PLOS ONE

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

--Manuscript Draft--

Manuscript Number:	PONE-D-21-10240
Article Type:	Research Article
Full Title:	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
Short Title:	Automated lung sound analysis database
Corresponding Author:	Shang-Ran Huang Heroic Faith Medical Science Co Ltd Taipei, TAIWAN
Keywords:	adventitious sound; auscultation; convolutional neural networks; lung sound; recurrent neural networks; respiratory monitor
Abstract:	A reliable, remote, and continuous real-time respiratory sound monitor with automated respiratory sound analysis ability is urgently required in many clinical scenarios—such as in monitoring disease progression of coronavirus disease 2019—to replace conventional auscultation with a handheld stethoscope. However, a robust computerized respiratory sound analysis algorithm has not yet been validated in practical applications. In this study, we developed a lung sound database (HF_Lung_V1) comprising 9,765 audio files of lung sounds (duration of 15 s each), 34,095 inhalation labels, 18,349 exhalation labels, 13,883 continuous adventitious sound (CAS) labels (comprising 8,457 wheeze labels, 686 stridor labels, and 4,740 rhonchi labels), and 15,606 discontinuous adventitious sound labels (all crackles). We conducted benchmark tests for long short-term memory (LSTM), gated recurrent unit (GRU), bidirectional LSTM (BiLSTM), bidirectional GRU (BiGRU), convolutional neural network (CNN)-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU models for breath phase detection and adventitious sound detection. We also conducted a performance comparison between the LSTM-based and GRU-based models, between unidirectional and bidirectional models, and between models with and without a CNN. The results revealed that these models exhibited adequate performance in lung sound analysis. The GRU-based models outperformed, in terms of F1 scores and areas under the receiver operating characteristic curves, the LSTM-based models in most of the defined tasks. Furthermore, all bidirectional models outperformed their unidirectional counterparts. Finally, the addition of a CNN improved the accuracy of lung sound analysis, especially in the CAS detection tasks.
Order of Authors:	Fu-Shun Hsu
	Shang-Ran Huang
	Chien-Wen Huang
	Chao-Jung Huang
	Yuan-Ren Cheng
	Chun-Chieh Chen
	Jack Hsiao
	Chung-Wei Chen
	Li-Chin Chen
	Yen-Chun Lai
	Bi-Fang Hsu
	Nian-Jhen Lin

	Wan-Ling Tsai
	Yi-Lin Wu
	Tzu-Ling Tseng
	Ching-Ting Tseng
	Yi-Tsun Chen
	Feipei Lai
Additional Information:	
Question	**Response**
Financial Disclosure Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS ONE* for specific examples. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate. **Unfunded studies** Enter: *The author(s) received no specific funding for this work.* **Funded studies** Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? • **NO** - Include this sentence at the end of your statement: *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.* • **YES** - Specify the role(s) played. * typeset	This study was partially funded by the Raising Children Medical Foundation, Taiwan (http://http://www.raising.org.tw). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Competing Interests Use the instructions below to enter a	The authors have declared that no competing interests exist.

competing interest statement for this submission. On behalf of all authors, disclose any competing interests that could be perceived to bias this work—acknowledging all financial support and any other relevant financial or non-financial competing interests. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate and that any funding sources listed in your Funding Information later in the submission form are also declared in your Financial Disclosure statement. View published research articles from *PLOS ONE* for specific examples. **NO authors have competing interests** Enter: *The authors have declared that no competing interests exist.* **Authors with competing interests** Enter competing interest details beginning with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests: [insert competing interests here]* * typeset	
Ethics Statement Enter an ethics statement for this submission. This statement is required if the study involved: • Human participants • Human specimens or tissue • Vertebrate animals or cephalopods • Vertebrate embryos or tissues • Field research	The recordings were approved by the Research Ethics Review Committee of Far Eastern Memorial Hospital (case number: 107052-F). Written informed consent was obtained from the 18 patients. This study was conducted in accordance with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Write "N/A" if the submission does not require an ethics statement.

General guidance is provided below. Consult the submission guidelines for detailed instructions. **Make sure that all information entered here is included in the Methods section of the manuscript.**

Format for specific study types

Human Subject Research (involving human participants and/or tissue)
- Give the name of the institutional review board or ethics committee that approved the study
- Include the approval number and/or a statement indicating approval of this research
- Indicate the form of consent obtained (written/oral) or the reason that consent was not obtained (e.g. the data were analyzed anonymously)

Animal Research (involving vertebrate animals, embryos or tissues)
- Provide the name of the Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board that reviewed the study protocol, and indicate whether they approved this research or granted a formal waiver of ethical approval
- Include an approval number if one was obtained
- If the study involved *non-human primates*, add *additional details* about animal welfare and steps taken to ameliorate suffering
- If anesthesia, euthanasia, or any kind of animal sacrifice is part of the study, include briefly which substances and/or methods were applied

Field Research

Include the following details if this study involves the collection of plant, animal, or other materials from a natural setting:
- Field permit number
- Name of the institution or relevant body that granted permission

Data Availability Authors are required to make all data underlying the findings described fully available, without restriction, and from the time of publication. PLOS allows rare exceptions to address legal and ethical concerns. See the PLOS Data Policy and FAQ for detailed information. A Data Availability Statement describing where the data can be found is required at submission. Your answers to this question constitute the Data Availability Statement and **will be published in the article**, if accepted. **Important:** Stating 'data available on request from the author' is not sufficient. If your data are only available upon request, select 'No' for the first question and explain your exceptional situation in the text box. Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	Yes - all data are fully available without restriction
Describe where the data may be found in full sentences. If you are copying our sample text, replace any instances of XXX with the appropriate details. • If the data are **held or will be held in a public repository**, include URLs, accession numbers or DOIs. If this information will only be available after acceptance, indicate this by ticking the box below. For example: *All XXX files are available from the XXX database (accession number(s) XXX, XXX.*). • If the data are all contained **within the manuscript and/or Supporting Information files**, enter the following: *All relevant data are within the manuscript and its Supporting Information files.* • If neither of these applies but you are able to provide **details of access elsewhere**, with or without limitations, please do so. For example:	All relevant data are within the manuscript and its Supporting Information files.

Data cannot be shared publicly because of [XXX]. Data are available from the XXX Institutional Data Access / Ethics Committee (contact via XXX) for researchers who meet the criteria for access to confidential data. *The data underlying the results presented in the study are available from (include the name of the third party and contact information or URL).* • This text is appropriate if the data are owned by a third party and authors do not have permission to share the data. * typeset	
Additional data availability information:	

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

Fu-Shun Hsu[1,2,3], Shang-Ran Huang[3], Chien-Wen Huang[4], Chao-Jung Huang[3], Yuan-Ren Cheng[3,5,6], Chun-Chieh Chen[4], Jack Hsiao[7], Chung-Wei Chen[2], Li-Chin Chen[8], Yen-Chun Lai[3], Bi-Fang Hsu[3], Nian-Jhen Lin[3,9], Wan-Ling Tsai[3], Yi-Lin Wu[3], Tzu-Ling Tseng[3], Ching-Ting Tseng[3], Yi-Tsun Chen[3], Feipei Lai[1,*]

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

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

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

[4] Avalanche Computing Inc., Taipei, Taiwan

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

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

[7] HCC Healthcare Group, New Taipei, Taiwan

[8] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

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

Short Title: Automated lung sound analysis database

*Corresponding Author

E-mail: flai@csie.ntu.edu.tw

ABSTRACT

A reliable, remote, and continuous real-time respiratory sound monitor with automated respiratory sound analysis ability is urgently required in many clinical scenarios—such as in monitoring disease progression of coronavirus disease 2019—to replace conventional auscultation with a handheld stethoscope. However, a robust computerized respiratory sound analysis algorithm has not yet been validated in practical applications. In this study, we developed a lung sound database (HF_Lung_V1) comprising 9,765 audio files of lung sounds (duration of 15 s each), 34,095 inhalation labels, 18,349 exhalation labels, 13,883 continuous adventitious sound (CAS) labels (comprising 8,457 wheeze labels, 686 stridor labels, and 4,740 rhonchi labels), and 15,606 discontinuous adventitious sound labels (all crackles). We conducted benchmark tests for long short-term memory (LSTM), gated recurrent unit (GRU), bidirectional LSTM (BiLSTM), bidirectional GRU (BiGRU), convolutional neural network (CNN)-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU models for breath phase detection and adventitious sound detection. We also conducted a performance comparison between the LSTM-based and GRU-based models, between unidirectional and bidirectional models, and between models with and without a CNN. The results revealed that these models exhibited adequate performance in lung sound analysis. The GRU-based models outperformed, in terms of *F1* scores and areas under the receiver operating characteristic curves, the LSTM-based models in most of the defined tasks. Furthermore, all bidirectional models outperformed their unidirectional counterparts. Finally, the addition of a CNN improved the accuracy of lung sound analysis, especially in the CAS

detection tasks.

Keywords: adventitious sound, auscultation, convolutional neural networks, lung sound, recurrent neural networks, respiratory monitor

1. Introduction

Respiration is vital for the normal functioning of the human body. Therefore, clinical physicians are frequently required to examine respiratory conditions. Respiratory auscultation [1-3] using a stethoscope has long been a crucial first-line physical examination. The chestpiece of a stethoscope is usually placed on a patient's chest or back for lung sound auscultation or over the patient's tracheal region for tracheal sound auscultation. During auscultation, breath cycles can be inferred, which help clinical physicians evaluate the patient's respiratory rate. In addition, pulmonary pathologies are suspected when the frequency or intensity of respiratory sounds changes or when adventitious sounds, including continuous adventitious sounds (CASs) and discontinuous adventitious sounds (DASs), are identified [1, 2, 4]. Patients with coronavirus disease 2019 exhibit adventitious sounds [5]; hence, auscultation may be a useful approach for disease diagnosis [6] and disease progression tracking. However, auscultation performed using a conventional handheld stethoscope involves some limitations [7]. First, the interpretation of auscultation results substantially depends on the subjectivity of the practitioners. Even experienced clinicians might not have high consensus rates in their interpretations of auscultatory manifestations [8, 9]. Second, auscultation is a qualitative analysis method. Comparing auscultation results between individuals and quantifying the sound change by reviewing historical records are difficult tasks. Third, prolonged continuous monitoring of respiratory sound is almost impractical.

To overcome the aforementioned limitations, computerized methods for respiratory sound recording and analyses based on traditional signal processing and machine learning have been proposed and reviewed [4, 10-13]. With the advent of the deep learning era, studies have developed novel deep learning–based methods for respiratory sound analysis. However, many of such studies have focused on only distinguishing healthy participants from participants with respiratory disorders [14-18] and distinguishing various types of normal breathing sounds from adventitious sounds [19-25]. Only a few studies [26-29] have explored the use of deep learning for detecting breath phases and adventitious sounds. Moreover, most previous studies on computerized lung sound analysis have been limited by insufficient data. As of writing this paper, the largest reported respiratory sound database is ICBHI 2017 Challenge [30], which comprises 6,898 breath cycles and 10,775 events of wheezes and crackles acquired from 126 individuals.

