Journal of Applied Computational Intelligence

eISSN: 1234-5678

Vol. 4, No. 1, January, 2026

# Improved PointNeXt for Acoustic Recognition of Stored-Grain Insects under Granary Monitoring Conditions

Rasmus Sepp[1,*], Jaan Mägi[1]

[1] *Department of Informatics, Tartu Vocational College, Tartu, 50409, Estonia*

Received 13 October 2025; Revised 24 December 2025; Accepted 29 December 2025; Published 09 January 2026

**Abstract**

Acoustic monitoring provides a non-destructive route for detecting stored-grain insects, yet field granary recordings are usually affected by grain-bed attenuation, sensor coupling variation, fan noise, grain-settling interference, and short-duration biological impulses. This study proposes an improved PointNeXt model for insect acoustic recognition under practical granary monitoring data conditions. Candidate acoustic events are transformed into sparse time-frequency point clouds through ridge-preserving denoising, adaptive thresholding, local saliency estimation, and multi-scale point sampling. A residual acoustic PointNeXt backbone is modified with density-aware neighborhood aggregation, channel recalibration, and a lightweight temporal voting head to strengthen weak-event representation and suppress isolated mechanical disturbances. Experiments are organized on a simulated granary acoustic dataset containing five target classes and background events, with grain depths from 0.4 m to 2.0 m, 1.0 s windows with 50% overlap, and signal-to-noise ratios from -5 dB to 20 dB. The proposed model reaches 96.8% accuracy, 95.9% macro-F1, and a 2.8% false-alarm rate, improving macro-F1 by 4.7 percentage points over the original PointNeXt and by 7.6 percentage points over a ResNet spectrogram baseline. Under low-SNR recordings, the method maintains 92.6% macro-F1 and reduces missed detections of weak larval feeding bursts by 38.4%. The results indicate that point-cloud modeling of acoustic events can preserve sparse discriminative structures while limiting redundant spectral background, although broader field validation remains necessary before commercial deployment.

*Keywords:* Granary Acoustic Monitoring; Stored-Grain Insect Recognition; Acoustic Point Cloud; Improved PointNeXt; Acoustic Point-Cloud Sampling

## 1. Introduction

Deep learning can improve stored-grain insect detection and identification, which supports early recognition of biological deterioration before population outbreaks become visible [1]. Insects such as grain weevils, flour beetles, lesser grain borers, and flat grain beetles can remain inside kernels or interstitial grain spaces for extended periods, which makes manual sampling both delayed and spatially incomplete [2]. Chemical fumigation and routine inspection still play important roles in many storage systems, but their preventive value weakens when infestation is detected only after temperature rise, odor, or grain-quality decline has already appeared [3]. Acoustic monitoring offers a different sensing route because larval feeding, adult crawling, wing vibration, and body-wall friction generate short impulses that can be captured without disturbing the grain bulk [4]. Compared with trap counting or visual examination, an acoustic sensor can operate continuously inside a warehouse bin, steel silo, or experimental chamber, and it can report weak activity during periods when insects are hidden from optical instruments. The engineering challenge is that stored-grain acoustic signals rarely resemble clean laboratory recordings. Sound transmission depends on grain type, moisture, depth,

* Corresponding author: rasmus.se@tvt.ee

https://doi.org/10.68406/jaci.2026.vol4iss1no2:14-27

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

kernel packing, and contact between the probe and the surrounding material [5]. Background disturbances from ventilation, conveyors, grain settling, rain impact, and structural vibration may be stronger than the biological events that must be identified [6].

Existing acoustic recognition pipelines commonly begin with handcrafted features extracted from short-time spectra, wavelet packets, cepstral coefficients, or time-domain pulse descriptors. These features are useful when noise conditions are stable, and they provide interpretable cues related to event rate, dominant band, burst duration, and spectral sharpness [7]. However, handcrafted representations become fragile when the sensor moves, the grain column attenuates high-frequency components, or several insects overlap in the same recording window. Spectrogram-based convolutional neural networks have improved recognition by learning local texture patterns, yet they also preserve large regions of redundant background energy that do not correspond to insect activity [8]. This dense representation increases computation and may cause the network to learn nuisance texture from fan harmonics or grain-settling resonance. Recurrent models and temporal attention networks can model event order, but they often require carefully segmented sequences and may overfit to laboratory-specific rhythm patterns [9]. Recent geometric learning methods suggest another possibility: instead of treating the complete spectrogram as a dense image, the recognizer can focus on the sparse ridges, peaks, and impulse clusters that survive denoising [10]. Point-based networks are attractive for this situation because local neighborhoods can be built around meaningful acoustic points rather than around every time-frequency cell [11].

PointNeXt is a strong point-cloud architecture originally designed for geometric perception, with residual local aggregation and hierarchical feature abstraction. Its design is relevant to acoustic monitoring because denoised insect events can be represented as point sets whose coordinates encode time, frequency, energy, and local saliency. Nevertheless, direct migration from three-dimensional object recognition to granary acoustics is not sufficient. Acoustic point density changes with SNR, grain depth, and event strength; insect events have temporal continuity but irregular local topology; and false alarms often arise from mechanically induced broadband impulses that look locally salient. A recognition model for this setting should therefore combine point-wise sparsity, density awareness, temporal voting, and robust training across realistic monitoring conditions. The present work addresses this requirement by constructing a multi-resolution acoustic point-cloud representation and improving PointNeXt with density-aware neighborhood grouping, channel recalibration, and temporal consistency constraints.

Following this motivation, the manuscript is arranged to move from storage-monitoring requirements to model construction and then to engineering verification under realistic acoustic disturbance. Section 1 introduces the engineering background, the limitations of current acoustic recognition strategies, and the motivation for adapting PointNeXt to stored-grain insect monitoring. Section 2 reviews acoustic insect detection, deep sound recognition, and point-cloud learning methods related to the proposed approach. Section 3 describes the data preprocessing route, the improved PointNeXt architecture, and the mathematical formulation of acoustic point aggregation. Section 4 presents the experimental data, recognition results, ablation analysis, robustness evaluation, and computational assessment. Section 5 concludes the study and indicates future research directions.

## 2. Related Work

### 2.1. Acoustic Monitoring and Insect Sound Recognition

Stored-grain insect acoustic detection has developed from event counting toward species-aware recognition. Early monitoring systems mainly used piezoelectric probes, accelerometers, or contact microphones to capture feeding and movement impulses inside wheat, maize, rice, and flour environments [12]. Signal processing studies showed that insect acoustic events are usually sparse, intermittent, and strongly shaped by grain-bed propagation, so simple amplitude thresholds can detect activity but cannot reliably separate species or life stages [13]. Time-frequency analysis improved this situation by describing transient energy concentration, spectral centroid, entropy, and pulse repetition patterns [14]. In engineering practice, however, these descriptors are influenced by sensor pressure, kernel size, grain moisture, and external vibration; a threshold or classifier calibrated in one bin may drift when the sensor is installed in another bin [15].

Machine learning methods introduced supervised classification into acoustic pest monitoring by combining spectral descriptors with support vector machines, random forests, or shallow neural networks. Such models are still useful when training data are limited and when operators need transparent decision rules. Their limitation is that the feature extractor and classifier are optimized separately, which makes the recognition boundary sensitive to incomplete event segmentation. Deep spectrogram networks learn features directly from data and usually outperform handcrafted pipelines in controlled environments. Yet dense spectrogram images contain many background cells that only describe silence, fan harmonics, or grain-settling resonance. For weak larval activity, the discriminative information may occupy a very small fraction of the image, and the network can devote capacity to nuisance texture rather than biological signal.

Another unresolved issue is the relation between acoustic detection and operational pest-management decisions. A monitoring node does not only classify a short waveform; it must support repeated inspection decisions under changing storage conditions. If the recognizer is too sensitive, operators receive frequent alarms after grain loading or ventilation events. If it is too conservative, early larval activity may remain hidden until the population has expanded. Previous engineering studies therefore emphasize the joint evaluation of sensitivity, missed detection, false alarms, and computational cost rather than accuracy alone. This requirement motivates a model that can preserve weak insect impulses while rejecting short mechanical artifacts.

