npj primary care respiratory medicine

www.nature.com/npjpcrm

ARTICLE OPEN

Check for updates

The unreliability of crackles: insights from a breath sound study using physicians and artificial intelligence

Chun-Hsiang Huang[1], Chi-Hsin Chen[1], Jing-Tong Tzeng[2], An-Yan Chang[3], Cheng-Yi Fan[1], Chih-Wei Sung[1,4], Chi-Chun Lee[2,3,6]✉ and Edward Pei-Chuan Huang[1,4,5,6]✉

BACKGROUND AND INTRODUCTION: In comparison to other physical assessment methods, the inconsistency in respiratory evaluations continues to pose a major issue and challenge.
OBJECTIVES: This study aims to evaluate the difference in the identification ability of different breath sound.
METHODS/DESCRIPTION: In this prospective study, breath sounds from the Formosa Archive of Breath Sound were labeled by five physicians. Six artificial intelligence (AI) breath sound interpretation models were developed based on all labeled data and the labels from the five physicians, respectively. After labeling by AIs and physicians, labels with discrepancy were considered doubtful and relabeled by two additional physicians. The final labels were determined by a majority vote among the physicians. The capability of breath sound identification for humans and AI was evaluated using sensitivity, specificity and the area under the receiver-operating characteristic curve (AUROC).
RESULTS/OUTCOME: A total of 11,532 breath sound files were labeled, with 579 doubtful labels identified. After relabeling and exclusion, there were 305 labels with gold standard. For wheezing, both human physicians and the AI model demonstrated good sensitivities (89.5% vs. 86.0%) and good specificities (96.4% vs. 95.2%). For crackles, both human physicians and the AI model showed good sensitivities (93.9% vs. 80.3%) but poor specificities (56.6% vs. 65.9%). Lower AUROC values were noted in crackles identification for both physicians and the AI model compared to wheezing.
CONCLUSION: Even with the assistance of artificial intelligence tools, accurately identifying crackles compared to wheezing remains challenging. Consequently, crackles are unreliable for medical decision-making, and further examination is warranted.

npj Primary Care Respiratory Medicine (2024)34:28 ; https://doi.org/10.1038/s41533-024-00392-9

INTRODUCTION

Breath sounds manifest with varying pitches, durations, and characteristics depending on the pathophysiologies affecting airflow within the respiratory tract. Abnormal breath sounds are observed in over 25% of adults upon auscultation, with prevalence increasing with age[1]. Auscultation thus serves as a valuable tool for diagnosing and assessing the disease severity in a real-time, non-invasive, and cost-effective manner. While studies have validated its reproducibility and reliability[2–4], the inter-observer agreement of breath sounds remains uncertain and heavily reliant on physicians' experience[5,6]. Additionally, physicians' preferences and auscultatory skills contribute to classification discrepancies[7,8]. Consequently, the clinical importance of conventional auscultation is waning. Luca Arts et al. conducted a meta-analysis on the diagnostic accuracy of lung auscultation, suggesting its limited contemporary role and advocating for its replacement with superior diagnostic modalities such as ultrasound or radiography[9].

Nevertheless, advancements in technology have revitalized auscultation by facilitating more consistent differentiation of breath sounds. The advent of digital stethoscopes offers enhanced resolution and noise cancellation. They inherently excel in sound acquisition compared to traditional bell and diaphragm stethoscopes, particularly for high-frequency sounds like wheezing[10,11]. Spectrograms provide visual assistance, enhancing breath sound classification and increasing inter-rater agreement[12,13]. Crucially, the rapid progress of machine learning has significantly enhanced the accuracy and objectivity of breath sound analysis[8,14].Through spectrogram and digital analysis, we gain a better understanding of adventitious sounds: wheezing presents as high-pitched and "musical" sounds, typically lasting more than 100 ms with a dominant frequency of 400 Hz or higher[15,16]. In contrast, crackles are discontinuous and explosive, with durations less than 20 ms and frequencies ranging from 60 to 2000 Hz[16,17].

It is widely acknowledged that human subjectivity impedes auscultation. Bohadana et al. observed that physicians' ability to describe lung sounds was superior for wheezes compared to crackles[7]. Whether the characteristics of adventitious sounds themselves contribute to classification difficulties remains unclear. Moreover, the robustness of deep learning models against different sound characteristics remains uncertain. Therefore, we established a database with breath sounds recorded in clinical field. By exploring the database with both conventional medical expertise and contemporary deep learning methodologies, our study aims to evaluate the difference in the identification ability of different breath sound.

