*Volume 2, 2026, Pages 239-252* *Journal of Smart and Sustainable Farming*
| Regular Article | Open Access

# Acoustic Sensing and Machine Learning for Early Detection of Insect Activity in Stored Grains

**Nina Cedergreen**
Department of Crop Production, S. Seifullin Kazakh Agrotechnical Research University, Astana 010000, Kazakhstan.
ninaceder.g@kazatu.edu.kz

**Sangi Olani Lemmi**
Department of Forestry, School of Agriculture and Natural Resource Management, Mattu University, Metu, Ethiopia.
sagni.olani@mau.edu.et

**Article Info**
Journal of Smart and Sustainable Farming
https://www.ansispublications.com/jssf/jssf.html

**© The Author(s), 2026.**
https://doi.org/10.64026/JSSF/2026020

Received 02 June 2026
Revised from 16 August 2026
Accepted 28 September 2026
Available online 22 October 2026
**Published by Ansis Publications.**

**Corresponding author(s):**
Sangi Olani Lemmi, Department of Forestry, Mattu University, Metu, Ethiopia.
**Email:** sagni.olani@mau.edu.et

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/ licenses/by/4.0/).

**Abstract -** Insects hidden inside stored grain are difficult to detect before damage becomes economically significant, and conventional sampling and trapping are slow, labour-intensive and insensitive at low insect densities. This article presents an acoustic sensing and machine-learning framework for early detection of insect activity in grain storage. The framework combines band-limited filtering and spectral subtraction with a fused vector of 113 descriptors (MFCC statistics, log-mel statistics, spectral-shape measures and pulse-train timing features) and a lightweight classifier that separates clean grain from three insect classes. No labelled field corpus was available for this study, so the framework is evaluated on a physically motivated synthetic benchmark in which insect feeding and movement are modelled as trains of damped sinusoidal pulses embedded in grain-bin noise with impulsive distractors. All numerical results are therefore simulated and should be read as a controlled comparison of methods, not as field performance. Under five-fold cross-validation at clip-level signal-to-noise ratios (SNR) of −24 to −4 dB, the proposed pipeline reached 87.9% four-class accuracy against 78.8% for an MFCC-only baseline, and detected insect activity with an F1 score of 0.961. In an independent noise sweep, the proposed pipeline outperformed the baseline by 8.5 to 14.5 percentage points between −24 and −12 dB. Feature extraction took about 15 ms per two-second clip on one CPU core. Limitations of simulation and the field validation required are discussed.

**Keywords -** Acoustic Sensing, Stored Grain, Insect Detection, Machine Learning, Mel-Frequency Cepstral Coefficients, Early Warning, Precision Grain Storage.

## I. INTRODUCTION

Grain is produced seasonally, but eaten all year round and therefore needs to be stored – perhaps for months – in bins, silos, warehouses or in household containers. Insects, like weevils, grain borers and flour beetles, feed on the kernels during storage, increasing the moisture and temperature in the area and setting up conditions for moulds to form. Many of these species form the most injurious stage within the kernel and so infestations will not be detected until after the adults have emerged, damage has been done or the kernel has been heated. Losses due to insects are estimated to be between 12% to 40% of the stored rice according to one investigation of stored rice [4]. Regardless of the exact number, the important point is that the early warning is the most economical way of controlling the disease, since the treatment is more economical, less chemical and more effective when the infestation is restricted to a small area of the grain mass.

Presently, probe sampling, pitfall traps and pheromone traps are used in addition to visual inspections. A recent review on stored product monitoring indicates that conventional insecticides are still the most important means of pest control, and although some stored product monitoring methods are less sensitive and more expensive, newer methods are more sensitive and specific and can provide input to decision support tools [1]. One such method is acoustic sensing. Short length of vibration, broadband frequency, produced by feeding, chewing and movement of insects, can be detected by microphones or contact (piezoelectric) sensors installed in the grain. The signal is self-generated and therefore in principle,

acoustic methods are able to detect hidden immature stages as well as adults. Signal processing and machine learning can be used to distinguish insect sounds in the background noise and, a commercial machine, can be used to identify rice weevil (Sitophilus oryzae) adults and larvae in grain [2].

Implementing this approach into a field system is, however, hard. The sounds produced by the ventilation fans, conveyors, rain on metal roofs and grain settling are similar in frequency and, in the instance of grain settling, impulsive, and therefore cannot be separated from insect sounds in a storage facility. These are the issues that have been addressed by those who develop an early-warning sensor deployed in the field: detecting a weak insect sound in an agricultural background noise, and controlling data volume and analysis speed to achieve timely alerts [3]. For a practical detector, the following three properties are needed: (1) low SNR property, (2) discrimination of insect pulse trains from isolated impulsive noise, and (3) computation complexity sufficiently low to allow continuous processing on low-cost equipment.

With a feature-fusion pipeline that utilizes a simple observation: that insect activity is not limited to the spectral domain, but is also temporal, this article tackles those requirements. The spectra of a single click when settling grain and of a series of repeated clicks from a feeding insect are similar, but the repetition rate and the regularity of the inter-pulses are different. The following is a list of the contributions. (i) Processing chain with band-pass filtering, spectral subtraction with noise profile self-estimated, and the 113-dimensional fused feature vector - the combination of cepstral, spectral and pulse-timing information. (ii) A ‘synthetic benchmark’ in which insect pulse trains, grain-bin noise and impulsive distractors are used with a clip-level SNR as a controlled variable, and which is transparent and parameterisable. (iii) Ablation study and noise-sweep creating and measuring the contribution of each processing and feature block. (iv) Gap between simulated and field performance and a clear action plan to mitigate any gap.

The remainder of the article is organised as follows. The related work is reviewed in section 2. The methodology is presented in section 3, with details on the signal model, pre-processing, the features used and the classifiers and evaluation protocol. The simulated results are presented and discussed in Section 4. This section ends where it began, and summarizes future work.

## II. RELATED WORKS

### *Acoustic and Vibro-Acoustic Sensing of Stored-Product Insects*

Recent studies concentrate on low-cost sensors and data-driven classification. adapted a MEMS microphone to record sounds of three major stored-rice pests, the lesser grain borer, the rice weevil and the red flour beetle, converted recordings to spectrograms and classified them with a convolutional neural network (CNN), reporting an average accuracy of 84.51% [4]. A companion study characterised the sound signatures of the same pests using waveforms, RMS energy, frequency content and spectrograms, and found that the two primary pests could be distinguished from the secondary pest [5]. These studies show that inexpensive sensing is feasible, but they also show that accuracy in species classification is moderate when recordings are made in realistic conditions.

The developed the Acoustic Stored Product Insect Detection System (A-SPIDS), which places piezoelectric sensors inside and microphones outside a sound-insulated container. Because the sensors were uncalibrated, they normalised signals against ambient sensor noise and used a pulse-amplitude measure and an average-energy measure; signals in flour were markedly weaker than in other commodities, and the authors present the results as a first step towards a detection algorithm [6]. The Beetle Sound Tube described by remains in the grain, records continuously and alerts the storekeeper by e-mail, and its authors discuss the difficulty of faint signals, noise and data volume [3]. Mankin *et al.* review the field and emphasise the potential for detecting immature stages and for automation as devices become cheaper [2].

### *Machine-Learning Classifiers for Insect and Pest Sounds*

Deep learning has been applied to a growing range of insect sound problems. proposed an Internet-of-Things (IoT) system that classifies red palm weevil larvae sounds with a modified mixed depthwise convolutional network, using a public dataset of 146 infested and 351 clean samples; they report a best accuracy of 97.38% and a ten-fold cross-validated mean of 95.90 ± 1.46% [7]. Added a dual-frequency and spectral fusion module to a pretrained EfficientNet-b7 and reported 80.26% accuracy on the InsectSet32 dataset [8]. Describe an IoT system with a multi-feature, multi-network model trained on 9,600 pest audio samples and report an accuracy of 99.82% [9]. Trained a CNN on spectrograms of fruit fly wingbeats and deployed it on an ESP32 microcontroller, reporting up to 96.86% accuracy on live field audio [10].

General audio pattern recognition provides further tools. Large pretrained audio networks such as PANNs [11] and attention-based spectrogram models such as the Audio Spectrogram Transformer [12] have set strong baselines on general sound datasets and can be fine-tuned for specialised tasks. A systematic review of data augmentation for sound classification [13] documents the augmentation techniques that most often help when labelled data are scarce, a situation that applies directly to stored-grain acoustics.

