scientific reports

Check for updates

OPEN

An explainable and accurate transformer-based deep learning model for wheeze classification utilizing real-world pediatric data

Beom Joon Kim[1,6], Jeong Hyeon Mun[2,6], Dae Hwan Hwang[2], Dong In Suh[3], Changwon Lim[2,4,7]✉ & Kyunghoon Kim[3,5,7]✉

Auscultation is a method that involves listening to sounds from the patient's body, mainly using a stethoscope, to diagnose diseases. The stethoscope allows for non-invasive, real-time diagnosis, and it is ideal for diagnosing respiratory diseases and first aid. However, accurate interpretation of respiratory sounds using a stethoscope is a subjective process that requires considerable expertise from clinicians. To overcome the shortcomings of existing stethoscopes, research is actively being conducted to develop an artificial intelligence deep learning model that can interpret breathing sounds recorded through electronic stethoscopes. Most recent studies in this area have focused on CNN-based respiratory sound classification models. However, such CNN models are limited in their ability to accurately interpret conditions that require longer overall length and more detailed context. Therefore, in the present work, we apply the Transformer model-based Audio Spectrogram Transformer (AST) model to our actual clinical practice data. This prospective study targeted children who visited the pediatric departments of two university hospitals in South Korea from 2019 to 2020. A pediatric pulmonologist recorded breath sounds, and a pediatric breath sound dataset was constructed through double-blind verification. We then developed a deep learning model that applied the pre-trained weights of the AST model to our data with a total of 194 wheezes and 531 other respiratory sounds. We compared the performance of the proposed model with that of a previously published CNN-based model and also conducted performance tests using previous datasets. To ensure the reliability of the proposed model, we visualized the classification process using Score-Class Activation Mapping (Score-CAM). Our model had an accuracy of 91.1%, area under the curve (AUC) of 86.6%, precision of 88.2%, recall of 76.9%, and F1-score of 82.2%. Ultimately, the proposed transformer-based model showed high accuracy in wheezing detection, and the decision-making process of the model was also verified to be reliable. The artificial intelligence deep learning model we have developed and described in this study is expected to help accurately diagnose pediatric respiratory diseases in real-world clinical practice.

Keywords Artificial Intelligence, Classification, Deep learning, Respiratory sound, Wheezing

Abbreviations

AI	artificial intelligence
AST	audio spectrogram transformer
CNN	convolutional neural network
AUC	area under the curve
ICBHI	International Conference on Biomedical and Health Informatics
MLP	multilayer perceptron

[1]Department of Pediatrics, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea. [2]Department of Statistics and Data Science, Chung-Ang University, Seoul, Republic of Korea. [3]Department of Pediatrics, Seoul National University College of Medicine, Seoul, Republic of Korea. [4]Department of Applied Statistics, Chung-Ang University, Seoul, Republic of Korea. [5]Department of Pediatrics, Seoul National University Bundang Hospital, Seongnam, Republic of Korea. [6]Beom Joon Kim and Jeong Hyeon Mun are contributed equally. [7] Changwon Lim and Kyunghoon Kim are jointly supervised this work. ✉email: clim@cau.ac.kr; journey237@snu.ac.kr

ResNet	residual network

Lung disease is an extremely common problem worldwide and the third-most-common cause of death. Physicians rely upon a variety of strategies to diagnose lung disease, including arterial blood gas analysis, spirometry, and radiologic imaging, but a stethoscope—as a simple, non-invasive method—remains the best tool for diagnosis. The stethoscope is also more economical and safer than any other method[1,2]. However, the stethoscope is not an objective test method, as it requires subjective interpretation by the clinician, and it therefore has a limitation in that the accuracy may vary considerably depending on the doctor's experience and judgment. Doctors must be well trained to accurately distinguish the characteristics of respiratory sounds using a stethoscope[3].

Moreover, traditional stethoscopes have additional limitations that complicate their use in modern medical practice[4]. Specifically, they cannot be utilized for remote treatments, because physicians must physically place the stethoscope on the patient's body to perform auscultation. Recently, spurred by global health crises such as the COVID-19 pandemic and significant regional disparities in medical access, the demand for remote medical services has surged, highlighting the necessity of innovative auscultation techniques. Artificial intelligence-enhanced stethoscopes offer the capabilities of real-time monitoring and can be integrated into wearable devices, allowing for the continuous monitoring of patients with chronic or critical conditions[4]. The development of high-performance AI stethoscopes is thus crucial for advancing modern medicine, as it would allow for more adaptable and accessible diagnostics in diverse medical environments.

Wheezes are high-pitched, continuous adventitious sounds caused by airflow limitation due to narrowed or obstructed airways[5]. They typically have a frequency range of 100 to 5000 Hz, last for at least 80 to 100 milliseconds, and exhibit a sinusoidal pattern in sound analysis. In contrast, rhonchi, which are also continuous adventitious sounds, are low-pitched and are associated with the accumulation of mucus in larger airways, generally featuring a dominant frequency below 200 Hz[6]. From a machine learning perspective, analyzing these respiratory sounds involves two key aspects: developing predictive models using traditional machine learning methods (e.g., support vector machine (SVM), artificial neural network (ANN)) and advanced deep learning architectures (e.g., convolutional neural networks (CNN), residual networks (ResNet)), as well as extracting relevant features that describe sound characteristics (e.g., Mel-frequency cepstrum coefficient (MFCC), singular spectrum analysis (SSA)) from given data[4].