METHODS

Study design and patient selection

This cross-sectional comparative study was conducted at the emergency department (ED) of the National Taiwan University

[1]Department of Emergency Medicine, National Taiwan University Hospital Hsin-Chu Branch, Hsinchu City, Taiwan, RO.C.. [2]College of Semiconductor Research, National Tsing Hua University, Hsinchu City, Taiwan, RO.C.. [3]Department of Electrical Engineering, National Tsing Hua University, Hsinchu City, Taiwan, RO.C.. [4]Department of Emergency Medicine, College of Medicine, National Taiwan University, Taipei City, Taiwan, RO.C.. [5]Department of Emergency Medicine, National Taiwan University Hospital, Taipei City, Taiwan, RO.C.. [6]These authors contributed equally: Chi-Chun Lee, Edward Pei-Chuan Huang. ✉email: cclee@ee.nthu.edu.tw; edward56026@gmail.com; edwardhuang@ntu.edu.tw

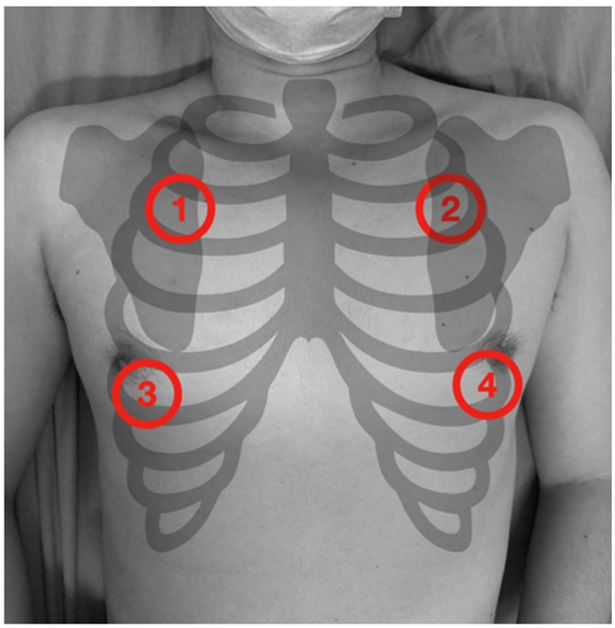

Fig. 1 The anterior view of the chest illustrating auscultation sites numbered 1 through 4 for data recording from patients. Auscultation recording was acquired at 4 sites of both lungs. The upper sites were located at the midclavicular line of the second intercostal space (area 1, 2), and the lower sites were at the anterior axillary line of the inferior scapular rim (area 3, 4).

Hospital Hsinchu Branch, a tertiary medical center with an average of 5000 monthly ED visits, between January 2021 and February 2022. Non-trauma patients aged over 20 years presenting to the ED were eligible for inclusion. Exclusion criteria comprised pregnant individuals, patients experiencing out-of-hospital cardiac arrest, those transferred to another medical facility, or those discharged against medical advice.

Outcomes

The outcome was the efficacy of the initial labeling physician and All-data AI model in the identifying different breath sounds. The sensitivity, specificity, and area under the receiver-operating characteristic (AUROC) curve were calculated. The definition of wheezing and crackles are inherently subjective. Hence, we considered there is still a gold standard but it should be determined by physicians with a majority rule.

Data collection

Patients' breath sounds were recorded at the ED with fidelity, including possible noises. Recording was performed using a CaRDIaRT Electronic Stethoscope DS101, with a frequency range from 0 to 8000 Hz. The 20-1000 Hz are specifically highlighted. We exported soundwaves into digital formats as ".wav" format at 16-bit depth, resampled all recordings into 16 kHz, and converted them into mel-spectrograms. A 10-second recording of breath sounds was obtained from four sites on both lungs. The upper sites were located at the midclavicular line of the second intercostal space, while the lower sites were located at the anterior axillary line of the inferior scapular rim (Fig. 1). Since each recording is exactly 10 seconds long, no partitioning or segmentation was performed.

The recordings were then uploaded to the online database "Formosa Archive of Breath Sound," with no post-processing or filtering applied. Each recording was assigned to a physician for labeling, who was blinded to the patient's clinical information. Physicians were allowed to replay the recordings multiple times and adjust the volume before labeling. Respiratory sound records from different chest locations of the same participant were presented concurrently during labeling. Breath sounds were categorized into five groups: normal, wheezing, crackles, unknown, and no breath sounds. Normal breath sounds were defined as unremarkable inspiration and expiration without adventitious sounds. Wheezing was characterized by high-pitched, "musical" sounds heard during either inspiration or expiration. Crackles were described as non-musical, brief, explosive sounds primarily occurring during inspiration[18]. Breath sounds were labeled as unknown if they could not be classified as wheezing or crackles but exhibited distinct inspiratory and expiratory phases. Recordings containing only ambient noise were labeled as no breath sounds. A pre-training course was provided for the labeling physicians to ensure inter-rater reliability. The pre-test demonstrated an acceptable Kappa value of 0.7 on the demo recordings.

Each breath sound was labeled by a single physician, with a total of five physicians involved in labeling the recordings. Additionally, six artificial intelligence (AI) breath sound interpretation models were developed: five AI models emulating the physicians, trained with their respective labeling data (referred to as AI doctors), and a final model trained with all available data (referred to as the All-data AI model). Every breath sound was then labeled by five AI doctors and the All-data AI model again. To ensure the robustness of labels in our database, labels were examined by three tests: physician's label, the All-data AI model's label and the majority opinion of the five AI doctors. Any discrepancy among these three tests would be considered as doubtful. Through above measures, we tried to identify as many doubtful labels as possible. Doubtful labels were then reassessed by two additional human physicians. The final label, determined by a majority rule among the three physicians (the initial labeling physician plus two additional physicians), served as the gold standard.

