# ACTIVITY-AWARE DENOISING FOR WEAKLY LABELLED INSECT SOUND CLASSIFICATION

Florian Krebs[1], Hannes Bradl[1], Stefan Grebien[1], and Franz Graf[1]

[1] *Joanneum Research, Graz, Austria*





## Abstract

Large-scale bioacoustic datasets are typically weakly labelled, assigning species labels to entire recordings without temporal annotations. When recordings are divided into short audio chunks for training, many chunks inherit the recording-level label despite containing no target species, introducing weak-label noise. In this work, we quantify this source of noise in the InsectSet459 dataset and investigate activity-aware denoising strategies for insect sound classification. We evaluate several approaches for incorporating activity information, including activity-based loss weighting, background class modelling, activity-aware pooling, and Multiple Instance Learning (MIL). Our results suggest that explicitly modelling inactive chunks consistently improves classification performance, with a macro F1-score improvement of 3.4 percentage points over the baseline. The findings demonstrate that inactive audio segments represent a measurable source of weak-label noise and that explicitly modelling them provides a simple and effective way to improve large-scale insect sound classification.

**Keywords:** *weak label noise, insect classification, multiple instance learning*

## 1. Introduction

Passive acoustic monitoring (PAM) has become an essential tool for large-scale biodiversity monitoring, enabling continuous ecosystem observation without disturbing wildlife [1]. Recent advances in deep learning and the availability of larger annotated datasets have substantially improved automatic species classification for birds [2], amphibians, and more recently insects [3], [4].

A major limitation of current bioacoustic datasets is that they are typically *weakly labelled*: recordings are annotated at the recording level without temporal information indicating when a species is vocalizing. During training, recordings are commonly divided into short audio chunks, each inheriting the recording-level label. As a result, many training samples contain only background sounds, silence, or other species while still receiving the target species label, introducing weak-label noise.

Weakly supervised learning has therefore received considerable attention in recent years. Multiple Instance Learning (MIL) has become one of the dominant approaches by allowing models to focus on the most informative audio segments instead of treating all segments equally [5], [6]. Similar ideas have been successfully applied to sound event detection using attention mechanisms [7], [8], alternative pooling strategies [9], or pseudo-label refinement [10]. While these methods improve learning from weakly labelled recordings, the extent to which inactive audio segments affect classification performance in large-scale insect sound classification datasets remains unclear.

In this work, we present a systematic analysis of activity-induced weak-label noise in large-scale insect sound classification. We use the recently introduced *InsectSet459* dataset [3], the first large-scale benchmark for insect sound classification, as a case study due to its recording-level annotations and large scale. We evaluate several activity-aware denoising strategies and analyse how the accuracy of activity estimation affects downstream species classification performance. The main contributions of this work are:

- We provide a quantitative analysis of activity-induced weak-label noise in the *InsectSet459* dataset.
- We compare several activity-aware denoising strategies under a common experimental framework, including loss weighting, background class modelling, activity-aware pooling, and MIL.
- We analyse how insect activity and activity estimation accuracy influence the effectiveness of

***Corresponding author:*** *florian.krebs@joanneum.at.*

activity-aware denoising.

## 2. Dataset

The *InsectSet459* dataset [3] contains 26,399 recordings (approximately 227 hours) from 459 species (310 Orthoptera and 149 Cicadidae), collected from xeno-canto, iNaturalist, and BioAcoustica. Recordings are annotated with a single species label indicating the dominant insect species present in the recording.

Following the original benchmark protocol [3], each recording is divided into overlapping 5 s audio chunks (2.5 s overlap) for training. To quantify weak-label noise and to train an insect activity detector, we manually annotated two subsets of the training data with chunk-level activity labels. Each chunk was labelled as either containing insect activity or not, irrespective of the insect species.

The first subset, *ActivityTest*, consists of 100 randomly selected recordings comprising 1,032 chunks and serves exclusively for evaluating the activity detector and estimating the prevalence of inactive chunks in the dataset. The second subset, *ActivityTrain*, contains 158 recordings (1,600 chunks) and is used to train the activity detector. To ensure sufficient representation of inactive audio segments, *ActivityTrain* was constructed in two stages. First, 50 recordings were selected uniformly at random. A preliminary activity classifier was then trained on these annotations and applied to the remaining recordings. Based on its predictions, additional 108 recordings were selected to cover the full range of predicted activity levels while preferentially including recordings with many low-activity chunks. As a result, *ActivityTest* contains 15.4% insect-free chunks, whereas the enriched *ActivityTrain* set contains 44.6%.

