# Edge-Server Collaborative System for Real-Time and In-Depth Damage Detection of Wind Turbine Blades Using Acoustic Signals

Zhi Zhu and Yoshinao Sato

*Fairy Devices Inc., Bunkyo-ku, Tokyo, 113-0034, Japan*
*zhu@fairydevices.jp, sato@fairydevices.jp*

## ABSTRACT

Efficient health monitoring is indispensable for the reliable operation of wind turbines. Damage to wind turbine blades, such as cracks and holes, typically generates whistle-like sounds during rotation. This study proposes a two-stage edge-server collaborative system for detecting blade damage using acoustic signals captured by arrays built from commodity microphones. The first stage employs a lightweight attention-based convolutional neural network to run on edge devices for the real-time binary classification to determine whether anomalous sounds are present. Suspicious time segments are stored for further analysis. The second stage uses a time-frequency sound event detection model that employs a detection transformer with an audio spectrogram transformer backbone to identify the time and frequency ranges of sound events via bounding boxes in the spectrograms. Owing to its high computational demand, this in-depth analysis is performed on a server. To validate the proposed system, acoustic signals were recorded intermittently for more than a year using micro-electromechanical system (MEMS) microphones externally attached to wind turbine towers. The models were trained and evaluated on a manually annotated dataset comprising 4,210 audio clips (15 s each) containing 14,420 sound events. The experimental results demonstrated that the binary classification model achieved an area under the receiver operating characteristic curve (AUC) of 0.920, whereas the sound event detection model attained an average precision at a 50% intersection-over-union threshold ($AP_{50}$) of 0.510. Furthermore, evaluations on test data under unseen conditions, comprising 496 clips with 135 sound events recorded by handheld recorders at different locations, yielded an AUC of 0.867 and an $AP_{50}$ of 0.440. The results highlight the robustness of the proposed system to variations in microphone types, recording locations, and environmental noise, demonstrating its strong potential for practical continuous automatic damage detection in wind power infrastructure.



## 1. INTRODUCTION

Reliable operation of wind power generation equipment requires health monitoring and predictive maintenance (M. Li et al., 2026; Cuesta, Leturiondo, Vidal, & Pozo, 2024). Cracks, holes, and other damages can occur in wind turbine blades due to various causes, including lightning strikes, object collisions, erosion, and material fatigue (Sun, Ooi, & Su, 2026; W. Wang, Xue, He, & Zhao, 2022). Such defects have been shown to reduce power generation efficiency (Algolfat, Wang, & Albarbar, 2023). If left unaddressed, they can potentially cause serious damage to the entire equipment, including blade fractures and tower collapse. Various methods have been investigated for detecting damage to wind turbine blades (Sun et al., 2026; W. Wang et al., 2022; Ding, Yang, & Zhang, 2023). These include strain measurements using resistive gauges and fiber-optic gratings (Fremmelev et al., 2022), acoustic emission monitoring (Van Dam & Bond, 2015), and visual inspection via drone-based imaging. Because different sensors capture distinct aspects of equipment health, damage detection methods for each modality are complementary.

This study addresses passive damage detection using acoustic signals. Cracks and holes on the surface of a wind turbine blade alter the airflow, generating whistle-like sounds in the audible frequency range when the blade rotates (Ding et al., 2023; Y. Zhu & Liu, 2023). Figure 1 shows examples of whistle-like sounds in the spectrograms. These sounds can be captured using commodity microphones placed on the surface of a wind turbine tower or in its vicinity (Kuo, Cheng, Lo, & Tu, 2023; Yang, Ding, & Zhou, 2025). Hence, there is no need to install sensors in the power generation equipment or mount external sensors on the blades (Ding et al., 2023). Sound classification (SC) models can be applied to determine the presence or absence of anomalous sounds in environmental noise (Y. Zhu, Liu, Li, Wan, & Cai, 2022; Y. Zhu & Liu, 2023). Damage can be continuously detected on edge devices using a lightweight SC model.

Notably, whistle-like sounds caused by blade damage have a distinctive spectral pattern, and a fair number of anomalous