Recent research has actively studied deep learning models that can classify normal and abnormal respiratory sounds with good performance[4,5]. However, many studies have used open databases, and there have not been many studies verifying the deep learning models that have been designed to classify respiratory sounds in actual clinical situations[7,8]. There have also been few studies examining pediatric patients, and the existing studies have either included too few subjects or too few specific types of abnormal respiratory sounds[2,9]. Moreover, most of the previous studies used Convolution Neural Network (CNN)-based models[7–10]. The CNN-based model adds a pooling layer after the convolution layer to increase computational efficiency. However, the pooling layer has the disadvantage of losing important information and not encoding the relative spatial relationship between feature maps.

A recent study proposed a new application of the Audio SpectrogramTransformer (AST) model to overcome these shortcomings and improve audio classification performance in breath sound classification[11]. In classifying crackle and wheezing using a publicly available adult breathing sound dataset, the proposed model outperformed the most recent CNN model[12,13]. The Transformer model converts all the input data to the same vector dimension regardless of its initial size, so there is no need to resize the input data even if it is large. Moreover, there is no problem with information loss because the dimension of the embedding vector is processed in the same dimension without the need for dimension reduction during the internal operation of the Transformer. We hypothesize that the longer the overall length and context required for accurate interpretation, the more suitable the Transformer model is for classifying respiratory sounds than the CNN model.

The purpose of the current study is to develop Transformer-based AI algorithms that can be applied to real-world clinical settings for the detection of wheezing and abnormal respiratory sounds in pediatric clinical practice. We also aim to compare the performance of these developed algorithms with the CNN model developed in the primary our study[10]. Finally, we intended to achieve optimal performance by applying the AST, which is the first convolution-free, purely attention-based model for audio classification[11].

Methods

Study design and data collection

This prospective study included children who visited the Department of Pediatrics at two university hospitals in Korea from 2019 to 2020. We recorded respiratory sounds from the patients who provided voluntary consent for such sounds to be recorded. The recordings were taken in an outpatient clinic by a pediatric pulmonologist using an electronic stethoscope (Jabes, GSTechnology, Seoul, Korea). The recorded auscultation sounds were classified as wheezing and other respiratory sounds based on the diagnosis made by the specialist. Four respiratory sounds were obtained from each patient by recording the anterior and posterior regions of both lungs for two cycles each. To verify the classifications, blinded validation was performed by two pediatric pulmonologists, and if one or more classifications matched the existing classification, they were tagged and stored in the database.

Evaluation of AI algorithm

We constructed a binary classification model to determine whether the recorded respiratory sounds contained wheezing sounds. We used 80% of the database as training data and 20% as test data. The mel spectrograms extracted from the audio data during the pre-processing process were used as input data. Our deep learning model was a pre-trained AST model containing a total of over 1 million data ImageNetconsisting of 1,000 classes[11]. We trained our data through this model and the multilayer perceptron (MLP) layer (Fig. 1). First,

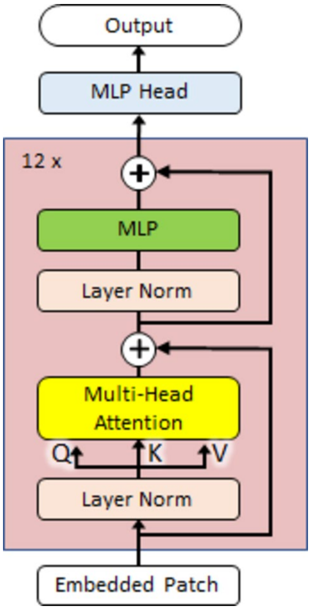

Fig. 1. Architecture of th*e AST model. After embedding the audio spectrogram into patches, we perform positional embedding to provide patch location information. It passes through a transformer encoder driven by a self-attention mechanism composed of a multi-head, and it calculates the final output through the MLP head.

we changed the audio input to 128-dimensional log mel filter bank (fbank), and the spectrogram was divided into 16 × 16 patches. Using a linear projection layer, each patch was flattened into a 1D vector with a size of 768. Positional embedding was then added. The outputs were (batch size × specified number of classes) and these were used to determine whether or not wheezing was present. Our experiments were conducted using Python version 3.6.5. We used the following AST Model structures to deal with the respiratory sounds in our applied model:

Audio Spectrogram Transformer

The AST model is a deep learning model that has a transformer structure through which it classifies audio. This model showed good performance on the task of classifying a dataset comprising 10-second audio clips into 527 classes. This model divides the mel spectrogram into patches, uses a linear projection layer to flatten each patch, and adds positional encoding to provide patch location information in the spectrogram. In the Transformer encoder, a multi-head self-attention mechanism is applied, and there are various performance differences depending on the patch size, the use of positional embeddings, and the pre-training[11]. This model also has the advantage of easy transfer learning.

Transformer uses self-attention, in which the model generates attention values between words in a sentence. The attention value is calculated through the scaled dot-product attention layer. In order to analyze the relationship between patches, the input values are composed of query, key, and value. All keys and queries are dot-multiplied, and the weight is obtained by applying the softmax function to the result. The equation for the attention score is as follows:

$$softmax\left(X\right)=e^{X}/\sum\nolimits_{n=1}^{N}e^{x_n},$$

$$Q,K,V=\ zU_Q,zU_K,zU_V$$

$$Attention\left(Q,K,V\right)=softmax\left(\frac{QK^T}{\sqrt{(d_k)}}\right)V$$

Softmax is a function used for multi-class classification that ensures the sum of the predicted probabilities for each class equals 1 for the input value.

