lung_V2 and HF_Tracheal_V1.

3.3 Statistics of training and test data sets

The compositions of the training and testing data sets of both the HF_Lung_V2 and HF_Tracheal_V1 databases are summarized in Table 3. Lung_V2_Train and Tracheal_V1_Train had 10,742 and 8700 15-s files, respectively, and Lung_V2_Test and Tracheal_V1_Test had 3403 and 2258 15-s files, respectively. The statistics of the I, E, and C labels in the training and test data sets of HF_Lung_V2 and HF_Tracheal_V1 are presented in Table 3. The mean durations of I, E and

C labels between all pairs of Lung_V2_Train, Tracheal_V1_Train, Lung_V2_Test, and Tracheal_V1_Test significantly differed (*p* < 0.001 for all).

Table 3. Training and test data sets of HF_Lung_V2 and HF_Tracheal_V1.

Recordings/	Attributes	Training datasets		Test datasets		p-value
Labels		HF_Lung_V2	HF_Tracheal_V1	HF_Lung_V2	HF_Tracheal_V1	
Recordings	No.	10735	8700	3403	2258	
	Total duration (min)	2683.75	2175	850.75	564.5	
I	No.	39134	18539	10245	4548	
	Total duration (min)	623.32	343.03	160.75	77.53	
	Mean duration (s)	0.96 ± 0.30	1.11 ± 0.34	0.94 ± 0.25	1.02 ± 0.29	<0.001*
E	No.	18359	13556	6217	3172	
	Total duration (min)	294.23	226.47	81.36	48.89	
	Mean duration (s)	0.96 ± 0.52	1.00 ± 0.41	0.79 ± 0.37	0.92 ± 0.33	<0.001*
C	No.	18278	5955	4190	919	
	Total duration (min)	254.87	106.89	52.40	15.26	
	Mean duration (s)	0.84 ± 0.48	1.08 ± 0.57	0.75 ± 0.34	1.00 ± 0.61	<0.001*

I: inhalation labels, E: exhalation labels, and C: continuous adventitious sound labels.

* indicates that mean durations of I, E, and C labels between all pairs of Lung_V2_Train, Tracheal_V1_Train, Lung_V2_Test, and Tracheal_V1_Test significantly differed (*p* < 0.001).

3.4 Model performance

The performance results of the trained models are presented in Table 4. The NCs had the worst F1 scores compared with other models. The models trained on a mixed set had best performance for inhalation (F1 score of 0.859) and exhalation detection (F1 score of 0.783) in lung sound analysis and for CAS detection (F1 score of 0.845) in the tracheal sound analysis. However, the models trained from the scratch had best performance in the inhalation (F1 score of 0.831) and exhalation detection (F1 score of 0.854) in tracheal sound analysis. Furthermore, the model first trained on Lung_V2_Train and then finetuned with Tracheal_V1_Train had the best performance (F1 score of 0.445) for CAS detection in tracheal sound analysis.

Table 4. F1 scores of the trained models.

Controls	Training Database/Strategy	Testing Database	F1 Score
Inhalation			
PC	Lung_V2	Lung_V2	0.845
NC	Tracheal_V1	Lung_V2	0.750
	Lung_V2+Tracheal_V1	Lung_V2	**0.859**
	Tracheal_V1→Lung_V2	Lung_V2	0.858
NC	Lung_V2	Tracheal_V1	0.621
PC	Tracheal_V1	Tracheal_V1	**0.831**
	Lung_V2+Tracheal_V1	Tracheal_V1	0.824
	Lung_V2→Tracheal_V1	Tracheal_V1	0.818
Exhalation			
PC	Lung_V2	Lung_V2	0.762
NC	Tracheal_V1	Lung_V2	0.546
	Lung_V2+Tracheal_V1	Lung_V2	**0.783**
	Tracheal_V1→Lung_V2	Lung_V2	0.773
NC	Lung_V2	Tracheal_V1	0.530
PC	Tracheal_V1	Tracheal_V1	**0.854**
	Lung_V2+Tracheal_V1	Tracheal_V1	0.840
	Lung_V2→Tracheal_V1	Tracheal_V1	0.838
Continuous adventitious sound			
PC	Lung_V2	Lung_V2	0.430
NC	Tracheal_V1	Lung_V2	0.367
	Lung_V2+Tracheal_V1	Lung_V2	0.428
	Tracheal_V1→Lung_V2	Lung_V2	**0.445**
NC	Lung_V2	Tracheal_V1	0.541
PC	Tracheal_V1	Tracheal_V1	0.829
	Lung_V2+Tracheal_V1	Tracheal_V1	**0.845**
	Lung_V2→Tracheal_V1	Tracheal_V1	0.740