### *Signal Characteristics and Noise Sources in Grain Storage*

Most of the acoustic emissions generated in grains by insects are associated with feeding and flight. They are not continuous tones, but rather, series of short, wide pulses with periods of silence between them, and the timing varies between species and life stages. Two immediate implications come next. Two concrete implications arise. The high energy of a pulse train causes the clip-level SNR to become significantly negative even when individual pulses are readily visible in a spectrogram,

and important information is contained in the timing of pulses which a long window spectrum averages out. The same is reflected in the signatures measured by, who found that the pulse-based and average energy amplitude signals were different and higher for the pulse-based signal [6].

Also the sensing medium is important. Signals in flour were less intense than in other stored products [6] and the rice experiments gave better results in a sound-insulated chamber than in open conditions [4, 5]. In a working facility, the signal should not be confused with the ventilation, conveying equipment, weather events or grain settling as described in the section of a deployed monitoring tube [3]. The sources are distributed in terms of both their time and frequency characteristics: The noise from machinery is generally low in frequency, low in intensity and stationary, it can be decreased by filtering and by subtracting it from a stationary noise estimate, and the impulsive events are typically high in intensity, high in frequency and not stationary, and must be rejected by the classifier. Hum, coloured noise and impulsive transients are therefore present in all classes, as is the simplified version of the two types of interference found in the benchmark in Section 3.

*Summary and Research Gap*

**Table 1** summarises representative studies. Reported accuracies are not directly comparable: they depend on the target (adults, larvae, wingbeats), the sensor, the medium, the number of classes and, importantly, on how data were split and whether recordings were made in a quiet chamber or in the field. Three gaps motivate the present work. First, most studies evaluate species classification rather than the operationally more important question of whether insects are present at all in noisy conditions. Second, noise robustness is rarely quantified as a function of SNR, so it is unclear at what noise level a method stops being useful. Third, many deep models require more computation than low-cost storage monitors can provide, whereas the temporal structure of insect pulse trains, which is cheap to compute, is under-used. This article targets these gaps with an explicit SNR-controlled evaluation of a low-cost feature-fusion pipeline.

**Table 1.** Representative Recent Studies on Acoustic Detection of Insects

| Ref. | Target / medium | Sensing | Method | Reported outcome |
|---|---|---|---|---|
| [4] | Three stored-rice pests | MEMS microphone | Spectrogram + CNN | Average accuracy 84.51% |
| [5] | Three stored-rice pests | MEMS microphone | Waveform, RMS, spectra | Primary pests separable from secondary pest |
| [6] | Three pests in stored products | Piezo sensors + microphones in insulated box | Noise-normalised pulse and energy metrics | Signals weaker in flour; first step to a detector |
| [3] | Stored grain (silos, flat stores) | In-grain sensor | Continuous recording, e-mail alert | Noise and data-volume challenges reported |
| [7] | Red palm weevil larvae (date palm) | Sound sensor, IoT | MixConvNet | 97.38% best; 95.90 ± 1.46% (10-fold) |
| [8] | Insect sound corpus (InsectSet32) | Recordings | EfficientNet-b7 + fusion module | 80.26% accuracy |
| [9] | Field pests (IoT) | Sensors, IoT | Multi-feature, multi-network | 99.82% accuracy reported |
| [10] | Oriental fruit fly (orchard) | Acoustic trap | Spectrogram CNN on ESP32 | Up to 96.86% (field) |

## III. PROPOSED METHODOLOGY

The proposed system takes a short audio or vibration clip from a sensor in the grain mass, suppresses stationary noise, extracts a fused feature vector and classifies the clip as clean grain or as insect activity of a given class. The design favours quantities that are inexpensive to compute and that capture both the spectral and the temporal signature of insect pulses.

The concept of acoustic sensing and machine-learning for early detection of insect activity in stored grain is presented in **Fig. 1**. The sound signals received from the grain storage areas are subjected to band-pass filtering, noise reduction, framing and windowing operations. The pre-processed signals are converted to an 113-dimensional feature vector consisting of MFCC statistics, spectral descriptors, and pulse-train timing features. The machine-learning classifier is trained using labelled acoustic clips of three classes of insects and clean grain. When operating, the trained model predicts insect activity from the incoming acoustic signals. The classifications and confidence scores that follow aid in early warning and timely grain storage management decisions.

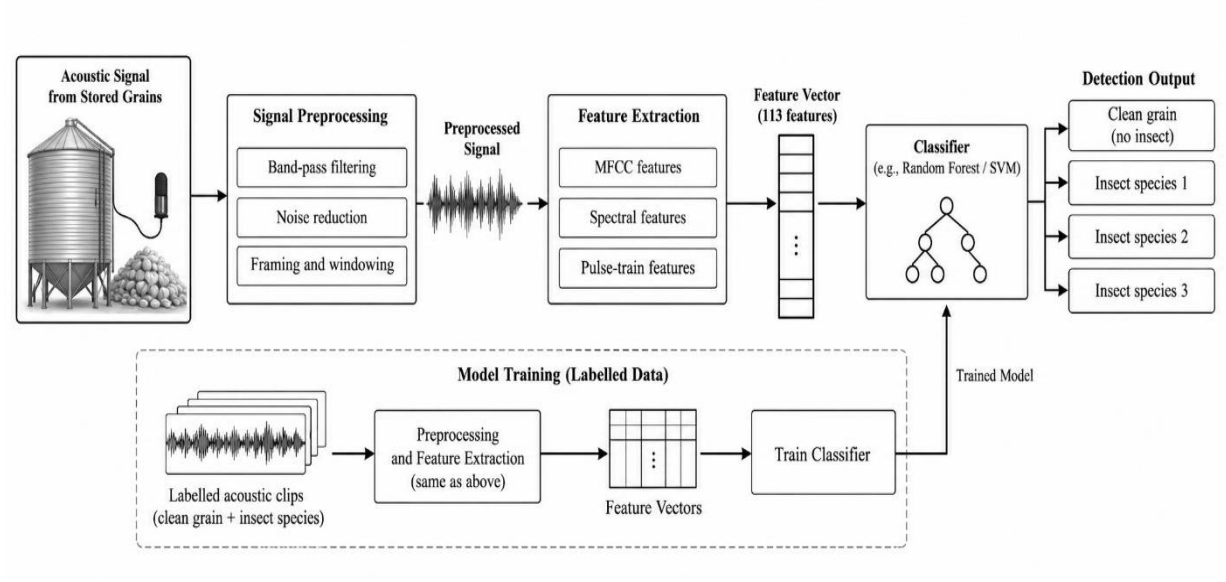



**Fig 1.** Block Diagram of The Proposed Acoustic Detection Pipeline.

*Signal Model and Synthetic Benchmark*

Because a labelled recording corpus from stored grain was not available for this study, the evaluation uses a synthetic benchmark. The benchmark is a controlled test bed, not a substitute for field data, and every result derived from it is labelled as simulated. A clip of duration T = 2 s is sampled at $f_s$ = 16 kHz and is modelled as

$$x[n] = w[n] + g \cdot \sum_k a_k \cdot s(n - n_k;\ f_k, \tau) \tag{1}$$

where w[n] is background noise, $n_k$ are pulse onset times, $a_k \in [0.4, 1.6]$ are random amplitudes, and each pulse is a damped sinusoid

$$s(t;\ f, \tau) = exp\left(-\frac{t}{\tau}\right) \cdot sin(2\pi\ ft\ + \phi), 0 \leq\ t \leq\ 5\tau \tag{2}$$

with random phase φ, centre frequency f drawn per pulse from a Gaussian, and decay constant τ. Inter-pulse intervals follow a Gamma distribution with shape 4 and mean 1/ρ, where ρ is the species pulse rate, which produces irregular but class-dependent timing. The gain g sets the clip-level SNR,

$$g\ =\ sqrt\left(\left(\frac{P_w}{P_{s0}}\right) \cdot 10^{\frac{SNR}{10}}\right) \tag{3}$$

where $P_w$ is the mean noise power, $P_{s0}$ the mean power of the unit-gain pulse train, and $P_s = g^2 P_{s0}$. The SNR is computed over the whole clip, including the silence between pulses, so the instantaneous SNR within a pulse is much higher than the clip-level value. Three insect classes (A, B, C) are defined as hypothetical surrogates and not as measured species; their parameters (**Table 2**) are chosen so that the classes overlap in centre frequency and pulse rate, as real species are expected to do. They were not fitted to any recording.

