*Sensors and Materials*, Vol. 38, No. 6 (2026) 3271–3286
MYU Tokyo



S & M 4502

# Robust Lung Sound Anomaly Detection Model for Wearable Devices Using Unsupervised Learning and Noise Distinction Algorithm

Takehiro Hirasawa,[1*] Wataru Noguchi,[2] Yasumasa Tamura,[3]
Kaoruko Shimizu,[4] Satoshi Konno,[4] and Masahito Yamamoto[3]

[1]Graduate School of Information Science and Technology, Hokkaido University,
Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido 060-0814, Japan
[2]Education and Research Center for Mathematical and Data Science, Hokkaido University,
Kita 12, Nishi 7, Kita-ku, Sapporo, Hokkaido 060-0812, Japan
[3]Faculty of Information Science and Technology, Hokkaido University,
Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido 060-0814, Japan
[4]Faculty of Medicine, Hokkaido University
Kita 15, Nishi 7, Kita-ku, Sapporo, Hokkaido 060-8638, Japan

(Received January 5, 2026; accepted May 22, 2026)

***Keywords:*** anomaly detection, unsupervised learning, autoencoder, lung sounds, Mel-spectrogram

The early detection of lung diseases in community healthcare settings is crucial, yet clinicians face challenges such as equipment limitations and heavy workloads. While AI-based automated screening using wearable devices offers a solution, developing robust models is difficult owing to the scarcity of abnormal data and the prevalence of environmental noise in real-world settings. We propose a noise-robust unsupervised anomaly detection framework for lung sounds acquired from wearable devices. This framework relies on continuous acoustic data acquired through a microphone sensor embedded in a prototype wearable device. The model is trained exclusively on normal lung sounds using a U-Net–based autoencoder with a composite loss function (mean squared error and structural similarity loss) to learn the manifold of normal respiration. The primary contribution of this work is a novel anomaly detection algorithm designed to distinguish between pathological anomalies and environmental noises. While pathological sounds (e.g., fine crackles) are stationary and synchronized with respiration, environmental noises (e.g., coughing and friction) are typically transient. The proposed algorithm evaluates the temporal persistence of reconstruction errors across sliding windows to calculate an anomaly detection rate. Experimental results demonstrate that while standard autoencoders struggle with noise, the proposed method successfully differentiates between persistent disease-related anomalies and transient noise artifacts. In this work, we demonstrate the feasibility of robust, unsupervised lung sound screening in noisy, real-world environments.

*Corresponding author: e-mail: hirasawa.takehiro.e5@elms.hokudai.ac.jp
https://doi.org/10.18494/SAM6160

ISSN 0914-4935 

## 1. Introduction

In regions with aging populations and limited medical resources, the burden on healthcare caregivers in community and home-visit settings is becoming increasingly significant.[1] The early detection of lung diseases, which have a high prevalence among the elderly, is a pressing issue. However, home-visit care is often constrained by the lack of portable diagnostic equipment and the limited time medical professionals can spend with each patient. Consequently, there is a growing demand for systems that can automatically screen for respiratory abnormalities without requiring constant expert supervision.

The integration of wearable devices and AI into medical diagnostics offers a promising solution to these challenges.[2] Wearable devices capable of measuring vital signs, including lung sounds, can enable continuous monitoring in home-care settings. Recent advancements in acoustic sensing technology for wearable devices have enabled the capture of high-resolution respiratory sounds in real-world settings, providing the essential data for our proposed diagnostic system. By automatically detecting abnormalities and guiding patients to specialized medical institutions only when necessary, such systems can significantly optimize medical resource allocation. However, deploying AI models in real-world environments presents distinct technical hurdles compared with controlled clinical settings.[3]

A primary challenge is the scarcity of labeled abnormal data[4] and the susceptibility to noise. Collecting large-scale, accurately labeled abnormal lung sound datasets is practically infeasible owing to the rarity of specific pathologies and the high cost of expert annotation. This necessitates an unsupervised learning approach trained solely on easy-to-collect normal data. However, data from wearable devices are inevitably corrupted by various types of noise,[5] such as environmental sounds, device friction, and coughing. Conventional unsupervised anomaly detection methods, which rely on reconstruction errors, often fail in this context because they flag both "pathological anomalies" and "noise anomalies" as defects. Specifically, standard methods lack the ability to distinguish between stationary abnormal sounds characteristic of lung diseases (e.g., fine crackles) and sudden, transient noise events.

To address the above challenges, we propose a robust anomaly detection algorithm tailored for wearable lung sound monitoring. We employ a convolutional autoencoder[6]-based architecture to learn the characteristics of normal breath sounds. Beyond simple reconstruction error-based anomaly detection, we introduce a novel post-processing algorithm that leverages the temporal characteristics of sound events. By analyzing the persistence of anomalies across sliding temporal windows, our method distinguishes between persistent pathological patterns and transient noise artifacts. This research demonstrates a practical approach to achieving robust lung sound screening in noisy, real-world environments.

## 2. Problem Formulation and Preliminaries

In this section, we formalize the problem of unsupervised anomaly detection in lung sounds collected via wearable devices and provide the theoretical background for the time–frequency representations used in this study. We first define the anomaly detection task under the constraint

of training exclusively on healthy data. Subsequently, we detail the signal processing pipeline, specifically the mathematical derivation and configuration of Mel-spectrograms tailored for interstitial lung disease (ILD) characteristics.

