Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Automated Wheezing Segmentation with Task-Specific Weighted Loss

Daehwan Hwang
Seoul National University Bundang Hospital
Wonwoo Shin
Seoul National University Bundang Hospital
Hyunwoo Kim
Seoul National University Bundang Hospital
June-woo Kim
Wonkwang University
Kyunghoon Kim
journey237@snu.ac.kr

Seoul National University Bundang Hospital

Article

Keywords: Audio Segmentation, Respiratory Sound, Wheezing

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10978827/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Automated Wheezing Segmentation with Task-Specific Weighted Loss

Authors: Daehwan Hwang[1‡], Wonwoo Shin[1], Hyunwoo Kim[1], June-Woo Kim, PhD[2*], and Kyunghoon Kim, MD, PhD[1,3*]

[1]Department of Pediatrics, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

[2]Department of Electronics Engineering, Wonkwang University, Iksan, Republic of Korea

[3]Department of Pediatrics, Seoul National University College of Medicine, Seoul, Republic of Korea.

[‡] these authors contributed equally to this work.

* these authors jointly supervised this work.

Address for Correspondence:

Kyunghoon Kim, MD, PhD

Department of Pediatrics, Seoul National University Bundang Hospital, Seongnam, Republic of Korea

Department of Pediatrics, Seoul National University College of Medicine, Seoul, Republic of Korea.

Phone number: +8210-9054-3064

E-mail: journey237@snu.ac.kr

June-Woo Kim, PhD

Department of Electronics Engineering, Wonkwang University, Iksan, Republic of Korea

Phone number +8210-2304-8514

E-mail: kaen2891@wku.ac.kr

Abstract

Accurate segmentation of wheezing events may support clinical decision-making and patient monitoring. Previous studies have largely focused on classifying abnormal respiratory sounds at the recording or respiratory-cycle level, whereas precise temporal localization of wheezing onset and offset remains challenging. We propose Wheezing Forecasting & Audio Segmentation Technology (W-FAST), an architecture for precise temporal segmentation of wheezing events. To address the severe class imbalance in clinical respiratory sound data, W-FAST employs a task-specific weighted loss incorporating a weighting factor α. W-FAST achieved a mean ICBHI score of 0.673 ± 0.017 on the ICBHI 2017 dataset. When the same ICBHI scoring scheme was applied to the SNUBH dataset, W-FAST achieved a mean score of 0.946 ± 0.016. In the SNUBH dataset, the ratio of wheezing to non-wheezing blocks was approximately 1:55 among valid 0.32-s blocks after zero-padded regions were masked. The proposed loss helped stabilize training under this severe imbalance and improved segmentation performance. These results suggest that W-FAST may provide an effective framework for fine-grained temporal localization of wheezing in clinical respiratory sound recordings.

Keyword: Audio Segmentation; Respiratory Sound; Wheezing

Introduction

Auscultation has long been utilized as a primary non-invasive diagnostic tool for respiratory diseases[1]; however, it is limited by the clinician's subjectivity and the difficulty of continuous monitoring[2]. Wheezing is a critical pathological breath sound observed in patients with conditions such as asthma and chronic obstructive pulmonary disease (COPD), yet its accurate identification necessitates a high level of clinical proficiency[3]. Furthermore, the accuracy of manual auscultation is susceptible to the examiner's physical and psychological conditions, including hearing acuity[4]. Consequently, cultivating specialized personnel and maintaining consistent, long-term monitoring in real-world clinical settings remains a significant challenge[5].

To overcome these limitations, research into AI-based respiratory sound classification using electronic stethoscopes has been actively conducted as an objective alternative[6]. Unlike human examiners, AI can provide consistent diagnostic results regardless of physical or psychological fatigue[7]. Moreover, deep learning models can learn a vast array of abnormal breath sound patterns more rapidly than humans, effectively addressing the shortcomings of traditional methods[8]. These automated systems have demonstrated high accuracy in identifying abnormal sounds, showcasing their potential to support or even augment clinician-led auscultation.

However, models trained exclusively on medical datasets exhibit clear limitations in real-world clinical applications, necessitating transfer learning and the construction of site-specific datasets[9], which inevitably increases the burden on medical experts. Most medical datasets are restricted by geographical limitations, unbalanced age demographics, and the use of a narrow range of medical devices[10]. Consequently, these datasets are insufficient to represent the diverse demographic distributions and varied acoustic characteristics of stethoscopes encountered globally[11]. To achieve practical utility, additional training tailored to the specific demographics and equipment of each clinical environment is essential. This continuous requirement for high-quality, expert-labeled datasets remains

a necessity, which fundamentally leads to a cumulative increase in clinician fatigue.

While AI can be employed to facilitate the labeling of respiratory sounds for training, current methodologies are insufficient, highlighting the urgent need for precise segmentation models. Visualization studies using Explainable AI techniques, such as Score-CAM[12], have revealed instances where models incorrectly respond to bio-signal or ambient noise[13]. This occurs because conventional studies typically classify an entire respiratory cycle as a single class, causing the model to learn normal segments and noise as if they were pathological signs. This fundamental issue undermines the accuracy and reliability of the models, explaining why their active adoption in clinical settings remains limited. To overcome these limitations, we conducted research on a precise, automated segmentation model designed to learn exclusively from accurate abnormal breath sound intervals.

In this study, we propose Wheezing Forecasting & Audio Segmentation Technology (W-FAST), an architecture optimized for precise interval detection of respiratory sounds to enhance diagnostic accuracy and clinical applicability. By automatically identifying and isolating fine-grained wheeze segments, W-FAST can be utilized for processing large-scale clinical data and monitoring. The primary contributions of this research are as follows:

- Precise Segmentation: We propose an architecture that detects accurate intervals of wheeze sounds within a recording, providing clinicians with localized data to facilitate rapid and accurate decision-making.
- Weight Optimization Strategy: We introduce a strategy to adjust class-specific weights within the loss function, accounting for the unique characteristics of medical data to derive the optimal trade-off between sensitivity and specificity based on clinical requirements.
- Validation of Clinical Efficacy: The robustness and practical utility of the model were extensively validated using both the ICBHI 2017 benchmark dataset and real-world data

from Seoul National University Bundang Hospital (SNUBH).

Related Works

Respiratory Sound Classification

Early studies primarily utilized in-house datasets and performed classification using traditional machine learning techniques, such as SVM[14], K-NN[15], Random Forest[16], and GMM[17], after extracting signal processing-based features like Mel-frequency cepstral coefficients (MFCCs) and spectral energy[18]. Some studies directly utilized raw waveforms or applied RNN-based models[19]. Subsequently, an approach that transforms audio into mel-spectrograms to train CNN-based and CRNN-based models has become the mainstream[20,21]. This method offers the advantage of learning representations directly from the data without the need for manual feature engineering[22].

Recently, prior studys on respiratory sound classification has expanded into the use of state-of-the-art Transformer architectures, such as the Audio Spectrogram Transformer (AST)[23,24]. Although Transformers originated in the field of Natural Language Processing (NLP), they have demonstrated exceptional performance across various domains, such as image and audio, by leveraging self-attention mechanisms. Notably, the current top-performing models on the ICBHI benchmark dataset are predominantly based on the AST architecture, achieving state-of-the-art (SOTA) results.

