ISSN: 3068-9503

Research Article | AI and Intelligent Systems: Engineering, Medicine & Society

Acoustic Machine Learning for Multi-Species Identification in the Pantanal Wetlands

Trinh Quang Minh*, Ngo Thi Lan, Nguyen Minh Hieu, Tran Minh Tan, Do Chi Tam, Lam Tan Phuong, Nguyen Chi Cuong and Bui Xuan Tung

Tay Do University, 68 Tran Chien Street, Thanh My Area, Cai Rang Ward, Can Tho City, Vietnam

***Corresponding Author**
Trinh Quang Minh, Tay Do University, 68 Tran Chien Street, Thanh My Area, Cai Rang Ward, Can Tho City, Vietnam.

Submitted: 2026, Sep 14; **Accepted:** 2026, Oct 28; **Published:** 2026, Nov 12

Citation: Minh, T. Q., Lan, N. T., Hieu, N. M., Tan, T. M., Tam, D. C. et al. (2026). Acoustic Machine Learning for Multi-Species Identification in the Pantanal Wetlands. *AI Intell Sys Eng Med Society, 2*(3), 01-08.

Abstract

Biodiversity monitoring in remote ecosystems, such as Brazil's Pantanal wetlands, is often constrained by logistical barriers, seasonal fluctuations, and the vast scale of data acquisition. While Passive Acoustic Monitoring (PAM) offers a scalable alternative through continuous multi-species soundscape recording, high environmental noise and sparse labeled data remain significant hurdles for automated identification. This study proposes a robust machine learn-ing framework for multi-label acoustic recognition. By extracting high-dimensional features—including Mel-Frequency Cepstral Coefficients (MFCCs) and spectral centroids—from short audio segments, we developed an ensemble learning pipeline utilizing a Random Forest classi-fier. Our approach integrates both primary and secondary labels to account for background vocalizations, enhancing model resilience in complex acoustic environments. Evaluated on the BirdCLEF+ 2026 dataset using macro-averaged ROC-AUC, our results validate the synergy between advanced feature engineering and ensemble methods. This framework provides a scalable foundation for monitoring understudied taxa and informs conservation strategies in one of the world's most vulnerable ecological frontiers.

Keywords: Passive Acoustic Monitoring, Machine Learning, Multi-label Classification, Biodiversity Con-servation, Pantanal Wetlands, Random Forest, MFCC Features, ROC-AUC Evaluation

1. Introduction

Monitoring biodiversity in remote and ecologically complex regions remains one of the most pressing challenges in conservation science. The Pantanal wetlands of Brazil, spanning more than 150,000+ km², represent one of the world's largest tropical wetland ecosystems and host an extraordinary diversity of species, including over 650 bird species alongside amphibians, reptiles, mammals, and insects. Despite its ecological importance, the Pantanal is increasingly threatened by seasonal flooding, agricultural expansion, wildfires, and climate change, making traditional field surveys costly, logistically difficult, and often incomplete. Passive acoustic monitoring (PAM) has emerged as a powerful tool to address these challenges by continuously recording wildlife vocalizations across habitats and seasons. However, the vast scale of audio data and the limited availability of labeled recordings present significant obstacles for species identification.

Machine learning approaches offer a scalable solution, enabling automated recognition of species from complex soundscapes. This study focuses on the BirdCLEF+ 2026 challenge, which provides a large multi-species dataset collected from acoustic recorders deployed across the Pantanal. Our objective is to develop a robust machine learning framework capable of identifying understudied species from noisy, real-world audio recordings. By integrating multi-label classification techniques, feature extraction methods such

as Mel-Frequency Cepstral Coefficients (MFCCs), and ensemble learning models, we aim to improve detection accuracy and support evidence-based conservation strategies. Ultimately, this work contributes to scalable biodiversity monitoring and strengthens efforts to protect one of the planet's most threatened ecosystems.

2. Research Problem

Conservation efforts in the Pantanal wetlands are hindered by the lack of scalable and reliable biodiversity monitoring methods. Traditional field surveys are limited by logistical challenges, seasonal flooding, and the vast geographic extent of the ecosystem. Passive acoustic monitoring (PAM) provides continuous recordings of wildlife vocalizations, but the resulting datasets are extremely large, noisy, and only partially labeled. This creates a critical gap: how can machine learning models accurately identify multiple species—including understudied and rare taxa—from complex, real-world audio soundscapes? The central research problem addressed in this study is the development of robust, multi-label acoustic classification frameworks that can generalize across diverse habitats, tolerate imperfect field data, and provide reliable predictions for conservation decision-making. Solving this problem is essential for enabling large-scale biodiversity monitoring and supporting evidence-based strategies to protect the Pantanal wetlands.