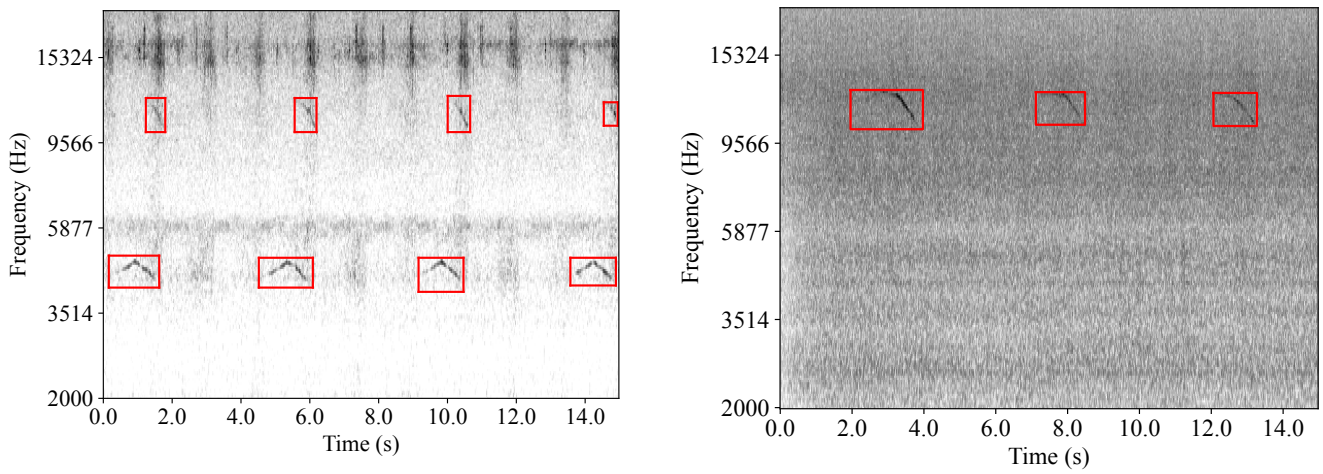



Figure 1. Examples of whistle-like sounds in spectrograms recorded using a custom-built device (left) and a commercial handheld recorder (right). Red boxes show the bounding boxes.

samples can be observed in practice. These characteristics contrast with those of conventional anomaly detection tasks, in which the patterns of anomalous signals are unknown, and only a few anomalous samples are available. This distinction allows us to treat blade damage detection as a supervised classification problem rather than an unsupervised anomaly detection problem. However, even if the blade is damaged, anomalous sounds may not necessarily be captured owing to various factors, such as wind speed, wind direction, blade rotation status, blade orientation, and the intensity and type of environmental noise.

While the presence of whistle-like sounds indicates damage to the blade, their shapes and positions in the spectrograms provide detailed information regarding the damage. Hence, sound event detection (SED) using bounding boxes in spectrograms, referred to as time-frequency SED, is indispensable for estimating the severity of cracks and holes that cause whistle-like sounds. Detection transformers (DETRs) for object detection in natural images (Carion et al., 2020; X. Zhu et al., 2021; F. Li et al., 2022; Liu et al., 2022; H. Zhang et al., 2023) can be adapted to SED with time-frequency bounding boxes in audio spectrograms (Z. Zhu & Sato, 2025). However, high-performance time-frequency SED models are computationally intensive and best executed on a server, making them unsuitable for edge devices.

Real-time anomaly screening at the edge significantly reduces data transmission overhead and optimizes server loads in industrial health management (F. Li, Li, & Peng, 2021; Zhou et al., 2021; C. Wang et al., 2023; von Däniken, Mikhaylov, Moallemi, Polonelli, & Magno, 2024; Lamdjad & Chaiter, 2026). This study proposes an edge-server collaborative system for wind turbine blade damage detection that combines two complementary models. In the first stage, a lightweight SC model continuously determines whether anomalous sounds are present for the real-time screening on edge devices. Suspicious time segments are forwarded to a server for further analysis. In the second stage, a time-frequency SED model identifies the bounding boxes of anomalous sounds in spectrograms on a server for in-depth analysis. The proposed system is designed for practical field deployment, enabling automatic and continuous damage detection using commodity sensors and edge-computing devices.

The key contributions of this study are as follows. (1) This study proposes an edge-server collaborative system comprising edge-side real-time continuous screening and server-side in-depth analyses for the health monitoring of wind power generation. (2) A comprehensive evaluation dataset was constructed, combining previously collected in-domain data using custom-built recording devices with newly collected out-of-domain data using commercial handheld recorders. (3) The effectiveness of the proposed system was experimentally demonstrated. (4) The proposed system can be implemented using commodity audio sensors and edge-computing devices for field deployment.

## 2. METHOD

### 2.1. System Architecture

The proposed system comprises two complementary models, as illustrated in Figure 2. In the first stage, a lightweight SC model determines whether input audio signals contain anomalous sounds, namely whistle-like sounds caused by cracks and holes in the wind turbine blades. Binary classification is performed continuously using sliding time windows in real time on edge devices installed on wind turbine towers. The time segments determined by the model to contain

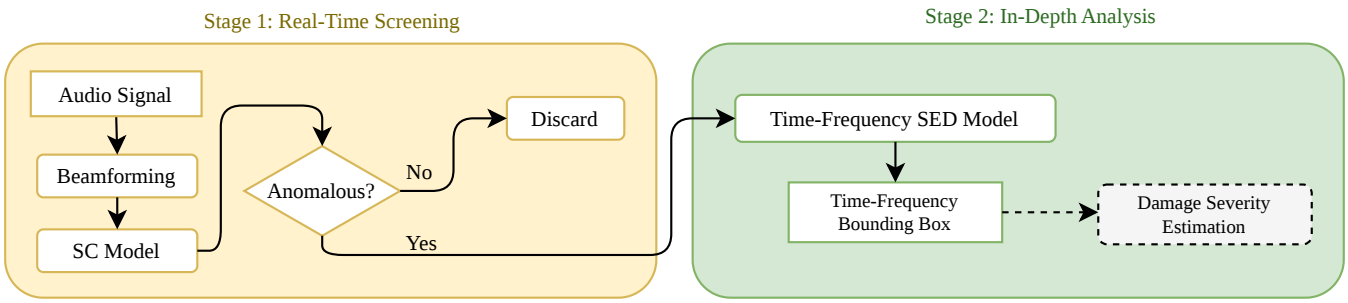



Figure 2. Overview of the proposed system. In the first stage, an SC model determines whether anomalous sounds are present for real-time screening on edge devices. In the second stage, a time-frequency SED model identifies the bounding boxes of the anomalous sounds in the spectrograms on a server for in-depth analyses. Estimating damage severity is beyond the scope of this study.

no anomalous sounds, referred to as normal, are discarded. The remaining time segments, referred to as anomalous, are forwarded to the second stage.