However, a critical limitation of the respiratory sound datasets used in these prior studies is that they are labeled at the respiration cycle level, meaning a single sample often contains both normal and abnormal breath sounds. Consequently, qualitative evaluations of these AI models using Score-CAM on real-world data have revealed that models frequently respond to normal breathing intervals or ambient noise rather than focusing exclusively on the abnormal sound segments[25]. This lack of diagnostic transparency and reliability has led researchers to investigate methods for precisely predicting the exact intervals of abnormal sounds. As a result, segmentation has emerged as a compelling alternative to overcome the inherent limitations of simple cycle-level classification[26].

Audio Segmentation

In the field of audio signal processing, segmentation is also widely known as Sound Event Detection (SED). Similar to how object detection in computer vision specifies locations using bounding boxes, SED predicts the onset (start) and offset (end) of specific sound classes along the time axis.

Traditional SED methodologies have primarily relied on AI models predicting the presence of a class at extremely fine frame-level intervals, which are typically around 10 ms, and subsequently merging these results into a continuous interval through post-processing[27,28]. However, such frame-byframe prediction is computationally intensive and has limitations in capturing long-term temporal context, making it susceptible to noise. To overcome these drawbacks, Satvik et al. proposed YOHO (You Only Hear Once), which migrates the algorithm of the YOLO (You Only Look Once) object detection algorithm into the audio domain[29]. YOHO simultaneously predicts the class, onset, and offset of sounds through a regression-based approach within relatively longer windows of approximately 0.3 seconds. This method offers the advantage of significantly reducing computational resources while enhancing both detection speed and accuracy compared to conventional models.

Despite these advancements, the application of audio segmentation in the field of respiratory sounds follows a notably different trend. Currently, the majority of research in respiratory sound segmentation focuses on isolating the respiratory cycle from background noise rather than identifying the abnormal sounds themselves. This is primarily due to the extreme scarcity of large-scale, publicly available datasets that provide precise temporal labels for fine-grained abnormal sounds (e.g., wheezing or crackles)[30]. Since most benchmark datasets only provide labels at the respiratory cycle level, it remains difficult for models to learn which specific intervals within a cycle represent actual pathological signs.

Consequently, to precisely identify subtle abnormal intervals hidden within a respiratory cycle, it is essential to optimize efficient detection architectures like YOHO for the respiratory sound domain. Especially considering the labeling fatigue of medical professionals, the development of an automated segmentation model capable of extracting high-resolution temporal information from limited data is a key challenge for ensuring clinical utility.

Methods

Architecture : W-FAST

The proposed W-FAST architecture builds upon the single-stage detection principles of YOHO, specifically redesigned for high-precision respiratory sound segmentation. The adoption of this architecture is primarily driven by the temporal characteristics of wheezing, which typically occurs over a very short duration of 50–250 ms[4]. While conventional frame-wise pipelines, which rely on localized 10-ms labeling, often suffer from temporal discontinuities, W-FAST treats audio segmentation as a unified regression problem within 320-ms blocks. By performing predictions within a temporal window that closely aligns with the actual duration of wheezing events, the framework ensures enhanced localization accuracy. Consequently, the model enables the simultaneous prediction of wheezing presence and its precise temporal boundaries, including onset and offset, within a single forward pass.

While the original YOHO framework was constrained to a MobileNet[31]-based backbone, the W-FAST framework is designed to be backbone-agnostic. We integrated diverse feature extractors, including ResNet[32], CNN[33] and EfficientNet[34], to capture the multi-scale spectral features inherent in pathological breath sounds (Fig. 1). To preserve the temporal resolution required for clinical diagnosis, we replaced the conventional dense layers in the output stage with 1D convolutional (Conv1D) layers. This modification ensures that the spatial temporal dependencies extracted by the backbone are directly mapped to the grid-based prediction cells without information loss.

A defining engineering characteristic of YOHO is its unified optimization framework, which employs Sum Squared Error (SSE) for both classification and temporal regression tasks. By adopting SSE instead of the conventional Cross-Entropy for classification, the model maintains a consistent gradient scale across class probability predictions and event boundary (onset and offset) estimations. This approach facilitates faster convergence and stable training in a single-stage detection architecture.

The total loss function, ℓ_{total}, is defined as the summation of the squared errors for class predictions and temporal coordinates as follows:

$$\ell_{total} = \begin{cases} (\hat{y}_{class} - y_{class})^2 + (\hat{y}_{onset} - y_{onset})^2 + (\hat{y}_{offset} - y_{offset})^2 & if\ y_{class} = 1 \\ (\hat{y}_{class} - y_{class})^2 & if\ y_{class} = 0 \end{cases}$$

where y and $\hat{y}$ represent the ground truth and the predicted values, respectively. This integrated SSE approach effectively mitigates potential imbalances between heterogeneous loss components during multi-task learning, providing a robust foundation for achieving precise millisecond-level segmentation of respiratory sounds.

While the original YOHO architecture was based on MobileNet blocks, we propose the W-FAST framework, which extends the backbone to ResNet, CNN, and EfficientNet variants to more effectively capture the complex spectral patterns of respiratory sounds (Fig. 1). By replacing the conventional dense layers with Conv1D layers in the output stage, the model preserves temporal dependencies and outputs both the existence of wheezing and its precise temporal location for each designated grid cell.

Task-specific Weighted Loss Function

The loss function of the W-FAST architecture is designed to simultaneously optimize both classification (for each grid block) and temporal regression (for onset and offset), following the multi-task learning paradigm of the YOLO framework. However, we observed that applying weights based solely on class frequency to address the extreme class imbalance led to a significant over-detection bias. In this state, the model tended to predict almost all segments as wheezing to minimize the total loss value. This optimization bias arises because the wheezing class involves three distinct loss components

($\ell_{class}, \ell_{onset}, \ell_{offset}$) whereas the "others" (non-wheezing) class is only associated with the classification loss (ℓ_{class}). Consequently, the network finds that predicting a positive result for wheezing yields a larger reduction in total loss compared to accurately identifying non-wheeze segments. To mitigate this imbalance and enable fine-grained control over performance based on clinical objectives, we propose a task-specific Weighted loss function that introduces an additional penalty constant, α, specifically for the others class. The total loss ℓ is formulated as follows:

$$\ell_{weighted} = \sum_{i \in all} w_{class-ratio} \cdot \ell_{class} + \sum_{i \in wheeze} (\ell_{onset} + \ell_{offset}) + \sum_{i \in non-wheeze} \alpha \cdot \ell_{class}$$

where ℓclass denotes the error between the predicted class probability and the ground truth, and ℓonset and ℓoffset represent the regression losses for the starting and ending points of the wheezing event, respectively. The term $w_{class-ratio}$ represents the base weight derived from the class distribution within the batch, while α acts as a control parameter to suppress excessive positive predictions by increasing the loss contribution of the others class. By empirically optimizing the value of α, we identified the optimal trade-off that maximizes diagnostic sensitivity for wheezing while effectively minimizing false positives[35].