Data size plays a major role in the creation of a robust and accurate deep learning–based respiratory sound analysis algorithm [31, 32]. Accordingly, the first aim of the present study was to establish a large and open-access respiratory sound database for training such algorithms for the detection of breath phase and adventitious sounds, mainly focusing on lung sounds. The second aim was to conduct a benchmark test on the established lung sound database by using eight recurrent neural network (RNN)-based models. RNNs [33] are effective for time-series analysis; long short-term memory (LSTM) [34] and gated recurrent unit (GRU) [35] networks, which are two RNN variants, exhibit superior performance to the original RNN model. However, whether LSTM models are superior to

GRU models (and vice versa) in many applications, particularly in respiratory sound analysis, is inconclusive. Bidirectional RNN models [36, 37] can transfer not only past information to the future but also future information to the past; these models consistently exhibit superior performance to unidirectional RNN models in many applications [38-40] as well as in breath phase and crackle detection [29]. However, whether bidirectional RNN models outperform unidirectional RNN models in CAS detection has yet to be determined. Furthermore, the convolutional neural network (CNN)–RNN structure has been proven to be suitable for heart sound analysis [41], lung sound analysis [19], and other tasks [39, 42]. Nevertheless, the application of the CNN–RNN structure in respiratory sound detection has yet to be fully investigated. Benchmarking can enable demonstrating the reliability and goodness of a database; it can also be applied to investigate the performance of the RNN variants in respiratory analysis.

In summary, the aims of this study are outlined as follows:

- Establish the largest open-access lung sound database as of writing this paper—HF_Lung_V1 (https://gitlab.com/techsupportHF/HF_Lung_V1).
- Conduct a performance comparison between LSTM and GRU models, between unidirectional and bidirectional models, and between models with and without a CNN in breath phase and adventitious sound detection based on lung sound data.
- Discuss factors influencing model performance.

2 Establishment of the lung sound database

2.1 Data sources and patients

The lung sound database was established using two sources. The first source was a database used in a datathon in Taiwan Smart Emergency and Critical Care (TSECC), 2020, under the license of Creative Commons Attribution 4.0 (CC BY 4.0), provided by the Taiwan Society of Emergency and Critical Care Medicine. Lung sound recordings in the TSECC database were acquired from 261 patients.

The second source was sound recordings acquired from 18 residents of a respiratory care ward (RCW) or a respiratory care center (RCC) in Northern Taiwan between August 2018 and October 2019. The recordings were approved by the Research Ethics Review Committee of Far Eastern Memorial Hospital (case number: 107052-F). Written informed consent was obtained from the 18 patients. This study was conducted in accordance with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

All patients were Taiwanese and aged older than 20 years. Descriptive statistics regarding the patients' demographic data, major diagnosis, and comorbidities are presented in Table 1; however, information on the patients in the TSECC database is missing. Moreover, all 18 RCW/RCC residents were under mechanical ventilation.

Table 1. Demographic data of patients.

	Subjects from RCW/RCC	Subjects in TSECC Database
Number (n)	18	261
Gender (M/F)	11/7	NA
Age	67.5 (36.7, 98.3)	NA
Height (cm)	163.6 (147.2, 180.0)	NA
Weight (kg)	62.1 (38.2, 86.1)	NA
BMI (kg/m^2)	23.1 (15.6, 30.7)	NA
Respiratory Diseases		
ARF	4 (22.2%)	NA
CRF	8 (44.4%)	NA
COPD AE	1 (5.6%)	NA
COPD	2 (11.1%)	NA
Pneumonia	4 (22.2%)	NA
ARDS	1 (5.6%)	NA
Emphysema	1 (5.6%)	NA
Comorbidity		
CKD	1 (5.6%)	NA
AKI	3 (16.7%)	NA
CHF	2 (11.1%)	NA
DM	7 (38.9%)	NA
HTN	6 (33.3%)	NA
Malignancy	1 (5.6%)	NA
Arrythmia	1 (5.6%)	NA
CAD	1 (5.6%)	NA

RCW: respiratory care ward, RCC: respiratory care center, ARF: acute respiratory failure, CRF: chronic respiratory failure, COPD AE: chronic obstructive pulmonary disease acute exacerbation, COPD: chronic obstructive pulmonary disease, ARDS: acute respiratory distress syndrome, CKD: chronic kidney disease, AKI: acute kidney injury, CHF: chronic heart failure, DM: diabetes, HTN: hypertension, CAD: cardiovascular disease. The mean values of the age, height, weight, and BMI are presented, with the corresponding 95% CI in parentheses.

2.2 Sound recording

Breathing lung sounds were recorded using two devices: (1) a commercial electronic stethoscope (Littmann 3200, 3M, Saint Paul, Minnesota, USA) and (2) a customized multichannel acoustic recording device (HF-Type-1) that supports the connection of eight electret microphones. The signals collected by the HF-Type-1 device were transmitted to a tablet (Surface Pro 6, Microsoft, Redmond, Washington, USA; Fig 1). Breathing lung sounds were collected at the eight locations (denoted by L1–L8) indicated in Fig 2a. The auscultation locations are described in detail in the caption of Fig 2. The two devices had a sampling rate of 4,000 Hz and a bit depth of 16 bits. The audio files were recorded in the WAVE (.wav) format.

Fig 1. Customized multichannel acoustic recording device (HF-Type-1) connected to a tablet.

Fig 2. Auscultation locations and lung sound recording protocol. (a) Auscultation locations (L1–L8): L1: second intercostal space (ICS) on the right midclavicular line (MCL); L2: fifth ICS on the right MCL; L3: fourth ICS on the right midaxillary line (MAL); L4: tenth ICS on the right MAL; L5: second ICS on the left MCL; L6: fifth ICS on the left MCL; L7: fourth ICS on the left MAL; and L8: tenth ICS on the left MAL. (b) A standard round of breathing lung sound recording with Littmann 3200 and HF-Type-1 devices. The white arrows represent a continuous recording, and the small red blocks represent 15-s recordings. When the Littmann 3200 device was used, 15.8-s signals were recorded sequentially from L1 to L8. Subsequently, all recordings were truncated to 15 s. When the HF-Type-1 device was used, sounds at L1, L2, L4, L5, L6, and L8 were recorded simultaneously. Subsequently, each 2-min signal was truncated to generate new 15-s audio files.

All lung sounds in the TSECC database were collected using the Littmann 3200 device only, where 15.8-s recordings were obtained sequentially from L1 to L8 (Fig 2b; Littmann 3200). One round of recording with the Littmann 3200 device entails a recording of lung sounds from L1 to L8. The TSECC database was composed of data obtained from one to three rounds of recording with the Littmann 3200 device for each patient.

We recorded the lung sounds of the 18 RCW/RCC residents by using both the Littmann 3200 device and the HF-Type-1 device. The Littmann 3200 recording protocol was in accordance with that used in the TSECC database, except that data from four to five rounds of lung sound recording were collected instead. The HF-Type-1 device was used to record breath sounds at L1, L2, L4, L5, L6, and L8. One round of recording with the HF-Type-1 device entails a synchronous and continuous recording of breath sounds for 30 min (Fig 2b; HF-Type-1). However, the recording with the HF-Type-1 device was occasionally interrupted; in this case, the recording duration was <30 min.

Voluntary deep breathing was not mandated during the recording of lung sounds. The statistics of the recordings are listed in Table 2.

Table 2. Statistics of recordings and labels of HF_Lung_V1 database.

	Littmann 3200	**HF-Type-1**	**Total**
Subjects			
n	261	18	261
Recordings			
Filename prefix	steth_	trunc_	NA
Rounds of recording	748	70	NA
No of 15-sec recordings	4504	5261	9765
Total duration (min)	1126	1315.25	2441.25
Labels			
No of I	16535	17560	34095

Total duration of I (min)	257.17	271.02	528.19
Mean duration of I (sec)	0.93	0.93	0.93
No of E	9107	9242	18349
Total duration of E (min)	160.25	132.60	292.85
Mean duration of E (sec)	1.06	0.86	0.96
No of C/W/S/R	6984/3974/152/2858	6899/4483/534/1882	13883/8457/686/4740
Total duration of C/W/S/R (min)	105.90/63.92/1.94/40.04	85.26/55.80/7.52/21.94	191.16/119.73/9.46/61.98
Mean duration of C/W/S/R (sec)	0.91/0.97/0.76/0.84	0.74/0.75/0.85/0.70	0.83/0.85/0.83/0.78
No of D	7266	8340	15606
Total duration of D (min)	111.75	55.80	230.87
Mean duration of D (sec)	0.92	0.87	0.89

I: inhalation, E: exhalation, W: wheeze, S: stridor, R: rhonchus, C: continuous adventitious sound, D: discontinuous adventitious sound. W, S, and R were combined to form C.

2.3 Audio file truncation

In this study, the standard duration of an audio signal used for inhalation, exhalation, and adventitious sound detection was 15 s. This duration was selected because a 15-s signal contains at least three complete breath cycles, which are adequate for a clinician to reach a clinical conclusion. Furthermore, a 15-s breath sound was be used previously for verification and validation [43] .

Because each audio file generated by the Littmann 3200 device had a length of 15.8 s, we cropped out the final 0.8-s signal from the files (Fig 2b; Littmann 3200). Moreover, we used only the first 15 s of each 2-min signal of the audio files (Fig 2b; HF-Type-1) generated by the HF-Type-1 device. Table 2 presents the number of truncated 15-s recordings and the total duration.

2.4 Data labeling

Because the data in the TSECC database contains only classification labels indicating whether a CAS or DAS exists in a recording, we attempted to label the event level of all sound recordings. Two board-certified respiratory therapists (NJL and YLW) and one board-certified nurse (WLT), with 8, 3, and 13 years of clinical experience, respectively, were recruited to label the start and end points of inhalation (I), exhalation (E), wheeze (W), stridor (S), rhonchus (R), and DAS (D) events in the recordings. They labeled the sound events by listening to the recorded breath sounds while simultaneously observing the corresponding patterns on a spectrogram by using customized labeling software [44]. The labelers were asked not to label sound events if they could not clearly identify the corresponding sound or if an incomplete event at the beginning or end of an audio file caused difficulty in identification. BFH held regular meetings to ensure that the labelers had good agreement on labeling criteria based on a few samples by judging the mean pseudo-κ value [27]. When developing artificial intelligence (AI) detection models, we combined the W, S, and R labels to form CAS labels (C). Moreover, the D labels comprised only crackles, which were not differentiated into coarse or fine crackles. The labelers were asked to label the period containing crackles but not a single explosive sound (generally less than 25 ms) of a crackle. Each recording was annotated by only one labeler; thus, the labels did not represent perfect ground truth. However, we used the labels as ground-truth labels for model training, validation, and testing. The statistics of the labels are listed in Table 2.

3. Inhalation, exhalation, CAS, and DAS detection

3.1 Framework

The inhalation, exhalation, CAS, and DAS detection framework developed in this study is displayed in Fig 3. The prominent advantage of the research framework is its modular design. Specifically, each unit of the framework can be tested separately, and the algorithms in different parts of the framework can be modified to achieve optimal overall performance. Moreover, the output of some blocks can be used for multiple purposes. For instance, the spectrogram generated by the preprocessing block can be used as the input of a model or for visualization in the user interface for real-time monitoring.

Fig. 3. Pipeline of detection framework.

The framework comprises three parts: preprocessing, deep learning–based modeling, and postprocessing. The preprocessing part involves signal processing and feature engineering techniques. The deep learning–based modeling part entails the use of a well-designed neural network for obtaining a sequence of classification predictions rather than a single prediction. The postprocessing part involves merging the segment prediction results and eliminating the burst event.