Where N is the number of patches, X is the input matrix, and x_n is the n-th input value. Q, K, and V are matrices representing the query, key, and value, respectively. $\mathrm{U_Q}$, U_K, U_V are the weights of the hidden layers for the query, key, and value. $\mathrm{d_k}$ is the dimensionality of query and key.

The encoder consists of 12 layers, each composed of multi-head attention, MLPs, and residual connections. Multi-head attention stacks several scaled dot-product attention layers and processes them in parallel. This enables analysis using various criteria through multiple heads simultaneously. The equation for passing through a multi-head attention composed of n attention layers is as follows:

$$MultiHead\left(Q,\ K,\ V\right)=Concat\left(head_1,\ldots\ ,head_n\right)W^{\mathrm{O}},$$

$$where\ head_i=Attention\left(QW_i^Q,\ KW_i^K,VW_i^V\right).$$

$QW_i^Q,\ KW_i^K,VW_i^V$ represent the weights of the hidden layers for the ***i***-th head corresponding to the query, key, and value, respectively. $\boldsymbol{W^O}$ denotes the weight of the final hidden layer in the Multi-Head Attention module.

Pre-processing

1) Data augmentation: We augmented the training data by applying the following six augmentation techniques: white noise addition, time shifting, stretching, reverse, minus, and fit transform (Supplement A). The librosa package was used to augment 580 pieces of training data into 4,060 pieces and then extract a mel spectrogram[13].
2) Feature Extraction: Mel spectrograms were extracted from the audio data of respiratory sounds. In this process, a 1024-point fast Fourier transform processor and a 64-bit mel filter bank were implemented. We converted all sound data to 44,100 Hz. Since 1-channel data and 2-channel data were mixed together, all data were 2-channelized to avoid data loss. We performed repeat padding for sound data of variable lengths. Shorter samples were repeated from batch to batch with the maximum sample size. The torchaudiopackage had been used in a previous work[14]. Moreover, time masking and frequency masking of SpecAugment were performed to avoid overfitting the train data to the model[14].

Optimal construction and validation of AI model

We used the cross-entropy loss function and Adam optimizer for deep learning. To identify the optimal hyper-parameters, we used five-fold cross-validation and the grid search method[15] (Supplement B). The applied model was learned over 150 epochs, the batch size was 10, and the learning rate was 0.000005. Table 1presents the hyper-parameters of all the models, including those selected for comparison. We applied the stochastic weight average technique, which updates the weight's average value every cycle to further boost the performance[16]. We evaluated the performance of the model using a test dataset; in this process, we obtained accuracy, precision, recall, F1-score, and area under the curve (AUC) values. We compared the performance of the proposed model with the following models: the model in the primary study[10], the AST model using previous study data, and ResNet34 + CBAM of this study data. We used a PyTorch framework that is compatible with the torchaudio used in pre-processing to prepare the deep-learning process.

Validation of AI model using Score-CAM

Score-CAM[17] is a tool to visualize how well a model makes image classification predictions, and it visualizes where in an image a model is particularly active. Since the proposed model conducts its analysis based on mel spectrogram images, we use Score-CAM to visualize the classification results and determine what noise affects the model and whether the model makes accurate predictions based on wheezing. The target layer is the first normalized layer from the last attention block.

Statistical analysis

To compare the characteristics of the recorded respiratory sounds, we used the Mann–Whitney U test. We also used box plots, histograms, and quartiles to compare the lengths of breathing sounds. Our metrics include accuracy, precision, recall, F1-score, and the area under the curve (AUC). Accuracy, precision, recall, and F1-score are calculated as follows:

Accuracy = $\frac{TP+TN}{N}$

Recall = $\frac{TP}{TP+FN}$

Precision = $\frac{TP}{TP+FP}$

F1 = $2\times\ \frac{recall\times\ precision}{recall+precision}$

N is number of sample, *TP* is number of true positive, *TN* is number of true negative, *FN* is number of false negative, *FP* is number of false positive predicted by the model. The AUC is a widely used metric for evaluating the performance of binary classification models. It measures the area under the Receiver Operating Characteristic (ROC) curve, providing a single scalar value that represents the model's ability to distinguish between positive and negative classes.

Models	Selected hyper-parameters with grid search	Accuracy	AUC	Precision	Recall	F1-score
Primary study results(ResNet34 + CBAM)	Epoch: 120 / Batch Size: 32 / Learning Rate: 1e-3	0.912	0.891	0.944	0.810	0.872
AST model using primary study data	Epoch: 100 / Batch Size: 16 / Learning Rate: 1e-4	0.930	0.944	0.840	1.000	0.913
ResNet34 + CBAM of follow-up study data	Epoch: 120 / Batch Size: 32 / Learning Rate: 1e-3	0.836	0.758	0.742	0.590	0.657
AST model using follow-up study data	Epoch: 150 / Batch Size: 10 / Learning Rate: 5e-6	0.911	0.866	0.882	0.769	0.822

Table 1. Performance for discriminating other respiratory sounds from wheezing. Abbreviations: AUC, area under the curve; CBAM, convolutional block attention module; ResNet, residual network; AST, audio spectrogram transformer.