**Table 2.** Parameters of the Simulated Insect Classes (Hypothetical Surrogates)

| Class | Centre frequency, mean (SD), Hz | Decay constant τ, ms | Pulse rate ρ, $s^{-1}$ |
|---|---|---|---|
| Species A | 3000 (800) | 2.5 | 8 |
| Species B | 4000 (900) | 2.0 | 11 |
| Species C | 5000 (1000) | 1.6 | 15 |
| Noise only | — | — | — |

Background noise is generated as coloured noise with a power spectrum proportional to $f^{-\beta}$ with β drawn uniformly from [0.5, 1.5], plus mains-like hum at 50 Hz and its first three harmonics with random amplitude, plus a Poisson number (mean 1.5) of impulsive transients lasting 5–30 ms with amplitudes of 2–6 noise standard deviations. The transients imitate grain settling and mechanical knocks and are present in the noise-only class as well, so that a detector cannot succeed simply by detecting impulses. **Fig. 2** shows one example clip per class.

*Pre-Processing*

Pre-processing has two stages. First, a fourth-order Butterworth band-pass filter (500–7900 Hz, applied forward and backward for zero phase) removes the low-frequency region dominated by hum and machinery. Second, spectral subtraction attenuates the remaining stationary noise. The short-time Fourier transform (STFT) X(m, k) is computed with a 512-sample Hann window (32 ms) and 256-sample hop. The noise power profile N(k) is estimated from the 20% of frames with the lowest total energy in the same clip, which avoids the need for a separate noise-only recording. The enhanced magnitude is

$$|\hat{s}(m,k)| = \sqrt{\max(|X(m,k)|^2 - \alpha \cdot N(k), \beta \cdot |X(m,k)|^2} \quad (4)$$

with over-subtraction factor α = 1.5 and spectral floor β = 0.05. The original phase is retained and the clip is reconstructed by inverse STFT. The floor limits the musical-noise artefacts typical of subtraction methods.

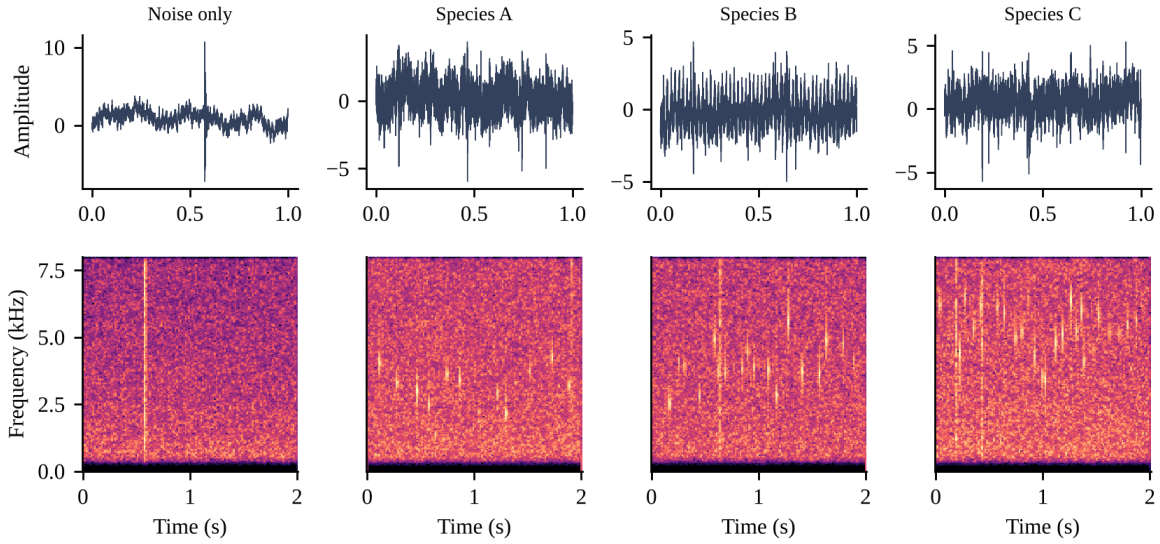



**Fig 2.** Simulated Examples at A Clip-Level SNR Of −12 Db. Top: First Second of The Raw Waveform. Bottom: Spectrogram of the Pre-Processed 2 S Clip. The Noise-Only Clip Contains One Impulsive Transient; The Insect Clips Contain Irregular Pulse Trains Whose Spectral Band and Rate Depend on the Class. These Are Synthetic Signals.

*Feature Extraction*

Four feature groups, totalling 113 values per clip, are extracted from the enhanced signal. Cepstral group (60 values). Power spectra are passed through a bank of 40 triangular mel filters $H_j(k)$ spanning 50–8000 Hz, giving log-mel energies

$$M(m,j) = \ln\left(\sum_k H_{j(k)} \cdot |X(m,k)|^2 + \varepsilon\right) \quad (5)$$

and the first 20 mel-frequency cepstral coefficients (MFCCs) are obtained with an orthonormal discrete cosine transform,

$$c_{q(m)} = \sum_{j=1}^{40} M(m,j) \cdot \cos\left[\frac{\pi q(j - ½)}{40}\right], q = 1, \ldots, 20 \quad (6)$$

For each coefficient the mean and standard deviation over frames and the mean absolute frame-to-frame difference (a delta measure) are retained. Log-mel group (40 values). The mean and standard deviation of every second log-mel band. Spectral-shape group (7 values). Spectral centroid, bandwidth, 85% roll-off frequency, spectral flatness, zero-crossing rate, waveform kurtosis and crest factor. Pulse-timing group (6 values). The signal is band-pass filtered (1–7.9 kHz), its Hilbert envelope e(n) is low-pass filtered at 400 Hz, and pulses are detected as peaks exceeding

$$e(n) > median(e) + 4 \cdot MAD(e), \text{ with minimum peak spacing 12 ms} \quad (7)$$

where MAD is the median absolute deviation. From the detected peaks the pulse rate (peaks per second), the mean and standard deviation of the inter-pulse interval, the mean peak prominence in MAD units, the envelope kurtosis and the envelope peak-to-median ratio are computed. These six values encode the timing structure that spectral features discard, and they are the main feature-level addition of this pipeline over cepstral baselines.

*Classification and Evaluation Protocol*

Features are standardised to zero mean and unit variance within each training fold. Five classifiers are compared: a support vector machine with radial basis function kernel ($SVM, K(x,x') = exp(-\gamma\|x - x'\|^2), C = 10, \gamma = 1/(d \cdot Var(X))$), a random forest (300 trees), a multilayer perceptron (two hidden layers of 128 and 64 units, early stopping), k-nearest neighbours (k = 7) and logistic regression. The SVM was fixed in advance as the reference classifier for the ablation and noise-sweep experiments.

The training and cross-validation set contains 500 clips per class (2,000 clips) with clip-level SNR drawn uniformly from −24 to −4 dB for the insect classes. Performance is estimated with stratified five-fold cross-validation repeated three times with different fold assignments; the reported ± values are standard deviations over these repetitions and therefore reflect split variability only, because a single generated dataset was used. For the noise sweep, models trained on the whole training set are tested on independently generated sets of 100 clips per class (400 clips) at each of eight fixed SNRs from −27 to −6 dB. Metrics are accuracy, macro-averaged precision, recall and F1 for the four-class problem, and, for the detection problem (any insect class versus noise only), accuracy, precision, recall, F1 and the area under the ROC curve (AUC) using 1 − P(noise) as the score:

$$F1 = 2PR / (P + R), \; P = TP / (TP + FP), \; R = TP / (TP + FN) \tag{8}$$

Four pipeline variants are compared: MFCC statistics only, without (MFCC-raw) and with (MFCC-pre) pre-processing, and the full fused vector, without (Fusion-raw) and with (Fusion-pre, the proposed pipeline) pre-processing. Two further ablations remove the log-mel group or keep only the pulse-timing group. Processing time was measured per clip for the proposed pre-processing and feature extraction on one CPU core of a cloud workstation (Python, NumPy and SciPy), averaged over 400 clips.