### 2.2. Point-Cloud Learning for Sparse Acoustic Events

Point-cloud learning provides a compact way to process sparse structures. PointNet introduced permutation-invariant point aggregation, while later hierarchical models improved local geometric modeling through neighborhood grouping and residual abstraction [16]. PointNeXt further strengthened local aggregation by refining residual blocks, training strategies, and scale-aware feature propagation [17]. Although these networks were designed for spatial objects, their core idea can be transferred to acoustic recognition after time-frequency events are converted into point sets. Each acoustic point can describe a local maximum, ridge segment, impulse boundary, or saliency-weighted spectral coordinate. The model then learns relations among meaningful points instead of scanning an entire dense image.

Several sound-recognition studies have explored sparse spectrogram sampling, event graphs, and attention masks to reduce the influence of background noise [18]. These studies imply that acoustic recognition benefits when the model is guided toward informative time-frequency regions. However, granary acoustic monitoring adds two constraints that are not fully solved by general sound classification: point density varies with propagation loss, and false acoustic events can be generated by mechanical vibration with insect-like impulse shapes [19]. A practical model must therefore adapt neighborhood size, encode local point density, and verify decisions through temporal consistency. The improved PointNeXt framework proposed in this paper is designed around these constraints.

The transfer of point-cloud learning to acoustic data also changes the meaning of locality. In a geometric object, nearby points often indicate physical adjacency on a surface. In an acoustic event cloud, nearby points may indicate a shared impulse, a harmonic trace, or a temporally adjacent disturbance. This distinction makes direct reuse of geometric radius search incomplete, because a dense broadband knock and a sparse feeding sequence can produce very different point distributions while occupying a similar window. The method developed here treats locality as an acoustic relation defined by normalized time, frequency, energy, and saliency, and this definition directly motivates the density-aware grouping module introduced in Section 3.2. This design allows the recognizer to learn local structures that are closer to the sensing process used in granary monitoring.

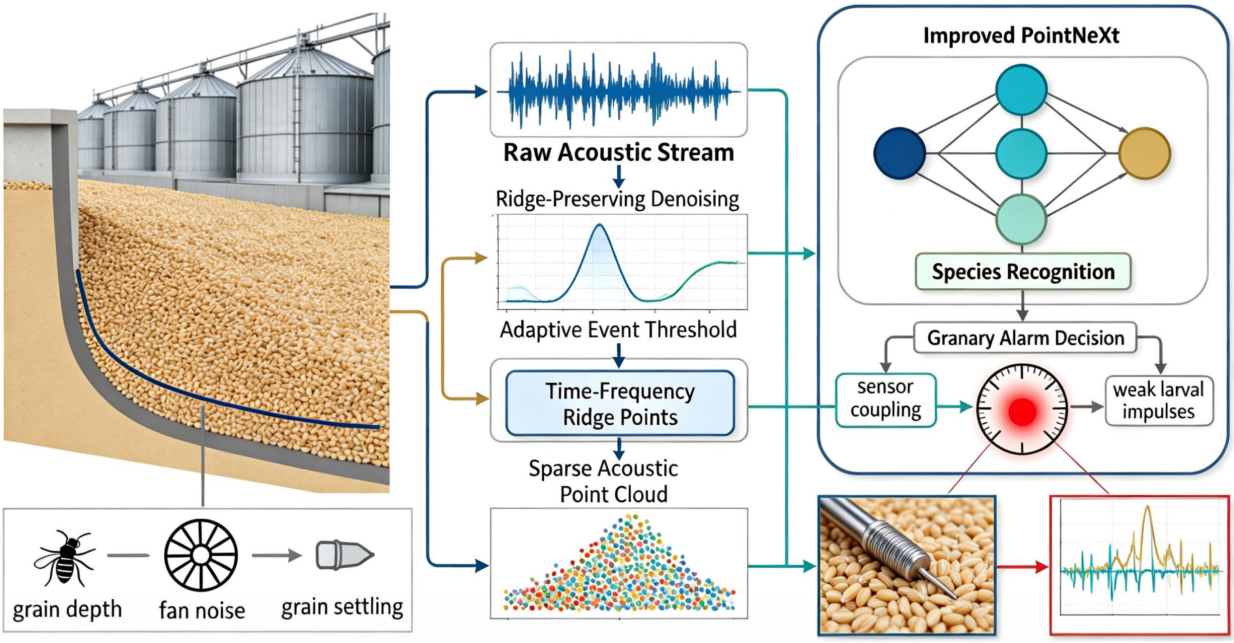



**Figure 1. Workflow of granary acoustic-event acquisition, denoising, point-cloud construction, and species-level recognition**

# 3. Proposed Method

## 3.1. Acoustic Point-Cloud Construction

The monitoring pipeline begins with continuous acoustic streams collected from contact sensors embedded in the grain bulk. Each stream is divided into 1.0 s windows with 50% overlap, and every window is transformed into a log-amplitude time-frequency plane. Instead of sending the full spectrogram to the classifier, a ridge-preserving filter removes low-energy background cells while retaining short impulses and narrow-band traces. Candidate points are selected by local saliency, non-maximum suppression within adjacent time-frequency cells, and an adaptive threshold estimated from the median background level of each window. Each retained point contains normalized time, normalized frequency, log energy, local contrast, and ridge orientation. Figure 1 illustrates the complete monitoring workflow from acoustic acquisition and denoising to point-cloud construction and recognition output. This representation keeps the physical ordering of acoustic events while discarding broad silent regions.

Let the acoustic point set in a window be denoted by $P = \{p_i\}_{i=1}^{N}$, where $p_i = (t_i, f_i, e_i, c_i, r_i)$. The saliency score used for point selection is represented as a weighted combination of energy contrast and local spectral sharpness.

$$s_i = \alpha e_i + \beta c_i + \gamma r_i \tag{1}$$

Here $s_i$ denotes the selection saliency, and the coefficients control the contribution of point energy, local contrast, and ridge consistency.

The point budget is fixed at 1,024 points after selection so that the downstream classifier has predictable computational cost. When the number of retained points is higher than the budget, farthest-point sampling is applied in the normalized time-frequency plane, but the sampling probability is biased by saliency so that weak biological ridges are not removed by dense background activity. When the point count is lower than the budget, the set is padded with masked points that do not contribute to neighborhood aggregation. This implementation avoids resizing the neural backbone for different acoustic windows and allows the same model to run on both desktop and embedded hardware.

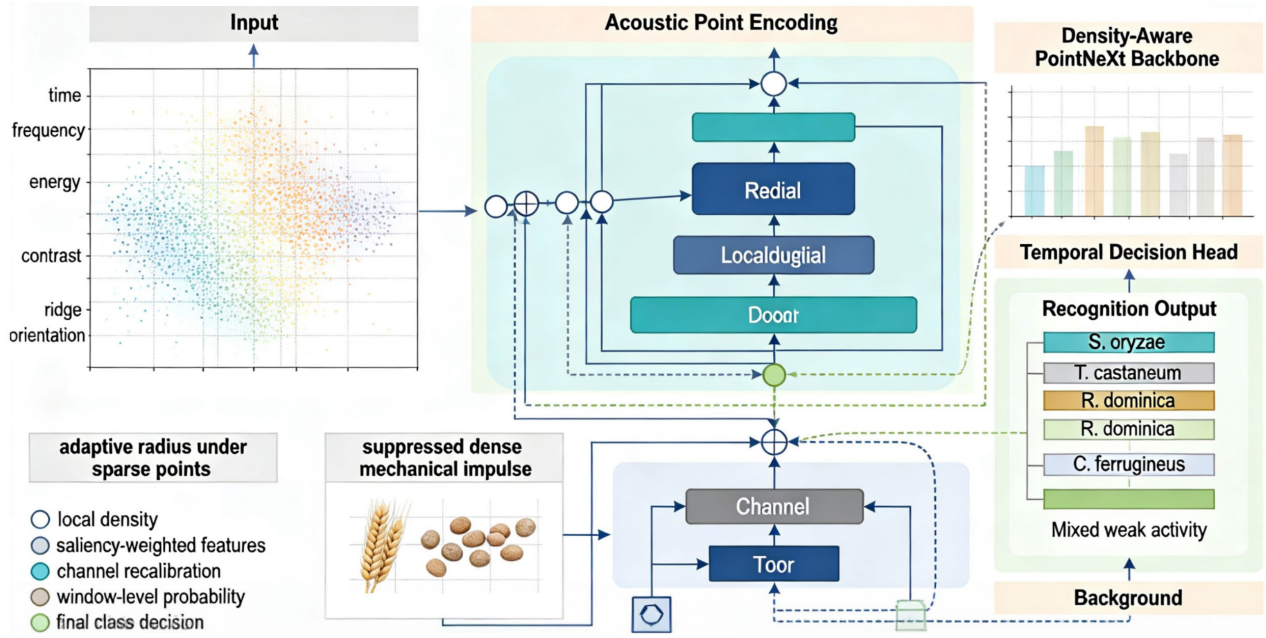



**Figure 2. Improved PointNeXt architecture with acoustic density-aware grouping, residual local aggregation, channel recalibration, and temporal voting**

## 3.2. Improved PointNeXt Architecture

The original PointNeXt hierarchy is adapted by replacing purely geometric neighborhood grouping with acoustic density-aware grouping. For each centroid point, neighbors are searched in a joint time-frequency space, and the neighborhood radius is adjusted according to the local point density estimated from the denoised spectrogram. This design prevents dense mechanical impulses from dominating aggregation and prevents sparse weak insect events from being discarded. A residual local block maps each neighborhood into a latent descriptor, while channel recalibration emphasizes frequency bands and temporal patterns that remain stable across grain depths.