ROC curve illustrates the trade-off between the true positive rate and the false positive rate at various classification thresholds. This provides a comprehensive view of the model's performance across all possible thresholds. And AUC is Area Under the ROC curve and is unbiased toward models that perform well on the minority class at the expense of the majority class—a property that is particularly beneficial when dealing with imbalanced data[18].

Ethics statement

This study was approved by the Institutional Review Boards (IRBs) of the Catholic University of Korea (IRB approval no. PC19OESI0045) and Seoul National University of Korea (IRB approval no. H-1907-050-1047). Written informed consent was obtained from at least one legal guardian for all participants. For children 7 years of age and older, the assent of the child was also obtained. All methods were performed while following the relevant guidelines and regulations.

Results

Characteristics of the respiratory sound database

In total, 194 wheeze sounds (11.85 ± 5.33 s) and 531 other respiratory sounds (11.64 ± 5.71 s) were collected. There were two sampling rates for the recorded files: 48,000 Hz and 44,100 Hz. To integrate these sampling rates, we down-sampled all the recorded files at 44,100 Hz. There were also two types of channels among the recorded files: Mono and Stereo. We converted all the sound data into 2-channels. The respiratory sounds had different lengths and were set equally to 21.72 s through repeat padding.

Performance of proposed model and comparison with other experiment models

Table 1. presents the performance comparison between the proposed model and other experimental models such as the model in the primary study, the AST model using the primary study data, and ResNet34 + CBAM of the follow-up study data model. The model in the primary study showed an accuracy of 91.2%, AUC of 89.1%, precision of 94.4%, recall of 81%, and F1-score of 87.2%. The AST model using primary study data showed an accuracy of 93%, AUC of 94.4%, precision of 84%, recall of 100%, and F1-score of 91.3%. The ResNet34 + CBAM of the follow-up study data model showed an accuracy of 83.6%, AUC of 75.8%, precision of 74.2%, recall of 59%, and F1-score of 65.7%.

The applied ImageNet pre-trained AST model outperformed the primary study model. The model had an accuracy of 91.1%, AUC of 86.6%, precision of 88.2%, recall of 76.9% and F1-score of 82.2%.

Validation results of proposed model using Score-CAM

Figure 2 presents a Score-CAM result showing a good example of a correct classification, while Fig. 3 shows a result of an incorrect classification.

Discussion

We constructed an AI model using the ImageNet pre-trained AST model to classify wheezes from recorded pediatric breath sounds. This model exhibited a high accuracy of 91.1% and an F1 score of 82.2%. It therefore outperformed both our previous ResNet-based model and prior models based on CNNs[10]. Another strength of our study is that it used breath sound data collected from pediatric patients in an actual clinical environment.

The advent of digital stethoscopes has facilitated the development of various machine learning methods that can overcome the limitations of traditional stethoscopes by presenting objective and quantitative results[19–21]. Shallow machine learning-based methods for lung sound classifications, such as SVM, KNN, and ANN, have only achieved around 80% accuracy[22]. However, studies involving pediatric patients have been fewer and have shown lower performance compared to those involving adults. Zhang et al.'s SVM-based model showed superior accuracy in detecting abnormal lung sounds compared to pediatricians[2]. However, its accuracy in wheeze classification was only 59.9%. Deep learning-based methods learn without manual feature extraction in an end-to-end learning method, and a reduced amount of data is required for pre-training through transfer learning[22,23]. A customized deep learning model can be constructed according to the specific input structure[22]. To date, various artificial intelligence breathing sound classification deep learning models have been developed, such as CNN, RNN, and FNN[5,24,25]. Among these, a model based on CNN, which extracts and analyzes features from the Mel spectrogram, was presented as the most basic and suitable model for classifying abnormal breathing sounds[26].Various CNN-attention hybrid models have recently been proposed, and these have shown advanced performance[27,28]. Deep learning applications in studies involving pediatric patients have also been fewer compared to those focused on adults, often facing limitations such as smaller sample sizes. Our previous research also developed a deep learning AI model for classifying wheezing in children using a convolutional block attention module in a CNN-based residual network structure, and this model showed an accuracy of 91.2%[10]. However, previous models tend to easily overfit and also require large amounts of training data[14]. Because CNN has the limitation of having to extract features close to the pixel reference through a filter of a certain size, it is necessary to use a deeper network or a larger convolution kernel to learn long-length information[12,29]. Recurrent Neural Networks (RNNs) are vulnerable to gradient vanishing and gradient explosion during training, making it difficult to efficiently update weights in RNNsusing training data[12,30].