Dynamic Undersampling

To further address the severe class imbalance in the SNUBH dataset, we implemented a Dynamic undersampling strategy. Although weighted loss functions can mitigate bias caused by class imbalance, they do not directly control the number or diversity of negative samples presented during training[36]. After preprocessing, the recordings containing wheezing events themselves comprised 475 wheezing blocks and 2,154 non-wheezing blocks, corresponding to an internal block-level ratio of approximately 1:4.5. Thus, these recordings already provided a substantial number of negative blocks before any

normal-only recordings were added. In preliminary validation experiments, file-level undersampling using equal numbers of wheezing-containing and normal-only recordings did not produce a satisfactory balance between sensitivity and specificity. We therefore retained all wheezing-containing recordings and randomly selected normal-only recordings at 10% of the number of wheezing-containing recordings, corresponding to a recording-level sampling ratio of 1:0.1. The subset of normal-only recordings was resampled at the beginning of each epoch. This procedure exposed the model to diverse normal examples across epochs while limiting the dominance of non-wheezing samples within each epoch.

Datasets

Table 1 summarizes the datasets used in this study: the ICBHI 2017 Challenge dataset[37], which is a representative benchmark in the field of respiratory sounds, and the SNUBH dataset[38], which was independently collected by our research team.

A. ICBHI2017 Dataset

The ICBHI 2017 dataset is a widely recognized public benchmark comprising respiratory sounds from patients of various ages, collected across four institutions in Portugal and Greece. It contains 920 audio files (approximately 5.5 hours), with cycle-level annotations categorized as Normal, Wheezing, Crackle, or Both. The overall class distribution consists of 46% Normal, 20% Wheezing, and 34% Crackle. In this study, we adhered to the official data split protocol to ensure an objective performance evaluation. This split is specifically designed to prevent data leakage by ensuring that audio recorded from the same institution or with the same

stethoscope model does not overlap between the training and test sets, thereby rigorously verifying the generalization capability of the proposed model.

B. SNUBH Dataset

The SNUBH dataset is real-world clinical data collected from pediatric patients with respiratory diseases at two hospitals in South Korea. To ensure the highest level of data reliability, two pediatric specialists performed independent annotation; sounds were only assigned a final Wheezing label when both specialists reached a consensus. The complete dataset consists of 2,598 audio files (approximately 5.4 hours) and exhibits a severe class imbalance, predominantly composed of Normal (85%), Wheezing (11%), Crackle (1%), Rhonchi (2%) and Stridor (1%). In this study, we constructed an analysis subset comprising 605 waveforms: 74 recordings containing wheezing events with precisely annotated onset and offset times and 531 normal recordings. During preprocessing, the recordings were divided into 0.32-s blocks, the temporal unit used by W-FAST, and zero-padded regions were masked out. The resulting valid data comprised 475 wheezing blocks and 26,023 non-wheezing blocks. Accordingly, the block-level ratio of wheezing to non-wheezing was approximately 1:54.8, indicating severe class imbalance.

Ethics statement

This study was approved by the Institutional Review Boards (IRBs) of The Catholic University of Korea (IRB No. PC19OESI0045) and Seoul National University Bundang Hospital (IRB No. B-2204-749-303). 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.

Audio Pre-processing

Each audio signal was segmented into 8-second windows with a hop length of 6 seconds to ensure temporal continuity between adjacent segments. For signals shorter than 8 seconds, zero-padding was applied. To prevent the padded regions from generating easy true negatives which could artificially inflate performance metrics, a masking strategy was implemented to exclude these regions from both loss calculation and performance evaluation.

For feature extraction, we utilized the Kaldi's Fbank implementation in Torchaudio, replacing the conventional Librosa-based Mel-spectrogram used in the original YOHO study. While the Librosa approach yields a time frames of 801, the Kaldi's Fbank method results in a frames of 798 due to differences in internal padding mechanisms. This transition was driven by empirical results, where the Kaldi's Fbank method demonstrated a 1–2% performance improvement over the conventional approach, leading to its adoption as the primary feature extraction algorithm in this work.

Evaluation Metrics

To objectively evaluate the model's performance, we adopted the ICBHI Score, the official evaluation metric of the ICBHI 2017 challenge, as the primary performance metric. This metric represents the balanced accuracy, defined as the arithmetic mean of sensitivity($\boldsymbol{S_e}$) and specificity($\boldsymbol{S_p}$), and is formulated as follows:

$$\boldsymbol{Score} = \frac{\boldsymbol{S_e + S_p}}{\boldsymbol{2}}$$

In addition, we employed secondary metrics, including the area under the receiver operating characteristic (AUROC) curve, to provide a comprehensive assessment of the model's diagnostic capabilities.

Performance evaluation was conducted following the protocol of the "**sed_eval**" library, a standard toolkit in audio event detection[40]. While sed_eval typically utilizes a 1-second segment resolution, we utilized a 0.32-second resolution, which corresponds to the model's native grid block size, to achieve higher temporal granularity. This refined evaluation framework is intended to ensure a rigorous validation of the model's temporal predictive accuracy at a fine-grained level, enabling a more precise assessment of the model's ability to localize abnormal respiratory sound events.

Training Details

To ensure statistical reliability, model training was conducted five times using different seeds, with each run lasting up to 100 epochs. An early stopping strategy was implemented to prevent overfitting, terminating the training if the validation loss failed to improve for 20 consecutive epochs. We employed the Adam optimizer[41] with a batch size of 32. Hyperparameter tuning for the learning rate was performed across a candidate set of 1e-5, 5e-5, 1e-6 selecting the value that yielded the best performance without a learning rate scheduler.

Dataset partitioning and repeated evaluation were performed differently for the two datasets. For the ICBHI 2017 dataset, we followed the official fixed train–test split. The model was trained five times using five different random seeds, while the official train–test split unchanged across runs.

For the SNUBH dataset, patient-wise five-fold cross-validation was performed. All recordings and segments from the same patient were assigned to a single fold to prevent patient-level data leakage. Each fold was used once as the held-out test set, while the remaining folds were used for model development. A validation subset derived from the training portion was used for early stopping and model selection, and a different random seed was used for each fold.

All performance metrics are reported as the mean ± standard deviation (SD). For the ICBHI 2017 dataset, the mean and SD were calculated from the five test-set scores obtained using different random seeds. For the SNUBH dataset, they were calculated from the five scores obtained on the non-overlapping held-out folds.

Code Availability

The source code and configuration files used to implement and evaluate W-FAST will be made publicly available upon publication at https://github.com/ppn0303/W-FAST. The code is available from the corresponding author for peer-review purposes upon reasonable request.

Use of generative artificial intelligence

During the writing of this manuscript, the authors used OpenAI Codex to assist with English-language and editing for clarity and readability. All outputs were reviewed and verified by the authors, who take full responsibility for the final content.

Results

Performance on Each Dataset

Fig. 2 presents the performance of the proposed W-FAST framework evaluated across various backbone architectures, with results sorted in descending order based on their mean scores. As illustrated in Fig. 2(a), when evaluated on the ICBHI 2017 dataset, the EffNetB2-backed W-FAST model achieved the highest score (0.673 ± 0.017). Notably, the EfficientNet family and high-capacity models overall demonstrated superior efficacy on this public dataset. In contrast, on the SNUBH dataset shown in Fig. 2(b), the ResNet34 configuration delivered the best diagnostic performance (0.946 ± 0.016), while the EfficientNet models exhibited relatively lower scores.