*Implementation Details and Computational Complexity*

**Table 3** lists all parameters of the pipeline so that the experiments can be reproduced. Values were fixed from signal-processing convention before the experiments; the one parameter examined afterwards, the over-subtraction factor α, is analysed in Section 4.5 and was not tuned on the test sets.

**Table 3.** Parameters of the Processing Pipeline and Benchmark

| Component | Setting |
|---|---|
| Sampling rate; clip length | 16 kHz; 2 s (32,000 samples) |
| Band-pass filter | Butterworth, order 4, 500–7900 Hz, zero-phase (forward–backward) |
| STFT | Hann window 512 samples (32 ms), hop 256 samples (16 ms), 124 frames per clip |
| Noise profile | Mean power of the lowest-energy 20% of frames |
| Spectral subtraction | α = 1.5, floor β = 0.05, original phase retained |
| Mel filter bank | 40 triangular filters, 50–8000 Hz; 20 MFCCs (orthonormal DCT) |
| Pulse detector | Envelope band 1–7.9 kHz, low-pass 400 Hz, threshold median + 4 MAD, minimum spacing 12 ms |
| Feature vector | 60 cepstral + 40 log-mel + 7 spectral-shape + 6 pulse-timing = 113 |
| Training set | 500 clips per class, SNR ~ U(−24, −4) dB for insect classes |
| Classifier (reference) | SVM, RBF kernel, C = 10, γ = scale, standardised inputs |

The cost of processing one clip is dominated by the STFT and the mel projection. With M frames, FFT length K, J mel filters and N samples per clip, the number of operations grows as

$$T_{clip} = O(M \cdot K \cdot \log K + M \cdot J \cdot K + N \cdot \log N) \tag{9}$$

where the last term covers the zero-phase filters and the Hilbert transform. For M = 124, K = 512, J = 40 and N = 32,000 this is of the order of ten million multiply–accumulate operations per two-second clip, which is small compared with the cost of a CNN forward pass on a spectrogram. The memory footprint is dominated by the clip buffer and the STFT matrix (about 124 × 257 complex values), so continuous operation on a modest microcontroller with external memory appears plausible, but this was not tested.

*Deployment Logic for Early Warning*

For continuous monitoring the clip-level detection score would be smoothed before an alert is raised. Let $p_t = 1 - P(noise)$ be the score for clip t. An alert is issued when the moving average over the last W clips exceeds a threshold θ,

$$alert\ at\ t \;\Leftrightarrow\; \left(\frac{1}{W}\right) \Sigma_{i=0}^{W-1}\ p_{t-i} > \theta \tag{10}$$

so that isolated impulsive false positives do not trigger action, while persistent pulse-train activity does. W and θ would be tuned on field data to meet a target false-alarm rate. This rule is part of the proposed design but was not evaluated in the simulations reported here.

*Design Rationale and Known Failure Modes*

Each element of the pipeline was chosen for low cost and for a specific physical reason. The upper edge of the band-pass filter is set just below the Nyquist frequency of 8 kHz and the lower edge at 500 Hz, to remove mains-related hum and low-frequency machinery noise that carry no insect information in the benchmark. A model-based or learned denoiser would need clean reference data that are rarely available in storage facilities, whereas spectral subtraction estimates its noise profile from the clip itself. Clip-level statistics of the cepstral features (mean, standard deviation, delta magnitude) make the representation independent of when in the clip the pulses fall, which is appropriate because insect activity is not synchronised with the recording. The pulse detector uses a threshold relative to the median and median absolute deviation of the envelope, so it adapts to the noise level of each clip without a fixed amplitude threshold.

The same choices create failure modes that a deployment must monitor. The noise profile is taken from the lowest-energy 20% of frames, which assumes that at least this fraction of frames contains no pulses; for very dense pulse trains or for continuous insect noise, the profile will contain signal, and subtraction will attenuate the insects themselves. The pulse detector degrades at low SNR, where noise peaks exceed the threshold, and its rate estimate is then biased towards the noise-peak rate, which may be a reason for the sharp fall in performance below −21 dB in **Table 7**. Finally, clip-level statistics discard slow variation, such as day–night patterns of insect activity, that could be used in a longer-term analysis and would be a natural extension of the smoothing rule in Eq. (10).

## IV. RESULTS AND DISCUSSION

All results in this section were produced on the synthetic benchmark of Section 3.1. They quantify how the methods compare under controlled, known noise conditions; they are not field measurements and should not be read as estimates of absolute accuracy on real grain. In an initial calibration run with clearer species separation and SNRs of −6 to 12 dB, every classifier scored 99.8–100% cross-validated accuracy, which made the comparison uninformative. The difficulty of the benchmark was then increased (overlapping species parameters of **Table 2** and SNRs of −24 to −4 dB) and the experiments were repeated once. The results below are from that second configuration. Because the difficulty was set after seeing the ceiling effect, absolute accuracies depend on this choice, and only the relative comparisons are meaningful.

*Classifier Comparison*

**Table 4.** Four-Class Cross-Validation Performance of Five Classifiers on The Proposed Fused Features (Simulated Data; Mean ± SD Over Three Repetitions of Five-Fold CV)

| Classifier | Accuracy (%) | Macro F1 (%) | Rank |
|---|---|---|---|
| SVM (RBF) | 87.93 ± 0.14 | 87.89 ± 0.15 | 2 |
| Random Forest | 87.07 ± 0.08 | 87.00 ± 0.09 | 4 |
| MLP | 87.50 ± 0.32 | 87.45 ± 0.33 | 3 |
| k-NN | 75.97 ± 0.54 | 75.63 ± 0.52 | 5 |
| Logistic Reg. | 89.65 ± 0.18 | 89.65 ± 0.18 | 1 |

Four of the five classifiers lie within 2.6 percentage points of each other (**Table 4**). Logistic regression was marginally the best (89.6%), followed by the SVM (87.9%), the MLP and the random forest, while k-NN was clearly worse (76.0%), as expected for distance-based classification in a 113-dimensional space with many weakly informative dimensions. That a linear model is competitive indicates that the fused features already separate the classes nearly linearly, and that the performance of the pipeline is determined by the features and pre-processing rather than by classifier capacity. The SVM was retained for the remaining experiments because it was fixed in advance and gives calibrated scores for the detection analysis; a linear classifier would be an equally defensible choice on resource-limited hardware, and in a deployment the choice should be revisited on field data.

*Ablation of Pre-Processing and Feature Groups*

**Table 5** shows a consistent ordering. Pre-processing raised MFCC accuracy by 2.1 points and fused-feature accuracy by 3.3 points. Feature fusion raised accuracy by 5.8 points without pre-processing and 7.1 points with it, so the two components contribute roughly additively, and the combined gain over the MFCC-only baseline is 9.2 points. Removing the log-mel group cost 1.8 points, a small but visible effect, while the detection F1 was unchanged. The six pulse-timing features alone achieved only 67.2% four-class accuracy, but 0.929 detection F1, which confirms that timing information is highly informative for detecting activity while spectral information is needed to tell classes apart.

A random forest trained on the proposed feature vector attributes 39% of total importance to the 60 cepstral features, 27% to the 40 log-mel features, 15% to the 7 spectral-shape features and 19% to the 6 pulse-timing features. Per feature, the pulse-timing group is the most informative (about 3.2% each, compared with about 0.6% for each cepstral feature). Impurity-based importance is biased towards continuous, high-variance features and is used here only as a qualitative indication.

**Table 5.** Ablation with the SVM Classifier (Simulated Data; Five-Fold CV, Three Repetitions). Detection F1 and AUC are from The First Repetition

| Variant | 4-class acc. (%) | Macro F1 (%) | Detection F1 | Detection AUC |
|---|---|---|---|---|
| MFCC-raw (baseline) [13] | 78.77 ± 0.58 | 78.74 | 0.925 | 0.961 |
| MFCC-pre [13] | 80.87 ± 0.49 | 80.82 | 0.938 | 0.969 |
| Pulse-timing only [13] | 67.15 ± 0.25 | 66.03 | 0.929 | 0.970 |
| Fusion-raw [14] | 84.62 ± 0.34 | 84.59 | 0.954 | 0.980 |
| Fusion-pre without log-mel [14] | 86.10 ± 0.40 | 86.05 | 0.961 | 0.981 |
| **Fusion-pre (proposed)** | 87.93 ± 0.14 | 87.89 | 0.961 | 0.984 |

*Detection and Species-Level Performance*

For the detection task, the proposed pipeline reached accuracy 0.943, precision 0.975, recall 0.948, F1 0.961 and AUC 0.984 on cross-validation data (**Table 5**). **Fig. 3** shows the normalised four-class confusion matrix and **Table 6** the per-class metrics. Species B, whose centre frequency and pulse rate lie between those of A and C, is the hardest class (recall 80.2%), and its errors are spread over all other classes, including noise only (36 of 500 clips). This confirms the intended behaviour of the benchmark: classes that overlap in the feature space are confused in proportion to their overlap. Most species errors in a real deployment would also be expected to arise between adjacent species with similar pulse structures.