3.2 Preprocessing

We processed the lung sound recordings at a sampling frequency of 4 kHz. First, to eliminate the 60-Hz electrical interference and a part of the heart sound noise, we applied a high-pass filter to the recordings by setting a filter order of 10 and cut-off frequency of 80 Hz. The filtered signals were then processed using the short-time Fourier transform (STFT). In the STFT, we set a Hanning window size of 256 and hop length of 64; no additional zero-padding was applied. Thus, a 15-s sound signal could be transformed into a corresponding spectrogram with a size of 938 × 129. To obtain the spectral information regarding the lung sounds, we extracted the following features [29, 45]:

- Spectrogram: We extracted 129-bin log-magnitude spectrograms.
- Mel frequency cepstral coefficients (MFCCs): We extracted 20 static coefficients, 20 delta coefficients (Δ), and 20 acceleration coefficients (Δ^2). We used 40 mel bands within a frequency range of 0–4,000 Hz. The frame width used to calculate the delta and acceleration coefficients was set to 9, which resulted in a 60-bin vector per frame.
- Energy summation: We computed the energy summation of four frequency bands, namely 0–250, 250–500, 500–1,000, and 0–2,000 Hz, and obtained four values per time frame.

After extracting the aforementioned features, we concatenated them to form a 938 × 193 feature matrix. Subsequently, we conducted min–max normalization on each feature. The values of the normalized features ranged between 0 and 1.

3.3 Deep learning models

We investigated the performance of eight RNN models, namely LSTM, GRU, bidirectional LSTM (BiLSTM), bidirectional GRU (BiGRU), CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU, in terms of inhalation, exhalation, and adventitious sound detection. Fig 4 illustrates the detailed model structures. The outputs of the LSTM, GRU, BiLSTM, and BiGRU models were 938 × 1 vectors, and those of the CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU models were 469 × 1 vectors. An element in these vectors was set to 1 if an inhalation, exhalation, CAS, or DAS occurred within a time segment in which the output value passed the thresholding criterion; otherwise, the element was set to 0.

Fig. 4. Model architectures and postprocessing for inhalation, exhalation, CAS, and DAS segment and event detection. (a) LSTM and GRU models; (b) BiLSTM and BiGRU models; and (c) CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU models.

For a fairer comparison of the performance of the unidirectional and bidirectional models, we trained additional simplified (SIMP) BiLSTM, SIMP BiGRU, SIMP CNN-BiLSTM, and SIMP CNN-BiGRU models by adjusting the number of trainable parameters. Parameter adjustment was conducted by halving the number of cells of the LSTM and GRU layers.

We used Adam as the optimizer in the benchmark model, and we set the initial learning rate to 0.0001 with a step decay (0.2×) when the validation loss did not decrease for 10 epochs. The learning process stopped when no improvement occurred over 50 consecutive epochs.

3.4 Postprocessing

The prediction vectors obtained using the adopted models can be further processed for different purposes. For example, we can transform the prediction result from frames to time for real-time monitoring. The breathing duration of most humans lies within a certain range; we considered this fact in our study. Accordingly, when the prediction results obtained using the models indicated that two consecutive inhalation events occurred within a very small interval, we checked the continuity of these two events and decided whether to merge them, as illustrated in the bottom panel of Fig 4a. For example, when the interval between the *j*th and *i*th events was smaller than *T* s, we computed the difference in frequency between their energy peaks ($\left|\boldsymbol{p_j} - \boldsymbol{p_i}\right|$). Subsequently, if the difference was below a given threshold *P*, the two events were merged into a single event. In the experiment, *T* was set to 0.5 s, and *P* was set to 25 Hz. After the merging process, we further assessed whether a burst event existed. If the duration of an event was shorter than 0.05 s, the event was deleted.

3.5 Dataset arrangement and cross-validation

We adopted fivefold cross-validation in the training dataset to train and validate the models. Moreover, we used an independent testing dataset to test the performance of the trained models. According to our preliminary experience, the acoustic patterns of the breath sounds collected from one patient at different auscultation locations or between short intervals had many similarities. To avoid potential data leakage caused by our methods of collecting and truncating the breath sound signals, we assigned all truncated recordings collected on the same day to only one of the training, validation, or testing datasets; this is because these recordings might have been collected from the same patient within a short period. The statistics of the datasets are listed in Table 3. We used only audio files containing CASs and DASs to train and test their corresponding detection models.

Table 3. Statistics of the datasets and labels of the HF_Lung_V1 database.

	Training Dataset	**Testing Dataset**	**Total**
Recordings			
No of 15-sec recordings	7809	1956	9765
Total duration (min)	1952.25	489	2441.25
Labels			
No of I	27223	6872	34095
Total duration of I (min)	422.17	105.97	528.14
Mean duration of I (sec)	0.93	0.93	0.93
No of E	15601	2748	18349
Total duration of E (min)	248.05	44.81	292.85
Mean duration of E (sec)	0.95	0.98	0.96
No of C/W/S/R	11464/7027/657/3780	2419/1430/29/960	13883/8457/686/4740
Total duration of C/W/S/R (min)	160.16/100.71/9.10/50.35	31.01/19.02/0.36/11.63	191.16/119.73/9.46/61.98
Mean duration of C/W/S/R (sec)	0.84/0.86/0.83/0.80	0.77/0.80/0.74/0.73	0.83/0.85/0.83/0.78
No of D	13794	1812	15606
Total duration of D (min)	203.59	27.29	230.87
Mean duration of D (sec)	0.89	0.90	0.89

I: inhalation, E: exhalation, W: wheeze, S: stridor, R: rhonchus, C: continuous adventitious sound, D: discontinuous adventitious sound. W, S, and R were combined to form C.

3.6 Task definition and evaluation metrics

[4] clearly defined classification and detection at the segment, event, and recording levels. In this study, we performed two tasks. The first task involved performing detection at the segment level. The acoustic signal of each lung sound recording was transformed into a spectrogram. The temporal resolution of the spectrogram depended on the window size and overlap ratio of the STFT. The aforementioned parameters were fixed such that each spectrogram was a matrix of size 938 × 129. Thus, each recording contained 938 time segments (time frames), and each time segment was automatically labeled (Fig 5b) according to the ground-truth event labels (Fig 5a) assigned by the labelers. The output of the prediction process was a sequential prediction matrix (Fig 5c) of size 938 × 1 in the LSTM, GRU, BiLSTM, and BiGRU models and size 469 × 1 in the CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU models. By comparing the sequential prediction with the ground-truth time segments, we could define true positive (TP; orange vertical bars in Fig 5d), true negative (TN; green vertical bars in Fig 5d), false positive (FP; black vertical bars in Fig 5d), and false negative (FN; yellow vertical bars in Fig 5d) time segments. Subsequently, the models' sensitivity and specificity in classifying the segments in each recording were computed.

Fig 5. Task definition and evaluation metrics. (a) Ground-truth event labels, (b) ground-truth time segments, (c) AI inference results, (d) segment classification, (e) event detection, and (f) legend. JI: Jaccard index.

The second task entailed event detection at the recording level. After completing the sequential prediction (Fig 5c), we assembled the time segments associated with the same label into a corresponding event (Fig 5e). We also derived the start and end times of each assembled event. The Jaccard index (JI; [27] was used to determine whether an AI inference result correctly matched the ground-truth event. For an assembled event to be designated as a TP event (orange horizontal bars in Fig 5e), the corresponding JI value must be greater than 0.5. If the JI was between 0 and 0.5, the assembled event was designated as an FN event (yellow horizontal bars in Fig 5e), and if it was 0, the assembled event was designated as an FP event (black horizontal bars in Fig 5e). A TN event cannot be defined in the task of event detection.

The performance of the models was evaluated using the *F1* score, and that of segment detection was evaluated using the receiver operating characteristic (ROC) curve and area under the ROC curve (AUC). In addition, the mean absolute percentage error (MAPE) of event detection was derived. The accuracy, positive predictive value (PPV), sensitivity, specificity, and *F1* score of the models are presented in the section of Supporting information.

3.7 Hardware and software

We trained the baseline models on an Ubuntu 18.04 server that was provided by the National Center for High-Performance Computing in Taiwan [Taiwan Computing Cloud (TWCC)] and was equipped with an Intel(R) Xeon(R) Gold 6154 @3.00 GHz CPU with 90 GB RAM. To manage the intensive computation involved in RNN training, we implemented the training module by using the TensorFlow 2.10, CUDA 10, and CuDNN 7 programs to run the NVIDIA Titan V100 card on the TWCC server for GPU acceleration.

4 Results

4.1 LSTM versus GRU models

Table 4 presents the *F1* scores used to compare the eight LSTM- and GRU-based models. When a CNN was not added, the GRU models outperformed the LSTM models by 0.7%–9.5% in terms of the *F1* scores. However, the CNN-GRU and CNN-BiGRU models did not outperform the CNN-LSTM and CNN-BiLSTM models in terms of the *F1* scores (and vice versa).

Table 4. Comparison of *F1* scores between LSTM-based models and GRU-based models.

Models	n of trainable parameters	Inhalation		Exhalation		CASs		DASs	
		F1 score		*F1* score		*F1* score		*F1* score	
		Segment Detection	Event Detection	Segment Detection	Event Detection	Segment Detection	Event Detection	Segment Detection	Event Detection
LSTM	300,609	73.9%	76.1%	51.8%	57.0%	15.1%	12.2%	62.6%	59.1%
GRU	227,265	**76.2%**	**78.9%**	**59.8%**	**65.6%**	**24.6%**	**20.1%**	**65.9%**	**62.5%**
BiLSTM	732,225	78.1%	84.0%	57.3%	63.9%	19.8%	19.1%	69.6%	70.0%
BiGRU	552,769	**80.3%**	**86.2%**	**64.1%**	**70.9%**	**26.9%**	**25.6%**	**70.3%**	**71.4%**
CNN-LSTM	3,448,513	77.6%	81.1%	**57.7%**	**62.1%**	45.3%	42.5%	**68.8%**	64.4%
CNN-GRU	2,605,249	**78.4%**	**82.0%**	57.2%	62.0%	**51.5%**	**49.8%**	68.0%	**64.6%**
CNN-BiLSTM	6,959,809	**80.6%**	**86.3%**	60.4%	65.6%	47.9%	46.4%	**71.2%**	**70.8%**
CNN-BiGRU	5,240,513	**80.6%**	86.2%	**62.2%**	**68.5%**	**53.3%**	**51.6%**	70.6%	70.0%

The bold values indicate the higher *F1* score between the compared pairs of models.

According to the ROC curves presented in Fig 6a–d, the GRU-based models outperformed the LSTM-based models in all compared pairs, except for one pair, in terms of DAS segment detection (AUC of 0.891 for CNN-BiLSTM vs 0.889 for CNN-BiGRU).

Fig. 6. ROC curves for (a) inhalation, (b) exhalation, (c) CAS, and (d) DAS segment detection.

The corresponding AUC values are presented.

4.2 Unidirectional versus bidirectional models

As presented in Table 5, the bidirectional models outperformed their unidirectional counterparts in all the defined tasks by 0.4%–9.8% in terms of the *F1* scores, even when the bidirectional models had fewer trainable parameters after model adjustment.

Table 5. Comparison of *F1* scores between the unidirectional and bidirectional models.