In the second stage, a time-frequency SED model identifies time-frequency bounding boxes in the spectrogram. The shapes and locations of anomalous sounds contain essential information for estimating damage severity. However, high-performance SED models are not suitable for edge devices because of their high computational cost; therefore, they are executed on a server.

When Internet connections are available, anomalous segments can be transmitted immediately from the edge devices to the server. Otherwise, they may be stored locally and collected on a regular or irregular basis. Storing or transmitting continuous audio data places a significant burden on storage capacities and network communications. To address this bottleneck, recent trends in structural health monitoring (SHM) emphasize the necessity of on-device data processing, or tiny machine learning (TinyML) (von Däniken et al., 2024). The proposed system employs a lightweight SC model as the first stage on the edge devices, transmitting only anomalous time segments to the server for in-depth analysis. Discarding the normal time segments on the edge devices significantly reduces the burden on data storage and transmission and optimizes the server loads. If the Internet bandwidth or storage capacity is limited, the Mel-spectrograms would be sufficient to be forwarded to the second stage. Since the computation of the short-time Fourier transform and Mel-filterbanks is highly lightweight, the edge device has enough computational power to compute the high-resolution Mel-spectrograms required for the second-stage model. As a result, the load on data storage and transfer is further reduced compared to sending raw acoustic signals.

Taking all of the above into account, the proposed system is an edge-server collaborative system that combines complementary models: the SC model in the first stage and the time-frequency SED models in the second stage. The first stage serves as real-time screening at the edge side, whereas the second stage performs an in-depth analysis on the server side, targeting only the time segments classified as anomalous. Although estimating damage severity from the shapes and locations of whistle-like sounds is an essential task, it is not investigated in this study. Notably, the proposed system is intended to be attached to an existing wind power generation unit and to operate without receiving any data from the wind power generation unit, such as the orientation of the rotor plane and blade rotation speed. Retrieving real-time operational data typically requires modifications to the power generation unit, leading to high deployment costs. Therefore, the independent operation of the proposed system has a practical advantage.

### 2.2. Sound Classification

The SC model receives a Mel-spectrogram with a fixed duration and outputs the score (i.e., the posterior probability that it contains anomalous sounds). In the first stage, we employed an attention-based convolutional neural network (ACNN) as an SC model for real-time screening. The ACNN model comprises $N$ stacked convolutional layers and one multi-head self-attention layer, as shown in Figure 3.

ACNNs have been widely used in audio classification tasks owing to their ability to combine local feature extraction with global context modeling (Z. Zhang, Xu, Zhang, & Cao, 2021; Wei, Zhu, Benetos, & Wang, 2021). We adopted this model structure because it is known to be performative while remaining lightweight.

### 2.3. Time-Frequency Sound Event Detection

The time-frequency SED model receives a Mel-spectrogram of fixed duration and outputs bounding boxes for the detected sound events and their scores. A bounding box is a rectangle specified by the time center, duration, frequency center, and bandwidth. In this study, we assumed a single class of sound events because the class structure of whistle-like sounds for estimating damage conditions has not been established.

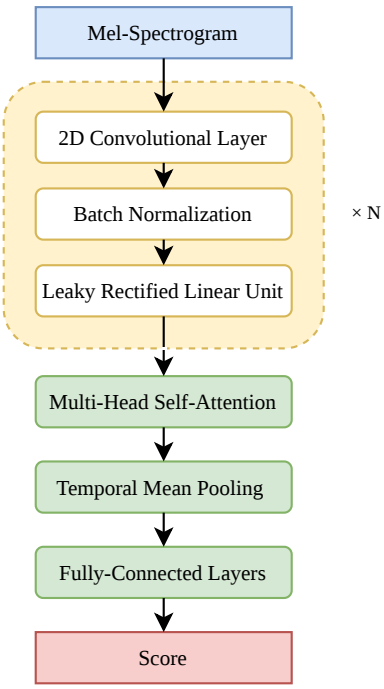



Figure 3. Model structure of the SC model. It comprises stacked convolutional layers, self-attention layers, temporal mean pooling, and fully-connected layers.

We adapted DETRs for object detection in natural images to the time-frequency SED in audio spectrograms. Audio spectrogram transformers (ASTs) (Gong, Chung, & Glass, 2021; Niizumi, Takeuchi, Ohishi, Harada, & Kashino, 2022; Huang et al., 2022; Gong, Lai, Chung, & Glass, 2022; Baade, Peng, & Harwath, 2022; Chong, Wang, Zhou, & Zeng, 2023; Chen, Liang, Ma, Zheng, & Chen, 2024) can be employed as a backbone to bridge the modality gap. Specifically, we adopted DETR with improved denoising anchor boxes (DINO) (H. Zhang et al., 2023) and an efficient audio transformer (EAT) (Chen et al., 2024). The EAT model was pretrained using a large-scale audio dataset, AudioSet (Gemmeke et al., 2017), to capture the intrinsic patterns of environmental sounds. Although most ASTs were trained at a 16 kHz sampling rate in prior studies, we used an EAT model pretrained at a 48 kHz sampling rate with 182 Mel-frequency bins because the whistle-like sounds of interest occurred in the 2–20 kHz range.