Model training

Based on the successful utilization of Mel spectrograms in breath sound classification as demonstrated in prior literature, all breath sound recordings underwent conversion into Mel spectrograms to facilitate efficient feature extraction and input[5,19,20]. Our model was constructed based on CNN14, which has exhibited promising performance in audio tagging tasks[21]. To further bolster performance and generalizability, we fine-tuned the network utilizing pre-trained weights from AudioSet, a comprehensive audio dataset comprising 2,063,839 training audio clips sourced from YouTube.

In order to tackle the challenges posed by data imbalance and scarcity, we implemented various data augmentation techniques, including SpecAugment and Mixup, alongside employing a batch-balancing strategy (Appendix 2). SpecAugment applies random time and frequency masks directly onto the Mel spectrograms, thereby enhancing model generalizability[22]. Additionally, Mixup blends two spectrograms in random ratios, effectively broadening the training distribution and resulting in enhanced performance[22]. Furthermore, the batch-balancing strategy mitigates data imbalance by oversampling the minority class within each batch[23]. By incorporating these methodologies, we optimized our model's accuracy and robustness in classifying respiratory sound recordings.

Statistical analysis

Dichotomous and categorical variables were presented as numbers (percentages). Sensitivity and specificity were calculated using standard formulas for a binomial proportion, with 95% confidence intervals (CIs) estimated using the Clopper-Pearson interval method. The calculation was performed with a "one v.s.

the others" approach (i.e., for wheezing, the calculation was performed comparing wheezing to non-wheezing cases). Comparison was conducted between the All-data AI model and the initial physician who labeled the breath sounds. Sensitivity and specificity comparisons were conducted using McNemar's test, as all breath sound files were labeled by both physicians and the All-data AI model. The performance in differentiating breath sounds was evaluated using the AUROC. Comparisons of AUROC values were performed using the Delong test. A p-value < 0.05 was considered statistically significant. Considering that breath sound recordings from the upper auscultation sites are clearer than those from the lower sites due to interference from the female breast, we performed the sensitivity analysis using recordings only from the upper chest. Moreover, a subgroup analysis by sex was also conducted. Finally, unweighted Cohen's kappa was measured to evaluate the agreement between physicians and the All-data AI model. All statistical analyses were conducted using the Statistical Package for the Social Sciences (SPSS), version 26.0 (IBM Corp., Armonk, NY, USA), Python, version 3.8 (Python Software Foundation), and R software, version 4.4.0 (R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

The Formosa Archive of Breath Sound has included 11,532 breath sound recordings at the time of writing, making it the largest breath sound database in Asia and one of the few databases dedicated to audio recording in clinical setting. Among them, there are 978 recordings of crackles, 277 recordings of wheezing, 4247 recordings of normal breath sound. Comparison of demographics between different breath sounds was not performed as a single patient could present with different breath sounds at various chest sites. Each of the 11,532 sound files was initially labeled by one of the five human physicians. Following the labeling by physicians and AI interpretation, 579 doubtful labels requiring further evaluation were identified. These doubtful labels were relabeled by two additional physicians, who were randomly selected from the rest four physicians. After a majority vote, those with undifferentiated or non-conclusive agreements were excluded. Ultimately, 305 final labels with definitive classifications (normal, wheezing, or crackles) were established as the gold standard. Among the 305 recordings, there were 199 patients and their characteristics are shown in Table 1. The median age was 68.8 years, and 105 (52.76%) patients were male. 19 (9.55%) patients have congestive heart failure. 11 (5.53%) patients have chronic obstructive pulmonary disease and 13 (6.53%) patients have asthma. The distribution of labels from physicians, the All-data AI model, and the gold standard are displayed in Fig. 2 and further illustrated in a confusion matrix (Supplementary Figure 1, Appendix 1). Notably, we found that many normal breath sounds were misclassified as crackles by both physicians and the All-data AI model.

Based on the 305 gold standard labels, a comparison of breath sound identification ability was conducted between the All-data AI model and the initial physician who labeled the breath sounds. For calculation of sensitivity and specificity, contingency table was provided (Supplementary Table 1, Appendix 1). The result was shown in Table 2: For wheezing, both human physicians and the All-data AI model exhibited good sensitivities (89.5% vs. 86.0%, $p = 0.480$) and good specificities (96.4% vs. 95.2%, $p = 0.248$). There was no significant difference observed in AUROC between the two (0.951 vs. 0.934, $p = 0.438$) (Fig. 3a). Regarding crackles, both human physicians and the All-data AI model demonstrated good sensitivities (93.9% vs. 80.3%, $p = 0.001$) but poor specificities (56.6% vs. 65.9%, $p = 0.023$). Again, there was no significant difference noted in AUROC between the two (0.728 vs. 0.721, $p = 0.168$) (Fig. 3b). For normal breath sound, both human physicians and the All-data AI model exhibited poor sensitivities (30.2% vs. 44.0%, $p = 0.020$) but good specificities (94.2% vs. 85.2%, $p = 0.002$). No significant difference in AUROC was observed between them (0.698 vs. 0.695, $p = 0.334$) (Fig. 3c).