PC: positive control, and NC: negative control. Bold values indicate better performance. Lung_V2+Tracheal_V1 denotes mixed set training. Tracheal_V1→Lung_V2 denotes pretrained tracheal sound model that was fine-tuned with the lung sound data. Lung_V2→Tracheal_V1 denotes the pretrained lung sound model that was fine-tuned with the tracheal sound data.

Table 5 displays the weighted mean values of F1 scores for event detection derived by averaging the two scores in the Lung_V2_Test and Tracheal_V1_Test. The results clearly demonstrate that models trained with a mixed set had the best performance when they were used in lung and tracheal sound analysis at the same time.

Table 5. Weighted mean values of F1 scores for inhalation, exhalation, and CAS event detection derived by averaging of the two scores in Lung_V2_Test and Tracheal_V1_Test.

Training set/strategy	Label Type		
	I	E	C
Lung_V2	0.810	0.732	0.499
Tracheal_V1	0.775	0.650	0.464
Lung_V2+Tracheal_V1	**0.861**	**0.839**	**0.566**
Tracheal_V1→Lung_V2	0.827	0.779	0.524
Lung_V2→Tracheal_V1	0.812	0.750	0.545

I: inhalation labels, E: exhalation labels, and C: continuous adventitious labels. Bold values indicate the best performance among the five models. Lung_V2+Tracheal_V1 denotes mixed set training. Tracheal_V1→Lung_V2 denotes the pretrained tracheal sound model that was fine-tuned with the lung sound data. Lung_V2→Tracheal_V1 denotes the pretrained lung sound model that was fine-tuned with the tracheal sound data.

4. Discussion

Our results reveal that NCs performed the worst compared with other models. It is because lung and tracheal sounds have different frequency ranges, energy drops, inhalation–exhalation duration ratios, and pause periods [2]. Besides, the mean durations for the I, E, and C labels significantly differed between HF_Lung_V2 and HF_Tracheal_V1 (see Table 2). The majority of feature distribution differences resulted from innate differences in the physical and physiological mechanisms underlying the generation of lung and tracheal sounds [31]. Thus, future studies must consider lung and tracheal sound as two distinct domains when building computerized analysis models. However, it is undeniable that the different recording devices used to record the lung [15, 23] and tracheal sounds may also generate some feature differences. In addition, the setting of the patients, such as use of mechanical ventilation, depth of sedation, body position, etc., may contribute to the divergence of feature distribution.

As presented in Table 4, mixed set training and domain adaptation can improve the performance of inhalation, exhalation and CAS detection in lung sound analysis and CAS detection in the tracheal sound; however, it did not bring benefits to inhalation and exhalation in tracheal sound analysis. The extent of benefits may hinge on the composition of the lung and tracheal sound data; using different lung and tracheal sound data sets should produce slightly different results. Nevertheless, as clearly indicated in Table 5, the model that was trained on a mixed set was capable of processing a testing set comprising mixed lung and tracheal sound data. Mixed set training is an attractive option for developing an all-purpose respiratory monitor; users are not required to pick a specific channel or switch to a specific algorithm for lung or tracheal sound analysis.