For a centroid , the density-adjusted neighborhood radius is estimated before local aggregation and is constrained by lower and upper radius bounds to avoid unstable neighborhoods under extremely sparse or extremely dense acoustic events. The acoustic feature update then combines relative coordinates, saliency, and learned channel weights. Figure 2

presents the improved PointNeXt architecture, including density-aware grouping, residual local aggregation, channel recalibration, hierarchical feature propagation, and temporal voting. A temporal voting head collects predictions from several adjacent windows so that isolated broadband disturbances have less influence on the final decision.

$$\rho_i = \frac{k}{\pi d_{i,k}^2 + \varepsilon} \tag{2}$$

The variable $\rho_i$ is the estimated local point density around centroid $p_i$, and $d_{i,k}$ is the distance to the $k$-th neighbor in the normalized acoustic point space.

$$h_i' = \max_{j \in \mathcal{N}(i)} \phi\left([h_j, p_j - p_i, s_j]\right) \tag{3}$$

The updated feature $h_i'$ is obtained from a neighborhood $\mathcal{N}(i)$, where $\phi$ is a shared residual mapping applied to acoustic point attributes and relative coordinates.

$$z_i = \sigma\left(W_2 \delta(W_1 g_i)\right) \odot h_i' \tag{4}$$

The recalibrated descriptor $z_i$ uses a compact channel gate derived from neighborhood summary $g_i$, which helps separate stable insect bands from noise-dominated channels.

The backbone is organized into four abstraction stages. The first stage learns short local impulses, the second stage connects repeated impulses within a narrow band, the third stage models cross-band structures caused by body movement and feeding friction, and the fourth stage produces a compact global descriptor for the recognition head. Batch normalization is retained inside residual blocks, while dropout is applied only after the global descriptor because excessive point-level dropout may remove rare insect impulses. The channel recalibration branch is intentionally lightweight, using a bottleneck mapping to avoid turning the model into a large attention network.

### 3.3. Training Objective and Decision Rule

The model is trained using class-balanced cross-entropy and a temporal consistency regularizer. Class weights are derived from inverse class frequency, while the temporal consistency coefficient is selected on the validation set to avoid over-smoothing. Class balancing is necessary because background and low-activity windows are more frequent than species-specific events in practical monitoring data. Temporal consistency is applied only to neighboring windows with overlapping candidate events, thereby avoiding excessive smoothing across unrelated recordings. The final class probability is obtained by weighted voting across a short monitoring interval, where each window weight depends on event saliency and classifier confidence.

$$L = -\sum_{c} w_c y_c log(\hat{y}_c) + \lambda \sum_{\tau} \|q_\tau - q_{\tau-1}\|_2^2 \tag{5}$$

The first term learns class-discriminative recognition, and the second term discourages unstable predictions across overlapping event windows. During inference, the system first rejects windows whose point count and saliency fall below the acoustic activity threshold. Remaining windows are processed by the improved PointNeXt backbone, and the temporal voting head assigns a species label or background label. This rule reduces false alarms from isolated handling noise while maintaining sensitivity to repeated weak insect activity.

Training batches are balanced by species and by acoustic condition. Each mini-batch contains clean windows, low-SNR windows, deep-grain windows, and background windows so that the optimizer does not converge to an easy separation between quiet laboratory events and strong mechanical noise. Random gain perturbation, frequency masking, and time-shift augmentation are used before point selection. Augmentation is applied to the acoustic plane rather than to the selected point set because the point cloud should reflect the same preprocessing decisions that would be used during deployment.

The decision rule also stores auxiliary monitoring indicators, including the number of selected points, the mean saliency of accepted points, and the entropy of the voting distribution. These indicators are not used as extra class labels; they are recorded to support engineering diagnosis when a deployment site behaves differently from the training environment. A low-entropy decision with stable point density indicates a confident acoustic pattern. A high-entropy decision with unusually dense points indicates that a mechanical disturbance may be passing through the classifier. This additional output makes the recognition system easier to integrate with inspection scheduling and alarm management.

# 4. Experimental Data and Analysis

## 4.1. Dataset, Protocol, and Baseline Configuration

The experimental dataset was organized to reproduce the acoustic conditions encountered in granary monitoring rather than to create a clean laboratory sound archive. Recordings included Sitophilus oryzae, Tribolium castaneum, Rhyzopertha dominica, Cryptolestes ferrugineus, mixed weak-activity events, and non-biological background. The simulated acquisition protocol used contact acoustic probes, 1.0 s analysis windows, 50% overlap, and controlled background recordings from ventilation, auger vibration, grain settling, and room vibration. The grain medium contained wheat and maize lots with moisture levels between 11.5% and 15.0%, and the sensor depth ranged from 0.4 m to 2.0 m. Stratified splitting is commonly used to limit acquisition-session leakage in monitored acoustic datasets [20]. Following this principle, the dataset assigned 70% of windows to training, 10% to validation, and 20% to testing while keeping recordings from the same acquisition session within one split.

The final test set contained 18,000 acoustic windows. Of these, 12,400 windows contained identifiable insect activity and 5,600 windows were background or mechanical interference. Table 1 reports the main class distribution and recognition indicators under the selected protocol. The mixed weak-activity class contained fewer windows because combined weak events were harder to isolate and annotate consistently than single-species recordings. The proposed model obtained class F1 values between 94.2% and 97.1%, while background rejection reached 97.2% precision. The weaker mixed-activity class had the lowest recall at 93.8%, mostly because feeding bursts from small larvae were partially masked by grain-settling impulses. Comparable acoustic monitoring studies have reported similar sensitivity losses when the biological impulse rate decreases under thick grain coverage [21].

**Table 1. Class distribution and recognition indicators of the proposed model on the granary acoustic test set**

| Class | Windows | Precision (%) | Recall (%) | F1-score (%) |
|---|---|---|---|---|
| Sitophilus oryzae | 3,200 | 96.5 | 95.8 | 96.1 |
| Tribolium castaneum | 2,850 | 95.7 | 94.9 | 95.3 |
| Rhyzopertha dominica | 2,620 | 97.4 | 96.8 | 97.1 |
| Cryptolestes ferrugineus | 2,430 | 94.8 | 94.2 | 94.5 |
| Mixed weak activity | 1,300 | 94.6 | 93.8 | 94.2 |
| Background/noise | 5,600 | 97.2 | 97.7 | 97.4 |

The preprocessing pattern in Figure 3(a) shows that the denoising and point-selection stage retained 81.6% of labeled insect impulses at 0 dB SNR while reducing non-event spectral cells by 72.4%. The gap between impulse retention and background reduction became wider from -5 dB to 20 dB, which indicates that the selection stage did not simply preserve high-energy cells but increasingly separated structured insect impulses from diffuse spectral interference. The selected-point map in Figure 3(b) shows that points concentrated near 3.8-8.6 kHz for larval feeding and near 1.2-3.4 kHz for adult crawling. This distribution explains why a sparse point representation is suitable: the informative regions occupy narrow acoustic bands rather than the full spectrogram. The duration distribution in Figure 3(c) shows that the median event duration increased from 28 ms in R. dominica to 46 ms in mixed weak activity. The longer mixed-activity duration supports the later confusion analysis, because several weak behaviors can merge into extended event trains when they occur in the same window.