Table 1. The demographics, vital signs, and laboratory data of the enrolled patients.

Variables	All patients ($n = 199$)
Age (years)	68.82 ± 16.03
Sex (male)	105 (52.76%)
BMI (%)	23.92 ± 4.68
Pre-existing Disease	
Hypertension	92 (46.23%)
Coronary artery disease	26 (13.07%)
Congestive heart failure	19 (9.55%)
Diabetic mellitus	56 (28.14%)
Chronic kidney disease	19 (9.55%)
Cerebrovascular accident	18 (9.05%)
COPD	11 (5.53%)
Asthma	13 (6.53%)
Lung cancer	15 (7.54%)
Other cancer	31 (15.58%)
Smoking status	
Never smoker	96 (48.24%)
Triage Vital Signs	
Body temperature	36.80 ± 0.90
Pulse rate	93.91 ± 25.65
Respiratory rate	22.10 ± 3.76
Systolic blood pressure	145.79 ± 35.32
Diastolic blood pressure	78.84 ± 17.59
SpO_2	94.95 ± 5.17
Laboratory Data	
White blood cell(K)	9.80 ± 5.63
Neutrophilic granulocyte (%)	74.38 ± 16.15
Hemoglobin (mg/dL)	11.80 ± 2.54
Creatinine (mg/dL)	1.61 ± 1.80
hsCRP (mg/dL)	6.05 ± 8.53
Lactic acid (mmol/L)	2.03 ± 1.57
NTproBNP (pg/mL)	4423.38 ± 8094.75

BMI body-mass-index, *COPD* chronic obstructive pulmonary disease, *hsCRP* high-sensitivity C-reactive protein, *NT-proBNP* NT-proB-type natriuretic peptide, *SpO2* oxyhemoglobin saturation by pulse oximetry.

The results of subgroup analysis by sex are presented in Supplementary Table 2a and 2b. For both male and female, subgroup analyses demonstrated a higher AUROC for wheezing compared to crackles, consistent with our main findings. No significant difference in AUROC was found between human physicians and the All-data AI model, except for crackle identification in female patients. Furthermore, after excluding breath sound recordings from the lower chest, 180 gold-standard recordings remained for the sensitivity analysis. The result was shown in Supplementary Table 3. The AUROC for wheezing identification remained above 0.9, while the AUROC for crackle identification stayed between 0.7 and 0.8. Lastly, Cohen's kappa was calculated to evaluate the agreement between human physicians and the All-data AI model. In wheezing identification, the Kappa value was the highest at 0.948 while for crackles and normal breath sound identification, the Kappa values were only