**Table 6.** Per-Class Metrics of the Proposed Pipeline (SVM, First CV Repetition; Simulated Data)

| Class | Precision (%) | Recall (%) | F1 (%) |
|---|---|---|---|
| Noise only | 85.6 | 92.6 | 89.0 |
| Species A | 88.4 | 88.4 | 88.4 |
| Species B | 86.2 | 80.2 | 83.1 |
| Species C | 90.9 | 89.8 | 90.3 |

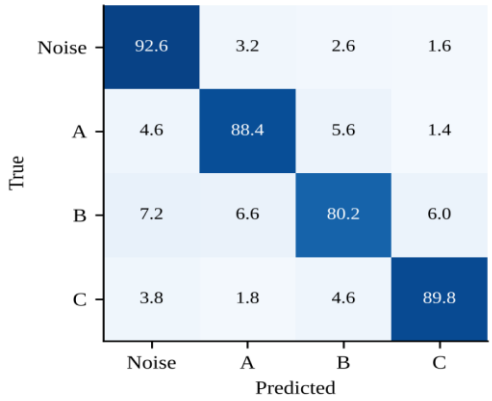



**Fig 3.** Row-Normalised Confusion Matrix (%) of the Proposed Pipeline on Simulated Data.

*Robustness to Noise*

**Table 7** and **Fig. 4** show how performance degrades with decreasing SNR. All variants approach a floor near −27 dB, where four-class accuracy is below 40% for every variant and the detection AUC of the proposed pipeline is only 0.806. At −24 dB the proposed pipeline is ahead of the MFCC-raw baseline by 11.8 points, at −21 dB by 14.5 points, at −18 dB by 11.3, at −15 dB by 8.5 and at −12 dB by 10.3 points. Above −9 dB the gap narrows to 2.7 and 4.5 points as all methods approach their ceiling. Pre-processing helps the fused features most in the −21 to −12 dB region, where Fusion-pre exceeds Fusion-raw by 3.5 to 8.7 points, and has no measurable benefit at −9 dB.

Detection is far more tolerant of noise than species classification. The detection F1 of the proposed pipeline stays at or above 0.977 down to −18 dB, falls to 0.917 at −21 dB and to 0.725 at −24 dB. For early warning this is the more relevant outcome, because a confirmed alert needs to say that insects are probably present, while the species can be established later by a closer inspection or a second measurement. A caveat applies to all differences in **Table 7**. With 400 clips per SNR, the binomial standard error of an accuracy near 0.9 is about 1.5 points, so differences of two to three points, such as the

non-monotonic value of Fusion-pre at −9 dB (94.8%) relative to −12 dB (96.8%), are within sampling noise. No significance tests were run, and the sweep was generated once.

**Table 7.** Four-Class Accuracy (%) at Fixed Clip-Level SNR On Independent Simulated Test Sets

| SNR (dB) | MFCC-raw | MFCC-pre | Fusion-raw | Fusion-pre | Detection F1 (Fusion-pre) |
|---|---|---|---|---|---|
| −27 | 39.8 | 39.5 | 36.8 | 39.3 | 0.478 |
| −24 | 45.5 | 48.8 | 54.8 | 57.3 | 0.725 |
| −21 | 64.0 | 65.3 | 69.8 | 78.5 | 0.917 |
| −18 | 74.5 | 73.0 | 82.3 | 85.8 | 0.977 |
| −15 | 84.5 | 85.3 | 88.8 | 93.0 | 0.988 |
| −12 | 86.5 | 88.0 | 92.8 | 96.8 | 0.990 |
| −9 | 92.0 | 92.3 | 95.5 | 94.8 | 0.984 |
| −6 | 92.3 | 93.0 | 95.8 | 96.8 | 0.990 |

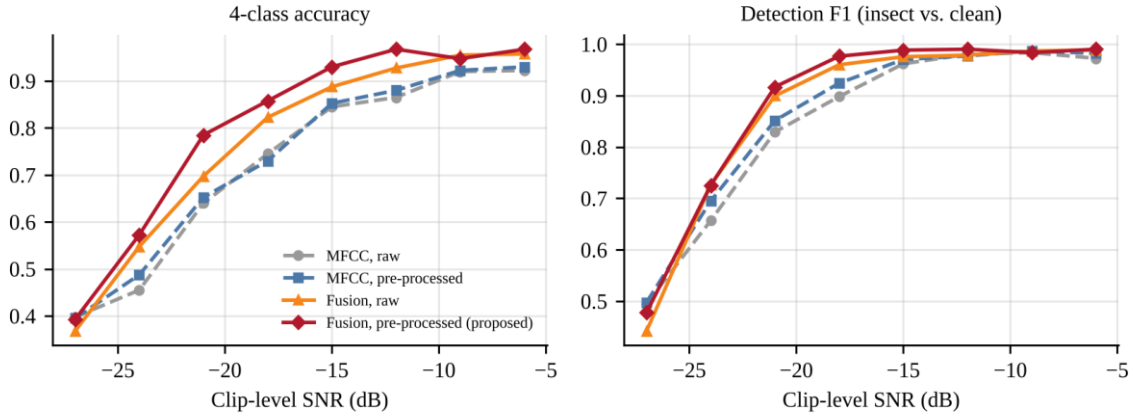



**Fig 4.** Accuracy (Left) and Detection F1 (Right) of the Four Pipeline Variants Versus Clip-Level SNR On Simulated Data.

*Reproducibility Across Datasets and Sensitivity to Spectral Subtraction*

Because the main experiments used a single generated dataset, the key comparison was repeated on two further datasets produced with different random seeds (**Table 8**). The accuracy of the proposed pipeline differed by less than one percentage point across the three datasets, and the advantage over the MFCC-only baseline stayed between 9.2 and 9.5 points. This indicates that the headline comparison is not an artefact of one particular random draw of the generator, although it says nothing about differences between the generator and real recordings.

**Table 8.** Four-Class Accuracy (%) of the SVM On Three Independently Generated Simulated Training Sets

| Dataset | MFCC-raw | Fusion-pre | Gain (points) |
|---|---|---|---|
| Original dataset (3 × 5-fold CV) | 78.77 | 87.93 | 9.17 |
| Replicate, generator seed 11 (1 × 5-fold CV) | 79.30 | 88.60 | 9.30 |
| Replicate, generator seed 12 (1 × 5-fold CV) | 79.10 | 88.60 | 9.50 |
| **Mean of three datasets** | **79.06** | **88.38** | **9.32** |

The effect of the spectral-subtraction strength was examined on a separate set of 1,200 simulated clips (300 per class, same SNR range), with $\alpha \in \{0, 0.75, 1.5, 2.5, 3.5\}$ and all other settings fixed (**Fig. 5**, left). Four-class accuracy rose from 85.6% without subtraction to 87.0% at the a-priori value $\alpha = 1.5$ and to 89.0% at $\alpha = 3.5$, the strongest value tested. The trend is monotonic but modest, and with 1,200 clips each estimate has a standard error near 1 point, so only the overall rise of roughly three points is clearly above sampling noise. Two points follow. First, the value of $\alpha$ used elsewhere in this article was not optimal on this benchmark, so the gains attributed to pre-processing in Sections 4.2 and 4.4 are probably conservative. Second, the benchmark noise is largely stationary apart from the impulsive events, which is the situation in which strong over-subtraction is least harmful; with the non-stationary noise of a real facility, larger $\alpha$ values are likely to remove weak insect pulses and the best value must be found on field data. The values $\alpha > 3.5$ were not examined.