		Inhalation		Exhalation		CASs		DASs	
Models	n of trainable parameters	*F1* score		*F1* score		*F1* score		*F1* score	
		Segment Detection	Event Detection	Segment Detection	Event Detection	Segment Detection	Event Detection	Segment Detection	Event Detection
LSTM	300,609	73.9%	76.1%	51.8%	57.0%	15.1%	12.2%	62.6%	59.1%
SIMP BiLSTM	235,073	**77.8%**	**84.1%**	**55.8%**	**62.4%**	**19.8%**	**17.9%**	**68.8%**	**68.9%**
GRU	227,265	76.2%	78.9%	59.8%	65.6%	24.6%	20.1%	65.9%	62.5%
SIMP BiGRU	178,113	**80.1%**	**86.1%**	**63.7%**	**70.0%**	**25.0%**	**22.2%**	**70.3%**	**71.3%**
CNN-LSTM	3,448,513	77.6%	81.1%	57.7%	62.1%	45.3%	42.5%	68.8%	64.4%
SIMP CNN-BiLSTM	3,382,977	**80.0%**	**85.8%**	**60.4%**	**66.2%**	**50.8%**	**50.2%**	**70.2%**	**70.2%**
CNN-GRU	2,605,249	78.4%	82.0%	57.2%	62.0%	51.5%	49.8%	68.0%	64.6%
SIMP CNN-BiGRU	2,556,097	**80.1%**	**85.9%**	**62.4%**	**68.4%**	**52.6%**	**51.5%**	**69.9%**	**69.5%**

The bold values indicate the higher *F1* score between the compared pairs of models. SIMP means the number of trainable parameters is adjusted.

4.3 Models with CNN versus those without CNN

According to Table 6, the models with a CNN outperformed those without a CNN in 26 of the

32 compared pairs.

Table 6. Comparison of *F1* scores between models without and with a CNN.

Models	n of trainable parameters	Inhalation		Exhalation		CASs		DASs	
		F1 score		*F1* score		*F1* score		*F1* score	
		Segment Detection	Event Detection	Segment Detection	Event Detection	Segment Detection	Event Detection	Segment Detection	Event Detection
LSTM	300,609	73.9%	76.1%	51.8%	57.0%	15.10%	12.20%	62.60%	59.10%
CNN-LSTM	3,448,513	**77.6%**	**81.1%**	**57.7%**	**62.1%**	**45.30%**	**42.50%**	**68.80%**	**64.40%**
BiLSTM	732,225	76.2%	78.9%	**59.8%**	**65.6%**	19.80%	17.90%	68.80%	68.90%
CNN-BiLSTM	6,959,809	**78.4%**	**82.0%**	57.2%	62.0%	**50.80%**	**50.20%**	**70.20%**	**70.20%**
GRU	227,265	78.1%	84.0%	57.3%	63.9%	24.60%	20.10%	65.90%	62.50%
CNN-GRU	2,605,249	**80.6%**	**86.3%**	**60.4%**	**65.6%**	**51.50%**	**49.80%**	**68.00%**	**64.60%**
BiGRU	178,113	80.3%	**86.2%**	**64.1%**	**70.9%**	25.00%	22.20%	**70.30%**	**71.30%**
CNN-BiGRU	2,556,097	**80.6%**	**86.2%**	62.2%	68.5%	**52.60%**	**51.50%**	69.90%	69.50%

The bold values indicate the higher *F1* score between the compared pairs of models.

The models with a CNN exhibited higher AUC values than did those without a CNN (Fig 6a–d), except that BiGRU had a higher AUC value than did CNN-BiGRU in terms of inhalation detection (0.963 vs 0.961), GRU had a higher AUC value than did CNN-GRU in terms of exhalation detection (0.886 vs 0.883), and BiGRU had a higher AUC value than did CNN-BiGRU in terms of exhalation detection (0.911 vs 0.899).

Moreover, compared with the LSTM, GRU, BiLSTM, and BiGRU models, the CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU models exhibited flatter and lower MAPE curves over a wide range of threshold values in all event detection tasks (Fig 7a–d).

Fig 7. MAPE curves for (a) inhalation, (b) exhalation, (c) CAS, and (d) DAS event detection.

5 Discussion

5.1 Benchmark results

According to the *F1* scores presented in Table 4, among models without a CNN, the GRU and BiGRU models consistently outperformed the LSTM and BiLSTM models in all defined tasks. However, the GRU-based models did not have superior *F1* scores among models with a CNN. Regarding the ROC curves and AUC values (Fig 6a–d), the GRU-based models consistently outperformed the other models in all but one task. Accordingly, we can conclude that GRU-based models perform slightly better than LSTM-based models in lung sound analysis. Previous studies have also compared LSTM- and GRU-based models [38, 46, 47]. Although a concrete conclusion cannot be drawn regarding whether LSTM-based models are superior to the GRU-based models (and vice versa), GRU-based models have been reported to outperform LSTM-based models in terms of computation time [38, 47].

As presented in Table 5, the bidirectional models outperformed their unidirectional counterparts in all defined tasks, a finding that is consistent with several previously obtained results [29, 36, 38, 40].

A CNN can facilitate the extraction of useful features and enhance the prediction accuracy of RNN-based models. The benefits engendered by a CNN are particularly vital in CAS detection. For the models with a CNN, the *F1* score improvement ranged from 26.0% to 30.3% and the AUC improvement ranged from 0.067 to 0.089 in the CAS detection tasks. Accordingly, we can infer that considerable information used in CAS detection resides in the local positional arrangement of the features. Thus, a two-dimensional CNN facilitates the extraction of the associated information. Notably, CNN-induced improvements in model performance in the inhalation, exhalation, and DAS detection tasks were not as high as those observed in the CAS detection tasks. The MAPE curves (Fig 7a–d) reveal that a model with a CNN has more consistent predictions over various threshold values.

In our previous study [26], an attention-based encoder–decoder architecture based on ResNet and LSTM exhibited favorable performance in inhalation (*F1* score of 90.4%) and exhalation (*F1* score of 93.2%) segment detection tasks. However, the model was established on the basis of a very small dataset (489 recordings of 15-s-long lung sounds). Moreover, the model involves a complicated architecture; hence, it is impossible to implement real-time respiratory monitoring in devices with limited computing power, such as smartphones or medical-grade tablets.

Few studies have performed event detection at the recording level by using a comparatively simple deep learning model. [29] used the BiGRU model and one-dimensional labels (similar to those used in the present study) for breath phase and crackle detection. Their BiGRU model exhibited comparable performance to our models in terms of inhalation event detection (*F1* scores, 87.0% vs 86.2%) and in terms of DAS event detection (*F1* scores, 72.1% vs 71.4%). However, the performance of the BiGRU model differed considerably from that of our models in terms of exhalation detection (*F1* scores: 84.6% vs 70.9%). One of the reasons for this discrepancy is that [29] established their ground-truth labels on the basis of the gold-standard signals of a pneumotachograph. Another reason is that an exhalation label is not always available following an inhalation label in our data. Finally, we did not specifically control the sounds we recorded; for example, we did not ask patients to perform voluntary deep breathing or keep ambient noise down. The factors influencing the model performance are further discussed in the next section.

5.2 Factors influencing model performance

The benchmark performance of the proposed models may have been influenced by the following factors: (1) unusual breathing patterns; (2) imbalanced data; (3) low signal-to-noise ratio (SNR); (4) noisy labels, including class and attribute noise, in the database; and (5) sound overlapping.

Fig 8 displays most of the breath patterns present in the HF_Lung_V1 database. Fig 8a illustrates the general pattern of a breath cycle in the lung sounds when the ratio of inhalation to exhalation durations is approximately 2:1 and an expiratory pause is noted [3, 4]. Fig 8b presents a frequent condition under which an exhalation is not completely heard by the labelers. However, because we did not ask the subjects to breath voluntarily when recording the sound, many unusual breath patterns might have been recorded, such as patterns caused by shallow breathing, fast breathing, and apnea as well as those caused by double triggering of the ventilator [48] and air trapping [49, 50]. These unusual breathing patterns might confuse the labeling and learning processes and result in poor testing results.

Fig 8. Patterns of normal breathing lung sounds. (a) General lung sound patterns and (b) general lung sound patterns with unidentifiable exhalations. “I” represents an identifiable inhalation event, “E” represents an identifiable exhalation event, and the black areas represent pause phases.

The developed database contains imbalanced numbers of inhalation and exhalation labels (34,095 and 18,349, respectively) because not every exhalation was heard and labeled. In addition, the proposed models may possess the capability of learning the rhythmic rise and fall of breathing signals but not the capability of learning acoustic or texture features that can distinguish an inhalation from an exhalation. This may thus explain the models’ poor performance in exhalation detection. However, these models are suitable for respiratory rate estimation and apnea detection as long as appropriate inhalation detection is achieved. Furthermore, for all labels, the summation of the event duration was smaller than that of the background signal duration (these factors had a ratio of approximately 1:2.5 to 1:7). The aforementioned phenomenon can be regarded as foreground–background class imbalance [51] and will be addressed in future studies.

Most of the sounds in the established database were not recorded during the patients performed deep breathing; thus, the signal quality was not maximized. However, training models with such nonoptimal data increase their adaptability to real-world scenarios. Moreover, the SNR may be reduced by noise, such as human voices; music; sounds from bedside monitors, televisions, air conditioners, fans, and radios; sounds generated by mechanical ventilators; electrical noise generated by touching or moving the parts of acoustic sensors; and friction sounds generated by the rubbing of two surfaces together (e.g., rubbing clothes with the skin). A poor SNR of audio signals can lead to difficulties in labeling and prediction tasks. The features of some noise types are considerably similar to those of adventitious sounds. The poor performance of the proposed models in CAS detection can be partly attributed to the noisy environment in which the lung sounds were recorded. In particular, the sounds generated by ventilators caused numerous FP events in the CAS detection tasks. Thus, additional effort is required to develop a superior preprocessing algorithm that can filter out influential noise or to identify a strategy to ensure that models focus on learning the correct CAS features. Furthermore, the integration of active noise-canceling technology [52] or noise suppression technology [53] into respiratory sound monitors can help reduce the noise from auscultatory signals.

The sound recordings in the HF_Lung_V1 database were labeled by only one labeler; thus, some noisy labels, including class and attribute noise, may exist in the database [54]. These noisy labels are attributable to (1) the different hearing abilities of the labeler, which can cause differences in the labeled duration; (2) the absence of clear criteria for differentiating between target and confusing events; (3) individual human errors; (4) tendency to not label events located close to the beginning and end of a recording; and (5) confusion caused by unusual breath patterns and poor SNRs. However, deep learning models exhibit high robustness to noisy labels [55]. Accordingly, we are currently working toward establishing better ground-truth labels.

Breathing generates CASs and DASs under abnormal respiratory conditions. This means that the breathing sound, CAS, and DAS might overlap with one another during the same period. This sound overlapping, along with the data imbalance, makes the CAS and DAS detection models learn to read the rise and fall of the breathing energy and falsely identify an inhalation or exhalation as CAS or DAS, respectively. This FP detection was observed in our benchmark results. In the future, strategies must be adopted to address the problem of sound overlap.

6 Conclusions

We established a large open-access lung sound database, namely HF_Lung_V1 (https://gitlab.com/techsupportHF/HF_Lung_V1), that contains 9,765 audio files of lung sounds (each with a duration of 15 s), 34,095 inhalation labels, 18,349 exhalation labels, 13,883 CAS labels (comprising 8,457 wheeze labels, 686 stridor labels, and 4,740 rhonchus labels), and 15,606 DAS labels (all of which are crackles).

We also investigated the performance of eight RNN-based models in terms of inhalation, exhalation, CAS detection, and DAS detection in the HF_Lung_V1 database. We determined that the bidirectional models outperformed the unidirectional models in lung sound analysis. Furthermore, the addition of a CNN to these models further improved their performance.