## 3. Methods

### 3.1. Overall Pipeline

Figure 1 summarizes the proposed workflow. An insect activity detector first estimates the activity of each 5 s chunk. These activity scores are subsequently incorporated into species classification using four different strategies: soft loss weighting, background class modelling, activity-aware top-$k$ pooling, and Multiple Instance Learning (MIL).

### 3.2. Insect Activity Estimation

To identify audio chunks containing insect vocalizations, we train a binary insect activity classifier on the manually annotated *ActivityTrain* dataset and evaluate it on the independent *ActivityTest* dataset.

Each chunk is represented by four spectral features. Specifically, we extract the (i) mean spectral flux, which measures temporal changes in spectral energy, the (ii) mean spectral flatness, which distinguishes tonal insect calls from broadband background noise, and two activity features that quantify the proportion of time–frequency bins exceeding an estimated noise floor. One activity feature uses a (iii) recording-level noise estimate (relative activity), while the other uses a (iv) chunk-specific noise estimate (absolute activity). These features are computed from STFT power spectrograms using frequencies above 2 kHz. A Random Forest classifier with 100 trees predicts an activity score corresponding to the probability that a chunk contains insect activity. This activity score is subsequently used by all proposed denoising methods.

### 3.3. Baseline Species Classifier

As baseline, we use the original InsectEffNet model proposed by Faiß et al. [3] without architectural modifications. The model employs an EfficientNetV2-S backbone initialized with ImageNet-21k weights and operates on 128-band mel spectrograms computed from audio sampled at 44.1 kHz.

During inference, chunk-level predictions are aggregated into a recording-level prediction via max pooling.

### 3.4. Activity-based Weak-label Denoising

#### *3.4.1. Top-k Pooling*

To incorporate activity information during inference, we replace max pooling with activity-aware top-$k$ pooling: Chunks predicted as background (activity score $< 0.5$) are discarded before aggregation. If no chunk satisfies this criterion, all chunks are retained. For each species, the final recording-level score is obtained by averaging the $k = 5$ highest-confidence chunk predictions. In case of background class weighting (Section 3.4.3), the output of the background class is used for determining the background.

#### *3.4.2. Soft Loss Weighting*

In this approach, the estimated insect activity is used to reduce the influence of potentially mislabelled training chunks. Specifically, the cross-entropy loss of each chunk is weighted by its predicted activity score:

$$\mathcal{L} = -w \sum_{s} Y_s \log(P_s), \tag{1}$$

where $w$ is the activity score estimated by the activity detector, $Y_s$ is the ground-truth label, and $P_s$ is the predicted probability for species $s$. Consequently, chunks with low predicted insect activity contribute less to the optimization, while highly active chunks receive a larger weight.

#### *3.4.3. Background Class Modeling*

To explicitly model inactive segments, we introduce an additional *background class* corresponding to “no insect activity” and a probabilistic labelling strategy which maps insect activity scores to a soft target vector $\tilde{Y}$ for each audio chunk. For a chunk associated with a ground-truth species $S$ and an activity score $a \in [0, 1]$, the target label $\tilde{Y}_s$ is

$$\tilde{Y}_s = \begin{cases} a, & \text{if } s = S, \\ 1 - a, & \text{if } s = b, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

.