### 2.1 Unsupervised anomaly detection in lung sounds

The primary objective of this research is to construct a binary classifier $f: \mathcal{X} \rightarrow \{0, 1\}$ capable of distinguishing between normal lung sounds ($y = 0$) and anomalous lung sounds ($y = 1$), where $\mathcal{X}$ denotes the space of audio signal segments. In a conventional supervised learning setting, the training dataset $\mathcal{D}_{train}$ consists of pairs $\{(x_i, y_i)\}_{i=1}^{N}$ covering both classes. However, in the context of community healthcare and home-visit settings, acquiring a comprehensive dataset of labeled abnormal sounds is practically infeasible owing to the scarcity of patients with specific anomalies and the high cost of expert annotation. Furthermore, almost all data collected in the field are normal.[3]

Therefore, we adopt an unsupervised anomaly detection framework.[7] The model is trained exclusively on a dataset of normal lung sounds.

$$\mathcal{D}_{train} = \{x_i \in \mathcal{X}_{normal} | i = 1, \cdots, N\} \tag{1}$$

The goal is to learn a reconstruction mapping $g_\theta(\cdot)$ that captures the manifold of normal respiratory patterns. During the inference phase, the model evaluates an unseen sample $x_{test}$. An anomaly score $\mathcal{A}(x_{test})$ is computed using the deviation from the learned normal manifold. The classification decision is determined by a threshold $\tau$.

$$\hat{y} = \begin{cases} 1 \text{ (anomaly) if } \mathcal{A}(x_{test}) > \tau \\ 0 \text{ (normal)} \quad \text{otherwise} \end{cases} \tag{2}$$

### 2.2 Challenge of environmental and transient noise

A critical challenge in using wearable devices for lung sound recording is the susceptibility to noise. Unlike the recordings obtained in controlled clinical environments or using simulation mannequins, real-world data are corrupted by the following.

1. Environmental Noise: Ambient speech, background sounds
2. Device Friction: Noise generated by the movement of the device against clothing or skin
3. Physiological Artifacts: Heart sounds, muscle movements, and coughing

Let $x(t)$ be the observed signal. Ideally, $x(t) = s(t)$, where $s(t)$ is the clean lung sound. In practice, the observed signal is

$$x(t) = s(t) + n_{stationary}(t) + n_{transient}(t), \tag{3}$$

where $n_{stationary}(t)$ represents continuous background noise and $n_{transient}(t)$ represents sudden, high-amplitude events such as coughing and friction.

Conventional autoencoder-based anomaly detection calculates the reconstruction error $\|x - \hat{x}\|^2$ as the anomaly score.[7] However, autoencoders trained only on clean breathing sounds will fail to reconstruct both pathological abnormal sounds (e.g., fine crackles in ILD) and transient noises, resulting in high anomaly scores for both cases. Thus, the problem formulation must extend beyond simple reconstruction error. We must distinguish between the following.

- Pathological Anomalies: Stationary or quasi-periodic signals synchronized with the respiratory cycle (e.g., fine crackles occurring repeatedly during inspiration[8])
- Noise Anomalies: Transient, nonperiodic events (e.g., a single cough and friction spike)

This distinction is the core motivation for the temporal consistency algorithm proposed later.

### 2.3 Characteristics of target lung sounds

To design an appropriate input representation, we must consider the acoustic characteristics of the target disease. We focus on ILD. The hallmark auscultatory finding in ILD is fine crackles (also known as Velcro rales). Fine crackles are discontinuous, explosive sounds.[9]

- Frequency Content: Fine crackles typically have a higher frequency content than normal vesicular breath sounds but are generally contained within the range of 100 to 2000 Hz. Normal breath sounds are dominant below 1000 Hz.[9]
- Duration: Individual crackles are extremely short (<20 ms).[9]

### 2.4 Mel-spectrogram representation

We utilize the Mel-spectrogram as the input feature for the convolutional neural network (CNN)[10,11]-based autoencoder. The Mel-spectrogram provides a time–frequency representation that aligns with human auditory perception and effectively captures the transient nature of crackles while maintaining frequency resolution. Figure 1 illustrates the Mel-spectrogram of respiratory sounds recorded using the wearable device employed in this study. Further details regarding the wearable device and the preprocessing methods are provided in Sects. 3.1 and 3.2.

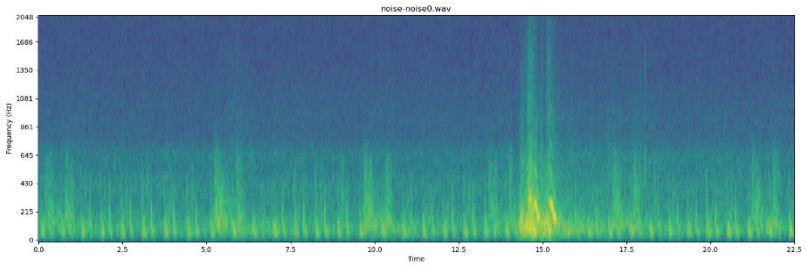



Fig. 1. (Color Online) Example of a Mel-spectrogram. The horizontal axis represents time, vertical axis represents frequency, and brightness of colors indicates amplitude.