Future studies can develop more accurate respiratory sound analysis models. First, highly accurate ground-truth labels should be established. Second, researchers should investigate the performance of RNN-based models containing state-of-the-art convolutional layers. Third, regional CNN variants can be adopted in lung sound analysis if the labels are expanded to two-dimensional bounding boxes [27]. Fourth, wavelet-based approaches, empirical mode decomposition, and other methods that can extract different features should be investigated [4, 56]. Finally, respiratory sound monitors should be equipped with the capability of tracheal breath sound analysis [52].

Acknowledgments

This study was partially funded by the Raising Children Medical Foundation, Taiwan. The authors thank the employees of Heroic Faith Medical Science Co. Ltd. who have ever partially contributed to developing the HF-Type-1 and establishing the HF_Lung_V1 database. This manuscript was edited by Wallace Academic Editing. We also thank the All Vista Healthcare Center, Ministry of Science and Technology, Taiwan for the support.

References

1. Bohadana A, Izbicki G, Kraman SS. Fundamentals of lung auscultation. New England Journal of Medicine. 2014;370(8):744-51.
2. Goettel N, Herrmann MJ. Breath Sounds: From Basic Science to Clinical Practice. Anesthesia & Analgesia. 2019;128(3):e42.
3. Sarkar M, Madabhavi I, Niranjan N, Dogra M. Auscultation of the respiratory system. Annals of thoracic medicine. 2015;10(3):158.
4. Pramono RXA, Bowyer S, Rodriguez-Villegas E. Automatic adventitious respiratory sound analysis: A systematic review. PloS one. 2017;12(5):e0177926.
5. Wang B, Liu Y, Wang Y, Yin W, Liu T, Liu D, et al. Characteristics of Pulmonary auscultation in patients with 2019 novel coronavirus in china. 2020.
6. Raj V, Renjini A, Swapna M, Sreejyothi S, Sankararaman S. Nonlinear time series and principal component analyses: Potential diagnostic tools for COVID-19 auscultation. Chaos, Solitons & Fractals. 2020;140:110246.
7. Sovijärvi A, Vanderschoot J, Earis J. Standardization of computerized respiratory sound analysis. Crit Care Med. 1997;156:974-87.
8. Berry MP, Martí J-D, Ntoumenopoulos G. Inter-rater agreement of auscultation, palpable fremitus, and ventilator waveform sawtooth patterns between clinicians. Respiratory care. 2016;61(10):1374-83.
9. Grunnreis FO. Intra-and interobserver variation in lung sound classification. Effect of training: UiT Norges arktiske universitet; 2016.
10. 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(9):1396-403.
11. Huq S, Moussavi Z. Acoustic breath-phase detection using tracheal breath sounds. Medical & biological engineering & computing. 2012;50(3):297-308.
12. Mesaros A, Heittola T, Virtanen T. Metrics for polyphonic sound event detection. Applied Sciences. 2016;6(6):162.
13. Pasterkamp H, Kraman SS, Wodicka GR. Respiratory sounds: advances beyond the stethoscope. American journal of respiratory and critical care medicine. 1997;156(3):974-87.
14. Chambres G, Hanna P, Desainte-Catherine M, editors. Automatic detection of patient with respiratory diseases using lung sound analysis. 2018 International Conference on Content-Based Multimedia Indexing (CBMI); 2018: IEEE.
15. Demir F, Sengur A, Bajaj V. Convolutional neural networks based efficient approach for classification of lung diseases. Health Inf Sci Syst. 2020;8(1):4.

16. Hosseini M, Ren H, Rashid H-A, Mazumder AN, Prakash B, Mohsenin T. Neural Networks for Pulmonary Disease Diagnosis using Auditory and Demographic Information. arXiv preprint arXiv:201113194. 2020.
17. Perna D, Tagarelli A. Deep Auscultation: Predicting Respiratory Anomalies and Diseases via Recurrent Neural Networks. 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)2019. p. 50-5.
18. Pham L, McLoughlin I, Phan H, Tran M, Nguyen T, Palaniappan R. Robust Deep Learning Framework For Predicting Respiratory Anomalies and Diseases. arXiv preprint arXiv:200203894. 2020.
19. Acharya J, Basu A. Deep Neural Network for Respiratory Sound Classification in Wearable Devices Enabled by Patient Specific Model Tuning. IEEE transactions on biomedical circuits and systems. 2020;14(3):535-44.
20. Aykanat M, Kılıç Ö, Kurt B, Saryal S. Classification of lung sounds using convolutional neural networks. EURASIP Journal on Image and Video Processing. 2017;2017(1).
21. Bardou D, Zhang K, Ahmad SM. Lung sounds classification using convolutional neural networks. Artif Intell Med. 2018;88:58-69.
22. Chen H, Yuan X, Pei Z, Li M, Li J. Triple-Classification of Respiratory Sounds Using Optimized S-Transform and Deep Residual Networks. IEEE Access. 2019;7:32845-52.
23. Grzywalski T, Piecuch M, Szajek M, Breborowicz A, Hafke-Dys H, Kocinski J, et al. Practical implementation of artificial intelligence algorithms in pulmonary auscultation examination. Eur J Pediatr. 2019;178(6):883-90.
24. Kochetov K, Putin E, Balashov M, Filchenkov A, Shalyto A. Noise Masking Recurrent Neural Network for Respiratory Sound Classification. Artificial Neural Networks and Machine Learning – ICANN 2018. Lecture Notes in Computer Science2018. p. 208-17.
25. Li L, Xu W, Hong Q, Tong F, Wu J, editors. Classification between normal and adventitious lung sounds using deep neural network. 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP); 2016: IEEE.
26. Hsiao C-H, Lin T-W, Lin C-W, Hsu F-S, Lin FY-S, Chen C-W, et al., editors. 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); 2020: IEEE.
27. 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(8):1798.
28. Liu Y, Lin Y, Gao S, Zhang H, Wang Z, Gao Y, et al., editors. Respiratory sounds feature learning with deep convolutional neural networks. 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl

Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech); 2017: IEEE.

29. Messner E, Fediuk M, Swatek P, Scheidl S, Smolle-Juttner F-M, Olschewski H, et al., editors. 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); 2018: IEEE.

30. Rocha B, Filos D, Mendes L, Vogiatzis I, Perantoni E, Kaimakamis E, et al., editors. A respiratory sound database for the development of automated classification. International Conference on Biomedical and Health Informatics; 2017: Springer.

31. 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.

32. Sun C, Shrivastava A, Singh S, Gupta A, editors. Revisiting unreasonable effectiveness of data in deep learning era. Proceedings of the IEEE international conference on computer vision; 2017.

33. Elman JL. Finding structure in time. Cognitive science. 1990;14(2):179-211.

34. Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735-80.

35. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:14061078. 2014.

36. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks. 2005;18(5-6):602-10.

37. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE transactions on Signal Processing. 1997;45(11):2673-81.

38. Khandelwal S, Lecouteux B, Besacier L. Comparing GRU and LSTM for automatic speech recognition. 2016.

39. Li L, Wu Z, Xu M, Meng HM, Cai L, editors. Combining CNN and BLSTM to Extract Textual and Acoustic Features for Recognizing Stances in Mandarin Ideological Debate Competition. INTERSPEECH; 2016.

40. Parascandolo G, Huttunen H, Virtanen T, editors. Recurrent neural networks for polyphonic sound event detection in real life recordings. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2016: IEEE.

41. Deng M, Meng T, Cao J, Wang S, Zhang J, Fan H. Heart sound classification based on improved MFCC features and convolutional recurrent neural networks. Neural Networks. 2020.

42. Zhao H, Zarar S, Tashev I, Lee C-H, editors. Convolutional-recurrent neural networks for speech enhancement. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2018: IEEE.

43. Pasterkamp H, Brand PL, Everard M, Garcia-Marcos L, Melbye H, Priftis KN. Towards the standardisation of lung sound nomenclature. European Respiratory Journal. 2016;47(3):724-32.

44. 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(,).

45. Chamberlain D, Kodgule R, Ganelin D, Miglani V, Fletcher RR, editors. Application of semi-supervised deep learning to lung sound analysis. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2016: IEEE.

46. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:14123555. 2014.

47. Shewalkar AN. Comparison of rnn, lstm and gru on speech recognition data. 2018.

48. Thille AW, Rodriguez P, Cabello B, Lellouche F, Brochard L. Patient-ventilator asynchrony during assisted mechanical ventilation. Intensive care medicine. 2006;32(10):1515-22.

49. Blanch L, Bernabé F, Lucangelo U. Measurement of air trapping, intrinsic positive end-expiratory pressure, and dynamic hyperinflation in mechanically ventilated patients. Respiratory care. 2005;50(1):110-24.

50. Miller WT, Chatzkel J, Hewitt MG. Expiratory air trapping on thoracic computed tomography. A diagnostic subclassification. Annals of the American Thoracic Society. 2014;11(6):874-81.

51. Oksuz K, Cam BC, Kalkan S, Akbas E. Imbalance problems in object detection: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020.

52. Wu Y, Liu J, He B, Zhang X, Yu L. Adaptive Filtering Improved Apnea Detection Performance Using Tracheal Sounds in Noisy Environment: A Simulation Study. BioMed Research International. 2020;2020.

53. Emmanouilidou D, McCollum ED, Park DE, Elhilali M. Computerized lung sound screening for pediatric auscultation in noisy field environments. IEEE Transactions on Biomedical Engineering. 2017;65(7):1564-74.

54. Zhu X, Wu X. Class noise vs. attribute noise: A quantitative study. Artificial intelligence review. 2004;22(3):177-210.

55. Rolnick D, Veit A, Belongie S, Shavit N. Deep learning is robust to massive label noise. arXiv preprint arXiv:170510694. 2017.

56. Pramono RXA, Imtiaz SA, Rodriguez-Villegas E. Evaluation of features for classification of wheezes and normal respiratory sounds. PloS one. 2019;14(3):e0213659.