An observation from this comparative analysis is the clear performance gap between the proposed framework and the baseline model. In the ICBHI 2017 dataset, the baseline YOHO model placed second to last. Furthermore, in the SNUBH dataset, YOHO recorded the lowest performance. In contrast, almost all variations of the W-FAST framework integrated with alternative backbone networks consistently achieved higher performance than the YOHO model across both datasets.

Ablation Study: Impact of W-FAST Module and Weighted Loss

Fig. 3 presents the performance distribution of various backbone architectures and the statistical significance of improvements achieved by the integration of the W-FAST module and the weighted loss strategy on the ICBHI 2017 and SNUBH datasets, respectively. On the ICBHI 2017 dataset, Fig. 3(a), the integration of the W-FAST module yielded significant performance improvements across most backbones, although these gains were less pronounced for EfficientNet and ResNet34. Furthermore, the subsequent addition of the weighted loss strategy demonstrated statistically significant

improvements for most models, with the exceptions of CNN6 and EfficientNet-B2.

In the SNUBH dataset, Fig. 3(b), the impact of the W-FAST module was most distinct in the CNN6 and ResNet18 architectures. Notably, the weighted loss strategy consistently demonstrated highly significant performance improvements ($p < 0.001$) across all tested architectures, underscoring its robustness in addressing the data imbalance characteristic of the clinical dataset.

Performance Evaluation via AUROC Curve

Fig. 4 illustrates the Receiver Operating Characteristic (ROC) curves and the corresponding AUROC values for the baseline, W-FAST, and W-FAST+WL models. A stepwise performance improvement is observed as the proposed architectural components and weighted loss strategy are integrated. On the SNUBH dataset, the AUROC increased from 0.789 (baseline) to 0.949 (W-FAST), reaching a peak of 0.960 upon the application of the weighted loss strategy. Similarly, on the ICBHI 2017 dataset, the AUROC improved from 0.603 to 0.686. These results demonstrate that the proposed framework effectively optimizes diagnostic sensitivity and specificity, particularly in addressing the severe class imbalances inherent in clinical respiratory sound data.

Sensitivity-Specificity Trade-off via Weight Optimization

To determine the optimal balance between sensitivity and specificity, we performed a sensitivity analysis on the additional weight factor (α) within our weighted loss strategy. Detailed results of this optimization process are provided in Supplementary Fig. 5. For the ICBHI 2017 dataset, the model achieved peak diagnostic performance at $\alpha = 1.75$. Beyond this value, we observed a pronounced trade-off where increasing the weight for the minority class led to a gradual decline in sensitivity. In

contrast, for the SNUBH dataset, the optimal performance was observed at $\alpha = 4$. Consistent with clinical expectations, increasing α beyond this threshold resulted in a clear inverse relationship between sensitivity and specificity, demonstrating the model's adaptability in tuning diagnostic priorities based on specific clinical requirements.

Qualitative Analysis via Score-CAM

Fig. 6 presents representative Score-CAM visualizations from the ICBHI 2017 and SNUBH datasets. In the successful cases shown in Fig. 6(a) and (d), the model predominantly focused on the 1–4 kHz range, consistent with the reported spectral characteristics of wheezing[4], and accurately localized the annotated wheezing intervals. In Fig. 6(b), the model exhibited spurious activation in a high-frequency noise region. In Fig. 6(e), the model detected only part of the annotated wheezing interval and also responded to a normal breathing region.

In the non-wheezing examples shown in Fig. 6(c) and (f), weak activations were observed over respiratory sound regions, but these responses did not result in positive wheezing predictions. Activation was also occasionally observed near the end of the audio signals across multiple models; however, it did not produce predicted wheezing intervals. These findings indicate that Score-CAM activation does not necessarily translate directly into a positive wheezing prediction.

Discussion

In this study, we proposed the W-FAST architecture for precise respiratory sound segmentation and implemented a task-specific weighted loss strategy to address severe class imbalance, a challenge in clinical datasets. Experimental results demonstrated that the proposed model achieves superior performance on both the public ICBHI 2017 benchmark and the real-world SNUBH dataset.

Our primary academic contributions: First, we demonstrated that framing respiratory sound analysis as a complex segmentation task, rather than relying on simple cycle-level classification, significantly enhances diagnostic performance. Conventional approaches predominantly train deep learning models using single labels assigned to an entire respiratory cycle. Consequently, these models often inadvertently learn normal respiratory intervals or ambient noise as pathological indicators, leading to spurious activations. In contrast, the W-FAST framework is designed to learn exclusively from fine-grained abnormal intervals, enabling highly precise block-level temporal localization. When integrated with diverse backbone networks, the W-FAST module effectively extracts intricate acoustic patterns, confirming that a grid-based output structure tailored to respiratory dynamics is more critical for temporal localization than merely increasing model depth[29].

Second, we presented a multi-layered integrated strategy to mitigate extreme data imbalance by combining weighted loss and dynamic undersampling. Methods in previous studies have mainly relied on simple data augmentation (such as time-shifting, pitch perturbation, etc.)[39] that does not consider class ratios, or have used a weighted loss approach that simply multiplies the inverse of class ratios by the loss[42]. However, these approaches show marginal efficacy in clinical datasets where the imbalance is extreme. Furthermore, a simple weighted loss fails to reflect the characteristics of medical data, where the risk values of False Positives and False Negatives are strictly different. There is also a concern of overfitting to a specific class or task in a task where two types of loss functions coexist.

To overcome these limitations, the proposed weighted loss strategy adaptively adjusts class-specific weights, enabling disease-detection-centered optimization that prioritizes clinical safety. In particular, to prevent the problem of overfitting to negative samples when weighted loss is applied alone, we combined the dynamic undersampling strategy, which resamples negative samples with different random seeds at the start of each epoch. This maximizes the precision of temporal localization by allowing the framework to learn more diverse and rich normal and abnormal respiratory sounds, serving as a quantitative basis for establishing an optimal trade-off that minimizes the most critical false negatives in clinical diagnostic environments[43].

Third, clinical trust was secured by conducting a qualitative evaluation alongside the use of a visualization tool. In initial experiments, at the point where the simple score was the highest, the Score-CAM visualization results responded across the entire frequency range, yielding meaningless results. However, through the integration of weighted loss and dynamic undersampling, responses in non-wheezing segments and frequency bands were reduced.

Consequently, the final Score-CAM visualization analysis confirms that the model accurately targets the characteristics of wheezing. Specifically, the network robustly isolates mid-to-low frequency bands and characteristically targets expiratory phase patterns during the occurrence of wheezing[44,45]. This alignment between the model′s visualization and established medical knowledge serves as strong empirical evidence supporting the robust clinical diagnostic validity of W-FAST.