### 2.4.1 Short-time Fourier transform

First, the discrete time-domain signal $x[n]$ is transformed into the time–frequency domain using the short-time Fourier transform (STFT). The signal is divided into overlapping frames and the discrete Fourier transform (DFT) is computed for each frame:

$$X(m,k) = \sum_{n=0}^{N_{FFT}-1} x[mH+n]w[n]e^{-j\frac{2\pi kn}{N_{FFT}}}, \tag{4}$$

where $m$ is the time frame index, $k$ is the frequency bin index, $N_{FFT}$ is the fast Fourier transform (FFT) window size, $H$ is the hop length (stride), and $w[n]$ is the window function (e.g., Hann window) to reduce spectral leakage. The power spectrogram is then obtained by taking the squared magnitude: $P(m,k) = \left|X(m,k)\right|^2$.

### 2.4.2 Mel-scale filter bank

The frequency resolution of the human ear is nonlinear; it is more sensitive to differences in lower frequencies than those in higher frequencies. The Mel scale approximates this perception. The conversion from hertz ($f$) to Mel-frequency ($f_{mel}$) is defined by Eq. (5).

$$f_{mel} = 2595 \cdot \log_{10}\left(1+\frac{f}{700}\right) \tag{5}$$

A Mel-scale filter bank consists of $N_{Mels}$ triangular filters spread equidistantly on the Mel scale. Let $H_j(k)$ be the $_j$-th filter bank. The Mel-spectrogram $S(m,j)$ is computed by applying these filters to the power spectrogram:

$$S(m,j) = \sum P(m,k) \cdot H_j(k). \tag{6}$$

Finally, a logarithmic scaling is applied to compress the dynamic range, matching the logarithmic perception of loudness (decibels):

$$S_{\log}(m,j) = \log(S(m,j)+\varepsilon), \tag{7}$$

where $\varepsilon$ is a small constant for numerical stability.

## 3. Data, Materials, and Methods

### 3.1 Data acquisition and materials

Data were recorded using a prototype wearable device developed by DENSEI COMTEC Inc., within the framework of a joint industry–academia–government collaboration involving

Hokkaido University and Ebetsu City Hospital.[12] The device is small, portable, and designed for vital sign measurement, with dimensions of 45 mm in length, 70 mm in width, and 30 mm in thickness. Lung sound data were acquired using the device's built-in microphone, held by an operator against the participant's chest, at a sampling frequency of 16,000 Hz. We collected "real data" from participants under a protocol approved by the Ethical Review Board for Life Science and Medical Research, Hokkaido University Hospital (Approval No. 022-0282). The dataset was partitioned such that there was no overlap of participants between the sets to ensure a rigorous evaluation. Data from 26 healthy volunteers were allocated for training, while those from one healthy volunteer were reserved for validation. The test set consisted of data from six patients diagnosed with ILD to represent the anomaly class and those from three healthy volunteers deliberately contaminated with noise (e.g., coughing) to evaluate robustness (Table 1). Among the primary medical abnormal sounds (e.g., coarse crackles, fine crackles, and wheezes), ILD is associated with fine crackles, and the test set is based on ILD-derived sounds.

### 3.2 Preprocessing and input representation

Raw audio signals, sampled at 16000 Hz, were segmented into fixed 5 s clips to generate model inputs. This fixed-length configuration was established to force the model to perform consistent reconstruction rather than variable-length processing, which avoids the negative effect of padding on reconstruction accuracy.[13] This specific duration was selected to ensure the capture of at least one complete phase of inspiration or expiration, where the characteristics of abnormal sounds are clearly represented. According to a physician's assessment, if neither phase is present within a 5 s interval, the segment should be categorized as a different type of anomaly and does not require representation in the training data. For the training phase, segments were extracted with a stride of 0.3 s to balance the trade-off between securing sufficient data volume and avoiding excessive overlap. This yielded a total of 3,891 segments. For validation and testing, segments were extracted sequentially with a stride of 2.5 s (resulting in a 50% overlap). This overlap strategy was employed to prevent the loss of critical features that might otherwise occur at the boundaries of the input window while also facilitating the time-series analysis described in the anomaly detection algorithm (Sect. 3.5). This resulted in 17 segments for validation, 66 segments for the anomaly test set, and 21 segments for the noise test set.

Subsequently, each 5 s segment was then converted into a Mel-spectrogram. A crucial step in this conversion was restricting the frequency range to 0–2048 Hz. This bandwidth covers the fundamental frequencies of breath sounds and fine crackles while excluding high-frequency

Table 1
Distribution of participants allocated for the anomaly detection model.

| Data type | Attribute | No. of participants |
|---|---|---|
| Train | Healthy | 26 |
| Validation | Healthy | 1 |
| Test | ILD Patient (anomaly) | 6 |
| Test | Healthy + Noise (noise) | 3 |

environmental noise. The STFT parameters were calibrated to map the 5 s clip to a final resolution of 64 × 128 (frequency bins × time steps).

### 3.3 Model architecture

The proposed anomaly detection model is built upon a convolutional autoencoder architecture derived from the U-Net[14] framework. To prevent the model from learning simple identity mappings, which would compromise the anomaly detection capability, skip connections between shallow layers of the encoder and decoder were removed. Deeper skip connections were retained to aid in reconstructing fine details. Furthermore, standard CNNs typically apply the same kernel across the entire image,[15,16] which can lead to a loss of frequency-dependent information in spectrograms. To address this, the model accepts a two-channel input. The first channel contains the standard 64 × 128 Mel-spectrogram, while the second channel encodes the frequency position, with the upper half set to 1 and the lower half to −1. This explicit frequency encoding compensates for the spatial invariance of CNN kernels, allowing the model to better learn frequency-specific features. The model outputs a single-channel 64 × 128 reconstructed Mel-spectrogram.