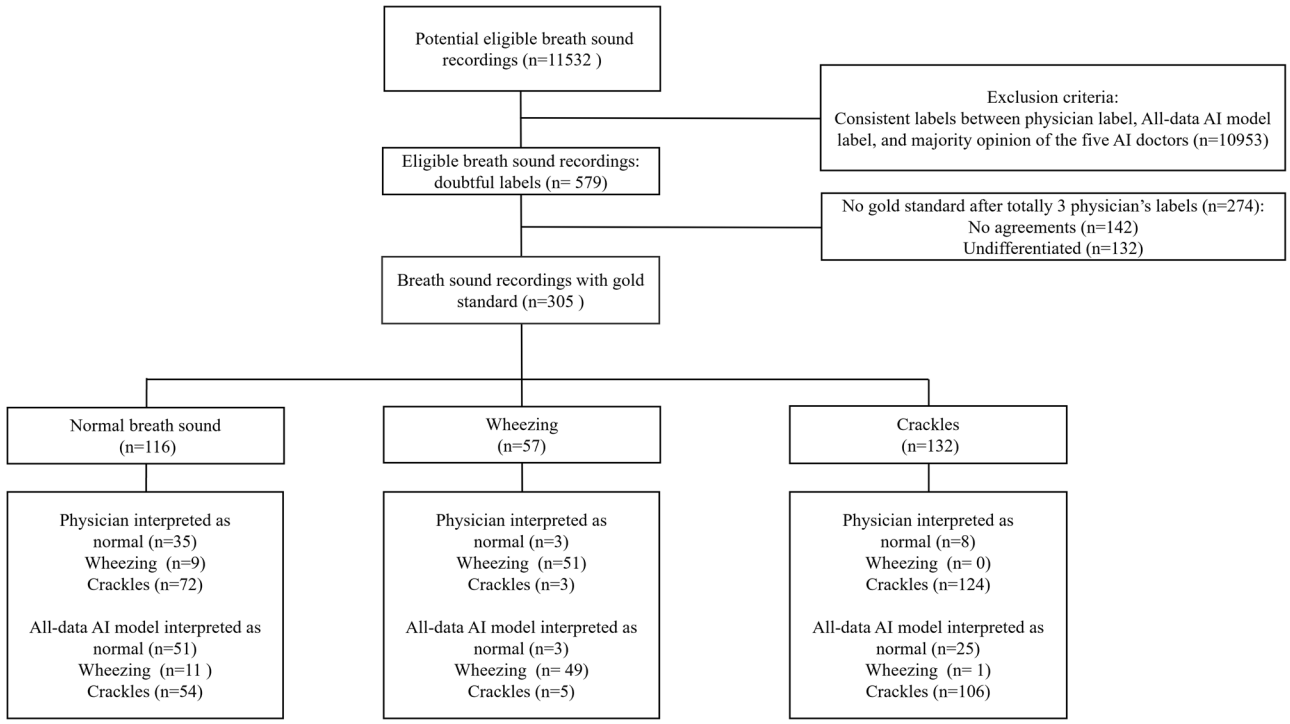

Fig. 2 Flow of data through the study.

Table 2. Comparison between human physician and All-data AI model in different breath sound identification.

	Wheezing			Crackles			Normal		
	Physician	All-data AI model	*p* value	Physician	All-data AI model	*p* value	Physician	All-data AI model	*p* value
Sensitivity	89.5 (78.5–96.0)	86.0 (74.2–93.7)	0.480	93.9 (88.4–97.3)	80.3 (72.5–86.7)	0.001	30.2 (22.0–39.4)	44.0 (34.8–53.5)	0.020
Specificity	96.4 (93.2–98.3)	95.2 (91.7–97.5)	0.248	56.6 (48.9–64.1)	65.9 (58.3–72.9)	0.023	94.2 (89.8–97.1)	85.2 (79.3–89.9)	0.002
AUROC	0.951 (0.920–0.972)	0.934 (0.901–0.959)	0.438	0.728 (0.674–0.777)	0.721 (0.667–0.771)	0.168	0.698 (0.643–0.749)	0.695 (0.640–0.746)	0.334

AUROC area under receiver operating characteristic curve.

fair to moderate (0.516 for crackles and 0.298 for normal breath sound) (Supplementary Table 4, Appendix 1).

DISCUSSION

Our study established the first breath sound database in the emergency department setting with clinical fidelity. Through double examination by both human experts and a deep learning model, the breath sound database was provided with relatively robust labels. Auscultation has long been criticized for its susceptibility to subjectivity and the observers' abilities, which limits its clinical utility[7,9,24]. While an increasing number of studies attempt to address these issues with artificial intelligence[25], it should be noted that different breath sounds not only influence human perception differently but also present varying complexities in signal processing and pattern recognition for machine learning. This study yielded two major findings: firstly, compared to wheezing, the identification of crackles proves to be more challenging and less prone to reaching a consensus. Secondly, despite employing multiple versions of deep learning models and adjustments in data sizes, the performance of crackle identification by deep learning still fell short compared to wheezing. Other auscultatory findings, such as heart murmurs, have been reported to exhibit good sensitivity and specificity, with AI achieving an AUC of 0.92 in detecting structural murmurs[26]. Therefore, despite a fair AUROC, crackles remain unreliable for its low specificity, low inter-rater agreement, and potential for confusion with normal breath sounds. Further examination is warranted for accurate diagnosis and proper management.

Our results are consistent with prior studies indicating that crackle identification is more inaccurate and unreliable than wheezing.[6,7,12,24,27]. The high-pitched musical tonal quality and longer duration of wheezing render it more distinctive and easier for human recognition compared to crackles, which are discontinuous and transient[7,28]. Additionally, louder background breath sounds also hinder the perception of crackles, particularly the coarse ones[29]. Previous studies have emphasized the importance of standardized terminology, auscultation training, and advanced equipment for improved breath sound classification[6,7,11,24,28]. A unified definition with common terminology can mitigate bias stemming from personal preferences or cultural differences, a factor crucial for crackles given its varied and vague manifestations.

In machine learning, our findings that wheezing identification rates surpass those of crackles are consistent with prior researches[5,30–32]. Emmanouilidou et al. developed signal