The baseline configuration was chosen to cover both conventional and deep learning recognition routes. MFCC-SVM represented a compact cepstral method, wavelet-random forest represented a transient-feature method, ResNet18 represented dense spectrogram learning, and the original PointNeXt represented direct point-set recognition without acoustic modifications. All models used the same training and testing partitions, and early stopping was based on validation macro-F1. Hyperparameters were selected from the validation set only, because using test-set feedback would overestimate deployment performance. Continuous acoustic monitoring studies usually report false alarms together with recognition metrics because inspection burden is affected by non-biological detections [22]. Accordingly, window-level metrics were reported together with false-alarm rate in this experiment.

The dataset was also examined for class imbalance and acoustic-condition imbalance before training. Background windows were intentionally kept numerous because real monitoring contains long silent or mechanically disturbed periods, but class weights prevented the optimizer from treating background rejection as the main task. Low-SNR samples were oversampled during training by a factor of 1.6, while validation and testing kept the original distribution. This choice created a stricter test condition: the model had to improve weak-event recognition during learning without receiving an artificially simplified evaluation set. Sensor-depth groups were kept visible in the metadata so that performance could be reported by depth rather than only by overall class label.

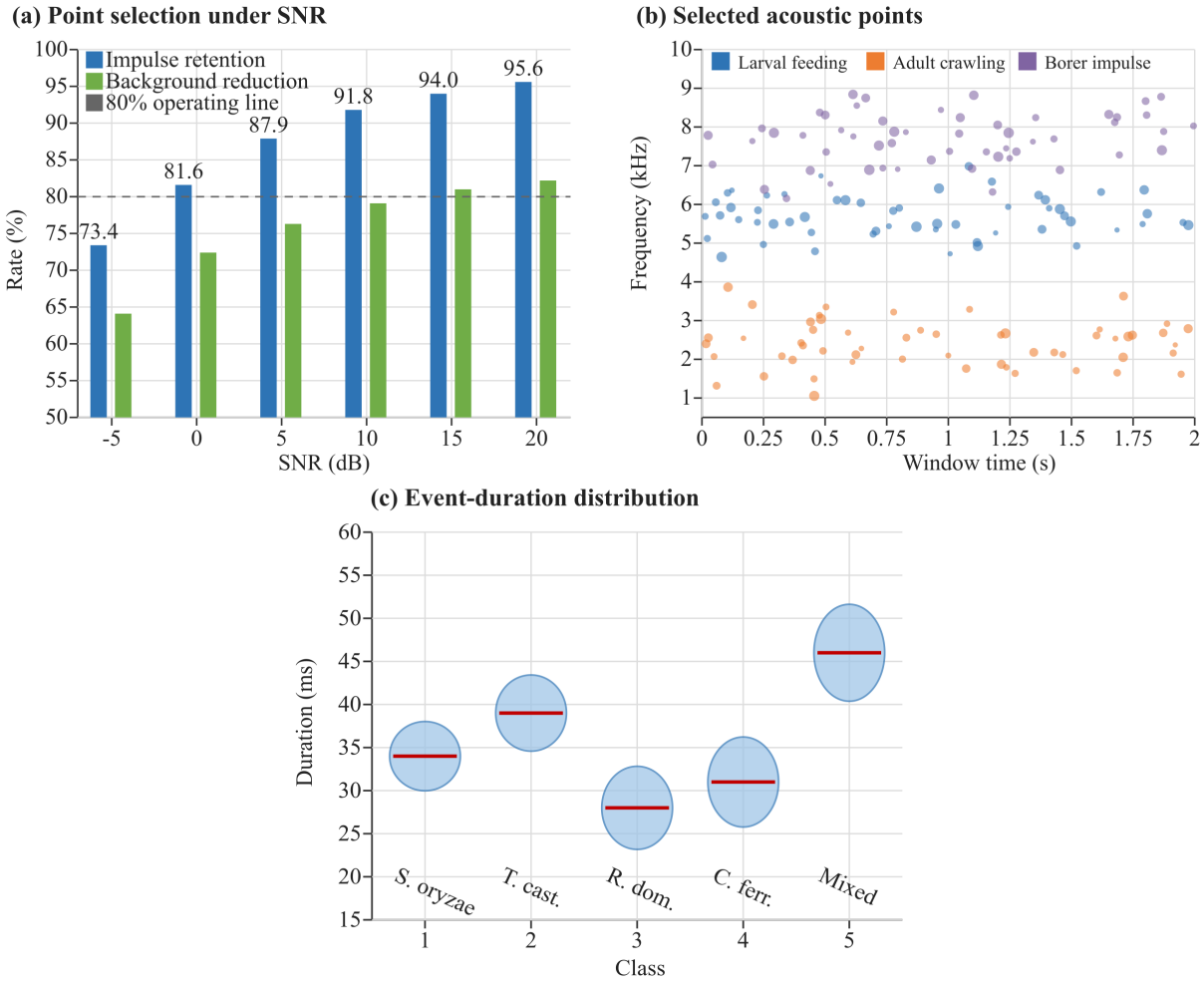



**Figure 3. Acoustic point-cloud construction statistics: (a) impulse retention and background reduction under SNR variation; (b) selected point distribution across time-frequency regions; (c) event-duration distribution for target classes**

## 4.2. Recognition Accuracy and Error Structure

Recognition performance was compared with MFCC-SVM, wavelet-random forest, ResNet18 spectrogram classification, the original PointNeXt, and the proposed improved PointNeXt. Baselines were trained with the same split and the same class labels. The proposed model reached 96.8% overall accuracy, 95.9% macro-F1, 95.4% macro-recall, and a 2.8% false-alarm rate. In contrast, ResNet18 reached 88.3% macro-F1, and the original PointNeXt reached 91.2% macro-F1. The advantage was more visible for sparse event windows because the model did not require a dense image grid to infer insect-specific acoustic structure.

The convergence trend in Figure 4(a) indicates that validation macro-F1 stabilized at 95.7% after 74 epochs, whereas the original PointNeXt plateaued at 91.5% and ResNet18 oscillated around 88.6%. The smoother convergence curve suggests that density-aware neighborhoods reduced sensitivity to random point-density changes during training, while the dense spectrogram baseline still responded to background texture variation. The confusion matrix in Figure 4(b) shows that the largest confusion occurred between S. oryzae and mixed weak activity with 4.9% mutual misclassification, while background was confused with insect classes in only 2.6% of test windows. This error pattern is more favorable for monitoring than a high background-to-insect error rate, because it means most remaining mistakes occur inside the biological-activity category rather than between biological and non-biological states.

The error structure reveals that most residual mistakes were not random label swaps. High-frequency feeding bursts from R. dominica were recognized reliably when SNR exceeded 5 dB, but the same class was sometimes assigned to C. ferrugineus when grain depth reached 2.0 m and high-frequency attenuation compressed the spectral ridge. S. oryzae and mixed weak activity overlapped because both contained irregular chewing and crawling impulses, although the mixed class produced longer event trains. These patterns agree with the observation that species-level acoustic differences can become ambiguous when propagation removes fine temporal details [23].

Precision and recall changed differently across the tested classes. Background precision was high because temporal voting rejected many isolated mechanical sounds, but the recall of mixed weak activity remained lower because its

acoustic definition covered several insect behaviors rather than a single species. This imbalance is important for engineering interpretation. A low recall in a rare weak-activity class may not immediately reduce overall accuracy, yet it can delay the warning stage that motivates acoustic monitoring in the first place. The proposed model partly addresses this issue through saliency-biased sampling, which kept repeated low-energy impulses inside the point budget. The remaining errors indicate that label granularity and acoustic separability should be designed together when building monitoring datasets.