However, several limitations remain to be addressed. First, the SNUBH dataset, while featuring highly precise temporal labels, has a relatively constrained data scale and is restricted to a South Korean pediatric population, which may not fully represent global demographic distributions and varied stethoscope characteristics. Second, this study did not extend the framework to recent Transformer-based models, such as the Audio Spectrogram Transformer, restricting the comparative

analysis to conventional CNN backbones. Future work will focus on scaling the W-FAST framework to advanced multi-task Transformer architectures to validate its generalizability. Additionally, we plan to enhance noise robustness for highly variable telemedicine environments and develop domain adaptation techniques to compensate for acoustic discrepancies across digital stethoscopes from different manufacturers.

Conclusion

In this study, we developed the W-FAST framework for precise respiratory sound segmentation. To resolve the persistent class imbalance inherent in this task, we integrated a task-specific weighted loss function and an epoch-based dynamic undersampling strategy. Through these methods, we established a medical AI engineering system capable of adjusting the optimal trade-off between sensitivity and specificity according to the diagnostic goal. Experimental results confirmed that the proposed model achieved significant performance improvements in both Score and AUROC metrics across both public and clinical datasets. Specifically, qualitative analysis using Score-CAM verified that the model accurately captures pathological signals consistent with the physiological patterns of respiration. This study provides a technical foundation for AI-based automated auscultation systems to serve as clinical decision support tools in real-world environments. We anticipate that these findings will contribute to enhancing the reliability of early diagnostic and remote monitoring systems for respiratory diseases.

Declarations

Authors' contributions: KH Kim conceptualized and designed the study, collected and analyzed the data, drafted the initial manuscript, and reviewed and revised the manuscript. DH Hwang and JW Kim conceptualized and designed the study, analyzed the data, drafted the initial manuscript, and reviewed and revised the manuscript. WW Shin and HW Kim collected and 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.

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

Competing interests: The authors declare that they have no competing interests.

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 Bundang Hospital (IRB approval no. B-2204-749-303). Written informed consent was obtained from at least one legal guardian for all participants. All methods were performed in accordance with the relevant guidelines and regulations and the Declaration of Helsinki.

Funding: This research was supported by the National Research Foundation of Korea(NRF) grant

funded by the Korea government(MSIT) (grant no. RS-2025-16066662). This research was also supported by the Regional Innovation System & Education(RISE) program through the Jeonbuk RISE Center, funded by the Ministry of Education(MOE) and the Jeonbuk State, Republic of Korea(2026-RISE-13-WKU).

List of abbreviations

AI: artificial intelligence

CNN: convolutional neural network

AUC: area under the curve

ICBHI: International Conference on Biomedical and Health Informatics

MLP: multilayer perceptron

ResNet: residual network

References

1. Sarkar, M., Madabhavi, I., Niranjan, N. & Dogra, M. Auscultation of the respiratory system. *Ann. Thorac. Med.* **10**, 158–168 (2015).

2. Gross, V., Dittmar, A., Penzel, T., Schüttler, F. & Von Wichert, P. The relationship between normal lung sounds, age, and gender. *Am. J. Respir. Crit. Care Med.* **162**, 905–909 (2000).

3. Meslier, N., Charbonneau, G. & Racineux, J. L. Wheezes. *Eur. Respir. J.* **8**, 1942–1948 (1995).

4. Pasterkamp, H., Kraman, S. S. & Wodicka, G. R. Respiratory sounds. *Am. J. Respir. Crit. Care Med.* **156**, 974–987 (1997).

5. Yang, S., Zhu, F., Ling, X., Liu, Q. & Zhao, P. Intelligent health care: applications of deep learning in computational medicine. *Front. Genet.* **12**, 607471 (2021).

6. İçer, S. & Gengeç, E. Classification and analysis of non-stationary characteristics of crackle and rhonchus lung adventitious sounds. *Digit. Signal Process.* **28**, 18–27 (2014).

7. Ravì, D. et al. Deep learning for health informatics. *IEEE J. Biomed. Health Inform.* **21**, 4–21 (2017).

8. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297 (2019).

9. Wiens, J., Guttag, J. & Horvitz, E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *J. Am. Med. Inform. Assoc.* **21**, 699–706 (2014).

10. Subbaswamy, A. & Saria, S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* **21**, 345–352 (2020).

11. Rennoll, V., McLane, I., Emmanouilidou, D., West, J. & Elhilali, M. Electronic stethoscope filtering mimics the perceived sound characteristics of acoustic stethoscope. *IEEE J. Biomed. Health Inform.* **25**, 1542–1549 (2021).

12. Wang, H. et al. Score-CAM: score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 111–119 (IEEE, 2020).

13. Choi, Y. & Lee, H. Interpretation of lung disease classification with light attention connected module. *Biomed. Signal Process. Control* **84**, 104695 (2023).

14. Sen, I., Saraclar, M. & Kahya, Y. P. A comparison of SVM and GMM-based classifier configurations for diagnostic classification of pulmonary sounds. *IEEE Trans. Biomed. Eng.* **62**, 1768–1776 (2015).

15. Naves, R., Barbosa, B. H. G. & Ferreira, D. D. Classification of lung sounds using higher-order statistics: a divide-and-conquer approach. *Comput. Methods Programs Biomed.* **129**, 12–20 (2016).

16. Mendes, L. et al. Detection of wheezes using their signature in the spectrogram space and musical features. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 5581–5584 (IEEE, 2015).

17. Lin, B.-S. & Lin, B.-S. Automatic wheezing detection using speech recognition technique. *J. Med. Biol. Eng.* **36**, 545–554 (2016).

18. Palaniappan, R., Sundaraj, K. & Ahamed, N. U. Machine learning in lung sound analysis: a systematic review. *Biocybern. Biomed. Eng.* **33**, 129–135 (2013).

19. Perna, D. & Tagarelli, A. Deep auscultation: predicting respiratory anomalies and diseases via recurrent neural networks. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)* 50–55 (IEEE, 2019).

20. Aykanat, M., Kılıç, Ö., Kurt, B. & Saryal, S. Classification of lung sounds using convolutional neural networks. *EURASIP J. Image Video Process.* **2017**, 65 (2017).

21. Bardou, D., Zhang, K. & Ahmad, S. M. Lung sounds classification using convolutional neural networks. *Artif. Intell. Med.* **88**, 58–69 (2018).

22. Polanco-Martagón, S. et al. Comparison of deep neural networks for the classification of adventitious lung sounds. *J. Clin. Med.* **14**, 7427 (2025).

23. Ariyanti, W., Liu, K.-C., Chen, K.-Y. & Tsao, Y. Abnormal respiratory sound identification using audio-spectrogram vision transformer. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* 1–4 (IEEE, 2023).

24. Bae, S. et al. Patch-mix contrastive learning with audio spectrogram transformer on respiratory sound classification. In *INTERSPEECH 2023* 5436–5440 (ISCA, 2023).

25. Kim, B. J. et al. An explainable and accurate transformer-based deep learning model for wheeze classification utilizing real-world pediatric data. *Sci. Rep.* **15**, 5656 (2025).

26. Suyanto, S. Performance improvement of COVID-19 cough detection based on deep learning with segmentation methods. *J. Appl. Data Sci.* **5**, 520–531 (2024).

27. Venkatesh, S., Moffat, D. & Miranda, E. R. Investigating the effects of training set synthesis for audio segmentation of radio broadcast. *Electronics* **10**, 827 (2021).