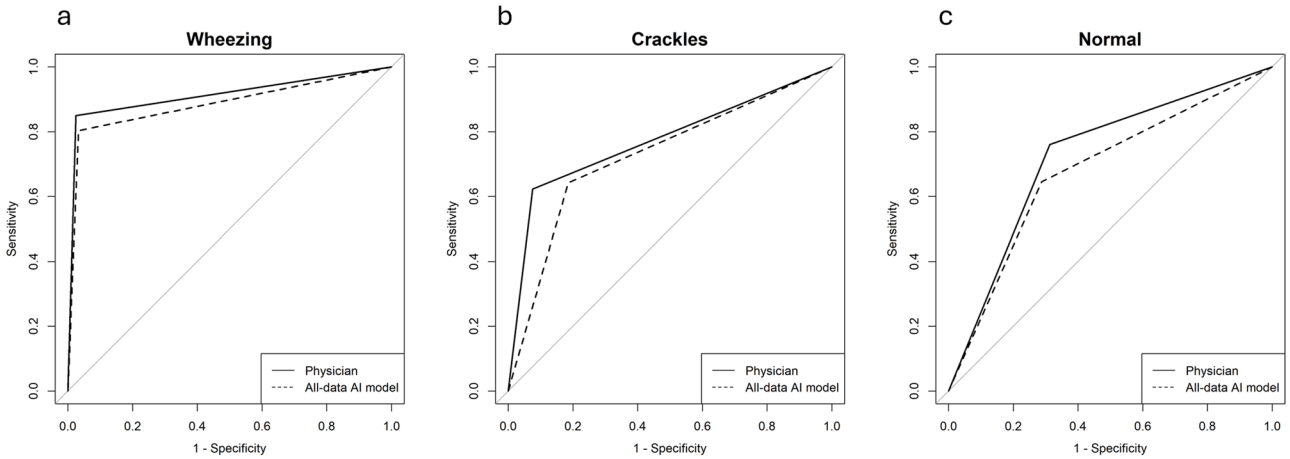

Fig. 3 Comparison of ROC curves between human and All-data AI model in different breath sound identification. No significant difference of AUROC was noted between physician and All-data AI model for wheezing, crackles and normal breath sound. 3a. Wheezing 3b. Crackles. 3c. Normal breath sound.

Table 3. Meaning and implications of the study.

	Wheezing	Crackles	Normal
For physician	The high sensitivity, specificity and AUROC rendered wheezing reliable for disease assessment, guiding management and clinical handover.	Normal breath sounds are often misclassified as crackles. The high sensitivity and low specificity of crackles suggest that physicians may have a low threshold for diagnosing pulmonary pathologies associated with crackles. However, the resulting low positive predictive value indicates that further examination is necessary before making management decisions and clinical judgments.	
For artificial intelligence	The unique sound feature of wheezing makes itself distinct to other sounds. It is easier to be recognized and for AI to learn.	The sensitivity and specificity of crackles and normal breath sounds are complementary due to frequent misclassification. Compared to wheezing, more effort is needed in AI training to improve the classification of crackles and normal breath sounds.	

AUROC area under receiver operating characteristic curve, *AI* artificial intelligence.

processing tools for the analysis of pediatric auscultations and found crackles to be more difficult to discriminate. From a signal processing perspective, the frequency of crackles is ill-defined and may overlap with wheezing. Its short duration comprises only a small proportion of the signal, rendering it susceptible to contamination by normal segments between crackles[32]. Previous studies have also indicated that crackles are more challenging to locate in the breath cycle or signal duration due to their brief existence. Shanthakumari's study confirmed that wheezing differs more from normal breath sounds compared to crackles in most first-order statistical features[33]. In summary, the difficulty in crackle detection stems from a low signal-to-noise ratio, crackle-like noise artifacts, and irregular loudness[34].

As shown in Table 2 and the confusion matrix, there were many normal breath sounds misclassified as crackles for both physicians and All-data AI model. This explained the high sensitivity of crackles and the high specificity of normal breath sounds. Other machine learning studies have also indicated that crackles are more likely to be confused with normal lung sounds[35,36]. Notably, our study recorded the breath sound with an electronic stethoscope. Peitao Ye, et al. had suggested that electronic stethoscope is prone to producing false crackles, potentially interfering with medical decision-making[37]. However, their definition of normal breath sounds was based solely on the fact that they came from healthy individuals, without further examination to rule out any pulmonary pathologies. Especially crackles is reported to be the most frequent adventitious sounds in healthy people[38]. Finally, the study recruited patients from ED. Since patients are assumed to be ill to visit the ED, there may be a subconscious effect that lower the physicians' thresholds to diagnose adventitious breath sound.

Our study has several limitations. Firstly, the inclusion of 579 breath sound files categorized as doubtful arose from discrepancies between physicians' labels and AI interpretations. Thus they were more ambiguous and much harder for classification essentially. Secondly, breath sound labeling was conducted by emergency physicians rather than pulmonologists. Although researches on the association between medical specialty and the ability to identify different breath sounds has yielded inconsistent results, we acknowledge that labeling physicians from a single specialty could be a limitation of our study, given their specific training and cultural practices[6,27,28,39]. Third, our breath sound files were recorded in ED environment and contaminated by ambient noise. Nevertheless, this setting offers a more clinically realistic scenario. Fourth, despite having fewer samples for wheezing, both human physicians and the All-data AI model demonstrated better wheezing recognition. This further strengthens our results, as fewer samples typically lead to poor machine learning training. Lastly, the use of Mel spectrogram as input may have contributed to the similar performance of the All-data AI model to human experts[16,40]. However, the current model was trained with our breath sound database, which comprises relatively large datasets and exhibited fair performance. Overall, while machine learning increasingly serves as a black box model for accomplishing more complex tasks, the decision-making process is often challenging for explanation and understanding by physicians[41]. Our study not only compared breath sound classification between humans and machines in different breath sound but also contributed to a better understanding of the black box model.