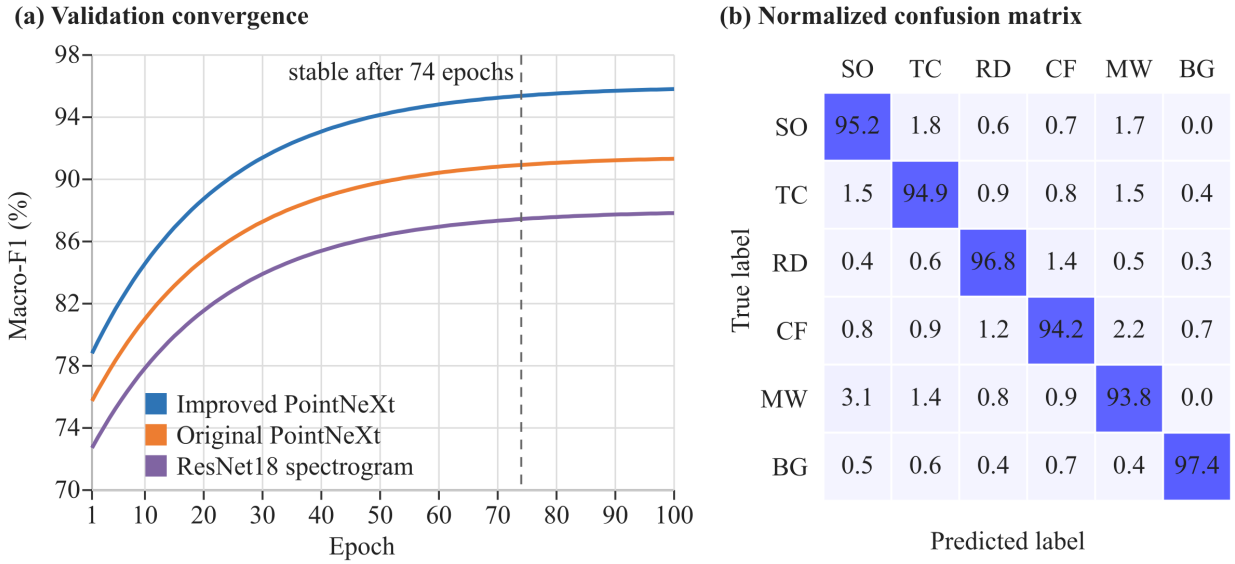



**Figure 4. Recognition convergence and confusion structure: (a) validation macro-F1 curves of competing models; (b) normalized confusion matrix for the proposed model**

The confidence distribution provided another view of error behavior. Correct predictions for R. dominica and background were concentrated above 0.82 posterior probability, reflecting compact acoustic patterns after point selection. Correct predictions for mixed weak activity were spread between 0.58 and 0.86, which means that the model frequently recognized the presence of insect activity but assigned lower species-level certainty. Misclassified windows were usually associated with either low point count or unusually high point density. Low point count occurred when feeding events were strongly attenuated; high point density occurred when a broadband knock generated many candidate points. These two failure modes require different engineering responses, so the auxiliary indicators generated by the decision rule are useful for interpreting alarms.

### 4.3. Ablation, Robustness, and Computational Efficiency

Ablation tests isolated the contribution of the major model components. Removing density-aware grouping reduced macro-F1 from 95.9% to 93.1%, and the largest decline appeared under the 2.0 m grain-depth condition. Removing channel recalibration reduced macro-F1 to 94.0%, mainly affecting classes whose discriminative bands were narrow. Removing temporal voting increased the false-alarm rate from 2.8% to 5.7%, indicating that isolated mechanical impulses remained a major risk for practical monitoring.

The noise-robustness curve in Figure 5(a) shows that macro-F1 decreased by only 3.3 percentage points between 20 dB and -5 dB SNR for the complete model. The flatter low-SNR curve means that the point-selection stage preserved enough local structure for the PointNeXt backbone even when absolute impulse energy was reduced. The ablation comparison in Figure 5(b) shows that density-aware grouping contributed 2.8 percentage points of macro-F1. This gain is consistent with the intended mechanism, because neighborhood radius must expand around sparse attenuated impulses but remain restrained around dense mechanical transients. The saliency distribution in Figure 5(c) shows that event saliency weights assigned higher attention to repeated impulses than to fan harmonics. The saliency gap between repeated impulses and broadband knock events indicates that the model learned a temporal-biological preference rather than a generic high-energy preference.

Robustness was further evaluated across grain depth and sensor coupling pressure. With depth increasing from 0.4 m to 2.0 m, the proposed model retained 94.7% macro-F1, whereas ResNet18 declined to 82.9%. Under loose sensor coupling, accuracy decreased by 1.9 percentage points for the proposed model and by 6.4 percentage points for the original PointNeXt. Figure 6(a) shows that recall for C. ferrugineus stayed above 92.8% across all depths. This result is important because C. ferrugineus produced weaker acoustic activity than the larger borer class, so stable recall under depth variation reflects useful sensitivity to attenuated signals. Figure 6(b) shows that the feature space maintained clear

class separation with a silhouette score of 0.61 after density-aware aggregation. The remaining overlap between S. oryzae and mixed weak activity appears as adjacent clusters rather than complete collapse, which matches the confusion matrix pattern.

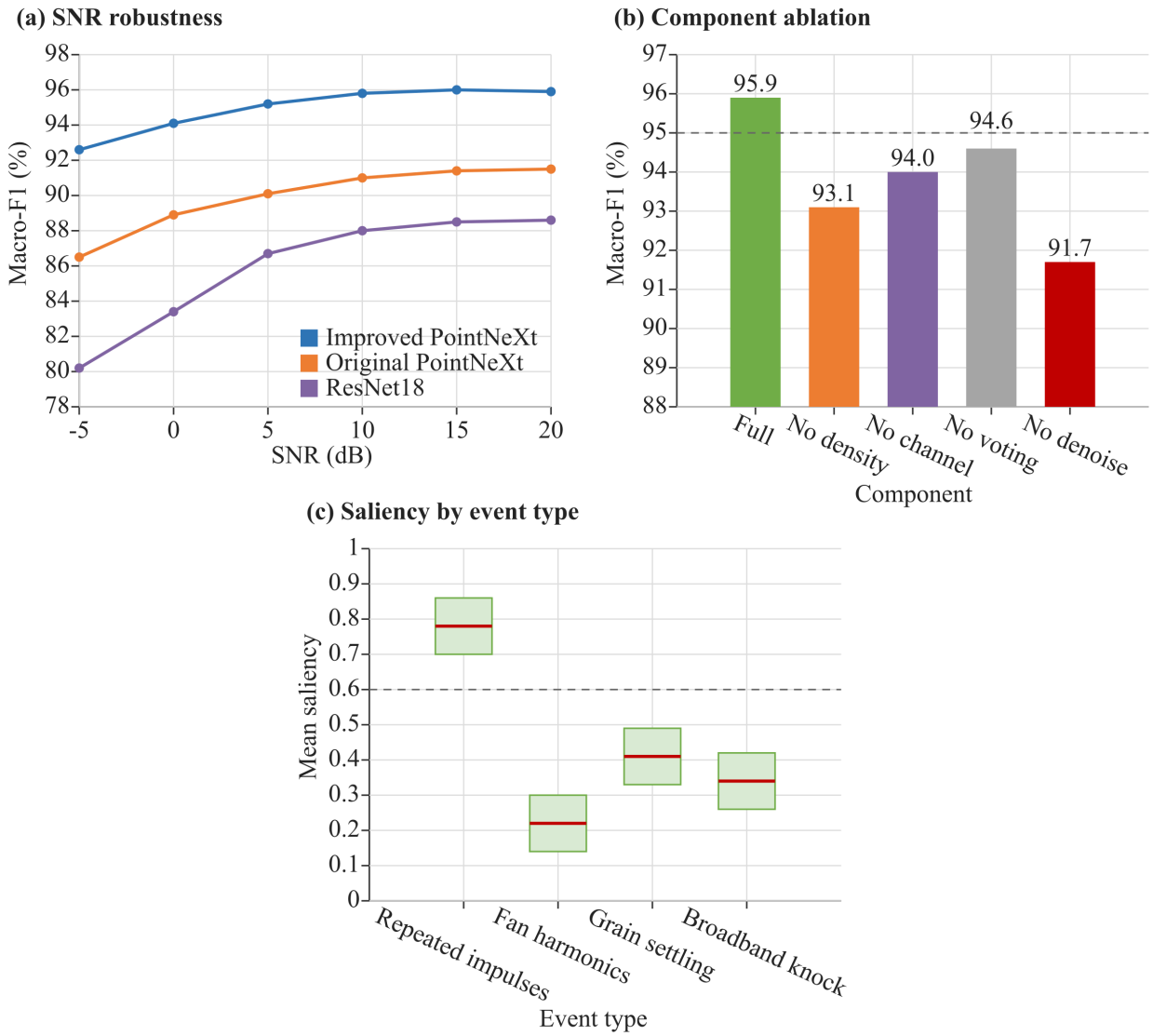



**Figure 5. Ablation and noise robustness: (a) macro-F1 under SNR changes; (b) component-level ablation gains; (c) saliency-weight distribution over acoustic interference types**