Fig 1

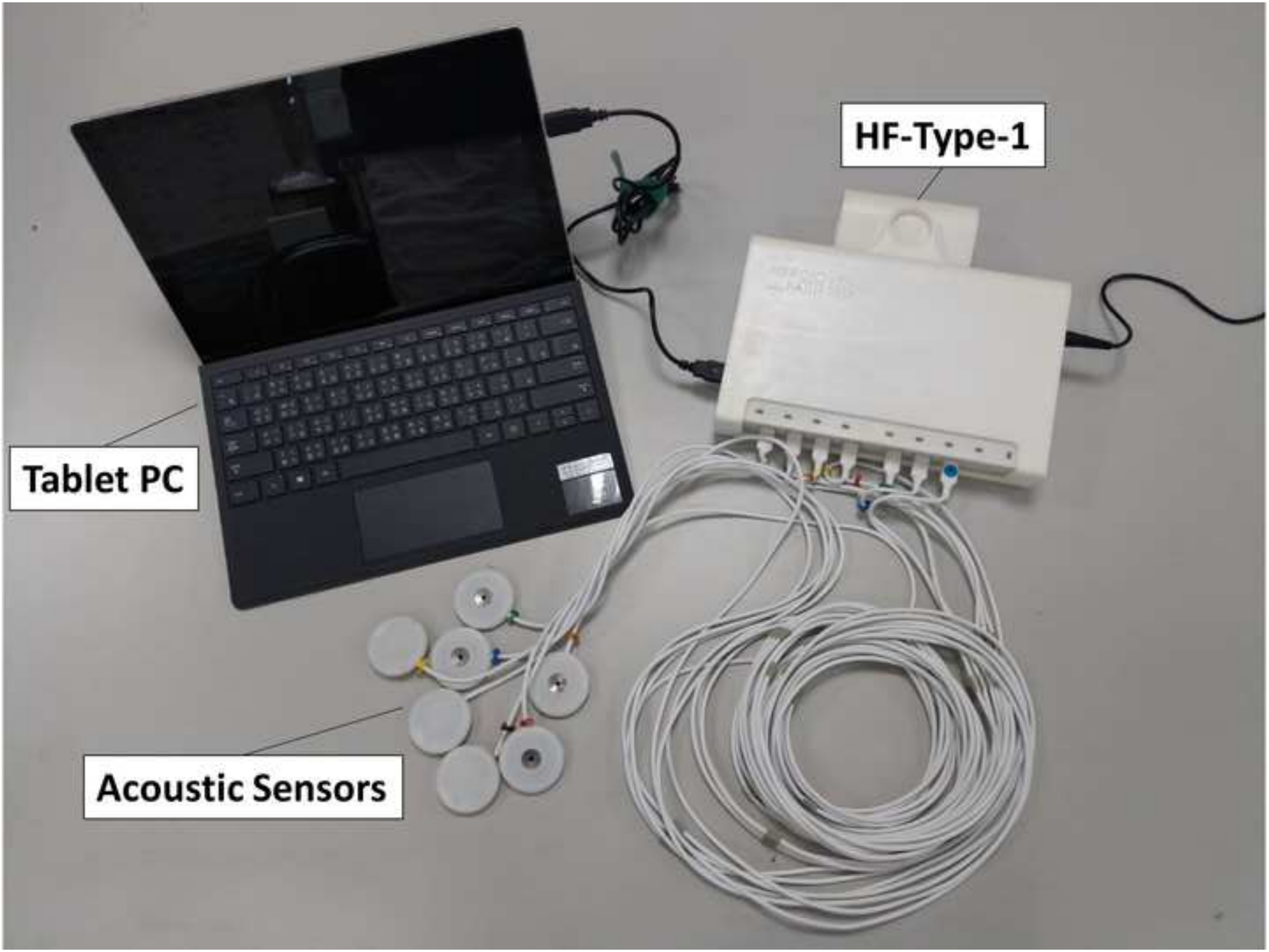


Fig 2

Click here to access/download;Figure;Fig2.tif

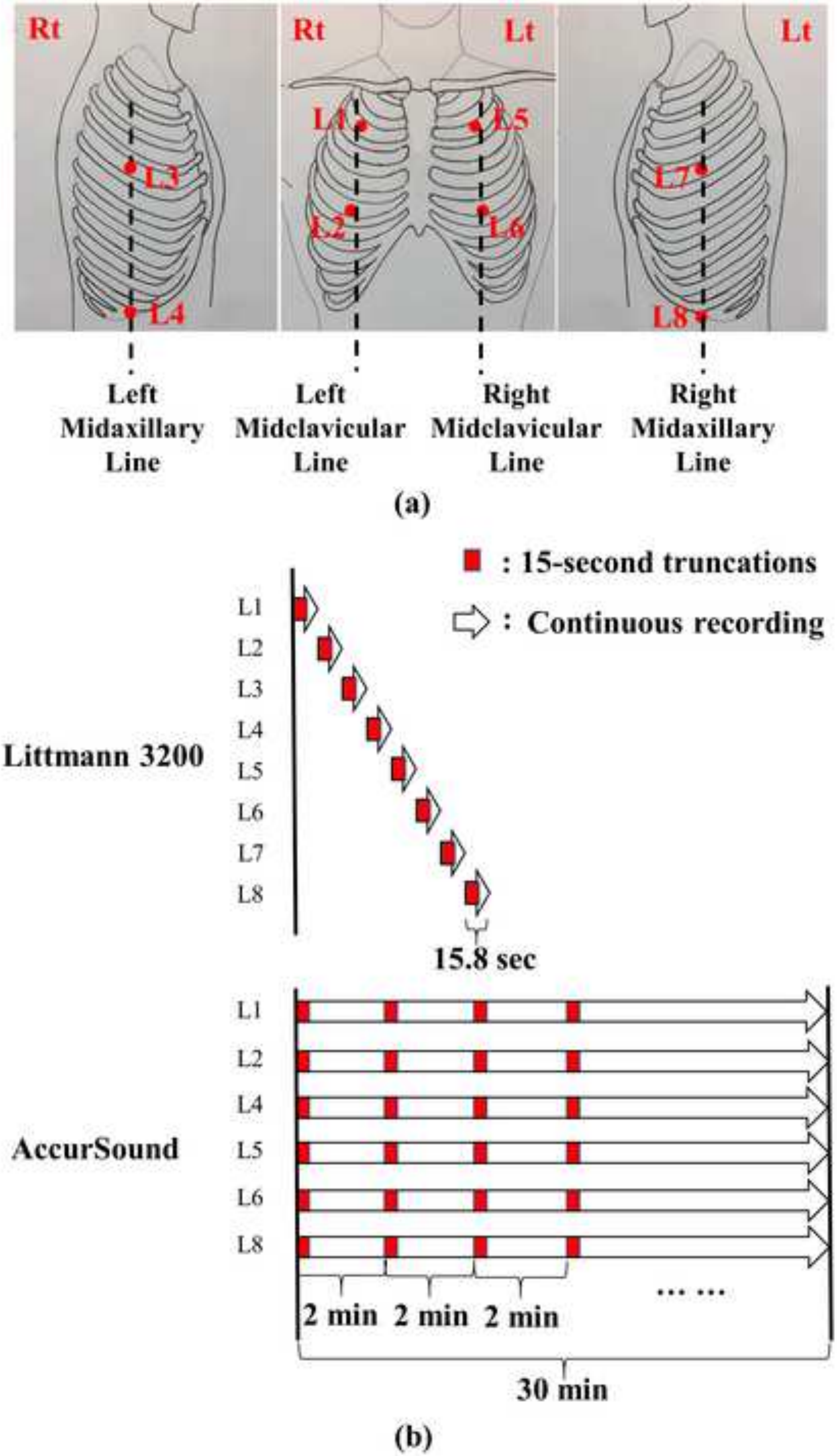
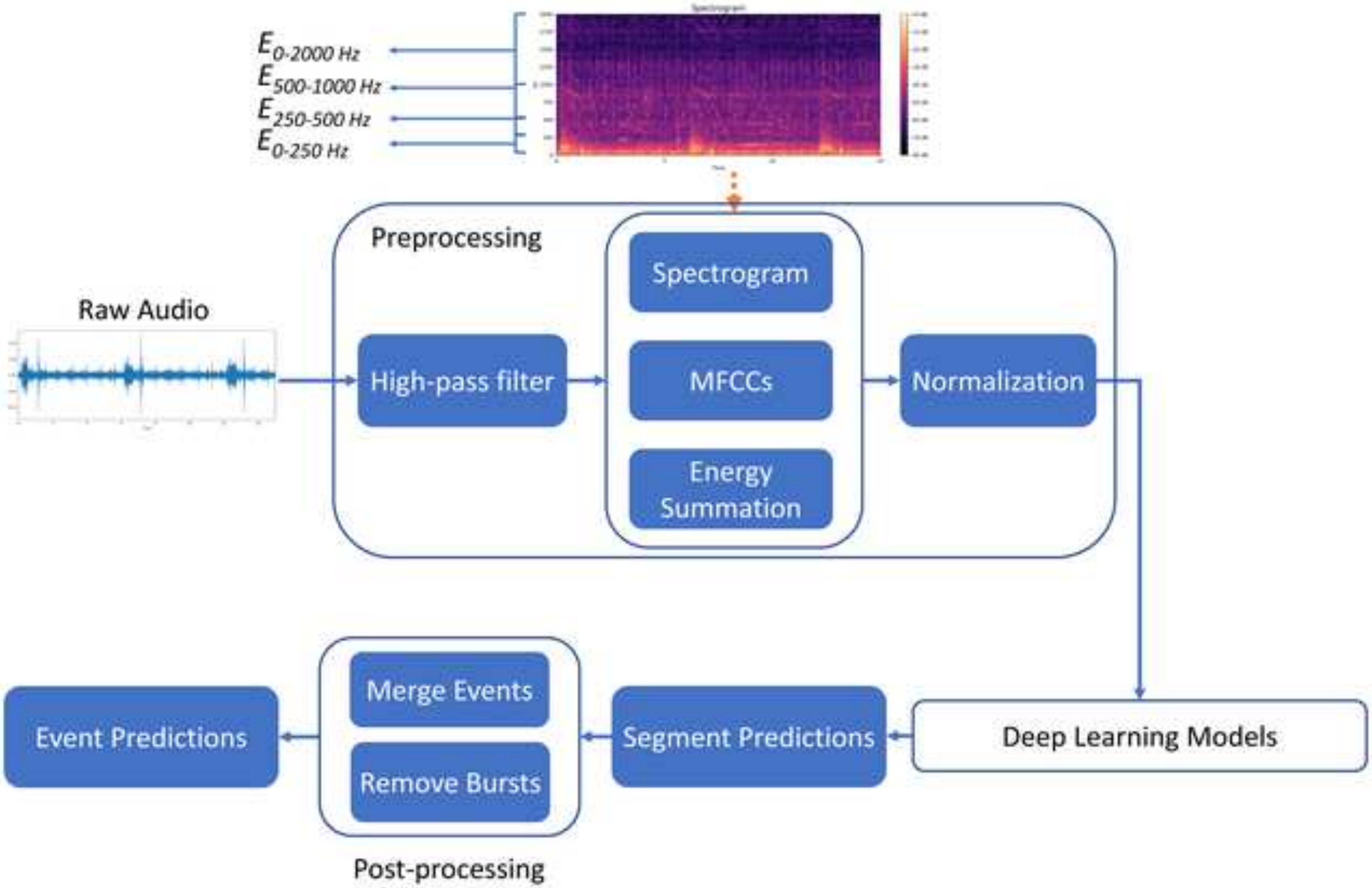


Fig 3


$E_{0\text{-}2000\ Hz}$
$E_{500\text{-}1000\ Hz}$
$E_{250\text{-}500\ Hz}$
$E_{0\text{-}250\ Hz}$
Preprocessing
Raw Audio
High-pass filter
Spectrogram
MFCCs
Energy Summation
Normalization
Event Predictions
Merge Events
Remove Bursts
Post-processing
Segment Predictions
Deep Learning Models