### 3.4 Model training

The model was trained exclusively on the 3891 normal segments from the train set. The objective is for the model to learn the manifold of normal lung sounds, such that any input deviating from this manifold will result in poor reconstruction.

We compared three loss function configurations: Model A using only mean squared error (MSE) loss,[17] Model B using a composite of MSE and structural similarity[18] (SSIM) loss[19] ($0.5 \cdot \mathcal{L}_{MSE} + 0.5 \cdot \mathcal{L}_{SSIM}$), and Model C using only SSIM loss. Here, $\mathcal{L}_{SSIM}$ is calculated as 1 − SSIM($x$, $y$). While $\mathcal{L}_{MSE}$ focuses on pixel-wise local errors, $\mathcal{L}_{SSIM}$ evaluates structural similarity considering luminance, contrast, and structure. On the basis of qualitative reconstruction results (Sect. 4.1), Model B was selected as the final model because it produced sharp, accurate reconstructions without the blurring seen in Model A or the convergence issues seen in Model C.

### 3.5 Anomaly detection algorithm

A simple anomaly score on an isolated 5 s window is insufficient, as transient, high-amplitude noises (e.g., coughing and device friction) can produce high reconstruction errors and be mistaken for anomaly. To solve this, we developed an algorithm that analyzes the temporal characteristics of the detected anomalies. This approach was conceptually inspired by the real-time event-counting framework proposed in Ref. 20, which analyzes sequences of classified segments to count wheezing events. However, our method differs fundamentally in both its architecture and objective. Whereas their method relies on supervised classification to identify discrete events, our model uses an unsupervised autoencoder's reconstruction error.

Furthermore, our goal is not to count discrete events, but to calculate the anomaly detection rate (*ADR*). This rate measures the persistence of the anomaly, allowing us to robustly separate stationary abnormal sounds (which occur periodically with respiration) from brief, transient noises.

The trained autoencoder (Model B) is used to detect anomalies based on reconstruction error. We developed a multistep algorithm to quantify this error and ensure robustness against noise.

First, the anomaly score is calculated using a weighted error metric. Since low-amplitude regions in a spectrogram typically contain less significant information, we emphasize errors in high-amplitude regions by weighting the pixel-wise difference by the input intensity.[21] The anomaly score $a_i$ for each pixel i is defined as the squared standardized weighted error:

$$a_i = \left( \frac{x_i(x_i - y_i) - \hat{\mu}}{\hat{\sigma}} \right)^2, \tag{8}$$

where $x_i$ and $y_i$ are the input and reconstructed values, and $\hat{\mu}$ and $\hat{\sigma}$ are the mean and standard deviation of the weighted error derived from the training set, respectively. Assuming that the standardized weighted error follows a normal distribution, this score follows a chi-squared ($\chi^2$) distribution with one degree of freedom. Pixels are flagged as "anomalous" if their score exceeds a threshold determined by the 99th percentile of this distribution.

Finally, to distinguish between transient high-amplitude noises, such as coughing, and persistent abnormal sounds such as fine crackles, we leverage the temporal consistency of these events. Abnormal sounds are typically stationary and linked to the respiratory cycle, whereas noise is often sudden and short-lived. The robust detection algorithm proceeds as follows.

1. A long measurement (e.g., 1 min) is segmented into n overlapping 5 s windows (2.5 s slide).
2. For each window j, the anomaly mask $M_j$ is computed, flagging pixels that exceed the $\chi^2$ threshold.
3. A window is defined as an "anomalous window" (*AW*) if the proportion of anomalous pixels exceeds 7% ($T_{pixel}$), which is based on empirical tuning.

$$AW_j = \begin{cases} 1 \text{ if } \left( \dfrac{\sum M_j}{total\ pixels} \right) > T_{pixel} \\ 0 \text{ otherwise} \end{cases} \tag{9}$$

4. According to Eq. (10), the final output is *ADR*, calculated as the proportion of anomalous windows over the total number of windows (*n*).

$$ADR = \frac{\sum_{j=1}^{n} AW_j}{n} \tag{10}$$

A high *ADR* indicates a persistent anomaly, while a low *ADR* suggests transient noise.

## 4. Results

### 4.1 Model training and reconstruction quality

The training performance varied significantly across the three loss function configurations. Model A ($\mathcal{L}_{MSE}$) and Model B ($0.5 \cdot \mathcal{L}_{MSE} + 0.5 \cdot \mathcal{L}_{SSIM}$) converged stably. However, Model C ($\mathcal{L}_{SSIM}$) failed to learn effectively and fell into a local minimum, likely because SSIM loss alone does not sufficiently penalize outputs that are all zero if the input contrast is low.

Figure 2 illustrates the reconstruction quality on a validation segment from a healthy participant. The reconstruction from Model A was notably blurry, suffering from a loss of detail typical of MSE-based autoencoders. Conversely, Model B produced a sharp reconstruction that faithfully reproduced the structure of the input, including low-frequency components. Model C failed to reconstruct key components, consistent with its poor convergence. Consequently, Model B was adopted for the subsequent anomaly detection tasks.

