Clean Data, Simple Models: A Practical Audio Preprocessing Approach for Multi-Species Sound Classification

Abstract

We present a lightweight audio-preprocessing pipeline that boosts simple classifiers for multi-species sound identification in Colombian soundscapes. Developed for BirdCLEF 2025 and evaluated on recordings from Reserva Natural El Silencio (Magdalena Medio Valley), the pipeline isolates vocalizations, removes silence, and filters noise to produce cleaner BirdNET embeddings. We train MLP and CNN models on raw vs. cleaned inputs. Results in multi-taxon species classification show that improving signal quality can offset model complexity, where a cleaned-input MLP matches or surpasses deeper baselines with modest compute. This underscores the value of preprocessing for bioacoustic monitoring in noisy, resource-limited settings and demonstrates that robust baselines can be built with accessible computing resources common in biodiversity-rich developing countries.

1 Introduction

Passive acoustic monitoring (PAM) has become a powerful and non-invasive tool for biodiversity assessment, enabling the continuous recording of animal vocalizations in diverse ecosystems [1; 2; 3]. In biologically rich tropical countries such as Colombia, PAM provides valuable insights into species presence, distribution, and ecological dynamics. The El Silencio Natural Reserve, one of the study sites for BirdCLEF 2025 [4], protects a diverse taxa, making it a valuable location for acoustic biodiversity monitoring. In this context, the BirdCLEF 2025 challenge aims to develop automated methods to identify various taxonomic groups, including birds, amphibians, mammals, and insects, from soundscape recordings collected in the reserve. However, analyzing the data from these soundscapes presents significant challenges due to overlapping vocalizations, background noise from anthropogenic sources, and strong class imbalances across species.

Despite the growing success of deep learning models such as BirdNET in large-scale bird identification tasks [5], most efforts in bioacoustics continue to rely on complex architectures or heavily supervised frameworks that require extensive annotated datasets. While these models often perform effectively within specific taxonomic groups (e.g., birds), their scalability and adaptability across other taxa (e.g., amphibians, mammals, or insects) remain limited [6; 7]. Furthermore, pipelines typically assume well-conditioned input data, placing limited attention on the signal preprocessing stage [8; 9].

In multi-species classification scenarios, especially those involving diverse vocal repertoires and heterogeneous sound environments, signal quality critically affects the separability of acoustic features. Although deep learning models offer powerful solutions, their performance can degrade in the presence of noise or misaligned inputs, conditions common in tropical soundscapes. As species vocalizations vary in frequency range, duration, and intensity, inadequate preprocessing may obscure biologically relevant information, ultimately limiting classification accuracy regardless of model complexity. We hypothesize that by enhancing signal quality through targeted preprocessing, it is possible to enable lightweight and general-purpose models to perform competitively, offering an efficient alternative for biodiversity-rich but computationally constrained contexts such as Colombia.

In this work, we propose a lightweight preprocessing pipeline tailored for the multi-species classification task of BirdCLEF 2025. By applying preprocessing steps such as silence removal, vocalization isolation, and controlled Gaussian noise addition, we enhance the clarity and informativeness of the

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

input representation. We extract acoustic embeddings using the BirdNET Analyzer [5] and pass them to simple classifiers, including multilayer perceptrons (MLPs) and generic convolutional neural networks (CNNs). Our results suggest that preprocessing is not merely a preliminary step but a critical enabler of robust and scalable species classification.

2 Materials and Methods

2.1 Materials: The dataset used in this study (taken from BirdCLEF 2025) includes labeled and unlabeled recordings from 206 species across four taxonomic groups: Aves, Amphibia, Mammalia, and Insecta. Recordings were sourced from Xeno-canto, iNaturalist, and the Colombian Sound Archive, all resampled to 32 kHz, and collected at El Silencio Natural Reserve (6°45'N, 74°12'W) [4]. Aves dominate with 146 species (70.9%) and 27,648 recordings, followed by amphibians (34 species, 16.5%, 583 recordings), insects (17 species, 8.3%, 155 recordings), and mammals (9 species, 4.4%, 178 recordings). Gathered in a tropical rainforest under ecological restoration, these recordings illustrate the value of passive acoustic monitoring (PAM) as a scalable tool for assessing biodiversity and conservation outcomes.