CONCLUSION

Both human physicians and deep learning model exhibited superior performance in identifying wheezing compared to crackles or normal breath sounds, indicating shared weaknesses in the classification of these categories. Therefore, medical decisions based on crackles should be made with caution and confirmed through additional examinations. For AI training, greater attention should be given to distinguishing between crackles and normal breath sounds (Table 3).

DATA AVAILABILITY

The datasets used during the current study are available from the corresponding author on reasonable request.

Received: 6 June 2024; Accepted: 6 October 2024;
Published online: 15 October 2024

REFERENCES

1. Aviles-Solis, J. C. et al. Prevalence and clinical associations of wheezes and crackles in the general population: the Tromsø study. *BMC Pulm. Med.* **19**, 1–11 (2019).
2. Sanchez, I. & Vizcaya, C. Tracheal and lung sounds repeatability in normal adults. *Respir. Med.* **97**, 1257–1260 (2003).
3. Jacome, C. & Marques, A. Computerized respiratory sounds are a reliable marker in subjects with COPD. *Respir. Care* **60**, 1264–1275 (2015).
4. Vyshedskiy, A., Ishikawa, S. & Murphy, R. L. Jr. Crackle pitch and rate do not vary significantly during a single automated-auscultation session in patients with pneumonia, congestive heart failure, or interstitial pulmonary fibrosis. *Respir. Care* **56**, 806–817 (2011).
5. Kim, Y. et al. Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning. *Sci. Rep.* **11**, 17186 (2021).
6. Hafke-Dys, H., Breborowicz, A., Kleka, P., Kocinski, J. & Biniakowski, A. The accuracy of lung auscultation in the practice of physicians and medical students. *PLoS One* **14**, e0220606 (2019).
7. Bohadana, A., Azulai, H., Jarjoui, A., Kalak, G., & Izbicki, G. Influence of observer preferences and auscultatory skill on the choice of terms to describe lung sounds: a survey of staff physicians, residents and medical students, *BMJ Open Respir. Res.* **7**, e000564 (2020).
8. Kim, Y. et al. The coming era of a new auscultation system for analyzing respiratory sounds. *BMC Pulm. Med.* **22**, 119 (2022).
9. Arts, L., Lim, E. H. T., van de Ven, P. M., Heunks, L. & Tuinman, P. R. The diagnostic accuracy of lung auscultation in adult patients with acute pulmonary pathologies: a meta-analysis. *Sci. Rep.* **10**, 7347 (2020).
10. Silverman, B. & Balk, M. Digital stethoscope-improved auscultation at the bedside. *Am. J. Cardiol.* **123**, 984–985 (2019).
11. Kevat, A. C., Kalirajah, A. & Roseby, R. Digital stethoscopes compared to standard auscultation for detecting abnormal paediatric breath sounds. *Eur. J. Pediatr.* **176**, 989–992 (2017).
12. Aviles-Solis, J. C., Storvoll, I., Vanbelle, S. & Melbye, H. The use of spectrograms improves the classification of wheezes and crackles in an educational setting. *Sci. Rep.* **10**, 8461 (2020).
13. Elphick, H. E. et al. Validity and reliability of acoustic analysis of respiratory sounds in infants. *Arch. Dis. Child* **89**, 1059–1063 (2004).
14. Palaniappan, R., Sundaraj, K. & Sundaraj, S. Artificial intelligence techniques used in respiratory sound analysis-a systematic review. *Biomed. Tech. (Berl.)* **59**, 7–18 (2014).
15. Meslier, N., Charbonneau, G. & Racineux, J. L. Wheezes. *Eur. Respir. J.* **8**, 1942–1948 (1995).
16. Faustino, P., Oliveira, J. & Coimbra, M. Crackle and wheeze detection in lung sound signals using convolutional neural networks. *Annu Int Conf. IEEE Eng. Med Biol. Soc.* **2021**, 345–348 (2021).
17. Sarkar, M., Madabhavi, I., Niranjan, N. & Dogra, M. Auscultation of the respiratory system. *Ann. Thorac. Med* **10**, 158–168 (2015).
18. Fouzas, S., Anthracopoulos, M. B., & Bohadana, A. Clinical usefulness of breath sounds. In *Breath Sounds: From Basic Science to Clinical Practice*, 33–52 (Springer, 2018).
19. Song, W., Han, J. & Song, H. Contrastive embeddind learning method for respiratory sound classification. in *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1275–1279 (IEEE, 2021).
20. Acharya, J., & Basu, A. Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning. *IEEE Trans. Biomed. Circuits Syst.* **14**, 535–544 (2020).
21. Kong, Q, et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE Signal Process Syst.* **28**, 2880–2894 (2020).
22. Park, D. S. et al. SpecAugment: A simple data augmentation method for automatic speech recognition. In *INTERSPEECH 2019*. 2019.
23. Shimizu, R. et al. Balanced mini-batch training for imbalanced image data classification with neural network. In *2018 First International Conference on Artificial Intelligence for Industries (AI4I)*, 2018, pp. 27–30. IEEE.