During training, the model calculates the Cross-

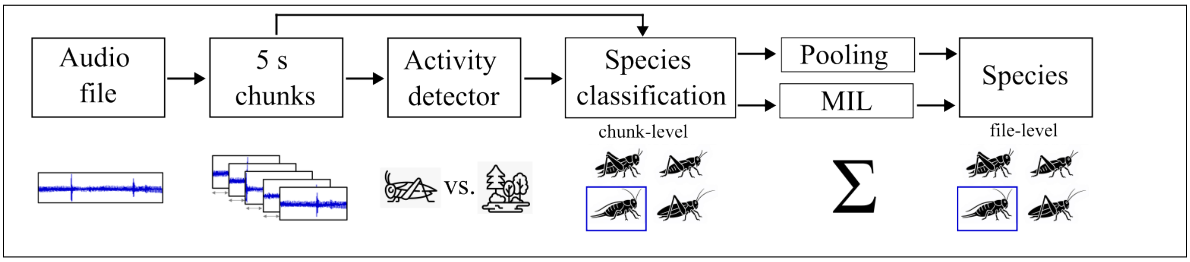


**Figure 1:** Overview of the proposed classification pipeline.

Entropy loss between its predicted log-probabilities and these soft targets:

$$\mathcal{L} = -\sum_{s} \tilde{Y}_s \cdot \log(P_s) \tag{3}$$

This approach allows the network to treat chunks with low insect activity primarily as background noise.

*3.4.4. Multiple Instance Learning*

We further evaluate an attention-based Multiple Instance Learning (MIL) model that aggregates chunk-level embeddings $h_i$ into a recording-level representation using learnable attention weights $w_i$,

$$H = \sum_{i=1}^{N} w_i h_i, \tag{4}$$

where $i$ is the chunk index. We use an attention layer consisting of a hidden layer with 128 units, tanh activation and a final output unit to compute chunk-level weights. We compare two initialization strategies: (i) training the MIL model from ImageNet-pretrained weights and (ii) fine-tuning from the proposed background-class model after removing the background output. In both cases, the attention module is first optimized while the CNN backbone is frozen before end-to-end fine-tuning of the complete network.

### 3.5. Experimental Setup

All experiments are implemented in PyTorch Lightning using the publicly available InsectEffNet implementation. The same data augmentation strategy and training hyperparameters as in [3] are employed unless explicitly stated otherwise.

Each experiment is repeated using three independent random seeds to account for stochastic variation during optimization. The original dataset split (60% training, 20% validation, and 20% test) is used throughout all experiments. Performance is evaluated on the test split using the macro-averaged F1-score at the recording level, which is well suited to the highly imbalanced class distribution of the dataset.

## 4. Results and Discussion

### 4.1. Quantifying weak-label noise

We first evaluate the insect activity detector, as all subsequent denoising strategies rely on it. On the manually annotated *ActivityTest* dataset, the detector achieves an F1-score of 96.1% and an accuracy of 93.4%, demonstrating that insect activity can be identified reliably.

Applying the detector to the complete InsectSet459 dataset using a threshold of 0.5 indicates that approximately 12.8% of all 5 s audio chunks contain no detectable insect activity. This estimate closely matches the 15.4% of inactive chunks observed in the manually annotated random subset, suggesting that the true proportion of inactive chunks lies between 12.8% and 15.4%.

### 4.2. Activity-based denoising

We next investigate whether exploiting the estimated insect activity improves downstream species classification. Table 1 compares all investigated denoising strategies with the standard weak-label baseline and reports the macro F1-score on the test set, averaged over three random seeds.

| Method | Macro F1 (%) |
|---|---|
| Baseline (max pooling) | 60.7 (0.5) |
| Baseline (top-k pooling) | 61.9 (0.2) |
| Soft loss weighting | 60.0 (0.3) |
| Background class | 63.2 (1.4) |
| MIL (attention pooling) | 62.3 (1.1) |
| Background class + top-k pooling | 63.9 (1.2) |
| Background class + mil finetuning | 64.1 (0.9) |

**Table 1:** Mean and standard deviation of the overall test classification performance (macro F1) over three seeds.

With the exception of soft loss weighting, all activity-aware approaches outperform the weak-label baseline. The largest improvement is obtained by combining the background-class formulation with either MIL fine-tuning or top-k pooling, increasing the macro F1-score by 3.4 and 3.2 percentage points respectively. This gain is substantially larger than the observed variation across random seeds, indicating that the improvement is robust.