2.2 Multiespecies classification methodology: This study focuses on the impact of preprocessing strategies on training robust audio classifiers [9], making the preprocessing pipeline a central component of the methodology. The overall structure follows the standard workflow in species identification tasks [10]: segmentation, preprocessing, feature extraction, and classification (Figure 1). All recordings were segmented into uniform five-second clips, which were then processed through steps such as Gaussian noise addition, silence removal, and voice cleaning to enhance signal clarity and consistency. From these refined segments, we extracted BirdNET embeddings, compact descriptors of species vocalizations, and compared models trained on embeddings from both raw and preprocessed audio. Finally, the embeddings were classified using lightweight models, including convolutional neural networks (CNNs) and multilayer perceptrons (MLPs), highlighting the often-underestimated importance of preprocessing in enabling efficient and scalable bioacoustic classification systems.

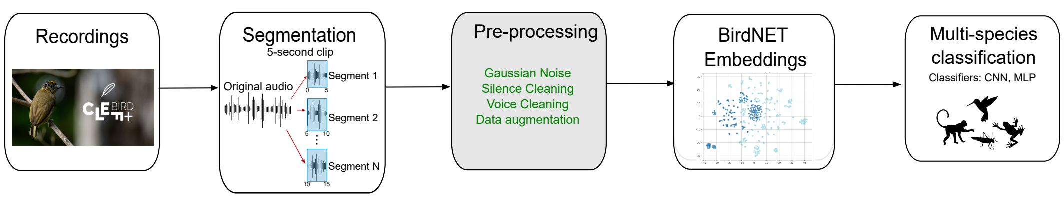


Figure 1: Multi-species classification methodology

2.2.1 Segmentation: Initially, all audio recordings are segmented into fixed-length clips of five seconds. This duration facilitates capturing complete vocalization events while ensuring uniform input dimensions for downstream processing. Segments shorter than five seconds but longer than two seconds are extended using a tiling technique, wherein the original audio content is repeated until the desired length is reached. This standardization is crucial for embedding generation and model training, as it enables batch processing and ensures consistent input size across the dataset.

2.2.2 Preprocessing: After obtaining the fixed-duration segments, a series of preprocessing techniques are applied to increase the quality and consistency of the data. These transformations, designed to improve signal clarity and standardization, aim to facilitate the learning of discriminative features by classifiers, ultimately optimizing performance across evaluation metrics. The complete pipeline is described below:

Band-Limited Gaussian Noise: The first preprocessing step handled inconsistencies in the source audio files, some of which had been up-sampled to 32 kHz or filtered (e.g., bandpass, lowpass), resulting in spectrograms with empty frequency bands that produced 'black' regions of zero energy. These irregularities, common in datasets from public repositories like Xeno-canto, can bias training and interfere with silence detection. To mitigate this, we applied a band-limited Gaussian noise augmentation strategy that injects low-level noise into inactive spectral regions, effectively filling the black bands, normalizing information content across samples, and improving the consistency of signal-based silence detection.

Silence Cleaning:

After correcting the zero-energy bands, we removed minimally informative audio, including silence and low-relevance background sounds. We manually filtered out non-bird classes, while automation was used for bird vocalizations. Each five-second segment was divided into ten 500-ms windows, and we calculated the variance of kurtosis across these windows. Flat signals, such as silence or steady noise, resulted in low variance, whereas bird calls exhibited higher variance. We then excluded the bottom 5% of segments based on kurtosis variance, effectively removing low-content audio while preserving the diversity of valid vocalizations.

Voice Cleaning: Improving the performance of both feature extraction and classification models requires incorporating a voice removal step. Human speech added by field recordists is present in several recordings and introduces noise into the feature extraction process. We used a pre-trained VGG-19 model to identify human voices in five-second segments. Therefore, it classified segments with human voice identifiers. For non-bird species, segments classified as voice were manually verified due to the lower data proportion. Ultimately, segments identified as human voice were removed from the labeled dataset.

Downsample: According to the labeled dataset, we observed imbalance in the number of audio segments for each species within each taxonomic group. To balance the dataset, we first calculated the median of audio segments for all non-bird and bird species separately. Then, we applied random downsampling to values slightly above the median: 500 for the bird group and 15 for the non-bird group. After this, the dataset was divided into train (60%), validation (20%), and test (20%).