### 4.2 Anomaly detection performance

We applied the trained Model B to the test set containing abnormal data. Figure 3 shows the reconstruction of a segment from a patient with ILD. Since the model was trained only on normal sounds, it failed to reconstruct the abnormal features (fine crackles), resulting in a discrepancy between the input and the output. When visualizing the pixel-wise anomaly scores,

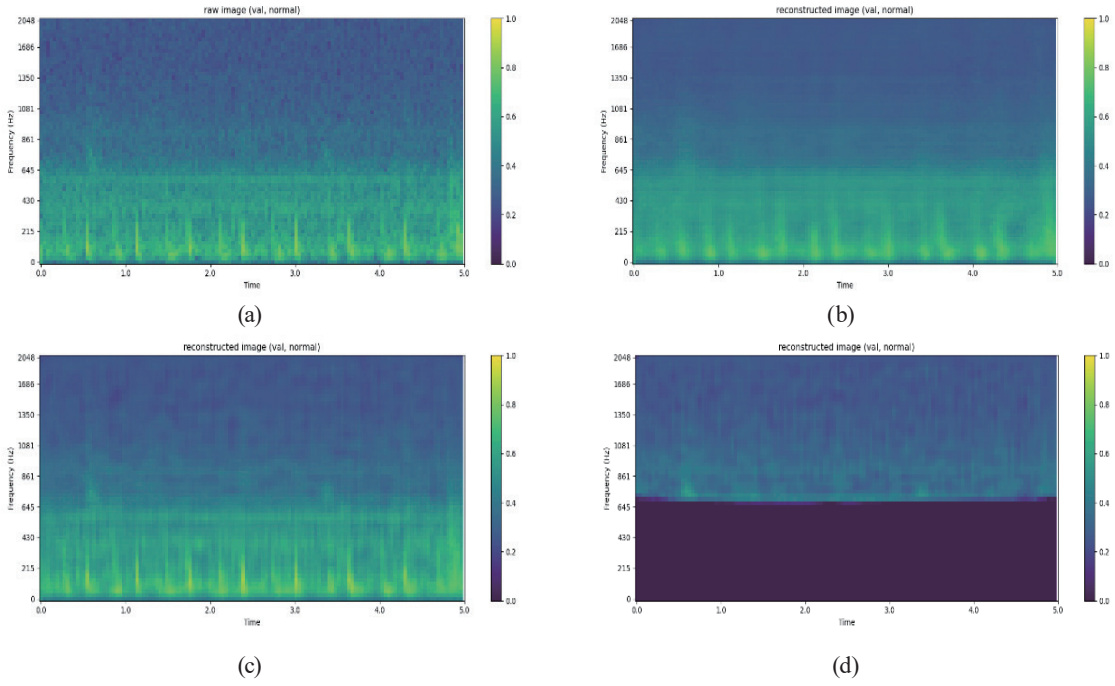
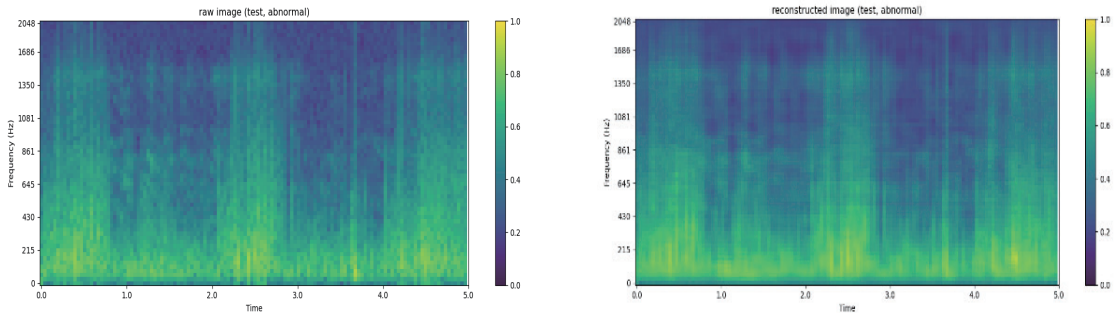



Fig. 2. (Color Online) Reconstruction results for a normal (healthy) lung sound. (a) Original input, (b) Model A (MSE only), (c) Model B (MSE + SSIM), and (d) Model C (SSIM only).




(a) (b)

Fig. 3. (Color Online) Reconstruction of an anomalous (ILD) lung sound. (a) Original input spectrogram and (b) reconstructed output from Model B.

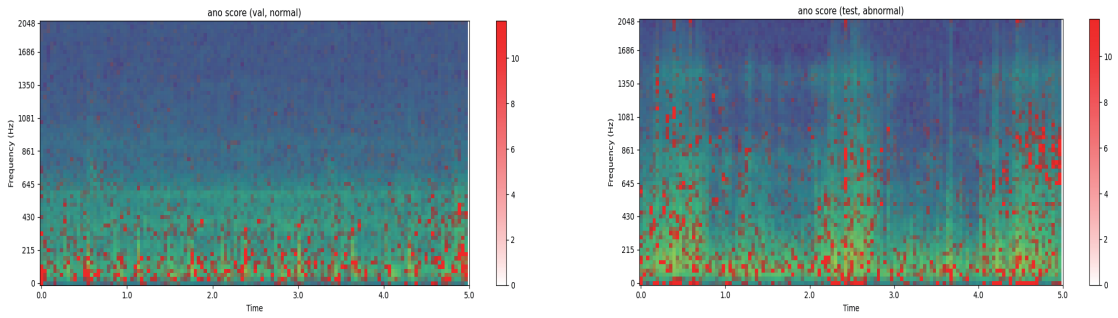
healthy participants showed universally low scores [Fig. 4(a)], whereas patients exhibited high anomaly scores concentrated in the mid-to-high frequency bands, corresponding to the timing of fine crackles [Fig. 4(b)].