28. Cakir, E., Parascandolo, G., Heittola, T., Huttunen, H. & Virtanen, T. Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**, 1291–1303 (2017).

29. Venkatesh, S., Moffat, D. & Miranda, E. R. You only hear once: a YOLO-like algorithm for audio segmentation and sound event detection. *Appl. Sci.* **12**, 3293 (2022).

30. Im, S. et al. Real-time counting of wheezing events from lung sounds using deep learning algorithms: implications for disease prediction and early intervention. *PLoS ONE* **18**, e0294447 (2023).

31. Howard, A. G. et al. MobileNets: efficient convolutional neural networks for mobile vision applications. Preprint at http://arxiv.org/abs/1704.04861 (2017).

32. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016).

33. Kong, Q. et al. PANNs: large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **28**, 2880–2894 (2020).

34. Freeman, I., Roese-Koerner, L. & Kummert, A. Effnet: an efficient structure for convolutional neural networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)* 6–10 (IEEE, 2018).

35. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollar, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 318–327 (2020).

36. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data* **6**, 27 (2019).

37. Rocha, B. M. et al. An open access database for the evaluation of respiratory sound classification algorithms. *Physiol. Meas.* **40**, 035001 (2019).

38. Kim, J.-W. et al. Adaptive metadata-guided supervised contrastive learning for domain adaptation on respiratory sound classification. *IEEE J. Biomed. Health Inform.* **29**, 5381–5393 (2025).

39. 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**, 22465 (2022).

40. Mesaros, A., Heittola, T. & Virtanen, T. Metrics for polyphonic sound event detection. *Appl. Sci.* **6**, 162 (2016).

41. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).

42. Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Cardoso, M. J. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* 240–248 (Springer International Publishing, 2017).

43. Salehi, S. S. M., Erdogmus, D. & Gholipour, A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *Machine Learning in Medical Imaging* 379–387

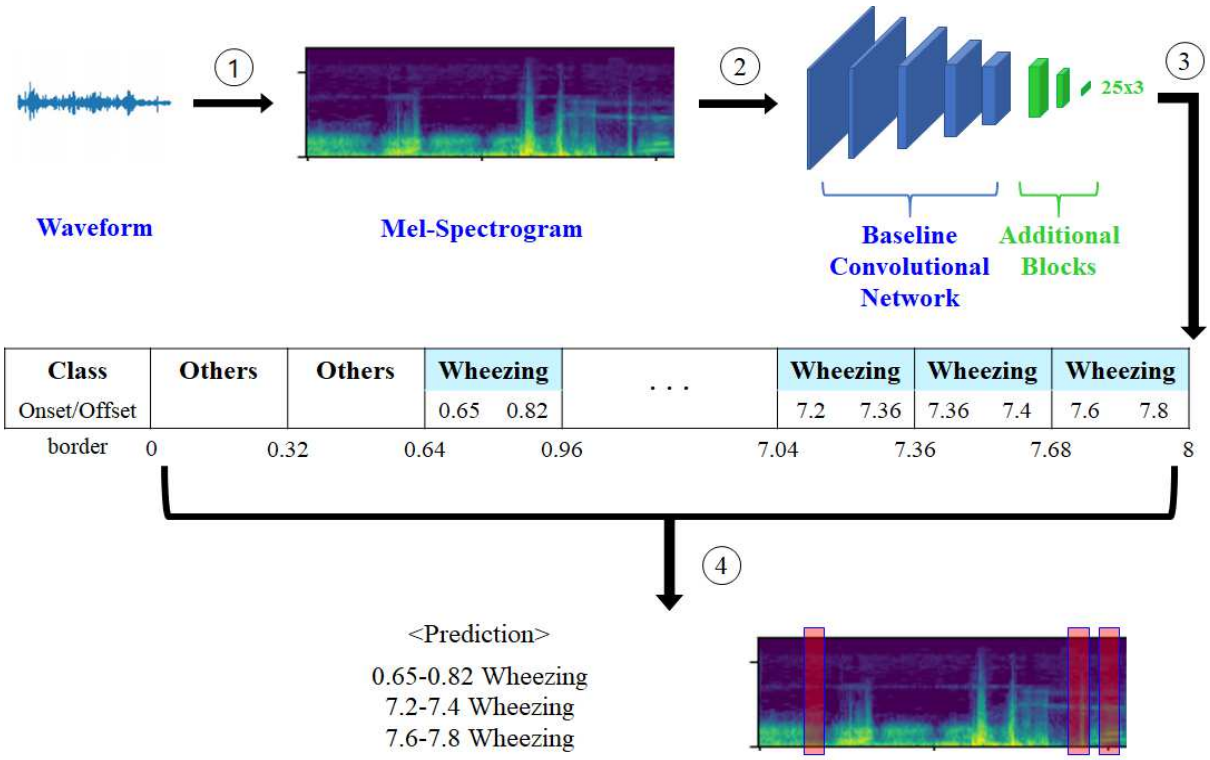
(Springer International Publishing, 2017).

44. Elphick, H. E., Ritson, S., Rodgers, H. & Everard, M. L. When a “wheeze” is not a wheeze: acoustic analysis of breath sounds in infants. *Eur. Respir. J.* **16**, 593 (2000).

45. Emmanouilidou, D., McCollum, E. D., Park, D. E. & Elhilali, M. Computerized lung sound screening for pediatric auscultation in noisy field environments. *IEEE Trans. Biomed. Eng.* **65**, 1564–1574 (2018).

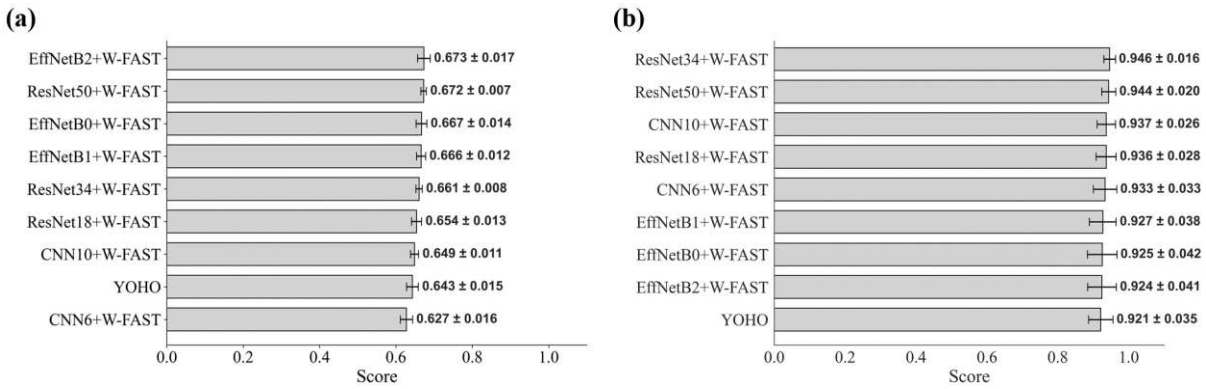
Figure legends

Figure 1. Schematic diagram of the W-FAST architecture for respiratory sound segmentation