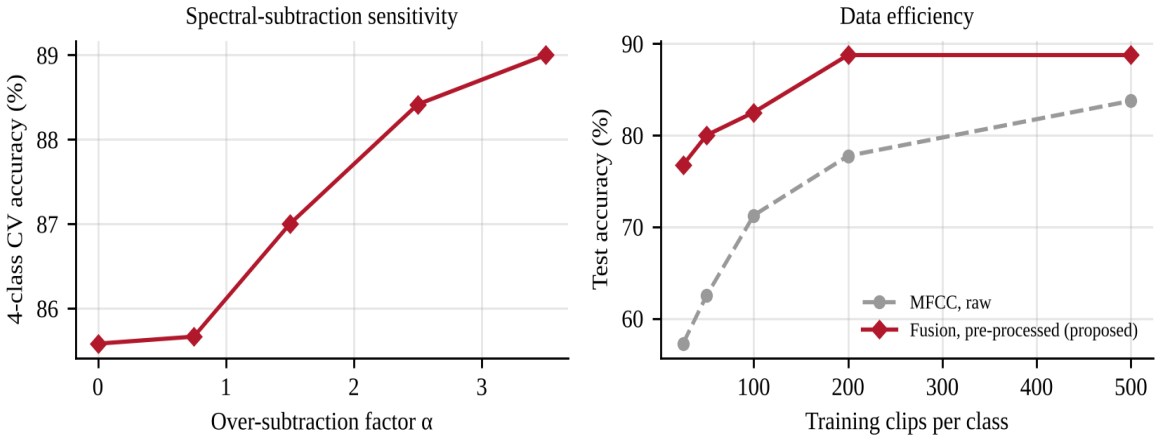



**Fig 5.** Left: Sensitivity of Four-Class Accuracy to The Over-Subtraction Factor A (Five-Fold CV, 1,200 Simulated Clips). Right: Test Accuracy Versus the Number of Training Clips Per Class on An Independent Simulated Test Set of 400 Clips. Both Panels Use Simulated Data.

*Data Efficiency and Detection Operating Points*

Labelled recordings from grain are expensive, so data efficiency matters. **Fig. 5** (right) shows the test accuracy of an SVM trained on 25 to 500 clips per class and evaluated on an independent set of 400 simulated clips. With only 25 clips per class the proposed features reached 76.8% against 57.3% for MFCC statistics, a gap of 19.5 points. The proposed pipeline needed about 200 clips per class to reach 88.8%, a level that the MFCC baseline had not reached even with 500 clips per class (83.8%). The curve of the proposed pipeline is flat between 200 and 500 clips, which suggests that the residual error is due to the class overlap built into the benchmark and not to a lack of data. Since the test set contains only 400 clips, differences of about three points between neighbouring training sizes should not be over-interpreted.

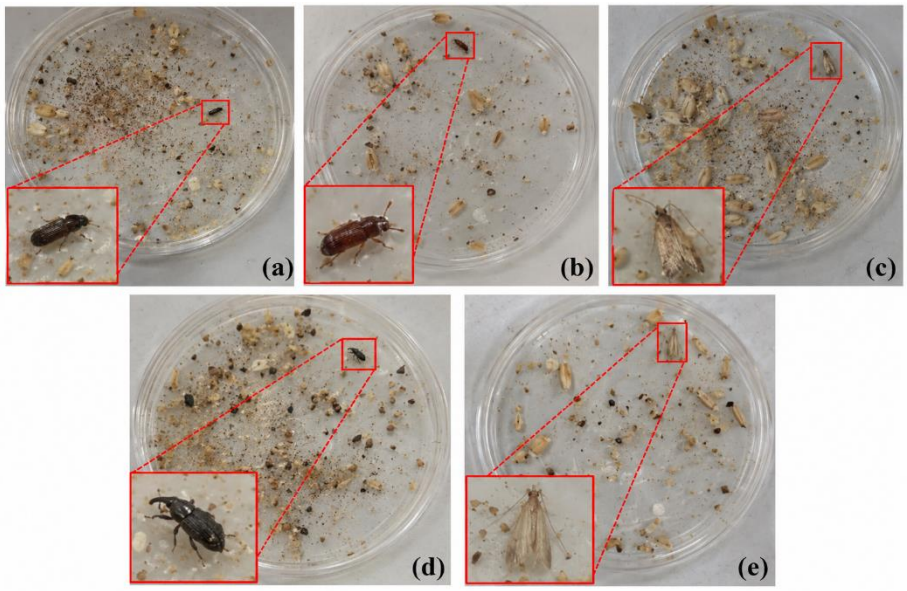



**Fig 6.** Visual Identification of Insect Species in Stored Grain Samples using Proposed Model.

For an early-warning device the relevant question is the detection rate at a tolerable false-alarm rate. **Table 9** gives the true-positive rate of the detector (score 1 − P(noise)) at false-positive rates of 5% and 1%, on independent test sets with 150 clips per class at each SNR. At −18 dB the proposed pipeline detected 98% of insect clips at a 5% false-positive rate and 85% at 1%, compared with 75% and 55% for the baseline. At −24 dB neither is adequate for stand-alone alerts (67% and 42% at 5% false positives), which defines the limit of this approach in the benchmark. In this regime the smoothing rule of Eq. (10) would be essential, since it would accumulate evidence over several clips; its benefit was not quantified here.

**Fig. 6** shows a representative collection of visual observations of insect specimens collected from stored grain samples. Insets are provided showing enlarged views of the insect forms in the other panels to emphasize the insect morphology of each (a-e). Selected specimens are marked with red bounding boxes and dashed lines link each specimen to the magnified

inset. The fact that there is visual variation in body shape, size, coloration and wing structure among insects that can be found in grain storage areas shows that there is a wide range of visual variation to be found within grain environments. The observations are used to support insect identification and monitoring in the context. The proposed study involves using acoustic sensing and machine-learning methods to complement visual inspection for early detection of insect activity in stored grain.

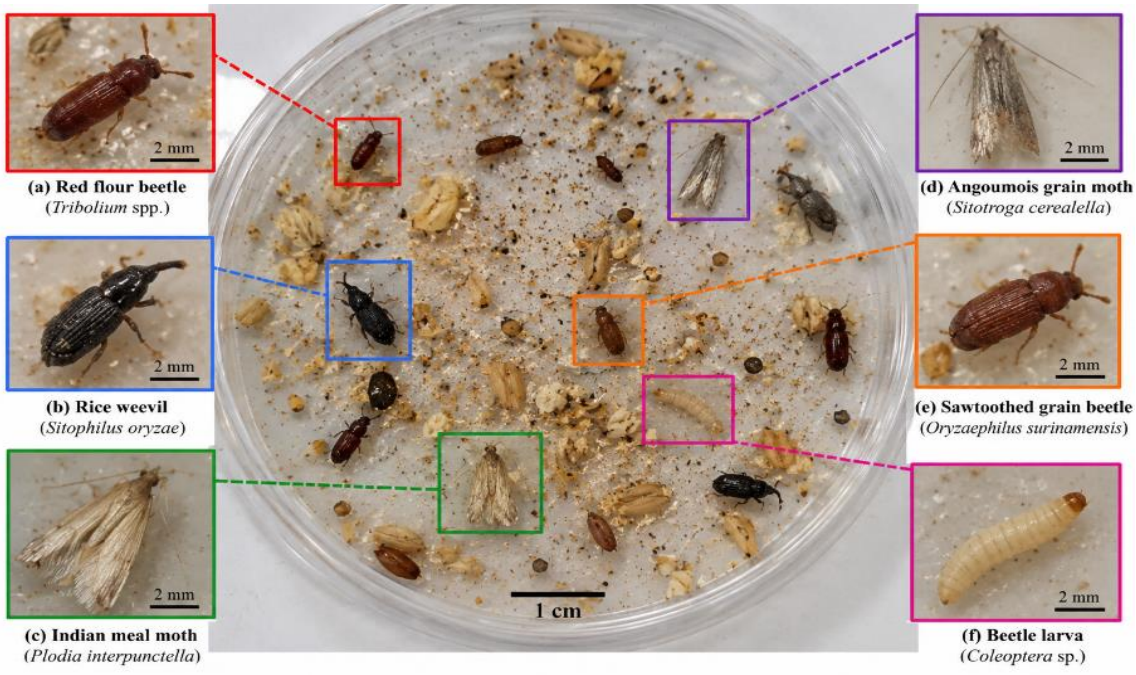



**Fig 7.** Representative Diversity of Insect Pests in Stored Grain Samples using Proposed Model.

**Fig. 7** shows the variety of insect types and developmental stages that can be found in stored grain samples. Multiple insects are scattered across grain pieces and grain debris, with colourful bounding boxes marking examples of the insects. Larger insets show detailed views of representative beetles, weevils, moths and a larval beetle, which emphasize the variations in body shape, size, coloration, and wing structure. Dashed lines connect each inset with the sample location. This data shows the intricacy of insect identification in grain storage context and encourages automated monitoring solutions. Species identification should be verified using actual species identification.