The model performance for CAS detection was significantly better for tracheal sounds than for lung sounds (Table 4). This result may be because CAS in tracheal sounds is louder, increasing the signal-to-noise ratio and facilitating CAS pattern identification. Additionally, ground truth labels in Tracheal_V1 were checked by four experts, which reduced the number of noisy labels; the C labels in Lung_V2 could be noisy and are currently undergoing reworking [15, 23]. Furthermore, CAS in Tracheal_V1 is thought to be a primarily monophonic event occurring in the inspiratory phase, characterized by extrathoracic upper airway obstructions [4] induced by anesthetic drugs. Thus, the features are not as diverse as those in lung sounds; CAS in lung sounds can be categorized as inspiratory, expiratory, and biphasic types and as monophonic and polyphonic events [2].

In contrast to the labeling in HF_Lung_V1 and HF_Lung_V2, DAS was not specifically labeled in HF_Tracheal_V1 because most diseases generating DAS (fine crackles, coarse crackles, and pleural friction rubs) do not occur in the upper airway close to the pretracheal region. However, DAS-like patterns were occasionally observed in the collected tracheal sounds. These patterns might be caused by air flowing through accumulations of fluids such as water, saliva, sputum, or blood in the upper airway. Fluid accumulation in the upper airway must be promptly managed by clinicians, such as those executing a dental procedure on a moderately or deeply sedated patient who is not able to voluntarily cough to expel fluids in the laryngeal region [32]; in such a case, the dental team must perform suction to prevent aspiration. Hence, a respiratory monitor capable of detecting fluid accumulation in the upper airway is of clinical importance. The labeling of DAS-like patterns in tracheal sounds worth consideration in the future research.

In clinical practice, capnography is more commonly used for pulmonary ventilation monitoring than tracheal sound auscultation. Moreover, a pulse oximeter is now required for blood oxygen monitoring in surgical procedures or in a diagnostic procedure involving anesthesia. However, both of these devices have some limitations. Capnographic accuracy can be compromised by the poor sampling of carbon dioxide due to open-mouth breathing [33, 34]; by the use of a face mask or nasal cannula [35-37]; or by procedures that cause airflow interference, such as esophagogastroduodenoscopy or bronchoscopy. Moreover, capnography is difficult to use in surgeries involving the facial or oral regions. Oxygen desaturation as measured by an pulse oximeter is a delayed response to abnormal pulmonary ventilation [38, 39]. Therefore, a tracheal sound monitor that automatically detects abnormal respiratory rates, upper airway obstructions, and apnea could have substantial clinical value and could complement capnography and oximetry [5, 8]. Therefore, more accurate tracheal sound analysis models should be developed.

5. Conclusion

The automated analysis of lung and tracheal sounds is clinically valuable. Lung sound and tracheal sound may have different acoustic features. Hence, the automated inhalation, exhalation, and CAS detection model trained on lung sounds performed poorly for tracheal sound analysis, and vice versa. However, mixed set training and domain adaptation can improve the performance of models 1) for inhalation, exhalation, and CAS detection in lung sound analysis and 2) for CAS detection in tracheal sound analysis relative to the PCs (lung models trained only by a lung sound and vice versa). In particular, a model derived from mixed set training has great flexibility allowing a user not to select a specific model for lung or tracheal sound analysis, which facilitates the setup of respiratory monitoring in a busy operating room, ward, or clinic.

Conflict of interest

F.-S. Hsu, S.-R. Huang, Y.-R. Cheng, Y.-C. Lai, T.-W. Cheng, C.-S. Lu, and C. Chen are full-time employees and N.-J. Lin, and W.-L. Tsai are part-time employees of Heroic Faith Medical Science Co., Ltd. C.-W. Hung and C.-C. Chen are with Avalanche Computing Inc., whom Heroic Faith Medical Science Co. Ltd. commissioned to train the deep learning models.

Acknowledgements

The sound collection was sponsored by Raising Children Medical Foundation, Taiwan. The authors thank the employees of Heroic Faith Medical Science for partially contributing to this study. This manuscript was edited by Wallace Academic Editing. The author would like to thank the National Center for High-performance Computing for providing computing resources.