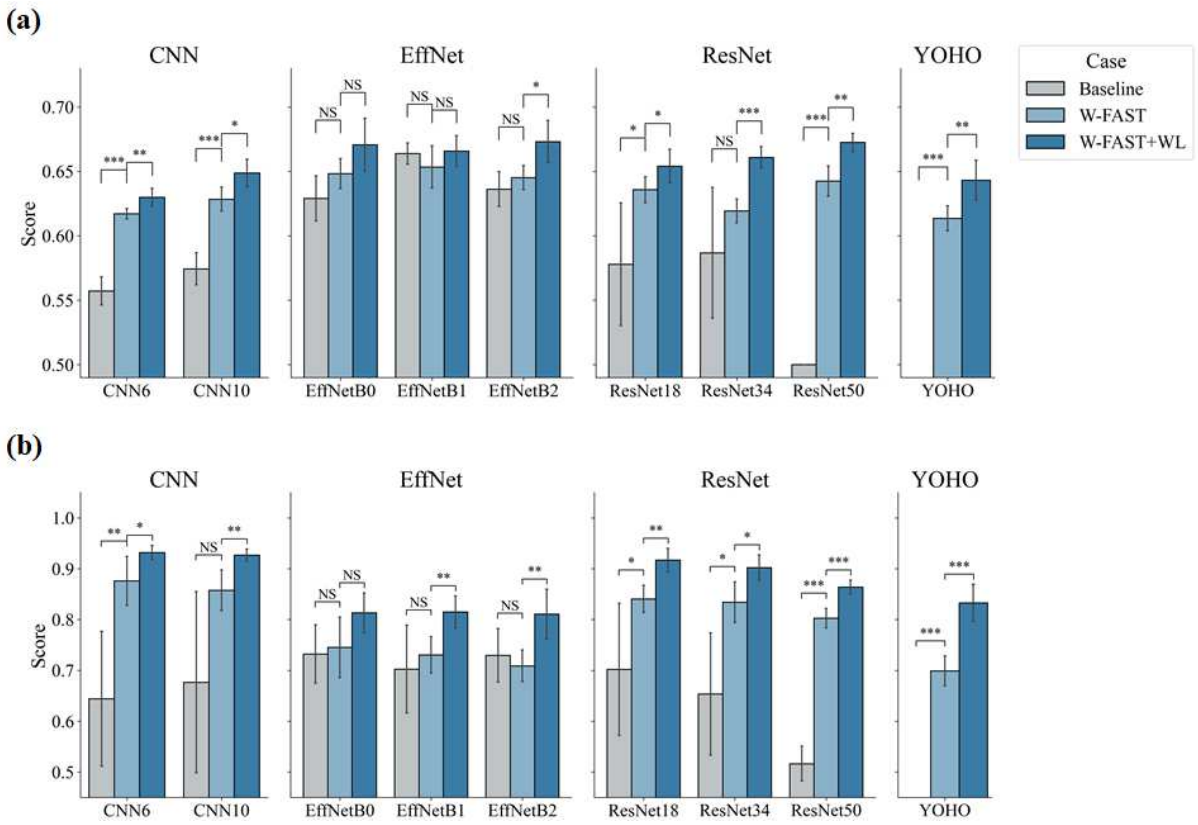
The framework consists of four main stages: (1) feature extraction using Kaldi's filterbanks to generate mel-spectrograms from raw audio waveforms; (2) input processing of the spectrograms into the deep learning backbone; (3) multi-task prediction of class probabilities and temporal coordinates (onset/offset) via sigmoid activation and coordinate scaling across 25 grid blocks; and (4) post-processing for aggregate prediction and visual representation of the segmented respiratory events.

Figure 2. Performance evaluation and comparative analysis across diverse backbone networks.



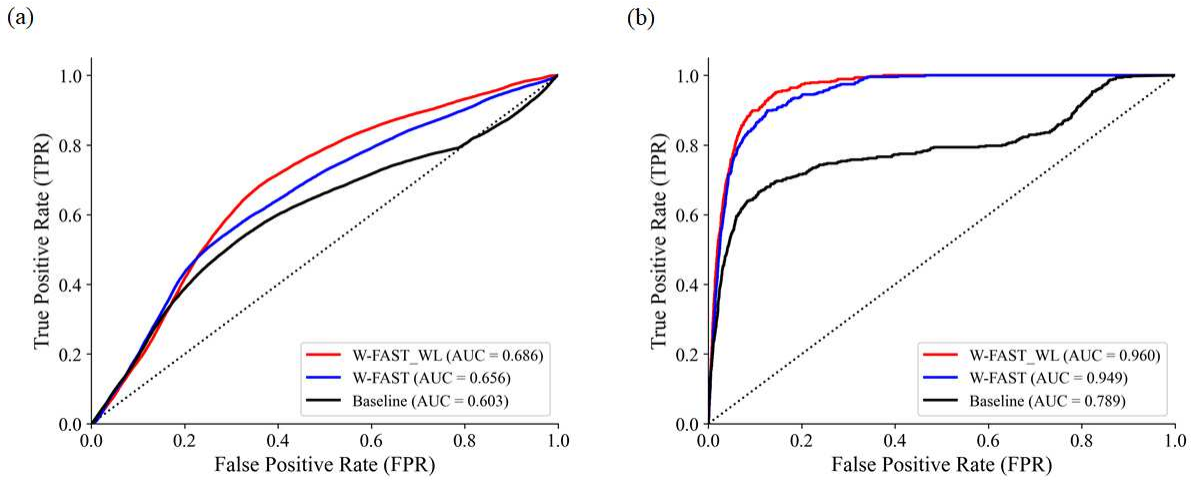
(a) Scores on the ICBHI 2017 dataset and (b) Scores on the SNUBH dataset. Each chart is sorted by descending order of their mean scores derived from five independent runs, with error bars indicating the standard deviation. All configurations incorporate the proposed weighted loss strategy.

Figure 3. Performance comparison and ablation study across diverse backbones on each dataset.