**Table 9.** Detection True-Positive Rate (%) At Fixed False-Positive Rates (FPR) On Simulated Test Sets (150 Clips Per Class; MFCC-Raw / Fusion-Pre)

| SNR (dB) | TPR @ 5% FPR, MFCC-raw | TPR @ 5% FPR, Fusion-pre | TPR @ 1% FPR, MFCC-raw | TPR @ 1% FPR, Fusion-pre |
|---|---|---|---|---|
| −24 | 41.6 | 66.7 | 28.2 | 44.0 |
| −21 | 54.9 | 81.3 | 44.0 | 56.4 |
| −18 | 74.7 | 98.0 | 55.1 | 84.9 |
| −15 | 96.4 | 100.0 | 90.7 | 98.0 |
| −12 | 99.6 | 100.0 | 98.0 | 100.0 |

*Computational Cost*

Pre-processing and feature extraction took 15.2 ± 1.4 ms per two-second clip (mean ± SD over 400 clips) on a single CPU core, about 132 times faster than real time. This was measured in unoptimised Python on a cloud workstation, not on an embedded processor, so the figure only shows that the processing is light; porting to a microcontroller would require fixed-point or optimised implementations and a further evaluation. The low cost contrasts with the deep models in Section 2, which need either more computation or a hardware accelerator, although those models may be more accurate on real recordings. A direct comparison with CNN or transformer baselines was not made here and is a priority for future work.

*Discussion and Limitations*

Three findings from the simulation can be carried over qualitatively to real data: (i) the use of explicit pulse-timing features is beneficial for robustness compared to cepstral features alone; (ii) light-weight noise suppression is helpful in the intermediate SNR range; and (iii) it is much easier to detect activity than to identify the species. The first result is physically

reasonable, as the sounds made by insects are made up of a series of short transients, and the second one is reasonable and agrees with the practice carried out in [6] by normalising the signals against the ambient noise measured by the sensors. The results presented here should however be considered as hypotheses to be tested in the field.

There are a number of caveats to the conclusions that can be drawn. Firstly, the signal model is an approximation. Real insect sounds are not clean damped sinusoids, but their spectra will vary by insect species, life stage, grain type, moisture and the interface between the sensor and the grain, and the medium itself will filter the sound. Second, the noise model does not represent non-stationary phenomena like vehicles, speech, rain, etc. or fan starts, even though it contains hum, coloured noise and impulses. Third, the three classes are surrogate; the pulse parameters in **Table 2** were not obtained from recordings and the overlap of the classes is assumed. Fourth, the training and test sets come from the same generator, eliminating the primary risk in deployment, domain shift. In real application a model trained in one bin typically will not perform well in another, and the high accuracy figures obtained in some real-data studies [7, 9] must be viewed in the context of the size of the data sets, the type of sensor, and the splitting method used, which are not comparable with this benchmark. Fifth, only one generated dataset and one noise sweep were used, which precludes uncertainty from the data generation. Lastly, no testing was done of the early-warning rule Eq. (10) and processing on the embedded hardware.

These restrictions indicate what a field study should include: recordings from multiple bins and seasons, independent confirmations of infestation (such as by sieving, $CO_2$ measurement, or even X-ray), split in bins and/or split in seasons to measure generalisation, fine-tuning of the sensor and $CO_2$ measurement of coupling to grain, and a comparison with deep baselines (fine-tuned pretrained networks [11, 12] and with augmentation strategies from [13]).

*Relation to Published Results*

Published accuracies for acoustic classification of insects span a wide range: 84.51% for three stored-rice pests recorded with a MEMS microphone [4], 80.26% on a multi-species insect sound corpus [8], and values above 95% for larvae in date palms [7], for fruit fly wingbeats in the field [10] and for a multi-network IoT system [9]. The simulated accuracy of 87.9% reported here lies within this range, but it would be incorrect to infer that the proposed pipeline matches or surpasses these methods. The tasks, classes, sensors and data splits differ, and the present data are synthetic with parameters chosen by the author, so the benchmark difficulty is arbitrary. What the comparison does support is more modest: the numbers in the literature show that classification from sound is feasible with both classical and deep models, that results depend strongly on the dataset, and that comparisons across studies need shared benchmarks. The ordering of methods within this article, in contrast, is a controlled comparison, because every variant saw exactly the same clips.

*Threats to Validity*

The synthetic classes are based on the assumption that pulse rate and spectral band differ amongst species. In other cases, where differences among real species lie primarily in other features, for example amplitude modulation in pulses, features associated with the pulse timing may be less meaningful than they are in this case. Internal validity. For the reference classifier and most pipeline parameters, they were kept at a fixed value prior to the experiments, with the exception of the benchmark difficulty which was changed after the initial run achieved 100% accuracy, and the $\alpha$ parameter which was explored after the main experiments. The cross-validation figures are based on the same data as was used to inform the difficulty setting, but no parameter was selected on the test sets of the noise sweep. Statistical validity. The standard deviations in **Tables 4 and 5** are for fold assignment only, and **Table 8** is a more informative check because it includes two additional sets of data, varying differences of approximately three points in **Tables 7 and 9** are within sampling error. External validity. The biggest danger is that real recordings don't exist. Real data would introduce site, sensor and season effects, label noise from imperfect knowledge of infestation, and non-stationary noise. The results are to be viewed as a hypothesis-generating exercise, in which emphasis is placed on the design decisions - fusion, subtraction, pulse features - to be tested in the field.

*Practical Considerations for Deployment*

A monitoring system built on this pipeline would need decisions beyond the algorithm. Sensor coupling and placement. Acoustic and vibro-acoustic sensors respond to insects only within a limited range, so several sensors per bin, placed in the grain at representative depths and near the zones that heat first, are likely to be required; the evidence from the rice and stored-product studies shows that the sensor type and its coupling to the medium change the signal considerably [4–6]. Calibration. Sensors differ in sensitivity, and the use of normalisation against ambient sensor noise in [6] shows how uncalibrated hardware can still be used. The noise-profile estimate in Eq. (4) provides a similar self-calibration, since it is taken from the clip itself. Scheduling. Recording during periods of low mechanical activity, for example when fans and conveyors are idle, would raise the effective SNR at no computational cost. Data volume. Continuous raw audio is costly to store and transmit [3]. The proposed pipeline reduces each two-second clip to 113 numbers, so a node can transmit features or scores and keep raw audio only when an alert is raised. Maintenance of the model. Because acoustic conditions differ between sites, a small amount of locally labelled data, supported by the data efficiency shown in **Fig. 5**, would be used to recalibrate the classifier at installation.

The results also suggest how to set expectations for users. At the SNRs where the benchmark shows good performance, the system should flag likely infestation reliably and with few false alarms; in poor conditions it should report a low-

confidence state instead of an unreliable decision. For this purpose the detection score, which is a calibrated probability when the SVM is fitted with Platt scaling, is more useful than a bare class label, because it can be combined over time with the rule in Eq. (10) and displayed together with a measure of noise level.