Following a previous study (Z. Zhu & Sato, 2025), a multiscale feature pyramid (Y. Li, Mao, Girshick, & He, 2022) originally proposed for vision transformers was used to enhance the adaptability of models that handle objects of various sizes. Henceforth, the entire model comprises an EAT backbone, multiscale feature pyramid, and DINO detection model, which has been demonstrated to be a state-of-the-art model (Z. Zhu & Sato, 2025). Figure 4 illustrates the model structure. In this paper, we refer to this model as DINO+EAT.

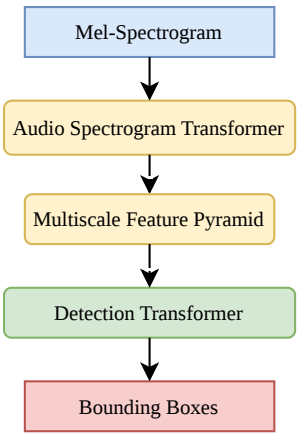



Figure 4. Model structure of the time-frequency SED model. It comprises an AST, multiscale feature pyramid, and DETR.

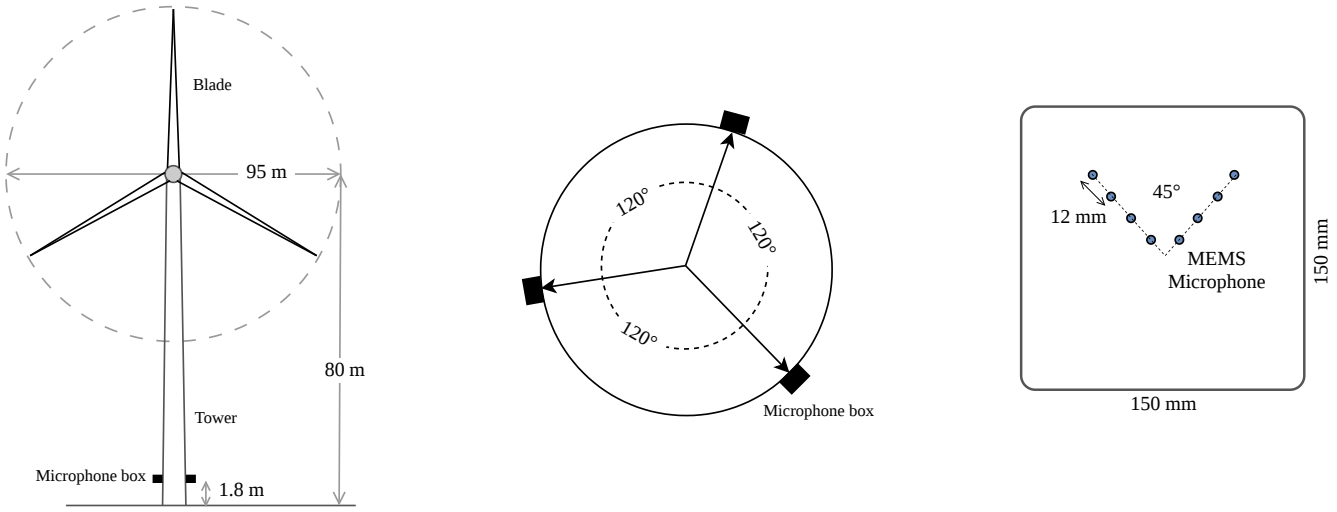
## 3. DATA

To evaluate the proposed system, we built two audio datasets of wind turbine sounds: in-domain and out-of-domain. The in-domain dataset is intended for continuous monitoring using permanently installed devices and is used for training and evaluation. The out-of-domain dataset is intended for temporary inspection using portable recorders and is used only for evaluation.

### 3.1. Device

The in-domain dataset was collected using custom-built recording devices. The recording device for a single tower comprised the main box and three microphone boxes. The main box contained a main processing board, data storage device, and power unit. Each microphone box is an Ingress Protection (IP) 65-rated waterproof enclosure measuring 150 mm × 150 mm × 100 mm, containing eight omnidirectional micro-electromechanical system (MEMS) microphones. The microphones were arranged in a V-shaped configuration with a 45° opening angle and 12-mm spacing, as illustrated in Figure 5. Acoustic signals can be transmitted through a waterproof filter integrated into the enclosure lid. Microphone boxes were installed on the surface of the tower at a height of 1.8 m from the base and spaced 120° apart (see Figure 5). Although the geometry of the microphone boxes was fixed, the orientation of the rotor plane swept by the blades was controlled based on wind direction. Hence, the microphone box closest to the blades captured whistle-like sounds with the highest intensity at any given moment. All boxes were daisy-chained via an automotive audio bus (A2B), with the main box as the master node. This custom-built device can record 24-channel audio signals at a sampling rate of 48 kHz with a 16-bit depth.

The out-of-domain dataset was recorded using a commercial handheld stereo recorder (TASCAM DR-05X) installed on a tripod near the tower. The sampling rate and bit depth were




(a) Front view (b) Top-down view (c) Microphone box

Figure 5. Custom-built device used for collecting the in-domain dataset. (a) Front view of a wind turbine with three microphone boxes installed on the tower surface. (b) Top-down view of the tower cross-section with three microphone boxes. (c) Overview of a microphone box.

the same as those used for the in-domain data.

### 3.2. Recording

The wind turbine audio dataset collected using the custom-built devices in a previous study (Z. Zhu & Sato, 2025) was employed as the in-domain dataset. It was collected from 46 land-based wind power generation units at two separate locations over 64 days, spanning approximately one year. Recordings were made intermittently, lasting 180–900 s at intervals of at least 15 min. In total, 6,026 audio clips totaling 1,147 h were collected.