References

[1] Bohadana A, Izbicki G, Kraman SS. Fundamentals of lung auscultation. New England Journal of Medicine. 2014;370:744-51.
[2] Pramono RXA, Bowyer S, Rodriguez-Villegas E. Automatic adventitious respiratory sound analysis: A systematic review. PloS one. 2017;12:e0177926.
[3] Sarkar M, Madabhavi I, Niranjan N, Dogra M. Auscultation of the respiratory system. Annals of thoracic medicine. 2015;10:158.
[4] Acres JC, Kryger MH. Upper airway obstruction. Chest. 1981;80:207-11.
[5] Ouchi K, Fujiwara S, Sugiyama K. Acoustic method respiratory rate monitoring is useful in patients under intravenous anesthesia. Journal of clinical monitoring and computing. 2017;31:59-65.
[6] Ramsay MA, Usman M, Lagow E, Mendoza M, Untalan E, De Vol E. The accuracy, precision and reliability of measuring ventilatory rate and detecting ventilatory pause by rainbow acoustic monitoring and capnometry. Anesthesia & Analgesia. 2013;117:69-75.
[7] Gaffey M. Upper Airway Obstruction. 2020.

[8] Boriosi JP, Zhao Q, Preston A, Hollman GA. The utility of the pretracheal stethoscope in detecting ventilatory abnormalities during propofol sedation in children. Pediatric Anesthesia. 2019;29:604-10.

[9] Yadollahi A, Giannouli E, Moussavi Z. Sleep apnea monitoring and diagnosis based on pulse oximetery and tracheal sound signals. Medical & biological engineering & computing. 2010;48:1087-97.

[10] Yu L, Ting C-K, Hill BE, Orr JA, Brewer LM, Johnson KB, et al. Using the entropy of tracheal sounds to detect apnea during sedation in healthy nonobese volunteers. Anesthesiology. 2013;118:1341-9.

[11] Liu J, Ai C, Zhang B, Wang Y, Brewer LM, Ting C-K, et al. Tracheal sounds accurately detect apnea in patients recovering from anesthesia. Journal of clinical monitoring and computing. 2019;33:437-44.

[12] Lu X, Azevedo Coste C, Nierat M-C, Renaux S, Similowski T, Guiraud D. Respiratory Monitoring Based on Tracheal Sounds: Continuous Time-Frequency Processing of the Phonospirogram Combined with Phonocardiogram-Derived Respiration. Sensors. 2021;21:99.

[13] Association AD. Guidelines for the use of sedation and general anesthesia by dentists. Adopted by the ADA House of Delegates. 2016.

[14] Lives SSS. WHO guidelines for safe surgery 2009. Geneva: World Health Organization. 2009.

[15] Hsu F-S, Huang S-R, Huang C-W, Huang C-J, Cheng Y-R, Chen C-C, et al. Benchmarking of eight recurrent neural network variants for breath phase and adventitious sound detection on a self-developed open-access lung sound database-HF_Lung_V1. PLoS One. 2021;16:e0254134.

[16] Earis J, Cheetham B. Current methods used for computerized respiratory sound analysis. European Respiratory Review. 2000;10:586-90.

[17] Gurung A, Scrafford CG, Tielsch JM, Levine OS, Checkley W. Computerized lung sound analysis as diagnostic aid for the detection of abnormal lung sounds: a systematic review and meta-analysis. Respiratory medicine. 2011;105:1396-403.

[18] Muthusamy PD, Sundaraj K, Abd Manap N. Computerized acoustical techniques for respiratory flow-sound analysis: a systematic review. Artificial Intelligence Review. 2020;53:3501-74.

[19] Messner E, Fediuk M, Swatek P, Scheidl S, Smolle-Juttner F-M, Olschewski H, et al. Crackle and breathing phase detection in lung sounds with deep bidirectional gated recurrent neural networks. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC): IEEE; 2018. p. 356-9.