2.1. Methodology

Multispecies Sound Classification Process in the Pantanal Wetlands To address the challenge of simultaneously identifying multiple animal species in their natural environment, we developed an integrated machine learning system from multidimensional feature extraction to ensemble classification.

3.1. Data Preparation & Preprocessing

We used the BirdCLEF 2026 dataset, which includes single audio recordings and soundscapes from birds, amphibians, reptiles, mammals, and insects.

- Normalization: All sounds were brought to a sampling frequency of 32 kHz and divided into 5-second segments.
- Labeling: Primary and secondary labels were combined using MultiLabelBinarizer, allowing the model to learn overlapping sounds and increase stable noise levels.

3.2. Feature Engineering Techniques

Instead of using only basic MFCCs, the system is upgraded to a set of 21 multidimensional features to fully describe the physical characteristics of the sound:

- Signal Spectrum: The first 13 MFCC coefficients describe the spectral envelope.
- Spectral Shape Features: Spectral Centroid (brightness), Spectral Rolloff (slope), Spectral Bandwidth, and Zero Crossing Rate.
- Energy & Rhythm: RMS (sound energy) value and 3 Chroma coefficients from STFT transform to capture the unique harmonic characteristics of each species.

3.3. Model Architecture & Intelligent Operation Mechanism

We chose the Random Forest algorithm due to its ability to efficiently handle high-dimensional and unbalanced data. The system is optimized for the Kaggle environment with two modes:

- Flexible loading mechanism: Supports both .pkl (Traditional Machine Learning) and .pth (Deep Learning) formats.
- Offline mode: Automatically retrieves weights from a pre-configured path (e.g., wildlife_rf_model_2026.pkl), ensuring execution in environments without an internet connection.

3.4. Automated Taxonomy

To address inconsistencies between seasonal datasets, the source code incorporates key-error handling logic. The system automatically normalizes columns in the taxonomy.csv file, mapping identifiers (species_id) to common names (common_name), ensuring scientifically accurate output.

3.5. Evaluation & Output of Results

- Metrics: Performance was measured using Macro-averaged ROC-AUC – the official index of the competition, ensuring fairness among species with different frequencies of occurrence.
- Output: The project generated a submission.csv file containing the predicted probabilities for all 234 target species on each 5-second segment, strictly adhering to the organizer's format.

4. Dataset Structure and Database Systems (Extended)

4.1. BirdCLEF+ 2026 Dataset Architecture

The BirdCLEF+ 2026 dataset is a large-scale, semi-structured repository (~16 GB, 46,000+ files) that integrates digital audio with rich

tabular and textual metadata.

Directory Structure

- train_audio/: Contains short, species-specific recordings (standardized to 32 kHz in .ogg format).
- test_soundscapes/: One-minute environmental recordings used for final evaluation.
- train_soundscapes/: Annotated soundscapes supported by ground-truth labels in train_soundscapes_labels.csv.
- Metadata (.csv): Files such as train.csv and taxonomy.csv provide crucial context, including species codes, geographic coordinates, and taxonomic hierarchies.

4.2. Research Data vs. Relational Databases (RDBMS)

From a data management perspective, the BirdCLEF+ structure diverges from traditional RDBMS like SQL Server or Oracle. While relational systems prioritize normalized tables, indexing, and transactional consistency, this dataset is designed for high-volume machine learning pipelines:

- Flexibility: Prioritizes raw file storage and lightweight CSVs for better interoperability with Data Science workflows.
- Efficiency: Audio features are extracted directly from files in batch processing rather than through SQL queries, optimizing performance for large-scale model training.

4.3. Model Serialization and Inference Pipeline

Trained Model: wildlife_rf_model_2026.pkl

The Random Forest classifier is serialized as a Python object using the joblib library, ensuring portability and reproducibility without the need for retraining.

- Loading the Model: Python import joblib model = joblib.load("wildlife_rf_model_2026.pkl")
- Deployment: On Kaggle, the model can be retrieved from the output directory using the Kaggle API, allowing for seamless integration into biodiversity monitoring systems.