The results further reveal clear differences between the investigated denoising strategies. Simply reducing the contribution of low-activity chunks during optimization (soft loss weighting) provides no benefit. In contrast, explicitly modelling inactive chunks as a separate background class consistently improves performance, indicating that an explicit representation of background is more effective than merely down-weighting uncertain samples. Likewise, MIL allows

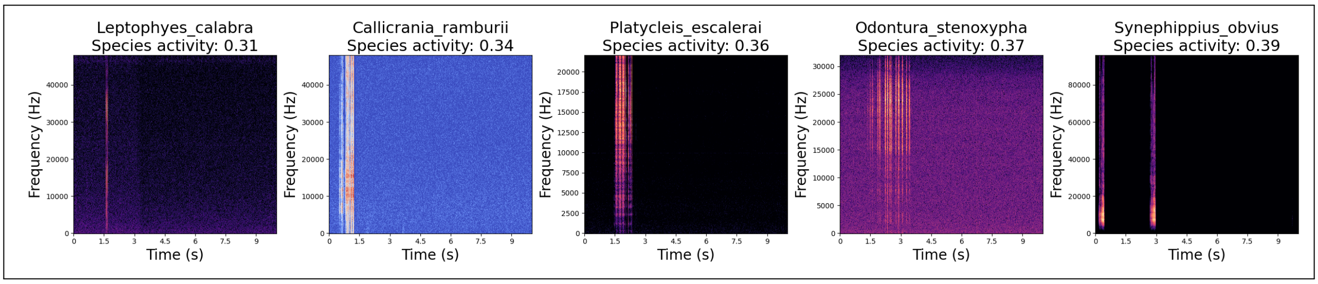


**Figure 2:** Example spectrograms (10 seconds) of the five species with the lowest average activity.

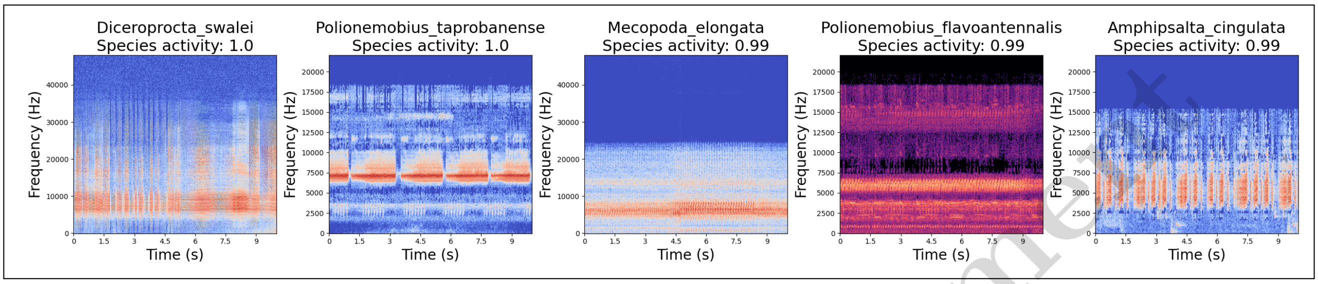


**Figure 3:** Example spectrograms (10 seconds) of the five species with the highest average activity.

the network to focus on informative chunks within a recording and partially compensates for weak-label noise without requiring explicit chunk-level annotations.

Overall, these results demonstrate that inactive audio segments negatively affect species classification and that explicitly modelling or selecting informative chunks provides an effective way to reduce this source of weak-label noise. Importantly, the proposed approach achieves this improvement without requiring additional species annotations (assuming the insect activity detector is available), making it practical for large weakly labelled bioacoustic datasets.

### 4.3. Insect activity per species

The amount of weak-label noise is expected to depend on the calling behaviour of individual species. While some insects produce nearly continuous acoustic signals, others vocalize only occasionally, resulting in many inactive audio chunks.

Figure 2 and 3 illustrate representative examples of the five species with the highest and lowest predicted activity. Highly active species produce almost continuous acoustic signals throughout the recording, whereas low-activity species emit only short calls separated by long silent intervals. Consequently, recordings of low-activity species contain substantially more inactive training chunks and therefore exhibit a higher degree of weak-label noise.

These observations suggest that the effectiveness of activity-based denoising is unlikely to be uniform across species. We therefore investigate this relationship quantitatively in the following section.

### 4.4. Relation between insect activity and denoising effectiveness

To better understand which species benefit from activity-based denoising, we analyze the relationship between the average insect activity of each species and the corresponding improvement in classification performance.