The out-of-domain dataset was collected from 24 units of the same type as the in-domain dataset at two separate locations on six days. The in-domain and out-of-domain datasets shared no locations.

### 3.3. Preprocessing

For the in-domain dataset, minimum variance distortion-free response (MVDR) beamforming was applied to eight-channel audio data from each microphone box to enhance sounds arriving vertically from above. For the out-of-domain dataset, stereo audio data were downmixed to a single channel. Subsequently, all recordings were split into 15-s segments. The sampling rate and bit depth were kept to 48 kHz and 16 bits, respectively.

### 3.4. Annotation

For the in-domain dataset, 4,210 segments were selected by balancing four criteria: recording date, average spectral flatness over time, standard deviation of loudness over time, and anomalous sound presence score. The anomalous sound scores were estimated using an attention-based convolutional recurrent neural network trained on a small dataset of 240 randomly selected segments with preliminary annotations. Nine crowd workers annotated the selected segments by drawing time-frequency bounding boxes around whistle sounds in spectrograms. Each segment was annotated by a single worker under the supervision of an overseer. If a segment contained anomalous sound events, it was considered anomalous; otherwise, it was classified as normal. The resulting dataset comprised 2,367 normal and 1,843 anomalous segments, containing 14,420 sound events with annotated bounding boxes.

Table 1. Composition of the annotated data.

| | **In-Domain** | **Out-of-Domain** |
|---|---|---|
| Normal segments | 2,367 | 449 |
| Anomalous segments | 1,843 | 47 |
| Anomalous sounds | 14,420 | 135 |

All segments in the out-of-domain dataset were annotated by the authors. The resulting dataset contained 449 normal and 47 anomalous segments and 135 anomalous sound events.

Table 1 lists the composition of the annotated data. Figure 1 shows examples of spectrograms with anomalous sound events in the in-domain and out-of-domain datasets recorded by a custom-built device and a handheld recorder.

## 4. EXPERIMENT: SOUND CLASSIFICATION

### 4.1. Setup

The waveforms were converted into Mel-spectrograms with 128 frequency bins using a 25-ms window with a 10-ms

hop. The SC model was trained and evaluated using five-fold cross-validation. The in-domain data were partitioned into five subsets at the recording level. For each fold, one subset was used for testing. The remaining subsets were used for training (87.5%) and validation (12.5%). The validation set was used to optimize the learning rate and select the best epoch.

We examined two models of different sizes: ACNN-small and ACNN-large. The ACNN-small model comprises two convolutional layers with 256 channels and four attention heads, whereas the ACNN-large model contains five convolutional layers with 512 channels and eight attention heads.

To improve the robustness of the model to unseen conditions, data augmentation was applied by adding environmental noise to the training data. The signal-to-noise ratios were randomly sampled from 0 to 30 dB in 10-dB steps. The noise clips were randomly sampled from the following datasets at ratios of 40%, 30%, and 30%: normal segments in the in-domain dataset, a subset of the Freesound Dataset 50K dataset (Fonseca, Favory, Pons, Font, & Serra, 2022) that consisted only of natural sounds (e.g., wind, rain, ocean, and birds), and field and forest recordings in the Advanced Telecommunications Research Institute International (ATR) ambient noise sound database 2 (ATR-Promotions, 2005). Notably, mixup augmentation was not applied to the SC model. In binary classification, superimposing two anomalous segments generates easy samples with dense anomalous sounds. The ratio of the original data to the augmented data was 1:1.

The small and large models were trained using an in-domain dataset with and without data augmentation. The adaptive moment estimation (Adam) optimizer and binary cross-entropy loss were employed for training. The initial learning rates were set to $2 \times 10^{-6}$ and $3.12 \times 10^{-6}$ for the small and large models, respectively.

The trained models were evaluated using the in-domain and out-of-domain datasets. Sound classification performance was measured using the area under the receiver operating characteristic curve (AUC). During the evaluation of the in-domain dataset, the results were averaged across five folds. The best model across the five folds when evaluated on the in-domain dataset was chosen for evaluation on the out-of-domain dataset.

### 4.2. Results

Table 2 lists the sound classification results. The ACNN-small and ACNN-large models with noise augmentation achieved in-domain AUCs of 0.910 and 0.920, respectively. The performance degraded to 0.726 and 0.867 out-of-domain. The performance degradation of the large model was less substantial than that of the small model. Whereas data augmentation did not significantly enhance the robustness of the large model to unseen conditions, the small model became less robust with data augmentation. A possible reason for the degradation of the small model is its limited capacity. When exposed to diverse and complex augmented audio data during training, a small model may struggle to effectively represent both the target anomalous sounds and noise variations simultaneously, resulting in interference that hinders its generalization to unseen out-of-domain environments. These results indicate that the ACNN models are robust to unseen microphones, recording locations, and ambient noise without data augmentation, regardless of their size. The classification threshold can be adjusted during practical deployment. Depending on operational priorities, a stricter threshold can be set to minimize the transmission of unnecessary audio segments, or a more permissive threshold can be applied to minimize the risk of missing blade defects.