Recently, the AST model has been proposed to overcome the limitations of existing lung sound classification models and ultimately improve audio classification performance[12]. AST is a pure attention-based audio classification model without convolution that can be directly applied to audio spectrogramsand capture long-range global context even in the lowest layers[11]. One drawback of a Transformer compared to CNN is that it requires more data for training, but as a pre-trained model was proposed on ImageNet, this drawback was

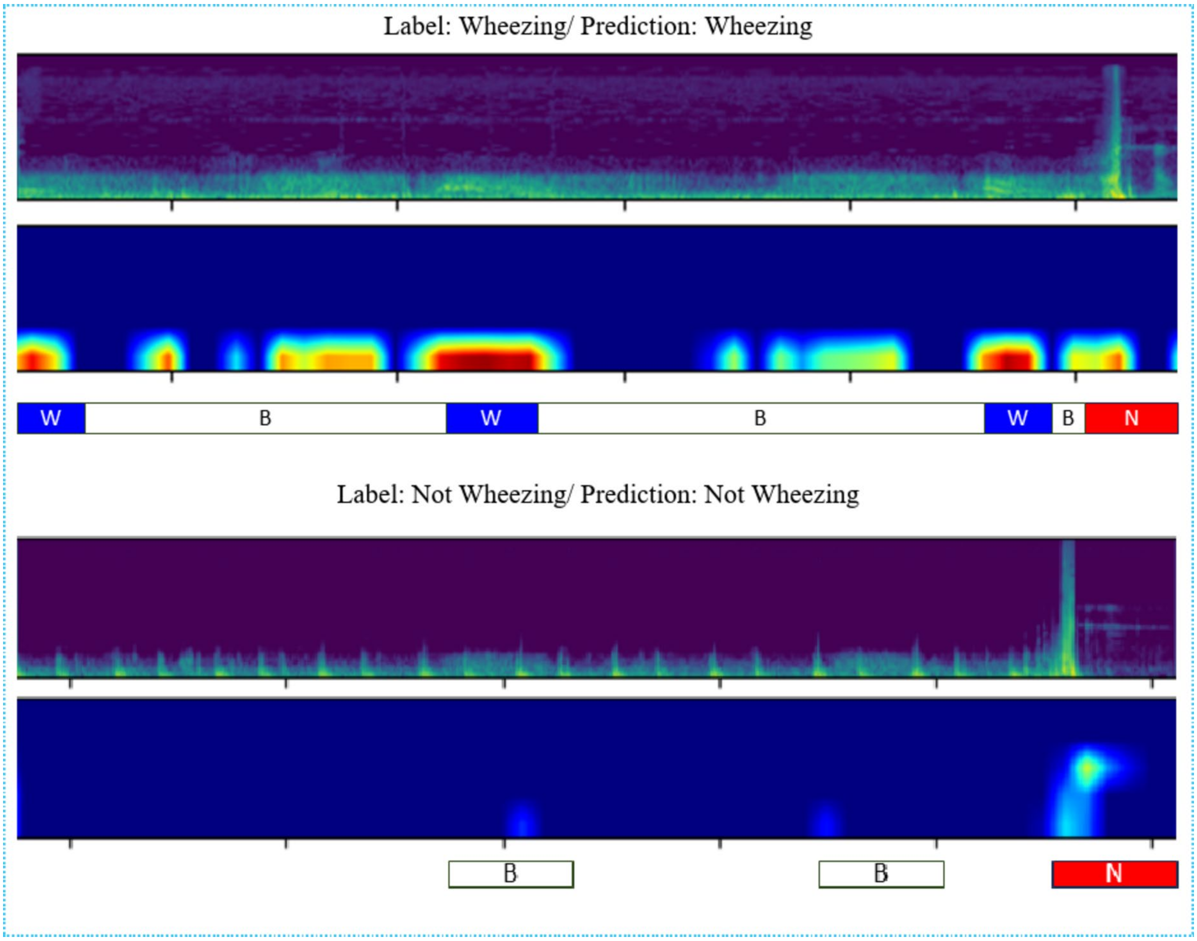

Fig. 2. Examples of well-classified samples with Score-CAM. Notations: B - Normal Breath, W - Wheezing, N - Noise.

compensated for, and it was confirmed to have higher performance than existing state-of-the-art models[11,31]. Moreover, because AST compresses all the information through self-attention, it can capture the global characteristics of the data well[12,31]. Therefore, it naturally supports variable-length input and can be applied to a variety of tasks without having to change the architecture[32]. This gives the model the advantage of classifying breath sounds of various lengths, making it highly useful in actual clinical practice. Although AST is the model that was most recently proposed in the field of audio classification, there has yet to be substantial research applying it to lung sound classification specifically. We applied the AST model pre-trained on ImageNet to classify pediatric wheezes. Our proposed model showed a high accuracy of 91.1% and an F1 score of 82.2% on our pediatric breath sounds dataset. This showed improvements of 7.5% in accuracy and 16.5% in F1 score compared to the classification of the same data with the ResNet34-based model from the previous study. Even when evaluated on our previous research dataset, the AST model showed performance improvements of 1.8% in accuracy and 4.1% in F1 score.