Across all species, the improvement in macro F1-score exhibits a weak negative correlation with average insect activity (Pearson correlation coefficient $r = -0.19$), indicating that species with lower activity generally benefit more from denoising. However, the weak correlation also suggests that insect activity alone cannot fully explain the observed improvements. To further investigate this relationship, species are grouped according to their average activity, and the mean improvement in F1-score is computed for each activity bin (Fig. 4).

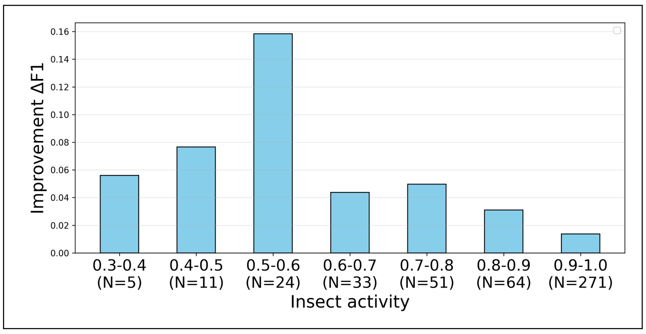


**Figure 4:** Mean species F1 improvement over mean insect activity. $N$ is the number of species in each bin.

The grouped analysis reveals a clear nonlinear relationship between insect activity and the effectiveness of denoising. Species with very high activity ($> 0.9$) benefit only marginally, with an average improvement of +1.4 percentage points. This is expected because their recordings already contain insect sounds in most audio chunks and therefore contain relatively little weak-label noise.

Interestingly, the largest improvements are observed for species with intermediate activity levels between 0.5 and 0.6, where denoising increases the F1-score by an average of +15.8 percentage points. This substantially exceeds the improvement observed for species with even lower activity, for example +5.6 percentage points for species with activity between 0.3 and 0.4.

A possible explanation is that species with intermediate activity provide the best balance between down-weighting mislabeled inactive chunks and retaining sufficient informative training data. In contrast, species with very low activity contain only few informative chunks, making them more sensitive to activity detection errors that incorrectly classify active chunks as background.

Overall, these findings show that insect activity is an important predictor of denoising effectiveness, although its influence is clearly nonlinear and may interact with other dataset characteristics.

### 4.5. Impact of activity classification performance

Finally, we investigate how the accuracy of the insect activity detector influences downstream species classification. To this end, we progressively reduce the detector performance by randomly flipping an increasing proportion of the activity predictions. Although this procedure does not reproduce the structure of real classification errors, it provides a controlled way to assess the sensitivity of the denoising framework to decreasing activity estimation accuracy.

Figure 5 shows the resulting species classification performance.

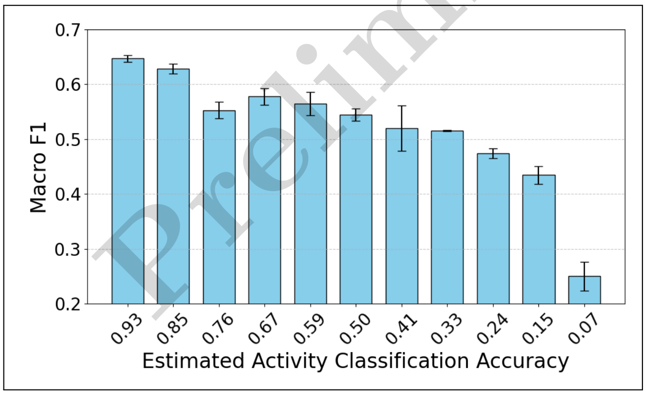


**Figure 5:** Influence of activity detector accuracy on downstream species classification performance. All but the first accuracy values of the x-axis are estimated from flip probabilities.

As expected, more accurate activity estimates consistently lead to better species classification performance. Consequently, the effectiveness of activity-based denoising depends not only on the chosen denoising strategy but also on the quality of the underlying activity detector.

However, the observed improvement is relatively modest. Extrapolating the approximately linear trend observed in Fig. 5 suggests that increasing the activity detector performance from 93.4% to an F1-score of 99.0% would improve the downstream species classification performance by only about 1.3 percentage points, corresponding to an estimated macro F1-score of approximately 65.4%.