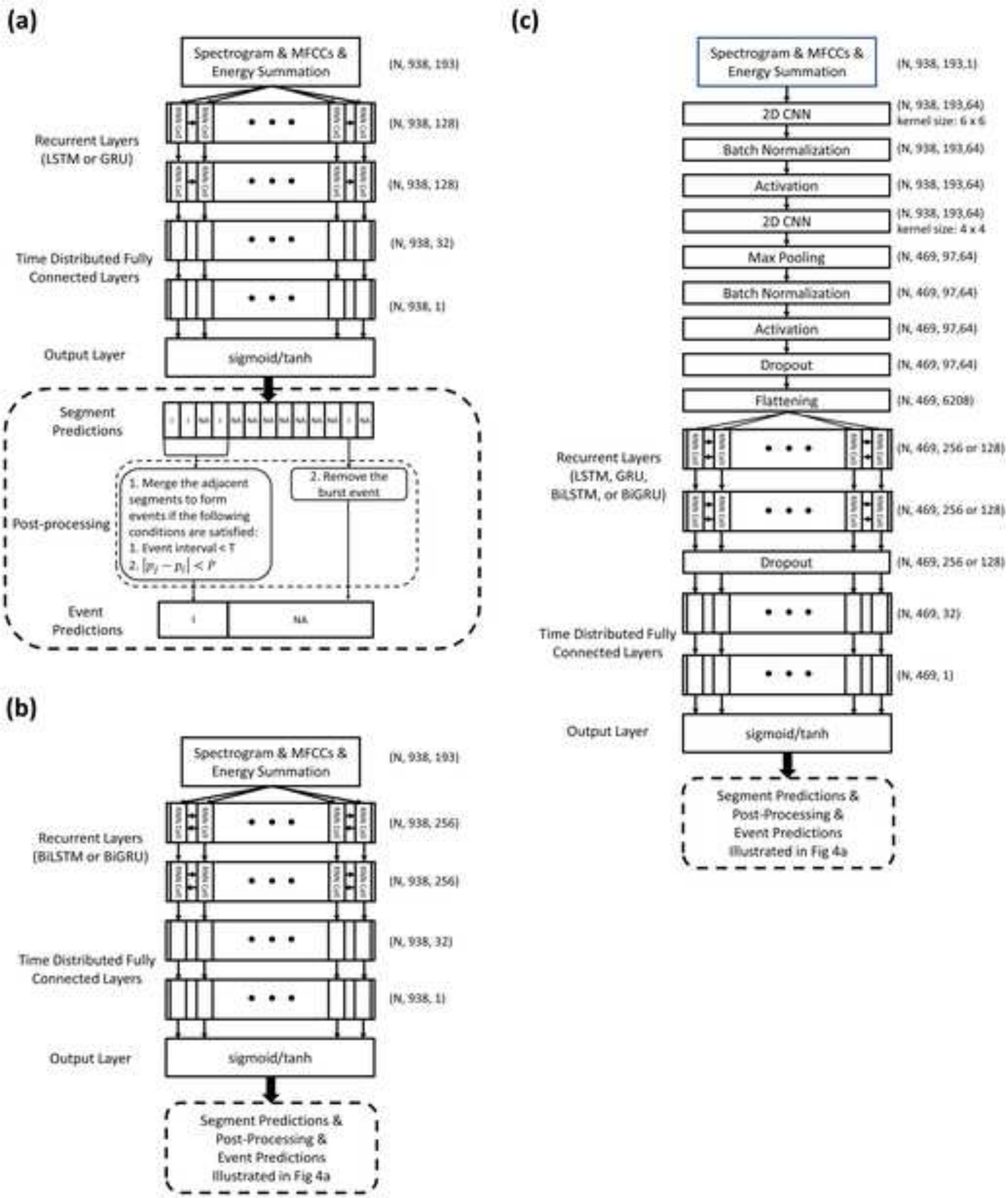
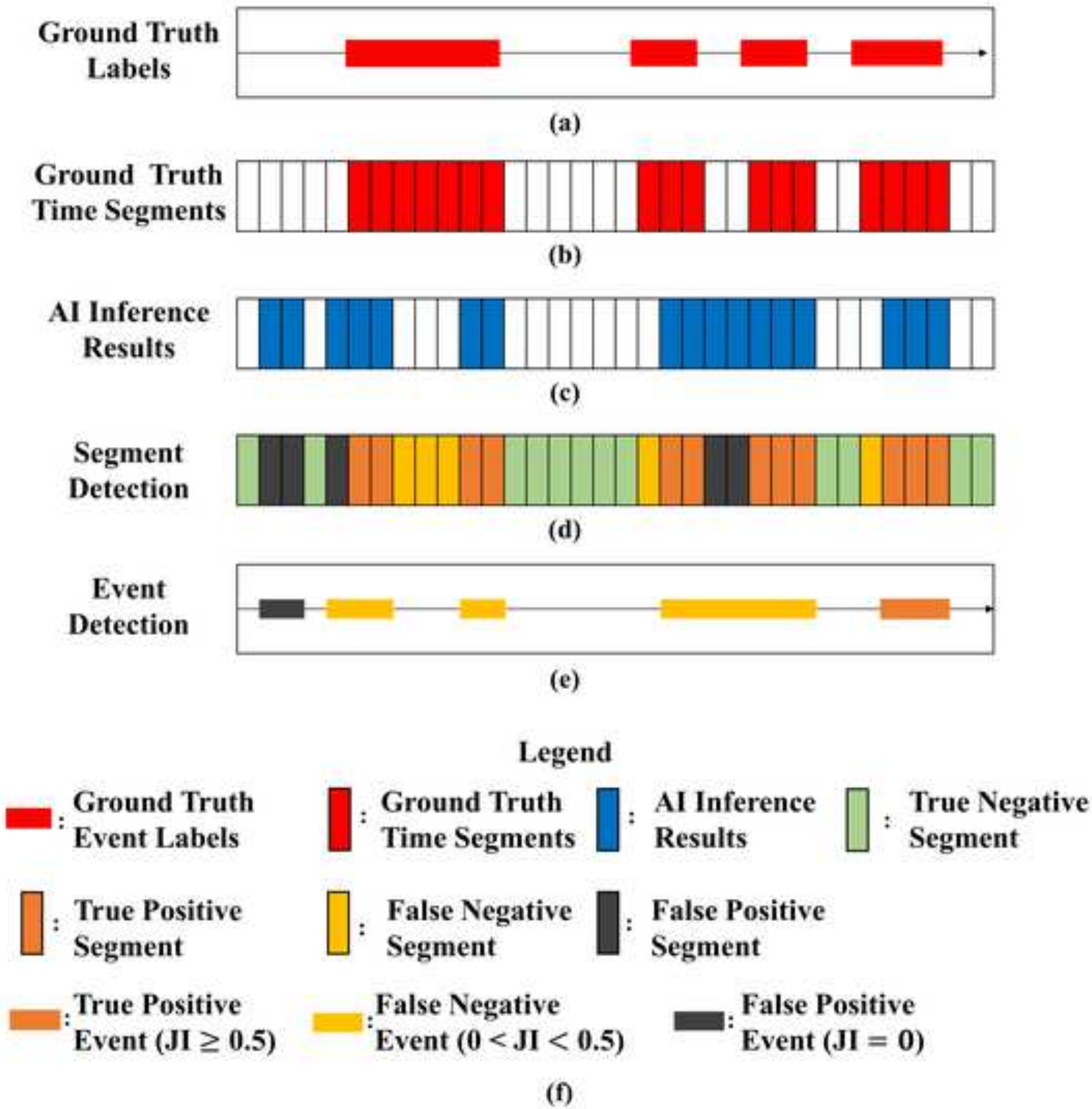
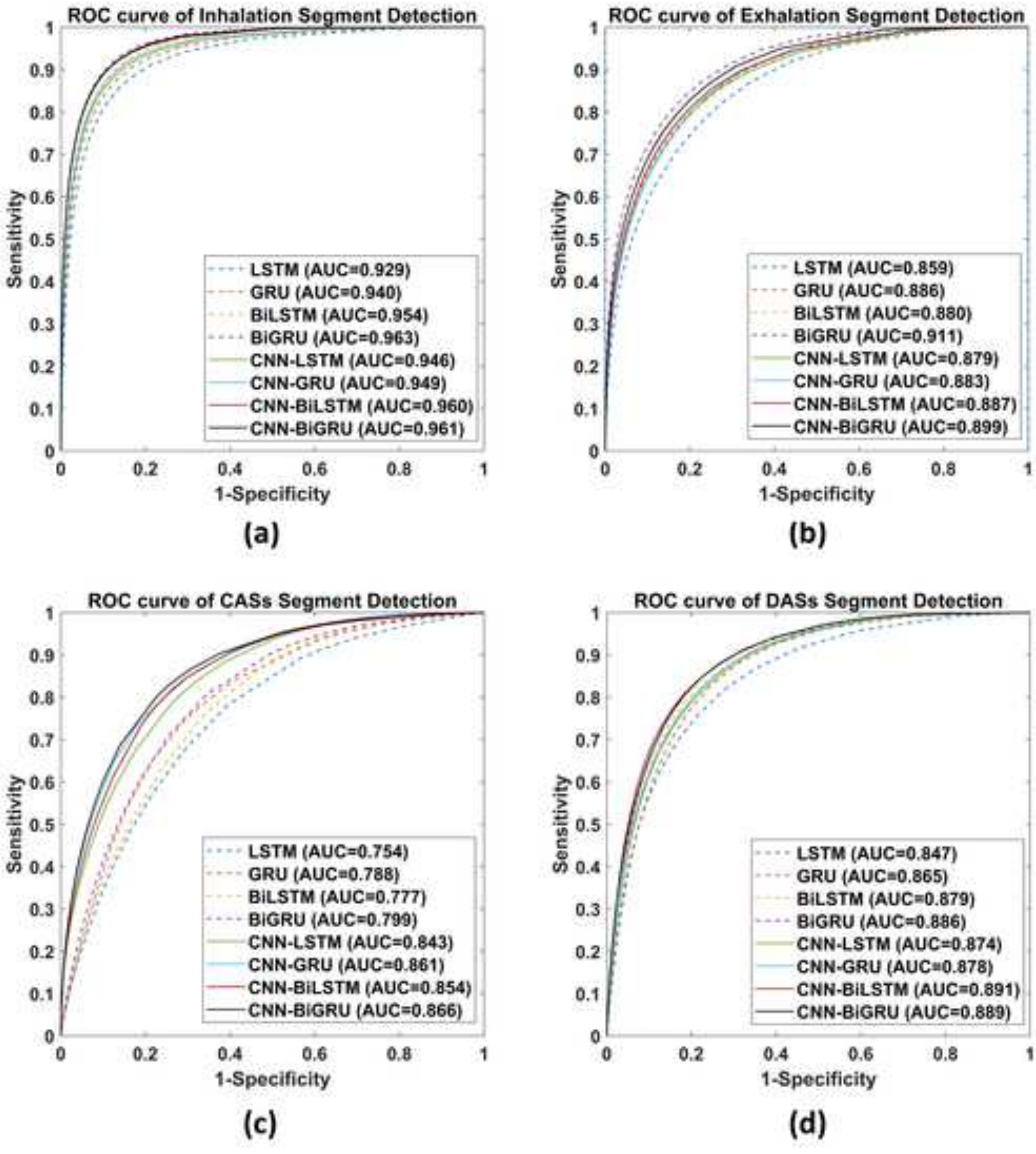

(a)
Spectrogram & MFCCs & Energy Summation
(N, 938, 193)
Recurrent Layers (LSTM or GRU)
(N, 938, 128)
(N, 938, 128)
Time Distributed Fully Connected Layers
(N, 938, 32)
(N, 938, 1)
Output Layer
sigmoid/tanh
Segment Predictions
Post-processing
1. Merge the adjacent segments to form events if the following conditions are satisfied:
1. Event interval < T
2. $|p_j - p_i| < P$
2. Remove the burst event
Event Predictions
NA
(b)
Spectrogram & MFCCs & Energy Summation
(N, 938, 193)
Recurrent Layers (BiLSTM or BiGRU)
(N, 938, 256)
(N, 938, 256)
Time Distributed Fully Connected Layers
(N, 938, 32)
(N, 938, 1)
Output Layer
sigmoid/tanh
Segment Predictions & Post-Processing & Event Predictions Illustrated in Fig 4a
(c)
Spectrogram & MFCCs & Energy Summation
(N, 938, 193,1)
2D CNN
(N, 938, 193,64) kernel size: 6 x 6
Batch Normalization
(N, 938, 193,64)
Activation
(N, 938, 193,64)
2D CNN
(N, 938, 193,64) kernel size: 4 x 4
Max Pooling
(N, 469, 97,64)
Batch Normalization
(N, 469, 97,64)
Activation
(N, 469, 97,64)
Dropout
(N, 469, 97,64)
Flattening
(N, 469, 6208)
Recurrent Layers (LSTM, GRU, BiLSTM, or BiGRU)
(N, 469, 256 or 128)
(N, 469, 256 or 128)
Dropout
(N, 469, 256 or 128)
Time Distributed Fully Connected Layers
(N, 469, 32)
(N, 469, 1)
Output Layer
sigmoid/tanh
Segment Predictions & Post-Processing & Event Predictions Illustrated in Fig 4a