Children's breath sounds have different characteristics than those of adults. For example, children's respiratory cycles are significantly faster than those of adults, and there are various deviations from the normal range depending on age[21]. Moreover, children have small rib cages and relatively large hearts, so there is notable interference with heart sounds[21,33]. Crying or other noises may also be auscultated. Auscultation is considered to have more clinical importance in children than it can in adults[2,9]. In general, respiratory diseases are more common in children, and rapid assessments of severity are essential, and it is also crucial to minimize radiation exposure or invasive tests[2]. However, because it may be difficult to have children cooperate with assessments, it may be difficult to obtain good quality breath sound samples[21]. Therefore, open databases often contain few or no pediatric breath sounds[26,34]. Further, there have been very few pediatric breath sound classification models using artificial intelligence compared to the importance of this task[9,35]. The Convolutional Recurrent Neural Network (CRNN) model, which combines CNN and RNN, showed higher accuracy than doctors in wheezing classification with an F1 score of 66.4%, but the performance was not overwhelmingly high[35]. Other researchers have confirmed a positive percent agreement of 0.90 for wheezing detection using the same model[9]. However, the dataset used to construct that model only included 40 wheezes. A recent study showed that an SVM-based model achieved higher accuracy than general pediatricians, with wheeze classification accuracy of

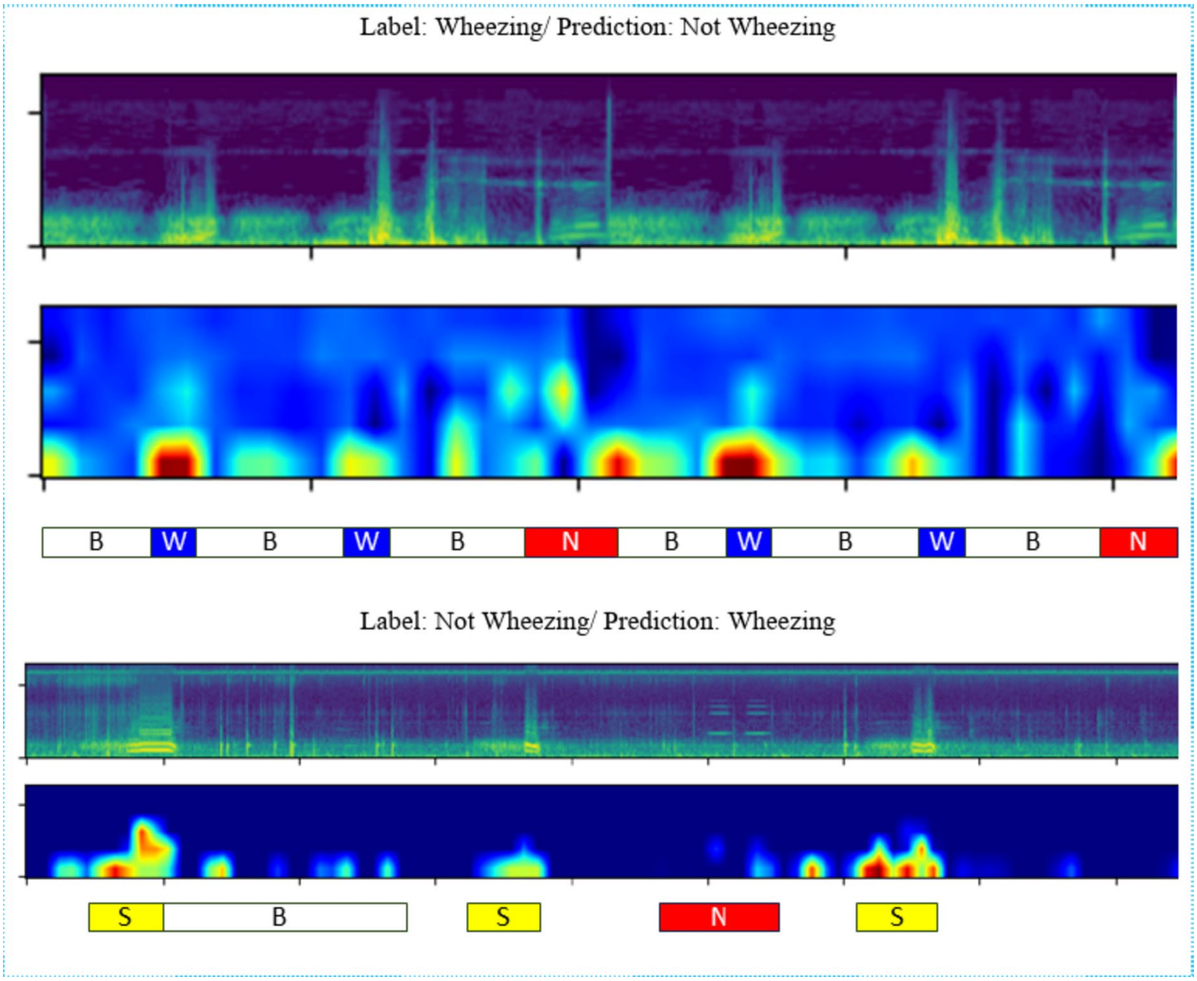

Fig. 3. Examples of misclassified samples with Score-CAM. B: Normal Breath, W: Wheezing, N: Noise, S: Strange sound like Wheezing.

90.42%[21]. However, more recent studies in children collected fewer wheezes than our study and showed poorer performance.

We applied the Score-Class Activation Map (CAM) to interpret the mechanism by which our model classifies wheezing[17]. Score-CAM is a tool that can interpret the model's decision-making process by visualizing the location where the model is activated in the image as well as evaluate whether it is influenced by external factors such as noise[17]. In cases where our model incorrectly predicted that there was no wheezing, there was too much noise in the audio file, and the wheezing was quiet compared to other sounds. The sample classified as wheezing included beeps from hospital medical equipment along with recordings of various unknown sounds that sounded like wheezing. However, the model did not activate to the medical device's notification sound, but it only weakly activated to sounds such as unknown sounds that sounded like wheezing. In the correctly classified case shown in Fig. 2, our model was strongly activated when wheezing was heard, and it was weakly activated by noise and normal breathing sounds. Even in the absence of wheezing, the model responded to noise and breathing, although with significantly less activation. Visualization using Score-CAM allowed us to confirm that our model was less responsive to noise and captured wheezing well.