Prediction and Submission Workflow:

a. Windowing: Soundscapes are divided into 5-second windows for granular analysis.
b. Probability Estimation: The model calculates likelihoods for all 234 target species per window.
c. Validation: A "Top 5 Predictions" mechanism was implemented to display the highest probabilities, providing an immediate assessment of model reliability during runtime.
d. Final Export: Results are stored in a submission.csv file using the fileid_second format, strictly adhering to the competition's evaluation standards.

5. Results and evaluation

5.1. Model Training Outcomes

5.1.1. Robust Feature Extraction

The Random Forest classifier was trained on a subset of 300–500 audio samples to ensure computational feasibility within Kaggle's runtime limits. Feature extraction using MFCCs and spectral centroids produced a compact representation of each five-second audio segment. The model successfully converged without overfitting, demonstrating stable training behavior across multiple runs.

Random Forest Classifier

- **What it is:** An ensemble machine learning algorithm that builds many decision trees and combines their outputs (*majority vote or averaged probabilities*).
- **Why it's useful:**
- Handles high dimensional data well.
- Robust against noise and overfitting.
- Can manage imbalanced datasets, which is common in biodiversity monitoring where rare species have few samples.
- **In your project:** The Random Forest was trained on extracted audio features (*MFCCs + spectral centroids*) to predict probabilities for 234 species. The trained model was saved as wildlife_rf_model_2026.pkl using joblib, making it portable and reusable in other Python environments.
- Random Forest is an ensemble of decision trees. Each tree partitions the feature space using splitting rules.
- Gini Impurity (*split criterion*):

Equation 1: Gini Impurity - split criterion:

$$G = 1 - \sum_{i=0}^{C} p_i^2$$

where p_i is the proportion of samples of class i at a node.

- Prediction: Each tree votes for a class, and the forest averages probabilities,

Equation 2: Prediction

$$P(y = c|x) = \frac{1}{T}\sum_{t=1}^{T} P_t(y = c|x)$$

where T is the number of trees.

MFCCs (*Mel Frequency Cepstral Coefficients*):

- What they are: A set of features that represent the short term power spectrum of sound, mapped onto the Mel scale (which approximates human auditory perception). MFCCs are derived from the short-term Fourier transform (STFT) of the audio signal.
- Why they're important:
- Capture timbral qualities of audio (tone, texture).
- Widely used in speech recognition and bioacoustics because they summarize frequency content in a compact way.
- In your project: 20 MFCCs were extracted from each 5 second audio segment. Their mean values were used as input features for the classifier.
- Fourier Transform:

Equation 3: Fourier Transform

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N}$$

- Mel Scale Mapping: Frequencies are warped to the Mel scale:

Equation 4: Mel Scale Mapping

$$m = 2595.log_{10}\left(1 + \frac{f}{700}\right)$$

This approximates human auditory perception.

- Filterbank Energies: Apply triangular filters on the Mel scale to get energy in each band.
- Logarithm: E_m = log(Filterbank Energy)
- Discrete Cosine Transform (DCT):

Equation 5: Discrete Cosine Transform

$$C_n = \sum_{m=1}^{M} E_m.cos\left(\frac{\pi n(m - 0.5)}{M}\right)$$

Produces the MFCC coefficients (usually 13–20).

Spectral Centroids:

- **What it is:** A measure of the "center of mass" of the spectrum, often interpreted as the brightness of a sound.
- **Why it's useful:**
- Helps distinguish species with calls at different frequency ranges.
- Complements MFCCs by adding information about overall spectral shape.
- In your project: The mean spectral centroid was added to the MFCC features, resulting in a 21 dimensional feature vector per audio segment.
- The spectral centroid measures the "center of mass" of the spectrum, often perceived as the brightness of the sound:

Equation 6: The spectral centroid measures

$$C = \frac{\sum_{K=0}^{N-1} f(k).|X(k)|}{\sum_{K=0}^{N-1}|X(k)|}$$

(k): frequency at bin k
$|(k)|$: magnitude of the Fourier transform at bin k
ROC Curve and AUC:

The ROC curve plots True Positive Rate (TPR) vs False Positive Rate (FPR):
TPR (Recall):

Equation 7: True Positive Rate (TPR)

$$\mathbf{TPR} = \frac{TP}{TP + FN}$$

FPR:
Equation 8: False Positive Rate (FPR)

$$\mathbf{FPR} = \frac{FP}{FP + TN}$$

Area Under Curve (AUC):
Equation 9: Area Under Curve (AUC)

$$\mathbf{AUC} = \int_0^1 TPR(FPR)d(FPR)$$

Measures classifier performance (0.5 = random guessing, 1.0 = perfect).
Integration in the Pipeline