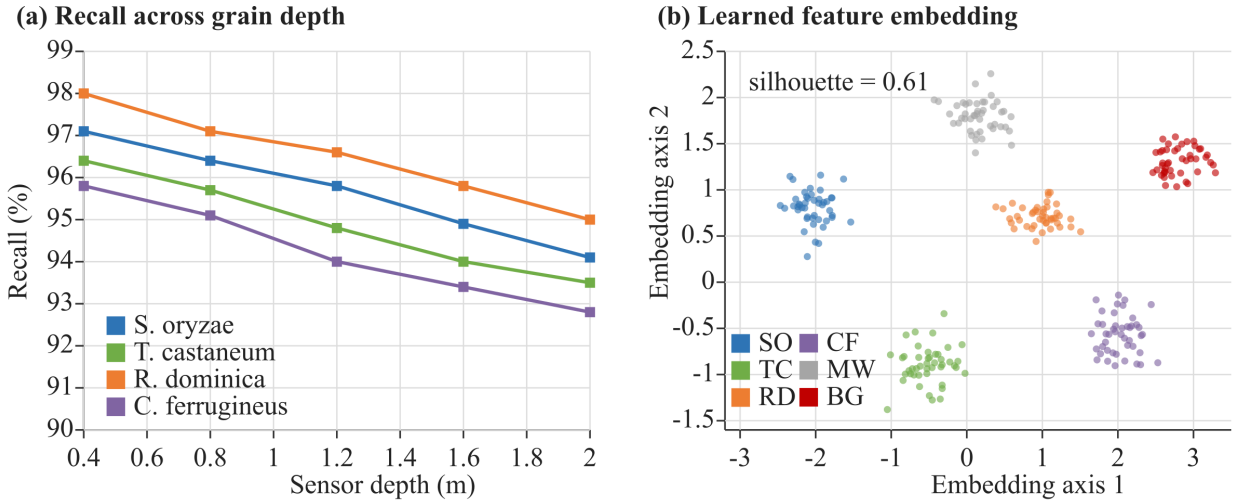



**Figure 6. Depth robustness and embedding behavior: (a) class recall across grain depths; (b) two-dimensional embedding of learned acoustic features**

The computational profile was measured on an embedded GPU and a desktop GPU in single-window inference mode. The improved PointNeXt used 2.34 million parameters, required 5.8 ms per window on the desktop GPU, and required 23.6 ms per window on the embedded device. The temporal voting head added 0.7 ms per monitoring interval because it operated on class probabilities rather than raw features. Figure 7(a) shows that the proposed model reduced inference latency by 41.8% compared with ResNet18 while improving macro-F1 by 7.6 percentage points. The result indicates that sparse acoustic points removed redundant spectrogram background before the classifier spent computation on feature extraction. Figure 7(b) shows that parameter growth over the original PointNeXt was limited to 0.42 million. Most of this increase came from channel recalibration and the voting head, so the recognition gain was not obtained by simply scaling the backbone. Figure 7(c) shows that the edge device processed 42.4 windows per second under the selected point budget. This throughput leaves enough margin for multi-sensor acquisition when windows overlap and alarms are aggregated over short monitoring intervals.

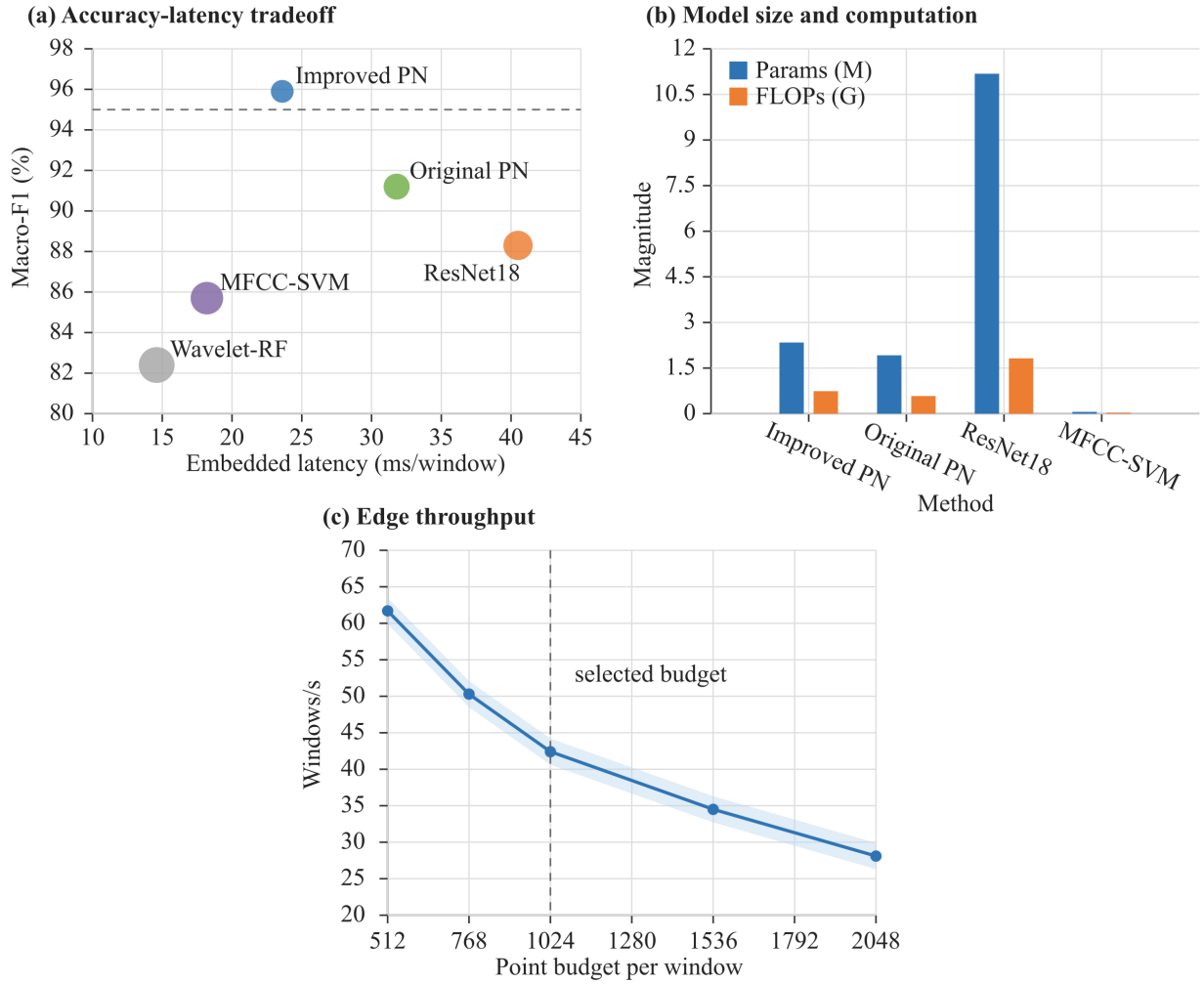



**Figure 7. Computational performance: (a) macro-F1 versus latency; (b) parameter and FLOP comparison; (c) edge-device throughput under different point budgets**

Calibration and threshold behavior were also examined because field systems must trade missed detection against unnecessary inspection alarms, a tension commonly considered in continuous biological monitoring [24]. Expected calibration error was used to measure the difference between predicted confidence and observed correctness across probability bins. With the decision threshold set to 0.62, the system produced 2.8% false alarms and 4.6% missed insect windows. Raising the threshold to 0.74 reduced false alarms to 1.9% but increased missed windows to 7.8%. Figure 8(a) shows that the probability calibration error was below 0.035 for four insect classes but reached 0.052 for mixed weak activity. The higher mixed-class error reflects the broader acoustic composition of that label and suggests that confidence should be interpreted more cautiously for weak combined events. Figure 8(b) shows that the operating point at threshold 0.62 achieved the best balance between detection sensitivity and inspection burden under the assumed monitoring protocol. The curve also indicates that thresholds above 0.68 would make the system more conservative than desirable for early warning because the missed-window rate rises faster than the false-alarm rate falls.