The emergence of open databases of respiratory sounds has motivated the performance of various respiratory sound classification studies[22,26,36]. The performance score using this data also serves as a comparison standard for breath sound classification models[22,36].However, these open databases feature a very large class imbalance[20,26]. Moreover, the length of the sample, the sampling rate of the recording, and the sound quality vary substantially[21,26]. Under these conditions, it is easy for overfitting to occur. Further, the non-strict labeling of samples may have negatively impacted the accuracy of learning[26,34]. The normal sound samples in the ICBHI data set al.so contain some mixtures of wheeze and rale[36]. Because our study was recorded in a double-blind manner by two experienced pediatric respiratory doctors following a specific protocol in an actual clinical setting, it is a dataset that is more effective for learning deep learning models and more useful for clinical application than open data. Furthermore, we collected breath sounds from real clinical patients while minimizing external noise, ensuring data quality. This makes our dataset more effective for training deep learning models and more suitable for clinical applications than open datasets. Previous models not only faced performance limitations but also failed to adequately reflect real clinical environments. Our study has the advantage of collecting breath sounds from actual patients across two medical institutions, gathering a larger number of cases than previous studies, and presenting a high-performance AI classification model.

This study has several limitations. First, despite the addition of more data than previous studies, the absolute sample size is still small[10]. Transformer models require more data for training than CNN models. We overcame the problem of insufficient data volume by using a pre-trained model. We also split the data by using 80% for training and 20% for validation. We used Audio Data Augmentation and SpecAugment as well. Second, there is an imbalance problem in the training dataset. In response to this, we used F1 score as the measurement standard, instead of accuracy. Later studies should use various techniques to solve data imbalance problems, such as the smote technique. Third, our model is a binary classification model that distinguishes wheezing sounds. The binary classification of wheezing is known to be the least difficult[21]. For better clinical utility, it is necessary to propose a model that shows high performance while simultaneously classifying not only wheeze but also rale and stridor. To address these limitations in future research, we will focus on building a more robust pediatric breath sound dataset by ensuring more accurate labeling and increasing the sample size. Additionally, we will explore various Transformer-based model combinations to maximize performance, portability, and usability, aiming to achieve optimal outcomes in clinical applications.

Conclusion

In this study, we confirmed that our AI model—which is designed based on AST, a simple and lightweight architecture—shows higher accuracy in wheezing classification than the CNN-based AI model that has been widely used in previous studies. We also confirmed that this model can be applied to children. AST is also advantageous for development on mobile devices because it stores all the weights of the model in the database. This development of our model is expected to be a particularly useful tool in decisions regarding the diagnosis and treatment of pediatric respiratory diseases.

Data availability

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Received: 6 February 2024; Accepted: 5 February 2025
Published online: 15 February 2025

References

1. Sarkar, M., Madabhavi, I., Niranjan, N. & Dogra, M. Auscultation of the respiratory system. *Ann. Thorac. Med.* **10** (3), 158–168 (2015).
2. Zhang, J. et al. Real-World Verification of Artificial Intelligence Algorithm-assisted auscultation of Breath sounds in children. *Front. Pediatr.* **9**, 627337 (2021).
3. Bardou, D., Zhang, K. & Ahmad, S. M. Lung sounds classification using convolutional neural networks. *Artif. Intell. Med.* **88**, 58–69 (2018).
4. Kim, Y. et al. The coming era of a new auscultation system for analyzing respiratory sounds. *BMC Pulm Med.* **22** (1), 119 (2022).
5. Pramono, R. X. A., Bowyer, S. & Rodriguez-Villegas, E. Automatic adventitious respiratory sound analysis: a systematic review. *PLoS One*. **12** (5), e0177926 (2017).
6. Meslier, N., Charbonneau, G., Racineux, J. L. & Wheezes *Eur. Respir J.* ;**8**(11):1942–1948. (1995).
7. Liu, R., Cai, S., Zhang, K. & Hu, N. Detection of Adventitious Respiratory Sounds based on Convolutional Neural Network. *International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*; 2019, 298–303. (2019).
8. Nguyen, T. & Pernkopf, F. Lung sound classification using Snapshot Ensemble of Convolutional neural networks. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2020**, 760–763 (2020).
9. Kevat, A., Kalirajah, A. & Roseby, R. Artificial intelligence accuracy in detecting pathological breath sounds in children using digital stethoscopes. *Respir Res.* **21** (1), 253 (2020).
10. Kim, B. J., Kim, B. S., Mun, J. H., Lim, C. & Kim, K. An accurate deep learning model for wheezing in children using real world data. *Sci. Rep.* **12** (1), 22465 (2022).
11. Gong, Y., Chung, Y-A. & Glass, J. Ast: Audio Spectrogram transformer. *arXiv preprint arXiv:210401778* 2021.
12. Wu, C. et al. Intelligent Stethoscope using Full Self-Attention Mechanism for Abnormal Respiratory Sound Recognition. *IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*; 2023, 1–4. (2023).
13. Ariyanti, W., Liu, K. C., Chen, K. Y. & Yu, T. Abnormal Respiratory Sound Identification Using Audio-Spectrogram Vision Transformer. *Annu Int Conf IEEE Eng Med Biol Soc.* ;2023:1-413. Wei S, Zou S, Liao F, lang w. A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification. *Journal of Physics: Conference Series* 2020;1453(1):012085. (2023).
14. Park, D. S. et al. Specaugment: a simple data augmentation method for automatic speech recognition. *arXiv Preprint arXiv* :190408779 (2019).
15. Zhang, Z. Improved Adam Optimizer for Deep Neural Networks. *IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*; 2018, 1–2. (2018).
16. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D. & Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:180305407* 2018.
17. Wang, H. et al. Score-CAM: Score-weighted visual explanations for convolutional neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*; 24 – 5. (2020).
18. He, H. Yunqian Ma. Imbalanced Learning: Foundations, Algorithms, and Applications 1st Edition. Wiley-IEEE Press, (2013).
19. Kim, Y. et al. Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning. *Sci. Rep.* **11** (1), 17186 (2021).
20. Siebert, J. N. et al. Deep learning diagnostic and severity-stratification for interstitial lung diseases and chronic obstructive pulmonary disease in digital lung auscultations and ultrasonography: clinical protocol for an observational case-control study. *BMC Pulm Med.* **23** (1), 191 (2023).
21. 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** (1), 1289 (2023).
22. Huang, D. M. et al. Deep learning-based lung sound analysis for intelligent stethoscope. *Mil Med. Res.* **10** (1), 44 (2023).