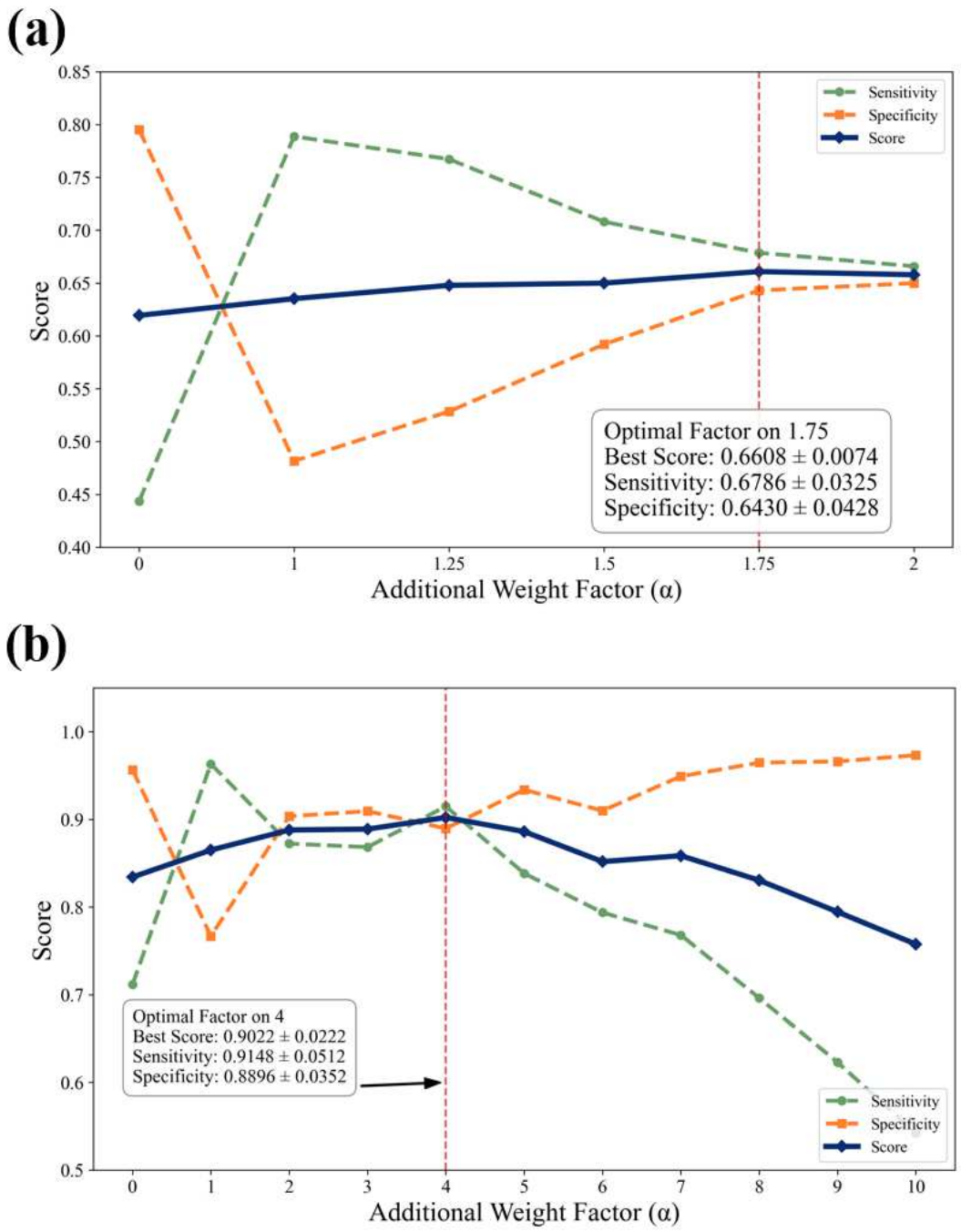
(A) on the ICBHI 2017 dataset, (B) on the SNUBH dataset. The bar charts represent the mean ICBHI scores for various architectures, including ResNet, CNN, EfficientNet, and the proposed W-FAST framework. Performance is evaluated across three configurations: (1) baseline, (2) the proposed model (W-FAST), and (3) the proposed model with Weighted Loss (W-FAST+WL). Error bars denote the standard deviation. Statistical significance was determined using a t-test for comparisons between the baseline vs. W-FAST and W-FAST vs. W-FAST+WL (*p < 0.05, **p < 0.01, ***p < 0.001; NS: not significant).

Figure 4. AUROC performance comparison of the baseline, W-FAST, and W-FAST with weighted loss (W-FAST_WL) models on the (a) ICBHI 2017 dataset and (b) SNUBH dataset.



The results indicate that the proposed W-FAST_WL (red) consistently outperforms the baseline (black) and standard W-FAST (blue) models.

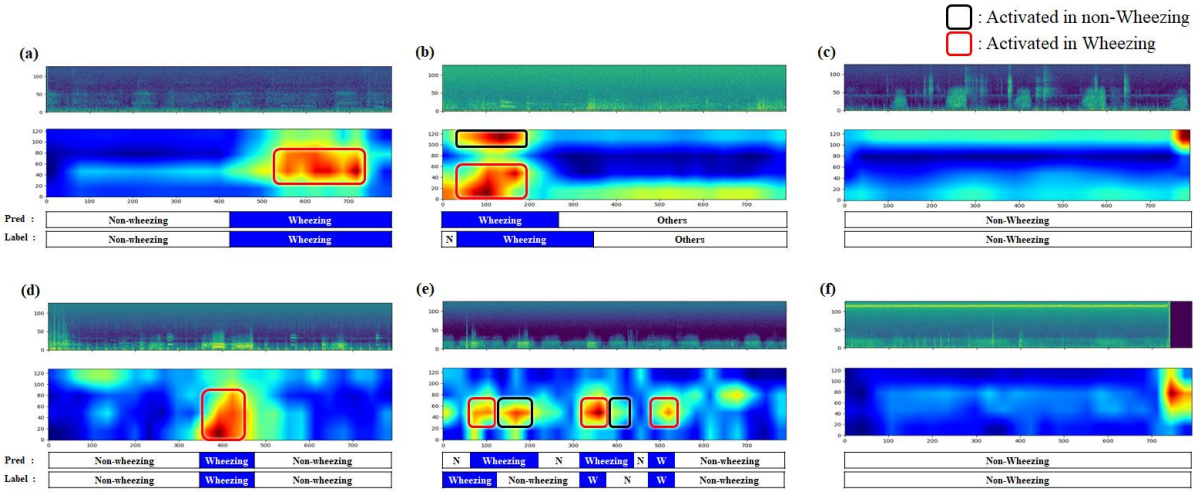
Figure 5. Effect of the additional weighting factor (α) on sensitivity–specificity trade-off.



Performance metrics according to the additional weighting factor (α) on the (a) ICBHI 2017 and (b) SNUBH datasets using a ResNet34 backbone. The score was calculated as the mean of

sensitivity and specificity. The red vertical dashed lines indicate the α values that produced the highest score: α = 1.75 for ICBHI 2017 and α = 4 for SNUBH. An α value of 0 represents the baseline condition without additional weighting.

Figure 6. Visual interpretability analysis of W-FAST predictions using Score-CAM



Representative results obtained using the ResNet34-based W-FAST+WL model. (a–c) show examples from the ICBHI 2017 test set, and (d–f) show examples from the held-out SNUBH test folds. (a) and (d) show accurate wheezing localization, (b) and (e) show partial localization or noise-related activation, and (c) and (f) show correctly predicted non-wheezing examples.

Table 1. Baseline Characteristics of Datasets

		ICBHI 2017	SNUBH
Nationality		Portugal and Greece	Republic of Korea
Patients		Adult and Children	Children
Age (MIN-MAX, Median)		(0-93, 61)	(0~19, 4)
Audio Files (n)		920	605
Total Length (h)		5.5	1.9
Class (n, ratio)	Normal	(3,642, 0.53)	(531, 0.88)
	Wheezing	(886, 0.13)	(74, 0.12)
	Crackle	(1,864, 0.27)	-
	Both	(506, 0.07)	-

Abbreviations: ICBHI, International Conference on Biomedical and Health Informatics; SNUBH, Seoul National University Bundang Hospital; MIN, Minimum; MAX, Maximum