### 4.3 Robustness to noise: algorithm results

We evaluated the complete algorithm (Sect. 3.5) on data of all participants. The algorithm demonstrated a strong capability to separate healthy, noisy, and abnormal recordings, though specific outliers were identified. Table 2 shows the *ADR* values for all participants. As shown, the separation was highly effective. 26 of the 27 healthy participants in quiet environments had *ADR*s near zero (0.00–5.26%). In contrast, five of the six ILD patients showed high *ADR* values (63.64–100.00%). The three healthy participants exposed to transient noise (e.g., coughing) produced a distinct, moderate *ADR* cluster (22.22–37.50%). If we set the diagnostic threshold at approximately 50%, the algorithm correctly classifies 29 of 30 healthy cases as normal and five of six anomaly cases as abnormal.

Two notable outliers were observed: one healthy participant registered an unusually high *ADR* of 60.00% and one patient registered an unusually low *ADR* of 25.00%. These outliers are analyzed in Sect. 5.

Figure 5 shows the visualizations of results of applying the proposed algorithm for each participant attribute and environment. The color-coded status bar is plotted above each spectrogram to visualize the temporal diagnosis based on the sliding windows. White indicates that the overlapping windows covering that duration were consistently classified as normal, while black indicates that at least one window was flagged as anomalous. Consequently, Fig. 5(a) displays a continuous white bar and Fig. 5(b) shows a continuous black bar reflecting persistent anomaly. In Fig. 5(c), the bar turns black only during the 12.5–17.5 s interval, demonstrating the algorithm's ability to temporally isolate the transient cough noise.




(a) (b)

Fig. 4. (Color Online) Heatmap of reconstruction errors transmitted through the Mel-spectrogram. (a) Normal and (b) anomaly spectrograms.

Table 2
*ADR* values for all participants, demonstrating separation between healthy, noise-affected, and abnormal recordings.

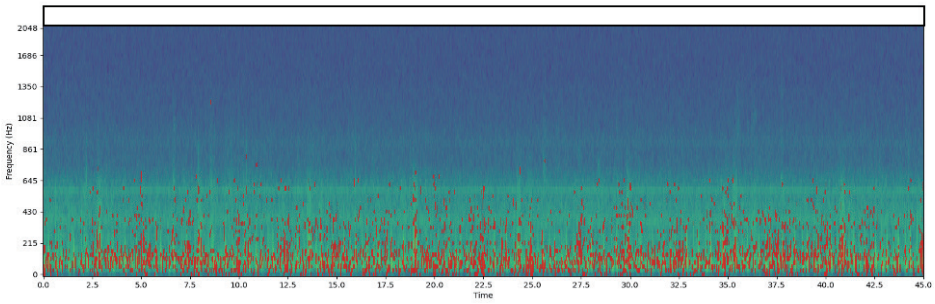
| Attribute | No. of Participants | *ADR* (%) | Classification (threshold @ 50%) |
|---|---|---|---|
| Patient | 5 | 63.64–100.00 | **Correct (anomaly)** |
| | 1 | 25.00 | Incorrect (normal) |
| Healthy + Noise | 3 | 22.22–37.50 | **Correct (normal)** |
| Healthy | 26 | 0.00–5.26 | **Correct (normal)** |
| | 1 | 60.00 | Incorrect (anomaly) |

## 5. Discussion

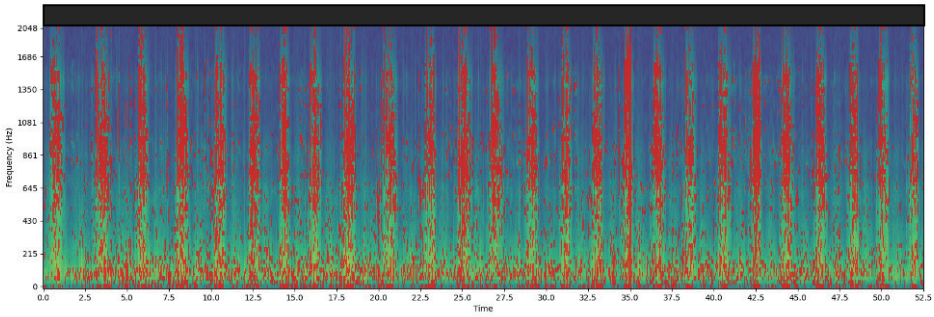
We successfully developed an unsupervised anomaly detection model for lung sounds using data from a novel wearable device. The choice of a composite MSE and SSIM loss function was critical; combining pixel-level accuracy with structural fidelity allowed the model to learn a robust representation of normal lung sounds.

The primary contribution of this work is the algorithm that distinguishes transient noise from stationary abnormal sounds. Simple pixel-error thresholding is insufficient for real-world data, as it flags both coughs and crackles as "anomalous". By introducing the *ADR* metric over a sliding window, we leverage the temporal nature of these sounds. Abnormal sounds such as fine crackles are persistent and linked to the respiratory cycle, leading to high *ADR* values. In contrast, transient noises such as coughs are sudden and short-lived, resulting in low-to-moderate *ADR* values. This distinction allows for a reliable classification of abnormal recordings while rejecting most noise-corrupted normal recordings.