23. Nguyen, T. & Pernkopf, F. Lung sound classification using Co-tuning and Stochastic normalization. *IEEE Trans. Biomed. Eng.* **69** (9), 2872–2882 (2022).
24. Ballİ, O., Kutlu, Y., COMPARISON OF ARTIFICIAL INTELLIGENCE PERFORMANCES & OBTAINED IN DATASET CLASSIFICATIONS USING RESPIRATORY DATA. *Bartın Univ. Int. J. Nat. Appl. Sci.* ;**5**(2):151–159. (2022).
25. Kim, Y. et al. Evolution of the stethoscope: advances with the adoption of machine learning and development of Wearable devices. *Tuberc Respir Dis. (Seoul).* **86** (4), 251–263 (2023).
26. Xia, T., Han, J. & Mascolo, C. Exploring machine learning for audio-based respiratory condition screening: a concise review of databases, methods, and open issues. *Exp. Biol. Med. (Maywood).* **247** (22), 2053–2061 (2022).
27. Petmezas, G. et al. Automated lung sound classification using a hybrid CNN-LSTM Network and focal loss function. *Sens. (Basel)* ;**22**(3). (2022).
28. 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** (3), 535–544 (2020).
29. Yamashita, R., Nishio, M., Do, R. K. G. & Togashi, K. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging.* **9** (4), 611–629 (2018).
30. Tsantekidis, A., Passalis, N. & Tefas, A. Chapter 5 - recurrent neural networks. In: (eds Iosifidis, A. & Tefas, A.) Deep Learning for Robot Perception and Cognition: Academic; 101–115. (2022).
31. Bae, S. et al. Patch-Mix Contrastive Learning with Audio Spectrogram Transformer on respiratory sound classification. *arXiv preprint arXiv:230514032* 2023.
32. Zhang, Y., Li, B., Fang, H. & Meng, Q. Spectrogram Transformers for Audio Classification. *IEEE International Conference on Imaging Systems and Techniques (IST)*; 2022, 1–6. (2022).
33. Habukawa, C. et al. Evaluation of airflow limitation using a new modality of lung sound analysis in asthmatic children. *Allergology Int.* **64** (1), 84–89 (2015).
34. Fraiwan, M., Fraiwan, L., Khassawneh, B. & Ibnian, A. A dataset of lung sounds recorded from the chest wall using an electronic stethoscope. *Data Brief.* **35**, 106913 (2021).
35. Grzywalski, T. et al. Practical implementation of artificial intelligence algorithms in pulmonary auscultation examination. *Eur. J. Pediatr.* **178** (6), 883–890 (2019).
36. Rocha, B. M. et al. An open access database for the evaluation of respiratory sound classification algorithms. *Physiol. Meas.* **40** (3), 035001 (2019).

Author contributions

KH Kim conceptualized and designed the study. KH Kim and BJ Kim collected and analyzed the data, drafted the initial manuscript, and reviewed and revised the manuscript. JH Mun and CW Lim conceptualized and designed the study, analyzed the data, drafted the initial manuscript, and reviewed and revised the manuscript. BS Kim, DH Hwang, and DI Suh designed the study, analyzed the data, and reviewed and revised the manuscript. All authors approved the final manuscript as submitted and agreed to be accountable for all aspects of the work.

Funding

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (No.2021R1I1A1A01052210). It was also supported by the New Faculty Startup Fund (No. 800-20230600) from Seoul National University.

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Review Boards (IRBs) of the Catholic University of Korea (IRB approval no. PC19OESI0045) and Seoul National University of Korea (IRB approval no. H-1907-050-1047). Written informed consent was obtained from at least one legal guardian for all participants.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-89533-9.

Correspondence and requests for materials should be addressed to C.L. or K.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.