Data Augmentation: In this step, species with fewer audio segments than the median were augmented up to that threshold, making the prior filtering of silence and voice cleaning essential to avoid introducing noise that could degrade performance. Augmentation was applied only to underrepresented bird and non-bird species in the training set, using techniques such as white noise addition, time-shifting, and background noise from the ff1010bird dataset [11; 12; 13], thereby increasing sample diversity and supporting better generalization of the classifier.

2.2.3 BirdNET Embeddings:

After preprocessing and cleaning, feature extraction was carried out using embeddings generated by the BirdNET Analyzer [5], a widely used CNN-based tool trained on spectrograms of bird vocalizations. This open-source framework offers a strong accuracy–efficiency trade-off [14]. We used v2.4, which outputs a 1024-dimensional embedding per three-second snippet. Since our inputs are five seconds, each segment yields two vectors that we aggregated into a single 1024-D representation via average pooling.

Embeddings Concatenation:

To add temporal context, for each five-second segment we concatenate its 1024-D vector with those of the next two segments (repeating the last available if needed), forming a 3,072-D input. This sliding window captures onsets, offsets, and inter-call gaps beyond single-segment scope, improving robustness and generalization in complex soundscapes (see Appendix A.2).

2.2.4 Training Classifier: Following the principle of "Clean Data and Simple Models," we prioritized lightweight architectures that balance efficiency and performance, utilizing the strength of the extracted embeddings to reduce training data and computational demands. To capture patterns within the embeddings, we fine-tuned Convolutional Neural Networks (ResNet-18 and ResNet-34 [15]), reshaping each embedding into a 32×96 matrix for compatibility with 2D convolutions and adapting input/output layers to the task. In parallel, we evaluated a Multi-Layer Perceptron (MLP), with hyperparameters (learning rate, hidden layers, neurons) optimized via Optuna, under the hypothesis that sufficiently informative embeddings allow even simple classifiers to achieve competitive performance.

Evaluation Metrics: We evaluated the performance of our multi-species classification models using standard metrics on a held-out test set comprising 30% of the dataset. F1-score, precision, and accuracy were reported to provide a broad overview of model performance and enable comparison across evaluation dimensions. However, our main evaluation metric was **Recall**, as it best reflects the model's ability to correctly identify each species and minimizes false negatives, an essential factor in biodiversity monitoring to avoid underestimating richness or overlooking rare taxa.

3 Results

Table 1 summarizes the performance of the models across multiple evaluation metrics. The MLP trained on raw embeddings achieved the highest internal scores in precision, accuracy, weighted F1, and weighted recall. However, when evaluated on the BirdCLEF 2025 Kaggle test set (unseen data), the MLP trained on balanced and preprocessed embeddings obtained the best score (0.756). Although the raw MLP appeared stronger on internal evaluation, the balanced MLP demonstrated superior generalization, underscoring the importance of addressing data imbalance in multi-species classification tasks. This conclusion is further supported by the learning curves presented in Appendix A.3, which show reduced overfitting and improved consistency for the balanced MLP.

Table 1: Comparison of models across evaluation metrics. While the MLP trained on raw data leads in internal metrics, the balanced MLP with cleaned data achieves the highest Kaggle score (0.756), indicating superior generalization.

Model	Precision - W	Accuracy	F1 - W	Recall - W	Kaggle Score
Resnet 18	0,77	0,76	0,76	0,7	0,657
MLP Raw	**0,89**	0,89	0,88	0,89	0,728
MLP Balanced	0,84	0,83	0,83	0,83	**0,756**

Figure 2 reports recall per species for a subset of birds, comparing cleaned data (blue) versus raw data (orange). We observe a consistent recall gain for most classes with cleaning. Overall, cleaning raises class-wise performance and supports the global idea that with better signal quality, a simple MLP may match or even surpass more complex models. Recall is particularly relevant in biodiversity monitoring because it measures the model's ability to correctly identify species and minimize false negatives; failing to detect a species that is actually present could lead to underestimating local richness or overlooking rare and conservation-critical taxa.