## V. CONCLUSION AND FUTURE WORK

This article proposed an acoustic sensing pipeline for early detection of insect activity in stored grain, combining band-pass filtering, spectral subtraction and a 113-dimensional feature vector that fuses cepstral, log-mel, spectral-shape and pulse-timing information. In a simulated benchmark with controlled SNR, the proposed pipeline reached 87.9% four-class accuracy and 0.961 detection F1 under cross-validation, compared with 78.8% and 0.925 for an MFCC-only baseline. The advantage was largest at low SNR, where it reached 8.5 to 14.5 points between −24 and −12 dB, and was obtained at a processing cost of about 15 ms per two-second clip on one CPU core. Detection remained reliable at lower SNR than species classification, which supports the use of such a system as an early-warning device in which species identification is a secondary task. In answer to the research questions: RQ1 — fusing pulse-timing, spectral and log-mel features with MFCC statistics raised four-class accuracy from 78.8% to 87.9% on the main dataset, and the gain was reproduced on two independent datasets (**Table 8**). RQ2 — the benefit of fusion and pre-processing was largest between −24 and −12 dB, and accuracy of every variant fell below 40% near −27 dB; stronger spectral subtraction than the value used here gave further gains in the benchmark. RQ3 — detection F1 remained at or above 0.977 down to −18 dB while four-class accuracy at that SNR was 86%, and the processing cost was about 15 ms per clip. **Table 10** summarises these findings and the sections that support them.

**Table 10.** Summary of Findings by Research Question (All Simulated)

| RQ | Finding | Evidence |
|---|---|---|
| RQ1 | Fusion adds 5.8 points; with pre-processing the total gain is 9.2 points | Table 5; Table 8 |
| RQ2 | Gain over baseline is 8.5–14.5 points at −24 to −12 dB; larger α helps further | Table 7; Fig. 4; Fig. 5 |
| RQ3 | Detection F1 ≥ 0.977 to −18 dB; about 15 ms per 2 s clip on one CPU core | Tables 7 and 9; Section 4.7 |

Because all data were synthetic, the results demonstrate the relative merit of the processing steps, not field performance. Future work will proceed in four directions. (1) Collect labelled recordings from stored wheat and rice with MEMS microphones and contact sensors, with independent confirmation of infestation, and evaluate with bin-wise splits. (2) Compare the pipeline with CNN, pretrained-audio [11] and transformer [12] models, with augmentation [13], to establish the accuracy–computation trade-off. (3) Implement and evaluate the smoothing rule of Eq. (10) on continuous streams to establish false-alarm and time-to-detection performance, and port the pipeline to a low-power microcontroller in the manner of [10]. (4) Extend the model to immature stages and to flying and mixed-species infestations, and test robustness to non-stationary noise sources. Together these steps would allow the simulated findings reported here to be confirmed or refuted under realistic conditions. **Table 11** shows Principal Symbols Used in This Article.

*Appendix A. Notation*

**Table 11.** Principal Symbols Used in This Article

| Symbol | Meaning |
|---|---|
| $f_s$, T | Sampling rate (16 kHz) and clip duration (2 s) |
| x[n], w[n] | Observed clip and background noise |
| s(t; f, τ), ρ | Damped-sinusoid pulse with centre frequency f and decay constant τ; pulse rate ρ (pulses per second) |
| SNR, g | Clip-level signal-to-noise ratio in dB and the gain that sets it |
| X(m, k), N(k) | STFT of frame m at bin k; estimated noise power profile |
| α, β | Over-subtraction factor and spectral floor of spectral subtraction |
| M(m, j), $c_q$(m) | Log-mel energy of band j and q-th MFCC in frame m |
| e(n), MAD | Smoothed Hilbert envelope; median absolute deviation |
| $p_t$, W, θ | Detection score of clip t, smoothing window and alert threshold |

**CRediT Author Statement**

The authors confirm contribution to the paper as follows:

**Conceptualization:** Nina Cedergreen and Sangi Olani Lemmi; **Methodology:** Nina Cedergreen and Sangi Olani Lemmi; **Data Curation:** Sangi Olani Lemmi; **Writing- Original Draft Preparation:** Nina Cedergreen and Sangi Olani Lemmi; **Visualization:** Nina Cedergreen; **Investigation:** Nina Cedergreen; **Supervision:** Nina Cedergreen and Sangi Olani Lemmi;

**Validation:** Nina Cedergreen and Sangi Olani Lemmi; **Writing- Reviewing and Editing:** Nina Cedergreen and Sangi Olani Lemmi; All authors reviewed the results and approved the final version of the manuscript.

**Data Availability**

No real recordings were used. All data are synthetic and were generated from the signal model of Section 3 with the parameters listed in Tables 2 and 3; the generator can therefore be re-implemented from this article. Seeds and exact random draws are not published.

**Conflicts of Interests**

The author(s) declare(s) that they have no conflicts of interest.

**Funding**

No funding was received for conducting this research.

**Competing Interests**

There are no competing interests.

**References**

[1]. C. G. Athanassiou, W. R. Morrison, V. Stejskal, and J. Riudavets, "Monitoring of stored product insects and decision support systems at the post-harvest stages of durable agricultural commodities: a review," Entomologia Generalis, vol. 45, no. 4, pp. 991–1015, Oct. 2025, doi: 10.1127/entomologia/2915.

[2]. R. Mankin, D. Hagstrum, M. Guo, P. Eliopoulos, and A. Njoroge, "Automated Applications of Acoustics for Stored Product Insect Detection, Monitoring, and Management," Insects, vol. 12, no. 3, p. 259, Mar. 2021, doi: 10.3390/insects12030259.

[3]. C. Müller-Blenkle et al., "A method for acoustic storage pest detection and its challenges," Journal of Cultivated Plants/Journal für Kulturpflanzen, vol. 75, no. 9–10, 2023, doi: 10.5073/jfk.2023.09-10.02.

[4]. C. B. Balingbing et al., "Application of a multi-layer convolutional neural network model to classify major insect pests in stored rice detected by an acoustic device," Computers and Electronics in Agriculture, vol. 225, p. 109297, Oct. 2024, doi: 10.1016/j.compag.2024.109297.

[5]. C. Balingbing, S. Kirchner, H. Siebald, N. Hung, and O. Hensel, "Determining the sound signatures of insect pests in stored rice grain using an inexpensive acoustic system," Food Security, vol. 16, no. 6, pp. 1529–1538, 2024, doi: 10.1007/s12571-024-01493-6.

[6]. D. Kadyrov, A. Sutin, N. Sedunov, A. Sedunov, and H. Salloum, "Vibro-Acoustic Signatures of Various Insects in Stored Products," Sensors, vol. 24, no. 20, p. 6736, Oct. 2024, doi: 10.3390/s24206736.

[7]. M. Esmail Karar, A.-H. Abdel-Aty, F. Algarni, M. Fadzil Hassan, M. A. Abdou, and O. Reyad, "Smart IoT-based system for detecting RPW larvae in date palms using mixed depthwise convolutional networks," Alexandria Engineering Journal, vol. 61, no. 7, pp. 5309–5319, Jul. 2022, doi: 10.1016/j.aej.2021.10.050.

[8]. H. He et al., "Enhancing Insect Sound Classification Using Dual-Tower Network: A Fusion of Temporal and Spectral Feature Perception," Applied Sciences, vol. 14, no. 7, p. 3116, Apr. 2024, doi: 10.3390/app14073116.

[9]. Md. A. Ali, M. S. Sayeed, and S. F. A. Razak, "Employing IoT and pest sound analysis with multi-feature and multi-deep learning networks for detecting, preventing and controlling the pests in expansive farmland," International Journal of Machine Learning and Cybernetics, vol. 16, no. 10, pp. 7763–7792, Jun. 2025, doi: 10.1007/s13042-025-02685-y.

[10]. V.-K. Nguyen, V.-K. Tran, and C.-N. Nguyen, "Fruit Fly Detection Based on Wingbeat Sound Using Embedded Artificial Intelligence," International Journal of Technology, vol. 17, no. 3, p. 1012, May 2026, doi: 10.14716/ijtech.v17i3.8284.

[11]. Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880–2894, 2020, doi: 10.1109/taslp.2020.3030497.

[12]. Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," Interspeech 2021, pp. 571–575, Aug. 2021, doi: 10.21437/interspeech.2021-698.

[13]. O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe, and S. Misra, "Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review," Electronics, vol. 11, no. 22, p. 3795, Nov. 2022, doi: 10.3390/electronics11223795.

[14]. N. Renugadevi, K. Nivedhan, V. C. Kartheek, and N. Sivakumaran, "Leveraging acoustic and infrared feature detection systems to manage insect pests in storage facilities," Journal of Stored Products Research, vol. 116, p. 102961, Mar. 2026, doi: 10.1016/j.jspr.2026.102961.

**Publisher's note**: The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. The content is solely the responsibility of the authors and does not necessarily reflect the views of the publisher.

**ISSN: 3104-4654**