Table 2. Performance of sound classification measured using AUCs. ID and OOD represent in-domain and out-of-domain, respectively.

| Model | Augmentation | ID | OOD |
|---|---|---|---|
| ACNN-small | | 0.919 | 0.839 |
| ACNN-small | ✓ | 0.910 | 0.726 |
| ACNN-large | | 0.913 | 0.862 |
| ACNN-large | ✓ | **0.920** | **0.867** |

## 5. EXPERIMENT: TIME-FREQUENCY SOUND EVENT DETECTION

### 5.1. Setup

The waveforms were converted into Mel-spectrograms with 182 frequency bins using a 25-ms window with a 10-ms hop. The time-frequency SED model was trained and evaluated using one of the cross-validation subset assignments used in the SC experiments. The model structure was determined based on a previous study (Z. Zhu & Sato, 2025). The EAT-base model contained 12 transformer layers with 768-dimensional embeddings. The output features were reshaped at four scale factors of 4, 2, 1, and 0.5 to construct a multiscale feature pyramid (Y. Li et al., 2022), with each feature projected to 256 dimensions. The DINO detection model comprised six deformable transformer encoders, six transformer decoders, and an auxiliary detection head. The number of object queries was set to 100.

We explored two data augmentation methods to enhance the model. First, we applied data augmentation by adding environmental noise to the training data using the same setup as that used in the SC experiments. Second, we applied mixup data augmentation (Z. Zhang et al., 2019), where the spectrograms of random pairs were averaged with equal weights, considering the union of all bounding box coordinates from

Table 3. Performance of time-frequency SED measured using $AP_{50}$. ID and OOD represent in-domain and out-of-domain, respectively.

| Model | Augmentation | ID | OOD |
|---|---|---|---|
| DINO+EAT | | 0.485 | **0.456** |
| DINO+EAT | ✓ | **0.510** | 0.440 |

two source audio clips. Because superimposing two spectrograms physically simulates the simultaneous occurrence of both sounds, combining their bounding boxes is an acoustically valid approach. By forcing the model to detect the union of bounding boxes from superimposed spectrograms, mixup generates more complex, diverse, and difficult training samples. The training dataset was augmented to three times its original size, consisting of original, noise-augmented, and mixup-augmented data in a 1:1:1 ratio.

The training hyperparameters were set based on a previous study (Z. Zhu & Sato, 2025). The model was trained for 12 epochs using the Adam with decoupled weight decay (AdamW) optimizer with an initial learning rate of $1 \times 10^{-4}$ for the detection head and $1 \times 10^{-5}$ for the backbone. The weight decay was set to $1 \times 10^{-4}$. The training loss was the weighted sum of the losses for the classification and box regression. We used focal loss for classification and $L_1$ and generalized intersection over union (IoU) loss for box regression. The trained models were evaluated using the in-domain and out-of-domain datasets. The performance of the time-frequency SED was measured using average precision at an IoU threshold of 50% ($AP_{50}$), calculated according to the Common Objects in Context (COCO) evaluation protocol (Lin et al., 2014).

### 5.2. Results

Table 3 lists the time-frequency SED results. Without augmentation, the DINO+EAT model achieved an $AP_{50}$ of 0.485 and 0.456 in-domain and out-of-domain, respectively. The augmentation significantly improved the performance to 0.510 in-domain, but slightly degraded it to 0.440 out-of-domain. These results indicate that noise and mixup augmentation are effective for improving performance in-domain; however, they do not necessarily improve the robustness to unseen environmental noise, microphones, or recording locations.

As only a single class of acoustic events is considered in this study, the score of an individual bounding box can be interpreted as the probability of the sound being anomalous. However, this score does not directly translate into the physical severity of the blade damage. Although a specific defect repeatedly generates anomalous sounds as the blade rotates, the recorded sound clarity fluctuates significantly with various environmental and operational factors, such as wind speed and direction, rotor speed, the relative angle between the rotor plane and the microphone box, and the location of the damage. Therefore, the inference score mainly reflects the acoustic clarity at a given moment rather than the actual severity of the damage. Consequently, directly using the precision-recall curve for sound event detection as a detection error trade-off curve for maintenance would be a suboptimal approach with limited practical validity. In practical deployments, rather than making decisions per acoustic event, it is essential to aggregate severity information estimated from multiple events over time to determine the priority of close inspections on a wind turbine basis. Developing a robust framework for such severity estimation is left for future work.

Table 4. Computational cost measured on the Raspberry Pi 5.

| Component | Memory (MiB) | RTF |
|---|---|---|
| MVDR | 37 | 0.048 |
| ACNN-small | 785 | 0.013 |
| ACNN-large | 1,142 | 0.051 |
| DINO+EAT | 2,246 | 2.334 |

## 6. EXPERIMENT: COMPUTATIONAL COST

To verify the feasibility of field-deploying the proposed system for real-time screening of edge devices, we measured the memory footprint and real-time factor (RTF) of the preprocessing pipeline and sound classification models using a Raspberry Pi 5. We also examined the time-frequency SED model on this device. As shown in Table 4, the ACNN models achieved RTFs well below 1.0, regardless of their size, confirming that edge-side real-time screening (i.e., preprocessing and sound classification) can be performed in real time.

In contrast, deploying the DINO+EAT model on an edge device is impractical not only in terms of processing speed but also memory footprint and power consumption. As shown in Table 4, it requires a high memory footprint of 2,246 MiB and exhibits an RTF of 2.334. Such a high computational cost would pressure other concurrent tasks, such as continuous data recording, preprocessing, and screening by the SC model. Additionally, running transformer-based models results in substantial power consumption on edge devices. More importantly, discarding normal segments at the edge is essential to avoid the continuous transmission of acoustic data, which places a significant burden on network bandwidth and server loads. Therefore, the edge-server collaborative architecture is reasonable for practical field deployment.