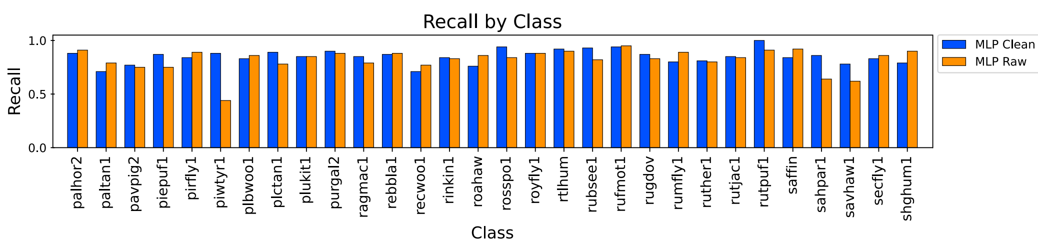


Figure 2: Recall by species for a subset of classes. Blue: cleaned data; orange: raw data. Cleaning improves recall for most species, reinforcing that better signal quality enables strong performance with simple models.

4 Conclusions

This study shows that targeted preprocessing including **silence removal, human voice filtering, and Gaussian noise compensation** is crucial for enabling lightweight models such as MLPs and basic CNNs to achieve competitive performance in multi-species classification. Although raw embeddings showed stronger results on internal metrics, the balanced and preprocessed datasets achieved better generalization on the BirdCLEF 2025 Kaggle test set, a dataset entirely unseen by the models. Demonstrating that data quality can outweigh model complexity, particularly in computationally constrained contexts. Learning curves further confirmed that preprocessing reduces overfitting and supports robustness across diverse acoustic environments, highlighting its value for bioacoustic monitoring (see Appendix A.3 for details). Nonetheless, challenges remain due to class imbalance and the underrepresentation of certain species. Future directions include refining automated noise and silence detection, exploring additional lightweight architectures such as compact transformers, and improving the pipeline to handle longer recordings and diverse ecological soundscapes, with the objective of increasing robustness against acoustic variability and ultimately advancing scalable, accessible, and reliable conservation tools.

References

[1] Rory Gibb, Ella Browning, Paul Glover-Kapfer, and Kate E. Jones. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, 10(2):169–185, 2019. doi: 10.1111/2041-210X.13101.

[2] Bryan C. Pijanowski, Luis J. Villanueva-Rivera, Sarah L. Dumyahn, Almo Farina, Bernie L. Krause, Brian M. Napoletano, Stuart H. Gage, and Nadia Pieretti. Soundscape ecology: The science of sound in the landscape. *BioScience*, 61(3):203–216, 2011. doi: 10.1525/bio.2011.61.3.6.

[3] V. Ramesh et al. Using passive acoustic monitoring to examine the impacts of ecological restoration on faunal biodiversity in the western ghats. *Biological Conservation*, 282:110071, 2023. doi: 10.1016/j.biocon.2023.110071.

[4] Holger Klinck, Juan Sebastián Cañas, Maggie Demkin, Sohier Dane, Stefan Kahl, and Tom Denton. Birdclef+ 2025. `https://kaggle.com/competitions/birdclef-2025`, 2025. Kaggle.

[5] Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021. doi: 10.1016/j.ecoinf.2021.101236.

[6] Daniel A. Nieto-Mora et al. Systematic review of machine learning methods applied to ecoacoustics and soundscape monitoring. *Heliyon*, 9(4):e20275, 2023. doi: 10.1016/j.heliyon.2023.e20275.

[7] Hao Wang et al. An efficient model for a vast number of bird species identification based on acoustic features. *Animals*, 12(18):2434, 2022. doi: 10.3390/ani12182434.

[8] Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152, 2022. doi: 10.7717/peerj.13152.

[9] Jialin Xie, Julien G. Colonna, and Jing Zhang. Bioacoustic signal denoising: a review. *Artificial Intelligence Review*, 54:3575–3597, 2021. doi: 10.1007/s10462-020-09932-4.

[10] Bas Ghani, Vincent J. Kalkman, Robert Planqué, and et al. Impact of transfer learning methods and dataset characteristics on generalization in birdsong classification. *Scientific Reports*, 15:16273, 2025. doi: 10.1038/s41598-025-00996-2.

[11] Anil Sampath Kumar, Thomas Schlosser, Stefan Kahl, and David Kowerko. Improving learning-based birdsong classification by utilizing combined audio augmentation strategies. *Ecological Informatics*, 82:102699, 2024. doi: 10.1016/j.ecoinf.2024.102699.