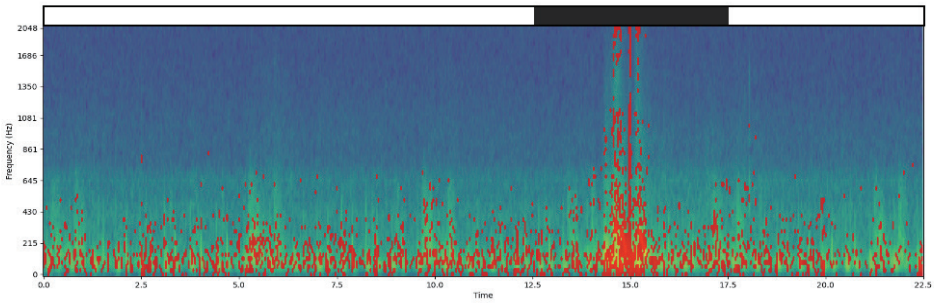
The algorithm's performance on the two outliers (Table 2) warrants discussion. The patient who registered a low *ADR* of 25.00% was noted to have a very short recording time of approximately 10 s. This brief duration is insufficient for the sliding-window algorithm to




(a)




(b)




(c)

Fig. 5. (Color Online) Results of the robust detection algorithm. (a) Healthy participant: 0/18 anomalous windows (ADR=0%). (b) Patient: 21/21 anomalous windows (*ADR* = 100%). (c) Noise (cough): 2/9 anomalous windows (ADR=22.2%).

reliably establish the persistence of the anomaly, leading to a skewed, low *ADR*. Conversely, the healthy participant with a high *ADR* of 60.00% was recorded in an environment with persistent, steady-state noise from the surrounding medical equipment. The algorithm correctly identified this sound as "persistent" but misclassified it as abnormal because its temporal characteristic (steady) was different from the transient noise (cough) it was designed to reject.

This highlights a key limitation and a direction for future work. The current algorithm effectively separates transient noise from persistent signals but does not distinguish between persistent abnormal sound (e.g., crackles) and persistent environmental noise (e.g., machine hum). Future iterations must incorporate more detailed medical features—such as specific frequency signatures, periodicity, or the inspiratory/expiratory phase timing of anomalies—to achieve a more granular and medically informed separation between different types of persistent sounds.

## 6. Conclusions

In this study, we validated a lung sound anomaly detection model designed for a new wearable device. We demonstrated that a U-Net-based autoencoder with a composite MSE-SSIM loss function, trained only on normal data, can effectively detect abnormal sounds from ILD patients.

Most importantly, the proposed robust algorithm that analyzes the temporal persistence of anomalies successfully distinguished between stationary abnormal sounds (high *ADR*) and transient noises (low *ADR*), addressing a key challenge for practical deployment. While the model showed high binary classification accuracy (correctly identifying 29 of 30 healthy and five of six anomalous cases), it was shown to be susceptible to steady-state, nonabnormal noise that can be mistaken for a persistent anomaly.

Future work will focus on integrating more detailed medical and signal characteristics to differentiate between abnormal and environmental persistent sounds. Additionally, we will continue to collect a larger volume of data to perform a more detailed and continuous verification of the model's performance. The integration of acoustic sensor technology with our robust unsupervised learning algorithm establishes an effective automated screening system capable of handling environmental noise. The results of this research confirm the potential of unsupervised AI models paired with wearable devices to serve as effective, low-cost screening tools for lung diseases in community healthcare.

## Acknowledgements

The authors express their sincere gratitude to DENSEI COMTEC Inc. and Manager Naoyuki Hasebe and the staff of Ebetsu City Hospital for their extensive cooperation in data acquisition and project administration.

## Author Contributions

T. H. conceptualized the study, developed the methodology and software, conducted the experiments, performed the formal analysis, and wrote the original draft of the manuscript. W. N., Y. T., and M. Y. supervised the technical direction of the research and contributed to the interpretation of results and the refinement of the study's claims. K. S. and S. K. administered

the overall joint research project and managed the ethical approval procedures at Hokkaido University Hospital. K.S. was also responsible for clinical data acquisition, specifically recording lung sounds from patients.