## 7. CONCLUSION

This study proposes a two-stage edge-server collaborative system for detecting damage to wind turbine blades based on acoustic signals captured by arrays built from commodity

MEMS microphones. To address the need for practical health monitoring, the proposed system combines real-time continuous SC at the edge side for screening and time-frequency SED at the server side for in-depth analysis. Initially, a lightweight ACNN model continuously identifies the presence of anomalous whistle-like sounds on the edge devices. The deployment of the lightweight ACNN model on the edge devices aligns with the emerging paradigm of on-device SHM using TinyML, which has proven effective in minimizing detection latency and data transfer congestion (von Däniken et al., 2024). Subsequently, suspicious time segments are forwarded to a server, where a time-frequency SED model comprising DETR with an AST backbone identifies the time and frequency ranges of anomalous sounds using bounding boxes in spectrograms. The effectiveness of the proposed system was investigated through extensive experiments using in-domain data collected over a year using custom-built microphone devices and out-of-domain data recorded using commercial handheld recorders. The experimental results demonstrate that the SC achieved an AUC of 0.920, whereas the time-frequency SED model attained an $AP_{50}$ of 0.510 on the in-domain dataset. Moreover, the first-stage real-time processing capability was confirmed using an edge device (Raspberry Pi 5). Furthermore, evaluations under unseen conditions yielded an AUC of 0.867 and $AP_{50}$ of 0.440, indicating that the proposed system exhibited a certain degree of robustness to unseen microphone types, recording locations, environmental noise, and the absence of spatial filtering (MVDR beamforming). These results highlight the effectiveness of the proposed system for practical automatic health monitoring of wind power generation equipment.

Although this study successfully achieved automatic detection of anomalous sounds, estimating the actual severity of blade damage from the shapes and locations of the time-frequency bounding boxes remains an essential task for future work. Previous studies have demonstrated that variations in damage severity, such as the progression of erosion or differences in crack size, manifest as distinct shifts in acoustic frequency bands and spectral characteristics (Solimine, Niezrecki, & Inalpolat, 2020; Y. Zhang, Avallone, & Watson, 2023). Based on these findings, analyzing the properties of the detected whistle-like sounds could provide a quantitative basis for assessing structural degradation. Moreover, while the proposed system operates independently of the wind power generation unit to be monitored to facilitate field deployment, integrating operational data from the wind turbine, such as the orientation of the rotor plane and blade rotation speed, could further enhance the effectiveness of the system at the expense of ease of deployment.

## ACKNOWLEDGMENT

The authors thank Yokogawa Electric Corporation for their cooperation in collecting the wind turbine audio data.

## REFERENCES

Algolfat, A., Wang, W., & Albarbar, A. (2023). Damage identification of wind turbine blades – a brief review. *Journal of Dynamics, Monitoring and Diagnostics*, *2*, 198–206. doi: 10.37965/jdmd.2023.422

ATR-Promotions. (2005). *ATR ambient noise sound database 2*. https://www.atr-p.com/products/esd.html.

Baade, A., Peng, P., & Harwath, D. (2022). MAE-AST: Masked autoencoding audio spectrogram transformer. In *Interspeech* (pp. 2438–2442).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision.* doi: 10.1007/978-3-030-58452-8_13

Chen, W., Liang, Y., Ma, Z., Zheng, Z., & Chen, X. (2024). EAT: Self-supervised pre-training with efficient audio transformer. In *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 3807–3815). doi: 10.24963/ijcai.2024/421

Chong, D., Wang, H., Zhou, P., & Zeng, Q. (2023). Masked spectrogram prediction for self-supervised audio pre-training. In *ICASSP*.

Cuesta, J., Leturiondo, U., Vidal, Y., & Pozo, F. (2024). A review of prognostics and health management in wind turbine components. In *Proceedings of the European Conference of the Prognostics and Health Management Society.* doi: 10.36001/phme.2024.v8i1.4093

Ding, S., Yang, C., & Zhang, S. (2023). Acoustic-signal-based damage detection of wind turbine blades – a review. *Sensors*, *23*(11), 4987. doi: 10.3390/s23114987

Fonseca, E., Favory, X., Pons, J., Font, F., & Serra, X. (2022). FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 829–852. doi: 10.1109/TASLP.2021.3133208

Fremmelev, M. A., Ladpli, P., Orlowitz, E., Bernhammer, L. O., McGugan, M., & Branner, K. (2022). Structural health monitoring of 52-meter wind turbine blade: Detection of damage propagation during fatigue testing. *Data-Centric Engineering*, *3*, e22. doi: 10.1017/dce.2022.20

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., & Moore, R. C. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.* doi: 10.1109/ICASSP.2017.7952261

Gong, Y., Chung, Y.-A., & Glass, J. (2021). AST: Audio spectrogram transformer. In *Proceedings of Interspeech* (pp. 571–575). doi: 10.21437/Interspeech.2021-698

Gong, Y., Lai, C.-I., Chung, Y.-A., & Glass, J. (2022).

SSAST: Self-supervised audio spectrogram transformer. In *AAAI* (Vol. 36, pp. 10699–10709).

Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., … Feichtenhofer, C. (2022). Masked autoencoders that listen. In *NeurIPS*.