[12] Lucas Ferreira-Paiva, Elizabeth Alfaro Espinoza, Vinicius Martins Almeida, Leonardo Felix, and Rodolpho Neves. A survey of data augmentation for audio classification. In *Congresso Brasileiro de Automática - CBA*, volume 3, 10 2022. doi: 10.20906/CBA2022/3469.

[13] Dan Stowell. freefield1010: An audio dataset of field recordings. `https://archive.org/details/freefield1010`, 2013. Accessed: 26-May-2025.

[14] A. Jana, M. Uili, J. Atherton, M. O'Brien, J. Wood, and L. Brickson. An automated pipeline for few-shot bird call classification, a case study with the tooth-billed pigeon, 2025. URL `https://arxiv.org/abs/2504.16276v2`. arXiv preprint.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL `https://doi.org/10.1109/CVPR.2016.90`.

A Appendix

A.1 Code and Data

The code and original datasets used in this study are openly available for reproducibility and further research at the following links: `https://anonymous.4open.science/r/NeurIPS-BirdCLEF-25/README.md` and `https://www.kaggle.com/competitions/birdclef-2025/data`.

A.2 Embeddings concatenation

For each five-second audio segment, we concatenated its corresponding embedding, with size 1024 according to the aggregation process, with those of the next two consecutive segments from the same audio file. This results in a single input vector of size 3072 (3×1024), tripling the temporal window considered by the model. For example, in Fig.1 there's an audio with 4 segments, and the first concatenated embedding is the union of *Audio0_0, Audio0_1, Audio0_2*. In cases where fewer than two additional segments were available (e.g., at the end of an audio file), the last available segment was repeated to maintain a consistent input size

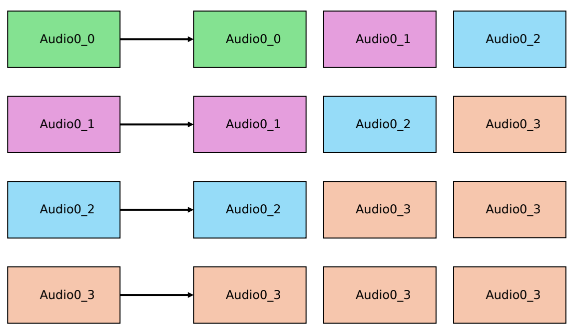
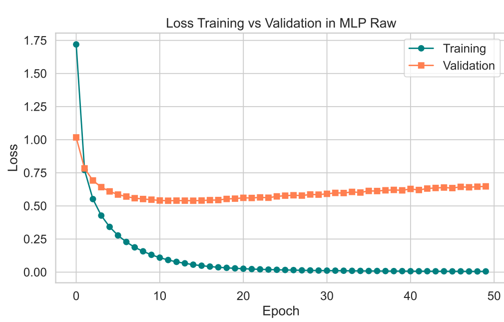


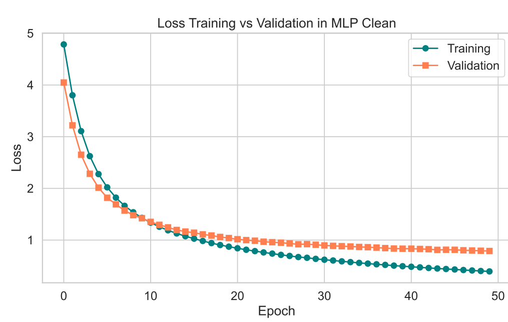
Fig. A 1: Embeddings concatenation description

A.3 Loss Train and Validation Curves in MLP

The following figure shows the comparison between the loss curves for both models: using raw and clean embeddings. It shows the training and validation loss curves, and models were trained up to 50 epochs. It can be noticed an overfitting effect in the case of the MLP raw model, and in the case of the MLP clean model, the difference between the train and validation curves in each epoch is lower than that of the MLP raw, showing better generalization.



(a) Loss curve for MLP model with raw embeddings. Validation curve indicates overfitting.



(b) Loss curve for MLP with clean embeddings. Validation and training curves indicates adecuated generalization.

Fig. A 2: Loss curve for MLP models