## References


1 K. Harai, H. Honda, and M. Kawaharada: J. Rural Med. **15** (2020) 16. https://doi.org/10.2185/jrm.3006
2 K. Perez, D. Wisniewski, A. Ari, K. Lee, C. Lieneck, and Z. Ramamonjiarivelo: Healthcare **13** (2025) 324. https://doi.org/10.3390/healthcare13030324
3 M. I. Ahmed, B. Spooner, J. Isherwood, M. Lane, E. Orrock, and A. Dennison: Cureus **15** (2023) e46454. https://doi.org/10.7759/cureus.46454
4 J. Saldanha, S. Chakraborty, S. Patil, K. Kotecha, S. Kumar, and A. Nayyar: PLoS One **17** (2022) e0266467. https://doi.org/10.1371/journal.pone.0266467
5 D. Emmanouilidou and M. Elhilali: Proc. 35th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society (IEEE, 2013) 2551. https://doi.org/10.1109/EMBC.2013.6610060
6 G. E. Hinton and R. R. Salakhutdinov: Science **313** (2006) 504. https://doi.org/10.1126/science.1127647
7 A. Desmet and M. Delore: Proc. Annu. Conf. Prognostics and Health Management Society (PHM Society, 2017). https://doi.org/10.36001/phmconf.2017.v9i1.2401
8 G. R. Epler, C. B. Carrington, and E. A. Gaensler: Chest **73** (1978) 333. https://doi.org/10.1378/chest.73.3.333
9 H. Pasterkamp, S. S. Kraman, and G. R. Wodicka: Am. J. Respir. Crit. Care Med. **156** (1997) 974. https://doi.org/10.1164/ajrccm.156.3.9701115
10 K. Simonyan and A. Zisserman: Proc. 3rd Int. Conf. on Learning Representations (OpenReview.net, 2015) 1.
11 B. F. P. Dossou and Y. K. S. Gbenou: Proc. 2021 IEEE/CVF Int. Conf. on Computer Vision Workshops (IEEE, 2021) 3526. https://doi.org/10.1109/ICCVW54120.2021.00393
12 Hokkaido University, Industry creation laboratories: https://www.mcip.hokudai.ac.jp/business/course/ (accessed November 2025) (in Japanese).
13 H. Lee and D. Shin: Sensors **25** (2025) 621. https://doi.org/10.3390/s25030621
14 O. Ronneberger, P. Fischer, and T. Brox: Proc. Medical Image Computing and Computer-Assisted Intervention (Springer, 2015) 234. https://doi.org/10.1007/978-3-319-24574-4_28
15 A. Ustubioglu, B. Ustubioglu, and G. Ulutas: Signal, Image Video Process. **17** (2023) 2211. https://doi.org/10.1007/s11760-022-02436-4
16 M. H. Tanveer, H. Zhu, W. Ahmed, A. Thomas, B. M. Imran, and M. Salman: Proc. 2021 Int. Conf. Computer, Control and Robotics (IEEE, 2021) 220. https://doi.org/10.1109/ICCCR49711.2021.9349416
17 D. E. Rumelhart, G. E. Hinton, and R. J. Williams: Nature **323** (1986) 533. https://doi.org/10.1038/323533a0
18 Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli: IEEE Trans. Image Process. 13 (2004) 600. https://doi.org/10.1109/TIP.2003.819861
19 P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger: Proc. 14th Int. Conf. Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP, 2019) 372. https://doi.org/10.5220/0007364503720380
20 S. Im, T. Kim, C. Min, S. Kang, Y. Roh, C. Kim, M. Kim, S. H. Kim, K. Shim, J. Koh, S. Han, J. Lee, D. Kim, D. Kang, and S. Seo: PLoS One **18** (2023) e0294447. https://doi.org/10.1371/journal.pone.0294447
21 N. Ito and M. Sugiyama: Proc. 2023 IEEE Int. Conf. Acoustics, Speech and Signal Processing (IEEE, 2023) 1. https://doi.org/10.1109/ICASSP49357.2023.10095988

## About the Authors

**Takehiro Hirasawa** received his B.S. degree from Hokkaido University, Japan, in 2025. He is currently a master's student at Hokkaido University. His research interests are in AI, image recognition, and image generation. (hirasawa.takehiro.e5@elms.hokudai.ac.jp)

**Wataru Noguchi** received his Ph.D. degree in information science and technology from Hokkaido University, Japan, in 2019. From 2019 to 2023, he was a postdoctoral researcher at Hokkaido University. From 2023 to 2025, he was a specially appointed assistant professor at the Education and Research Center for Mathematical and Data Science, Hokkaido University. His research interests include artificial intelligence, deep learning, and cognitive modeling. (w.noguchi@mdsc.hokudai.ac.jp)

**Yasumasa Tamura** received his Ph.D. degree from Hokkaido University, Japan, in 2015. From 2016 to 2017, he was a research fellow of the Japan Society for the Promotion of Science and a visiting researcher in Université Libre de Bruxelles, Belgium. In 2017, he became an assistant professor in Tokyo Institute of Technology (2017–2023). Currently, he is an assistant professor in Hokkaido University (2024–). His research interests include computational intelligence, swarm intelligence, combinatorial optimization, multi-agent systems, swarm robotics, and distributed systems. (ytamura@ist.hokudai.ac.jp)

**Kaoruko Shimizu** received her M.D. degree from Hokkaido University, Japan, in 2002 and her Ph.D. degree from the Graduate School of Medicine, Hokkaido University, Japan, in 2011. From 2011 to 2022, she was an assistant professor at Hokkaido University Hospital, Japan. Since 2024, she has been a lecturer at Hokkaido University. Her research interests are in radiology and physiology in respiratory disease. (okaoru@med.hokudai.ac.jp)

**Satoshi Konno** received his M.D. degree from Asahikawa Medical University, Japan, in 1995 and his Ph.D. degree from the Graduate School of Medicine, Hokkaido University, Japan, in 2001. From 2016 to 2018, he was an associate professor at Hokkaido University, Japan. Since 2019, he has been a professor at Hokkaido University. (satkonno@med.hokudai.ac.jp)

**Masahito Yamamoto** received his Ph.D. degree from the Graduate School of Engineering from Hokkaido University, Japan, in 1996. From 1996 to 1997, he was a research fellow of the Japan Society for the Promotion of Science. He has served as an assistant professor (1997–2000) and associate professor (2000–2012) at Hokkaido University. Since 2012, he has been a professor at the autonomous systems engineering laboratory, Hokkaido University, Japan. He is also a concurrent faculty member of the Center for Human Nature, Artificial Intelligence, and Neuroscience, Hokkaido University (2020–). His research interests include artificial life and intelligence, swarm intelligence, combinatorial optimization, and game and sports AI programming. (masahito@ist.hokudai.ac.jp)