[20] Jácome C, Ravn J, Holsbø E, Aviles-Solis JC, Melbye H, Ailo Bongo L. Convolutional neural network for breathing phase detection in lung sounds. Sensors. 2019;19:1798.

[21] Hsiao C-H, Lin T-W, Lin C-W, Hsu F-S, Lin FY-S, Chen C-W, et al. Breathing Sound Segmentation and Detection Using Transfer Learning Techniques on an Attention-Based Encoder-Decoder Architecture. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC): IEEE; 2020. p. 754-9.

[22] Nakano H, Furukawa T, Tanigawa T. Tracheal sound analysis using a deep neural network to detect sleep apnea. Journal of Clinical Sleep Medicine. 2019;15:1125-33.

[23] Hsu F-S, Huang S-R, Huang C-W, Cheng Y-R, Chen C-C, Hsiao J, et al. An Update of a Progressively Expanded Database for Automated Lung Sound Analysis. arXiv preprint arXiv:210204062. 2021.

[24] Hestness J, Narang S, Ardalani N, Diamos G, Jun H, Kianinejad H, et al. Deep learning scaling is predictable, empirically. arXiv preprint arXiv:171200409. 2017.

[25] Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. Proceedings of the IEEE international conference on computer vision2017. p. 843-52.

[26] Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. Journal of Big data. 2016;3:1-40.

[27] Xu W, He J, Shu Y. Transfer Learning and Deep Domain Adaptation. Advances in Deep Learning: IntechOpen; 2020.

[28] Hsu F-S, Huang C-J, Kuo C-Y, Huang S-R, Cheng Y-R, Wang J-H, et al. Development of a Respiratory Sound Labeling Software for Training a Deep Learning-Based Respiratory Sound Analysis Model. arXiv preprint arXiv:210101352. 2021.

[29] Cohen L. Time-frequency analysis: Prentice Hall PTR Englewood Cliffs, NJ; 1995.

[30] He K, Girshick R, Dollár P. Rethinking imagenet pre-training. Proceedings of the IEEE/CVF International Conference on Computer Vision2019. p. 4918-27.

[31] Goettel N, Herrmann MJ. Breath Sounds: From Basic Science to Clinical Practice. Anesthesia & Analgesia. 2019;128:e42.

[32] Hanamoto H, Sugimura M, Morimoto Y, Kudo C, Boku A, Niwa H. Cough reflex under intravenous sedation during dental implant surgery is more frequent during procedures in the maxillary anterior region. Journal of Oral and Maxillofacial Surgery. 2013;71:e158-e63.

[33] Maddox RR, Williams CK, Oglesby H, Butler B, Colclasure B. Clinical experience with patient-controlled analgesia using continuous respiratory monitoring and a smart infusion system. American journal of health-system pharmacy. 2006;63:157-64.

[34] Friesen RH, Alswang M. End-tidal PCO 2 monitoring via nasal cannulae in pediatric patients: accuracy and sources of error. Journal of clinical monitoring. 1996;12:155-9.

[35] Hardman J, Curran J, Mahajan R. End-tidal carbon dioxide measurement and breathing system filters. Anaesthesia. 1997;52:646-8.

[36] Patino M, Redford DT, Quigley TW, Mahmoud M, Kurth CD, Szmuk P. Accuracy of acoustic respiration rate monitoring in pediatric patients. Pediatric Anesthesia. 2013;23:1166-73.

[37] Ahmed I, Aziz E, Newton N. Connection of capnography sampling tube to an intravenous cannula. Anaesthesia. 2005;60:824-5.

[38] Cacho G, Pérez-Calle J, Barbado A, Lledó J, Ojea R, Fernández-Rodríguez C. Capnography is superior to pulse oximetry for the detection of respiratory depression during colonoscopy. Revista

espanola de enfermedades digestivas. 2010;102:86.
[39] Lam T, Nagappa M, Wong J, Singh M, Wong D, Chung F. Continuous pulse oximetry and capnography monitoring for postoperative respiratory depression and adverse events: a systematic review and meta-analysis. Anesthesia & Analgesia. 2017;125:2019-29.