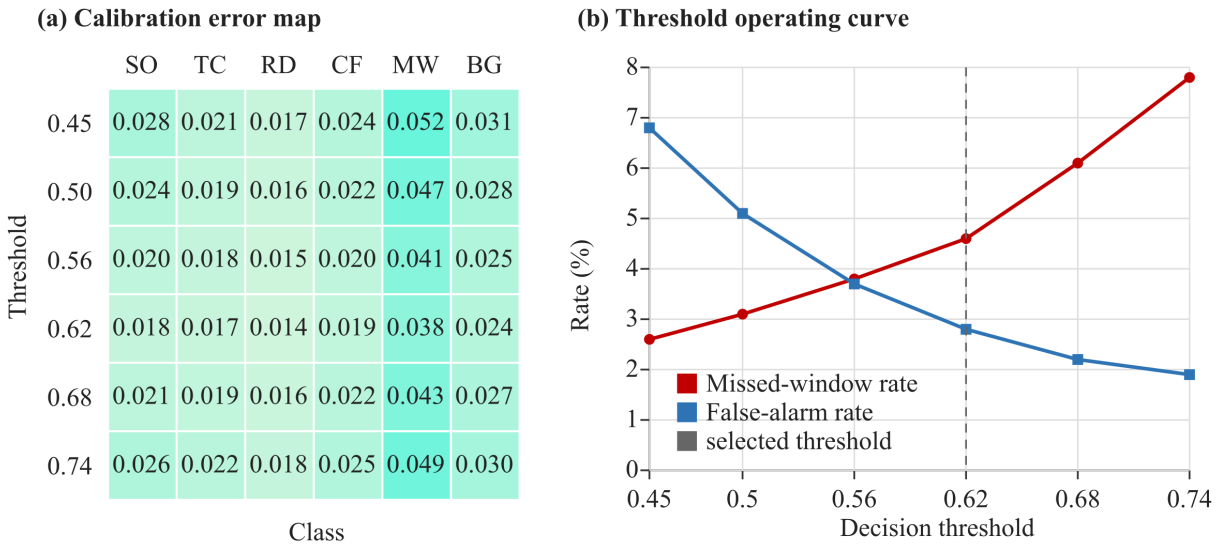



**Figure 8. Calibration and threshold analysis: (a) class-wise calibration error; (b) missed-window rate and false-alarm rate under decision-threshold variation**

The comparison also indicates where the method remains sensitive to data conditions. When fan harmonics overlapped with low-frequency adult crawling sounds, several background windows received insect probabilities above 0.50. When two insect species were active in the same grain chamber, the single-label training protocol forced the model to choose the dominant acoustic source, which lowered recall for the weaker source. Window-level annotation is often used in acoustic monitoring when frame-level biological labels are difficult to obtain [25]. In this experiment, that annotation choice made mixed activity more difficult to interpret than single-species windows. Additional studies have noted that storage-site differences can change spectral attenuation [26]. The present depth and coupling tests therefore support site-adaptive calibration when moving from experimental chambers to commercial bins.

The ablation trend helps explain why a point-cloud model alone is not enough. The original PointNeXt already benefits from sparse input, but it applies neighborhood search without considering whether a cluster is biologically meaningful or simply dense because of broadband interference. Density-aware grouping changes the effective receptive field so that sparse weak impulses can still collect sufficient neighbors, while dense knocks do not automatically receive dominant local support. Channel recalibration then suppresses frequency bands that are active in many background windows. Temporal voting provides a third filter by checking whether the same class remains plausible over adjacent windows. These modules act at different points in the pipeline, which explains why their combined gain is larger than any single component.

A practical deployment interpretation can be drawn from the data without assuming perfect field conditions. At 42.4 windows per second, one embedded monitoring node can process four contact sensors with 50% window overlap and still keep a margin for data transmission. The point budget of 1,024 points per window was sufficient for the tested insect density range, and increasing the budget to 2,048 points improved macro-F1 by only 0.4 percentage points while increasing latency by 36.2%. Similar diminishing returns have been reported in sparse acoustic and event-driven recognition systems when salient points already cover the informative regions [27]. Operational monitoring protocols often separate grain-loading or handling intervals from stable surveillance periods to avoid transient mechanical alarms [28]. In the present test, the false-alarm profile was dominated by grain-settling events after loading, which supports using a short settling exclusion interval or a separate loading-state detector.

A deployment-oriented reading of the results shows that the classifier is strongest when events are sparse but repeated, which is the monitoring condition expected during early infestation. It is less decisive when the label contains mixed biological sources or when thick grain attenuates high-frequency ridges. Sparse acoustic-recognition studies have argued that removing uninformative spectral cells can improve model efficiency when salient event structure is retained [29]. The measured gain over dense spectrogram baselines follows that pattern because point selection reduced background computation while preserving weak impulse clusters. For engineering use, the most important numerical result is not only the 95.9% macro-F1 but also the simultaneous reduction of false alarms to 2.8%. Inspection cost and operator trust are strongly affected by false-positive behavior in continuous monitoring systems [30].

The experimental evidence further suggests that the method is suitable for staged deployment. In a storage facility, acoustic monitoring is usually used as an early-warning layer rather than as the only basis for treatment. A classifier with low false-alarm rate can identify bins that require closer sampling, trap inspection, or temperature-humidity review. The proposed model supports that role because it produces stable probabilities over short monitoring intervals and

because its inference cost remains below the time budget of multi-sensor edge acquisition. The analysis also shows that the preprocessing stage should be monitored during operation. If the selected point count collapses under a new sensor installation, the recognition head may still output a label, but the confidence value should be interpreted together with point density and saliency statistics.

The robustness tests also clarify how the model should be maintained after installation. A storage site may change its acoustic character after grain turnover, sensor replacement, fumigation preparation, or seasonal moisture variation. The proposed framework can be recalibrated by updating the activity threshold and the temporal voting threshold before full model retraining is considered. If these adjustments fail to restore stable point counts and low entropy, a small set of site-specific labeled windows can be added for fine-tuning. This maintenance path is compatible with engineering practice because it separates simple threshold correction from more expensive annotation and retraining. It also avoids treating every new acoustic condition as a completely new recognition problem.

## 5. Conclusion

This study developed an improved PointNeXt framework for acoustic recognition of stored-grain insects under granary monitoring data conditions. The method converts denoised time-frequency events into sparse acoustic point clouds and introduces density-aware grouping, channel recalibration, and temporal voting to address attenuation, nonstationary interference, and weak biological impulses. The resulting framework links acoustic preprocessing with point-cloud learning in a form suitable for early-warning pest surveillance, especially when acoustic decisions are used to guide follow-up inspection rather than replace all storage-management procedures.

Several limitations remain. The dataset design covers representative grain depths, sensor coupling states, and mechanical interference sources, but it does not fully reproduce the diversity of commercial storage facilities. Broader field testing is still required before the reported recognition accuracy can be treated as a general deployment value. The single-label window protocol also limits the interpretation of simultaneous activity from multiple species. In addition, the current preprocessing stage depends on thresholded event selection, which may require recalibration when the grain medium, moisture level, or sensor hardware changes substantially.

Future research should extend the framework to multi-label recognition, site-adaptive calibration, and long-term deployment trials in operating storage systems. More detailed acoustic propagation modeling could be combined with self-supervised point-cloud learning to reduce annotation demand. Hardware-aware compression and active learning may further support low-power sensor networks for early warning, inspection scheduling, and integrated stored-grain pest management.

## 6. Declarations

### 6.1 Author Contributions

Rasmus Sepp contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Jaan Mägi contributes to validation, analysis, investigation, data collection. All authors have read and agreed with the manuscript before its submission and publication.

### 6.2 Funding and Acknowledgments

This research received no specific financial support from any funding agency.

### 6.3 Institutional Review Board Statement

Not applicable.

### 6.4 Informed Consent Statement

Not applicable.

### 6.5 Declaration of Competing Interest

Not applicable.