24. Mehmood, M., Abu Grara, H. L., Stewart, J. S., & Khasawneh, F. A. Comparing the auscultatory accuracy of health care professionals using three different brands of stethoscopes on a simulator. *Med. Devices (Auckl).* **7**, 273–281 (2014).
25. Gurung, A., Scrafford, C. G., Tielsch, J. M., Levine, O. S. & Checkley, W. Computerized lung sound analysis as diagnostic aid for the detection of abnormal lung sounds: a systematic review and meta-analysis. *Respir. Med.* **105**, 1396–1403 (2011).
26. Prince, J. et al. Deep learning algorithms to detect murmurs associated with structural heart disease. *J. Am. Heart Assoc.* **12**, e030377 (2023).
27. Mangione, S. & Nieman, L. Z. Pulmonary auscultatory skills during training in internal medicine and family practice. *Am. J. Respir. Crit. Care Med.* **159**, 1119–1124 (1999).
28. Moriki, D. et al. Physicians' ability to recognize adventitious lung sounds. *Pediatr. Pulmonol.* **58**, 866–870 (2023).
29. Kiyokawa, H., Greenberg, M., Shirota, K. & Pasterkamp, H. Auditory detection of simulated crackles in breath sounds. *Chest* **119**, 1886–1892 (2001).
30. Palaniappan, R., Sundaraj, K., & Lam, C. Reliable system for respiratory pathology classification from breath sound signals, in *2016 International Conference on System Reliability and Science (ICSRS)*, 152–156 (IEEE, 2016).
31. Chamberlain, D., Kodgule, R., Ganelin, D., Miglani, V., & Fletcher, R. R. Application of semi-supervised deep learning to lung sound analysis. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 804–807. (IEEE, 2016).
32. Emmanouilidou, D., Patil, K., West, J., & Elhilali, M. A multiresolution analysis for detection of abnormal lung sounds. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 3139–3142 (IEEE, 2012).
33. Shanthakumari, G. & Priya, E. Interpretation of lung sounds using spectrogram-based statistical features. In *International Conference on Futuristic Communication and Network Technologies*, 815–823. (Springer, 2020).
34. Grønnesby, M., Solis, J. C. A., Holsbø, E., Melbye, H., & Bongo, L. A. Bongo, Feature extraction for machine learning based crackle detection in lung sounds from a health survey. 2017.
35. Serbes, G., Ulukaya, S., & Kahya, Y. P. An automated lung sound preprocessing and classification system based onspectral analysis methods. In *Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18–21 November 2017*, 45–49 (Springer, 2018).
36. Park, J. S. et al. A machine learning approach to the development and prospective evaluation of a pediatric lung sound classification model. *Sci. Rep.* **13**, 1289 (2023).
37. Ye, P. et al. Regularity and mechanism of fake crackle noise in an electronic stethoscope. *Front. Physiol.* **13**, 1079468 (2022).
38. Oliveira, A. & Marques, A. Respiratory sounds in healthy people: a systematic review. *Respiratory Med.* **108**, 550–570 (2014).
39. Aviles-Solis, J. C. et al. International perception of lung sounds: a comparison of classification across some European borders. *BMJ Open Respiratory Res.* **4**, e000250 (2017).
40. X. Li, G. A. Ng, & F. S. Schlindwein, transfer learning in heart sound classification using mel spectrogram. In *2022 Computing in Cardiology (CinC)*, vol. 498, 1–4. (IEEE, 2022).
41. Loyola-Gonzalez, O. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*. **7**, 154096–154113 (2019).

ACKNOWLEDGEMENTS

This work was supported by Taiwan National Funds through National Science and Technology Council (grant number MOST 111-2320-B-002-054 and NSTC 112-2320-B-002 -044).

AUTHOR CONTRIBUTIONS

EPC Huang and C.C Lee conceived the study, designed the trial, and obtained research funding. EPC Huang, C.C Lee and C.H Sung supervised the conduct of the trial and data collection. C.Y Fan, C.H Chen, J.T Tzeng and A.Y Chang undertook recruitment of participating centers and patients and managed the data, including quality control. C.H Huang drafted the manuscript and conducted the revision. EPC Huang and C.C Lee take the responsibility for the paper as a whole. All authors read and approved the final manuscript.

COMPETING INTERESTS

The authors declared no competing interests.

ETHICS APPROVAL

The study was approved by the Institution Review Board of the National Taiwan University Hospital Hsinchu Branch (no. 109-129-E). Informed consents were written and obtained after explanation of all aspects of the study.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41533-024-00392-9.

Correspondence and requests for materials should be addressed to Chi-Chun Lee or Edward Pei-Chuan Huang.