Kuo, S.-F., Cheng, S., Lo, F.-C., & Tu, T.-H. (2023). Wind turbine blade damage detection and classification based on sound feature signal using machine learning. In *Proceedings of the Asia Pacific Conference of Sound and Vibration.* doi: 10.3397/IN_2023_0637

Lamdjad, B., & Chaiter, A. (2026, March). AI-powered predictive maintenance and prognostic health management using edge-based predictive algorithms for industrial operations. *Preprints*. doi: 10.20944/preprints202603.0010.v1

Li, F., Li, L., & Peng, Y. (2021). Research on digital twin and collaborative cloud and edge computing applied in operations and maintenance in wind turbines of wind power farm. *Advances in Transdisciplinary Engineering*, *17*, 80–92. doi: 10.3233/ATDE210263

Li, F., Zhang, H., Liu, S., Guo, J., Ni, L. M., & Zhang, L. (2022). DN-DETR: Accelerate DETR training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13619–13627). doi: 10.1109/TPAMI.2023.3335410

Li, M., Xu, Z., Li, S., Kikuchi, Y., Dong, Y., Gryllias, K. C., … Carroll, J. (2026). Health prognostics and maintenance decision-making for wind energy: A comprehensive overview. *Renewable and Sustainable Energy Reviews*, *226*, 116269. doi: 10.1016/j.rser.2025.116269

Li, Y., Mao, H., Girshick, R., & He, K. (2022). Exploring plain vision transformer backbones for object detection. In *Proceedings of the European Conference on Computer Vision.* doi: 10.1007/978-3-031-20077-9_17

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision.* doi: 10.1007/978-3-319-10602-1_48

Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., … Zhang, L. (2022). DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *Proceedings of the International Conference on Learning Representations.*

Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N., & Kashino, K. (2022). Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation. In *PMLR* (Vol. 166, pp. 1–24).

Solimine, J., Niezrecki, C., & Inalpolat, M. (2020). An experimental investigation into passive acoustic damage detection for structural health monitoring of wind turbine blades. *Structural Health Monitoring*, *19*(6), 1711–1725. doi: 10.1177/1475921719895

Sun, B., Ooi, K. T., & Su, M. (2026). Wind turbine blade damage: A systematic review of detection, diagnosis, performance impact, and lifecycle health management. *Renewable and Sustainable Energy Reviews*, *230*, 116668. doi: 10.1016/j.rser.2025.116668

Van Dam, J., & Bond, L. J. (2015). Acoustic emission monitoring of wind turbine blades. In *Proc. SPIE 9439, Smart Materials and Nondestructive Evaluation for Energy Systems* (p. 94390C). doi: 10.1117/12.2084527

von Däniken, E., Mikhaylov, D., Moallemi, A., Polonelli, T., & Magno, M. (2024). Tiny on-device structural health monitoring for wind turbines using mems pressure sensors. In *2024 IEEE Sensors Applications Symposium (SAS)* (p. 1-6). doi: 10.1109/SAS60918.2024.10636460

Wang, C., Guo, R., Yu, H., Hu, Y., Liu, C., & Deng, C. (2023). Task offloading in cloud-edge collaboration-based cyber physical machine tool. *Robotics and Computer-Integrated Manufacturing*, *79*, 102439. doi: 10.1016/j.rcim.2022.102439

Wang, W., Xue, Y., He, C., & Zhao, Y. (2022). Review of the typical damage and damage-detection methods of large wind turbine blades. *Energies*, *15*(15), 5672. doi: 10.3390/en15155672

Wei, W., Zhu, H., Benetos, E., & Wang, Y. (2021). Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports*, *11*, 21557. doi: 10.1038/s41598-021-01045-4

Yang, C., Ding, S., & Zhou, G. (2025). Wind turbine blade damage detection based on acoustic signals. *Scientific Reports*, *15*(1), 3930. doi: 10.1038/s41598-025-88276-x

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., … Shum, H.-Y. (2023). DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations.*

Zhang, Y., Avallone, F., & Watson, S. (2023). Leading edge erosion detection for a wind turbine blade using far-field aerodynamic noise. *Applied Acoustics*, *207*, 109365. doi: 10.1016/j.apacoust.2023.109365

Zhang, Z., He, T., Zhang, H., Zhang, Z., Xie, J., & Li, M. (2019). *Bag of freebies for training object detection neural networks.*

Zhang, Z., Xu, S., Zhang, S., & Cao, S. (2021). Attention based convolutional recurrent neural network for environmental sound classification. *Neurocomputing*, *453*. doi: 10.1016/j.neucom.2020.08.069

Zhou, X., Kang, Z., Canady, R., Bao, S., Balasubramanian, D. A., & Gokhale, A. (2021). Exploring cloud assisted tiny machine learning application patterns for phm scenarios. In *Annual Conference of the PHM Society*

(Vol. 13). doi: 10.36001/phmconf.2021.v13i1.3054

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2021). Deformable DETR: Deformable transformers for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations.*

Zhu, Y., & Liu, X. (2023). A lightweight CNN for wind turbine blade defect detection based on spectrograms. *Machines*, *11*(1), 99. doi: 10.3390/machines11010099

Zhu, Y., Liu, X., Li, S., Wan, Y., & Cai, Q. (2022). Wind turbine blade defect detection based on acoustic features and small sample size. *Machines*, *10*(12), 1184. doi: 10.3390/machines10121184

Zhu, Z., & Sato, Y. (2025). Sound event detection using time-frequency bounding boxes with a self-supervised audio spectrogram transformer. In *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)* (pp. 150–154). doi: 10.5281/zenodo.17251589