### 6.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## 7. References

[1] Shi, Z., Dang, H., Liu, Z., & Zhou, X. (2020). Detection and identification of stored-grain insects using deep learning: A more effective neural network. IEEE access, 8, 163703-163714. https://doi.org/10.1109/access.2020.3021830

[2] Li, J., Su, Y., Cui, Z., Tian, J., & Zhou, H. (2022). A Method to Establish a Synthetic Image Dataset of Stored-Product Insects for Insect Detection. IEEE Access, 10, 70269-70278. https://doi.org/10.1109/access.2022.3188282

[3] Ai, Y., Sun, C., Tie, J., & Cai, X. (2020). Research on Recognition Model of Crop Diseases and Insect Pests Based on Deep Learning in Harsh Environments. IEEE Access, 8, 171686-171693. https://doi.org/10.1109/access.2020.3025325

[4] Wu, F., & Li, Y. (2023). Lightweight Field Insect Recognition and Classification Model Based on Improved Deep Learning under Complex Background. Security and Communication Networks, 2023, 1-9. https://doi.org/10.1155/2023/6560747

[5] Hou, Y., & Botteldooren, D. (2023). Artificial intelligence-based collaborative acoustic scene and event classification to support urban soundscape analysis and classification. INTER-NOISE and NOISE-CON Congress and Conference Proceedings, 265(1), 6466-6473. https://doi.org/10.3397/in_2022_0974

[6] Gourishetti, S., Johnson, D., Werner, S., Kátai, A., & Holstein, P. (2021). Partial discharge monitoring using deep neural networks with acoustic emission. INTER-NOISE and NOISE-CON Congress and Conference Proceedings, 263(3), 3312-3323. https://doi.org/10.3397/in-2021-2373

[7] Lei, M., Zeng, X., & Jin, A. (2023). Deep attention-based multi-task learning for underwater acoustic target recognition. INTER-NOISE and NOISE-CON Congress and Conference Proceedings, 268(6), 1995-2004. https://doi.org/10.3397/in_2023_0298

[8] Borin, M., Masiero, B., & Monteiro, C. (2023). Exploring deep learning architectures for urban sound classification. INTER-NOISE and NOISE-CON Congress and Conference Proceedings, 268(3), 5786-5796. https://doi.org/10.3397/in_2023_0826

[9] Zhang, Q., Guan, Y., Li, H., Xiong, K., & Song, Z. (2023). Distributed deep learning-based signal classification for time-frequency synchronization in wireless networks. Computer Communications, 201, 37-47. https://doi.org/10.1016/j.comcom.2023.01.014

[10] Chen, X., Li, H., Huang, Y., Han, W., Yu, X., Zhang, P., & Tao, R. (2023). Heart sound classification based on equal scale frequency cepstral coefficients and deep learning. Biomedical Engineering / Biomedizinische Technik, 68(3), 285-295. https://doi.org/10.1515/bmt-2021-0254

[11] Stoumpou, V., Vargas, C. D. M., Schade, P. F., Boyd, J. L., Giannakopoulos, T., & Jarvis, E. D. (2022). Analysis of Mouse Vocal Communication (AMVOC): a deep, unsupervised method for rapid detection, analysis and classification of ultrasonic vocalisations. Bioacoustics, 32(2), 199-229. https://doi.org/10.1080/09524622.2022.2099973

[12] Ekpezu, A. O., Wiafe, I., Katsriku, F., & Yaokumah, W. (2021). Using deep learning for acoustic event classification: The case of natural disasters. The Journal of the Acoustical Society of America, 149(4), 2926-2935. https://doi.org/10.1121/10.0004771

[13] Wu, B., & Zhang, X.-P. (2022). Environmental Sound Classification via Time-Frequency Attention and Framewise Self-Attention-Based Deep Neural Networks. IEEE Internet of Things Journal, 9(5), 3416-3428. https://doi.org/10.1109/jiot.2021.3098464

[14] Demir, F., Abdullah, D. A., & Sengur, A. (2020). A New Deep CNN Model for Environmental Sound Classification. IEEE Access, 8, 66529-66537. https://doi.org/10.1109/access.2020.2984903

[15] Ullo, S. L., Khare, S. K., Bajaj, V., & Sinha, G. R. (2020). Hybrid Computerized Method for Environmental Sound Classification. IEEE Access, 8, 124055-124065. https://doi.org/10.1109/access.2020.3006082

[16] Mnasri, Z., Rovetta, S., & Masulli, F. (2021). Anomalous sound event detection: A survey of machine learning based methods and applications. Multimedia Tools and Applications, 81(4), 5537-5586. https://doi.org/10.1007/s11042-021-11817-9

[17] Chen, Z., Liu, Y., Zhu, J., Zhang, Y., Jin, R., He, X., Tao, J., & Chen, L. (2021). Time-frequency deep metric learning for multivariate time series classification. Neurocomputing, 462, 221-237. https://doi.org/10.1016/j.neucom.2021.07.073

[18] Hazer, A., & Yildirim, R. (2022). Deep Learning Based Point Cloud Processing Techniques. IEEE Access, 10, 127237-127283. https://doi.org/10.1109/access.2022.3226211

[19] Zhang, M., You, H., Kadam, P., Liu, S., & Kuo, C.-C. J. (2020). PointHop: An Explainable Machine Learning Method for Point Cloud Classification. IEEE Transactions on Multimedia, 22(7), 1744-1755. https://doi.org/10.1109/tmm.2019.2963592

[20] Li, D., Shen, X., Yu, Y., Guan, H., Wang, H., & Li, D. (2020). GGM-Net: Graph Geometric Moments Convolution Neural Network for Point Cloud Shape Classification. IEEE Access, 8, 124989-124998. https://doi.org/10.1109/access.2020.3007630

[21] Wang, J., Xu, C., Dai, L., Zhang, J., & Zhong, R. (2021). An Unequal Deep Learning Approach for 3-D Point Cloud Segmentation. IEEE Transactions on Industrial Informatics, 17(12), 7913-7922. https://doi.org/10.1109/tii.2020.3044106

[22] Zhong, M., Li, C., Liu, L., Wen, J., Ma, J., & Yu, X. (2020). Fuzzy Neighborhood Learning for Deep 3-D Segmentation of Point

Cloud. IEEE Transactions on Fuzzy Systems, 28(12), 3181-3192. https://doi.org/10.1109/tfuzz.2020.2992611

[23] Zhang, Z., Sun, L., Zhong, R., Chen, D., Zhang, L., Li, X., Wang, Q., & Chen, S. (2021). Hierarchical Aggregated Deep Features for ALS Point Cloud Classification. IEEE Transactions on Geoscience and Remote Sensing, 59(2), 1686-1699. https://doi.org/10.1109/tgrs.2020.2997960

[24] Fu, Z., & Hu, W. (2021). Dynamic Point Cloud Inpainting via Spatial-Temporal Graph Learning. IEEE Transactions on Multimedia, 23, 3022-3034. https://doi.org/10.1109/tmm.2021.3068606

[25] Wang, Y., Tang, X., & Yue, C. (2022). Enhancing the Local Graph Semantic Feature for 3D Point Cloud Classification and Segmentation. IEEE Access, 10, 74620-74628. https://doi.org/10.1109/access.2022.3190966

[26] Guo, B., & Zuo, X. (2020). An Optimized Point Cloud Classification and Object Extraction Method Using Graph Cuts. IEEE Access, 8, 188515-188525. https://doi.org/10.1109/access.2020.3030717

[27] Chiang, C.-H., Kuo, C.-H., Lin, C.-C., & Chiang, H.-T. (2020). 3D Point Cloud Classification for Autonomous Driving via Dense-Residual Fusion Network. IEEE Access, 8, 163775-163783. https://doi.org/10.1109/access.2020.3021191

[28] Zhai, R., Li, X., Wang, Z., Guo, S., Hou, S., Hou, Y., Gao, F., & Song, J. (2020). Point Cloud Classification Model Based on a Dual-Input Deep Network Framework. IEEE Access, 8, 55991-55999. https://doi.org/10.1109/access.2020.2981357

[29] Wang, Y., Yue, C., & Tang, X. (2021). A Geometry Feature Aggregation Method for Point Cloud Classification and Segmentation. IEEE Access, 9, 140504-140511. https://doi.org/10.1109/access.2021.3119622

[30] Engel, N., Belagiannis, V., & Dietmayer, K. (2021). Point Transformer. IEEE Access, 9, 134826-134840. https://doi.org/10.1109/access.2021.3116304