Fig 5



Ground Truth Labels
(a)
Ground Truth Time Segments
(b)
AI Inference Results
(c)
Segment Detection
(d)
Event Detection
(e)
Legend
: Ground Truth Event Labels
: Ground Truth Time Segments
: AI Inference Results
: True Negative Segment
: True Positive Segment
: False Negative Segment
: False Positive Segment
: True Positive Event (JI ≥ 0.5)
: False Negative Event (0 < JI < 0.5)
: False Positive Event (JI = 0)
(f)


ROC curve of Inhalation Segment Detection
Sensitivity
1-Specificity
LSTM (AUC=0.929)
GRU (AUC=0.940)
BiLSTM (AUC=0.954)
BiGRU (AUC=0.963)
CNN-LSTM (AUC=0.946)
CNN-GRU (AUC=0.949)
CNN-BiLSTM (AUC=0.960)
CNN-BiGRU (AUC=0.961)
(a)
ROC curve of Exhalation Segment Detection
Sensitivity
1-Specificity
LSTM (AUC=0.859)
GRU (AUC=0.886)
BiLSTM (AUC=0.880)
BiGRU (AUC=0.911)
CNN-LSTM (AUC=0.879)
CNN-GRU (AUC=0.883)
CNN-BiLSTM (AUC=0.887)
CNN-BiGRU (AUC=0.899)
(b)
ROC curve of CASs Segment Detection
Sensitivity
1-Specificity
LSTM (AUC=0.754)
GRU (AUC=0.788)
BiLSTM (AUC=0.777)
BiGRU (AUC=0.799)
CNN-LSTM (AUC=0.843)
CNN-GRU (AUC=0.861)
CNN-BiLSTM (AUC=0.854)
CNN-BiGRU (AUC=0.866)
(c)
ROC curve of DASs Segment Detection
Sensitivity
1-Specificity
LSTM (AUC=0.847)
GRU (AUC=0.865)
BiLSTM (AUC=0.879)
BiGRU (AUC=0.886)
CNN-LSTM (AUC=0.874)
CNN-GRU (AUC=0.878)
CNN-BiLSTM (AUC=0.891)
CNN-BiGRU (AUC=0.889)
(d)

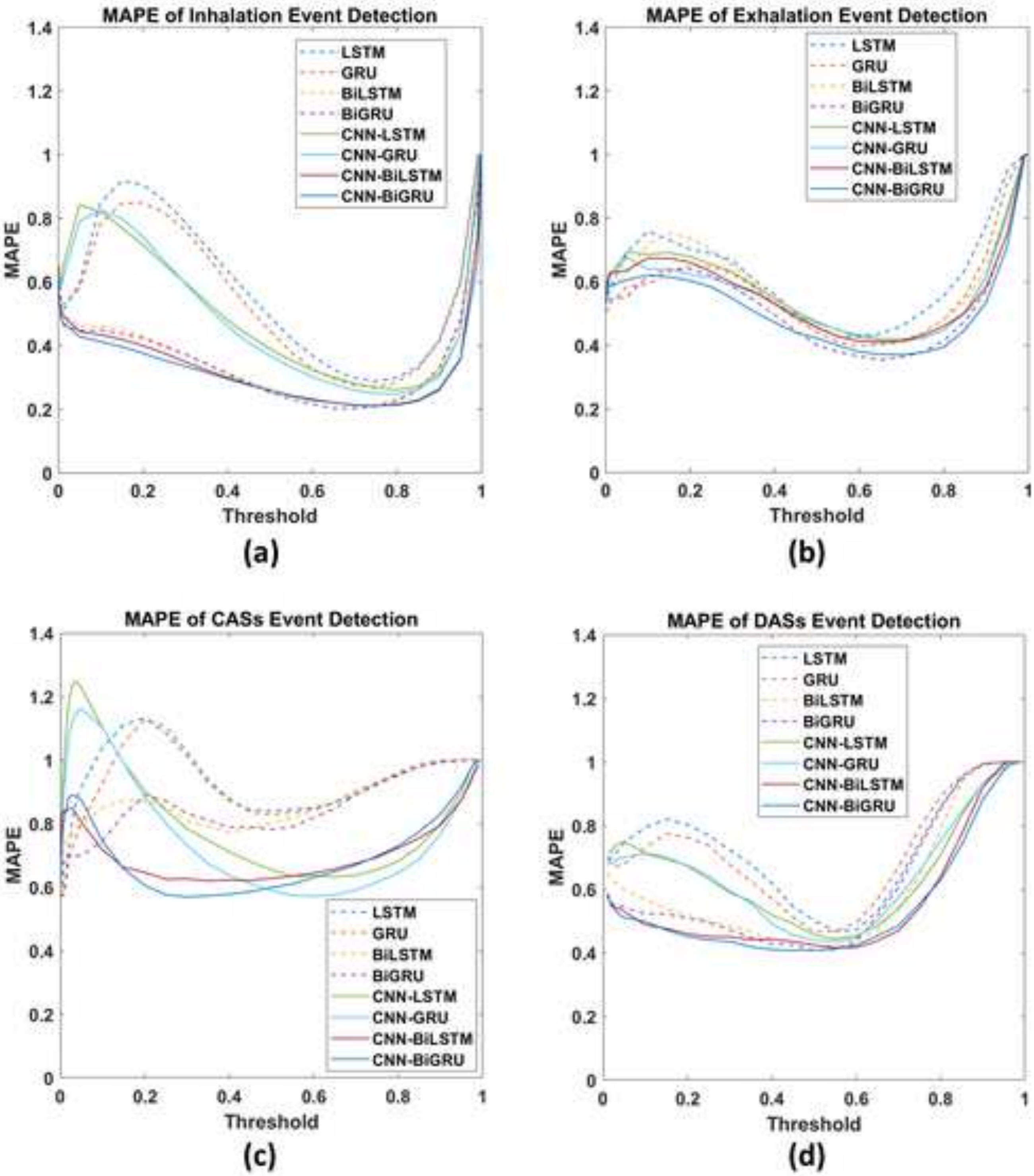
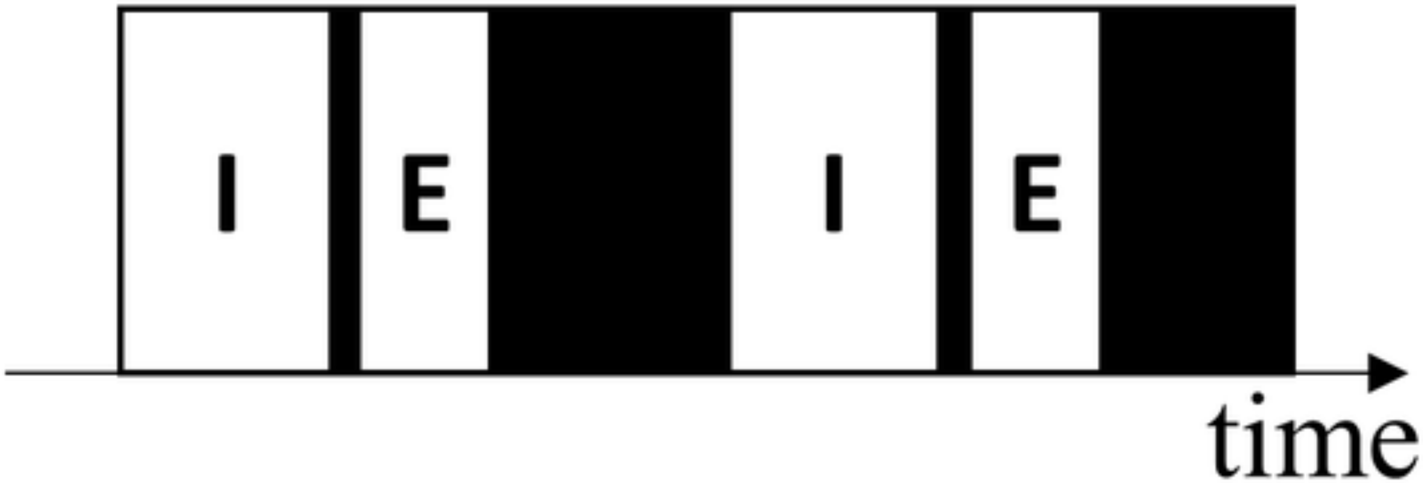
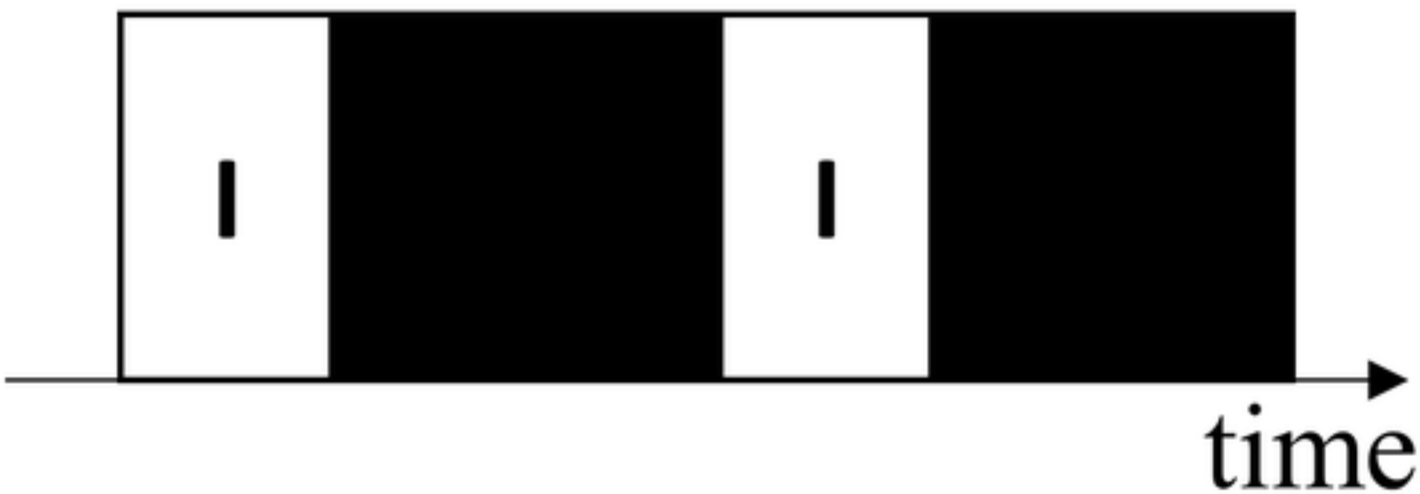

MAPE of Inhalation Event Detection
MAPE
Threshold
LSTM
GRU
BiLSTM
BiGRU
CNN-LSTM
CNN-GRU
CNN-BiLSTM
CNN-BiGRU
(a)
MAPE of Exhalation Event Detection
MAPE
Threshold
(b)
MAPE of CASs Event Detection
MAPE
Threshold
(c)
MAPE of DASs Event Detection
MAPE
Threshold
(d)

Fig 8



(a)



(b)

Click here to access/download
Supporting Information
Supporting information.docx