To put this result into perspective, we additionally trained both the baseline and the proposed denoising approach using an enlarged training set obtained by incorporating the original validation split. While this experiment was not the primary focus of this work, it increased the macro F1-score by 5.1 percentage points for the baseline and by 5.7 percentage points for the proposed denoising approach. Consequently, the advantage of activity-based denoising increased from 3.4 to 4.0 percentage points. Although this observation is based on only a single increase in training data, it suggests that activity-based denoising may become even more effective as larger weakly labelled bioacoustic datasets become available, since removing noisy samples is less likely to reduce the amount of informative training data.

Overall, these findings indicate that the current activity detector is already sufficiently accurate for the proposed denoising framework. While further improvements in activity estimation are expected to provide only limited additional gains, increasing the amount of training data appears to offer substantially greater potential. At the same time, the observed increase in denoising effectiveness suggests that the proposed approach scales well with larger datasets. Future improvements are therefore likely to benefit most from combining activity-aware learning with larger and more diverse training datasets while addressing additional sources of weak-label noise, such as overlapping species, label ambiguity, class imbalance, and recording variability.

## 5. Conclusion

In this work, we investigated the impact of weak-label noise caused by inactive audio chunks in the large-scale *InsectSet459* dataset and evaluated several activity-based denoising strategies for insect sound classification. Using an insect activity detector, we estimated that approximately 13–15% of all training chunks contain no insect sounds despite inheriting the recording-level species label, providing a quantitative estimate of this source of weak-label noise.

Explicitly incorporating activity information consistently improved classification performance, with the combination of a background class and MIL fine-tuning achieving the best results and improving the macro F1-score by 3.4 percentage points over the baseline. Additional analyses showed that denoising is most effective for species with moderate activity levels, while further improvements in activity detection alone are unlikely to substantially increase classification performance.

Overall, our results demonstrate that inactive audio segments represent a measurable and practically relevant source of weak-label noise in insect sound classification. More broadly, they show that explicitly identifying and modelling specific sources of weak-label noise can improve large-scale bioacoustic classification with a limited amount of additional activity annotation, avoiding the need for detailed species-level temporal annotations.

## 6. Acknowledgments

The authors used ChatGPT (OpenAI, GPT-5.5) to assist with proofreading and improving the grammar and readability of the manuscript. The authors reviewed and edited all generated text and take full responsibility for the final content.

## References

[1] S. Hoefer, D. T. McKnight, S. Allen-Ankins, E. J. Nordberg, and L. Schwarzkopf, "Passive acoustic monitoring in terrestrial vertebrates: A review," *Bioacoustics*, 2023, ISSN: 0952-4622.

[2] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, 2021, ISSN: 1574-9541.

[3] M. Faiß, B. Ghani, and D. Stowell, *InsectSet459: An open dataset of insect sounds for bioacoustic machine learning*, 2025.

[4] D. Funosas et al., *ECOSoundSet: A finely annotated dataset for the automated acoustic identification of Orthoptera and Cicadidae in North, Central and temperate Western Europe*, https://arxiv.org/abs/2504.20776v1, 2025.

[5] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, 2013, ISSN: 0004-3702.

[6] A. Kumar and B. Raj, "Audio Event Detection using Weakly Labeled Data," in *Proceedings of the 24th ACM International Conference on Multimedia*, USA: Association for Computing Machinery, 2016.

[7] Q. Kong, C. Yu, Y. Xu, T. Iqbal, W. Wang, and M. D. Plumbley, "Weakly Labelled AudioSet Tagging With Attention Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019, ISSN: 2329-9304.

[8] H. Ghaffari and P. Devos, "Robust Weakly Supervised Bird Species Detection via Peak Aggregation and PIE," *IEEE Transactions on Audio, Speech and Language Processing*, 2025, ISSN: 2998-4173.

[9] B. McFee, J. Salamon, and J. P. Bello, "Adaptive Pooling Operators for Weakly Labeled Sound Event Detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018, ISSN: 2329-9304.

[10] H. Van Truong, N. Nvn, A. To, D.-T. Le, and H. D. Nguyen, "Reliable Sound Re-labeling Approach for Bird Sound Classification," *Procedia Computer Science*, 2023, ISSN: 1877-0509.