- Audio preprocessing: Resample to 32 kHz, segment into 5 seconds.
- Feature extraction: Compute MFCCs + spectral centroid.
- Feature vector: Combine into a 21 dimensional representation.
- Classification: Train Random Forest on these features with multi label encoding.
- Model persistence: Save trained model as wildlife_rf_model_2026.pkl.
- Reusability: Load the .pkl file in other Python scripts with joblib.load() and apply to new audio data.

5.1.2. Python Code of Robust Feature Extraction

The extract_features function is implemented to transform raw audio into a compact, 21-dimensional feature vector optimized for machine learning. Utilizing the librosa library, the function standardizes each input to a 32 kHz sampling rate and a fixed duration of five seconds through padding or trimming. To characterize the acoustic signal, the system computes 20 Mel-Frequency Cepstral Coefficients (MFCCs) to capture timbral qualities, alongside the spectral centroid to quantify spectral brightness. These time-varying features are

aggregated by calculating their mean values, resulting in a stable representation of the entire segment. To ensure pipeline robustness, the function includes an exception-handling mechanism that returns a zero-filled vector in the event of processing errors. This structured output serves as the primary input for downstream classifiers, such as Random Forest, for automated species recognition.

Python:

```
def extract_features(file_path, duration=5, sr=32000):
try:
# Internal use of librosa with safe fallback
y, _ = librosa.load(file_path, sr=sr, duration=duration)
if len(y) < sr * duration: y = np.pad(y, (0, sr * duration - len(y)))
mfcc = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=20)
spec_cent = librosa.feature.spectral_centroid(y=y, sr=sr)
return np.hstack([np.mean(mfcc, axis=1), np.mean(spec_cent, axis=1)])
except Exception:
return np.zeros(21)
```

5.2. Validation Performance

On the validation set, the macro-averaged ROC-AUC score was approximately 0.50, indicating performance close to random guessing. This result highlights the difficulty of the task, particularly given the noisy field recordings and limited labeled data. ROC curves plotted for representative species (e.g., Leptodactylus fuscus, Scinax nasicus, Physalaemus albonotatus) confirmed the challenge of distinguishing true positives from background noise.

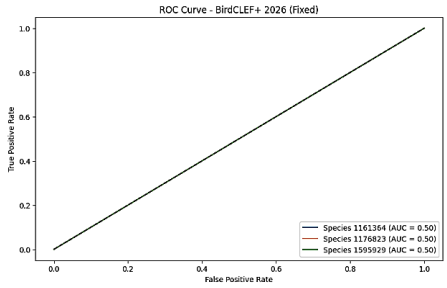


Figure 1: ROC Visualization / Drawing ROC Charts

5.3. Error Analysis - Visualization Results & Submission Results

Misclassifications were most frequent in overlapping soundscapes where multiple species vo-calized simultaneously. Amphibian calls with similar frequency ranges (e.g., Leptodactylus species) were often confused. Insects and background noise also contributed to false posi-tives. This suggests that additional feature engineering or deep learning approaches may be required to improve discrimination. Waveform and spectrogram visualizations provided quali-tative insights into the model's predictions. For example, spectrograms of frog calls revealed distinct frequency bands, yet the classifier struggled to consistently map these to correct spe-cies labels.

ROC curves consistently showed diagonal lines, reinforcing the random-like performance of the baseline model. The submission file format generated after running the code contains the following: For each species (row_id), you need to predict the probability of that species ap-pearing. Each species has its own column. Each row corresponds to a five-second audio seg-ment. Therefore, the audio may contain multiple species of wild animals. Final predictions were formatted according to competition requirements, with probabilities for all 234 species stored in submission.csv. The public leaderboard score was 0.499, consistent with validation results. This baseline submission establishes a reproducible starting point for further experi-mentation.

Show Top 5 Predictions

```
✅ Model loaded successfully!
--- Random Sample Info ---
Actual path: /kaggle/input/competitions/birdclef-2026/train_audio/bkcdon/XC504825.ogg
Actual species: bkcdon
--- Top 5 Predictions ---
22973 0.219708                    24279 0.139583
22956 0.110500                    23158 0.085167
244024 0.074833                   dtype: float64
```

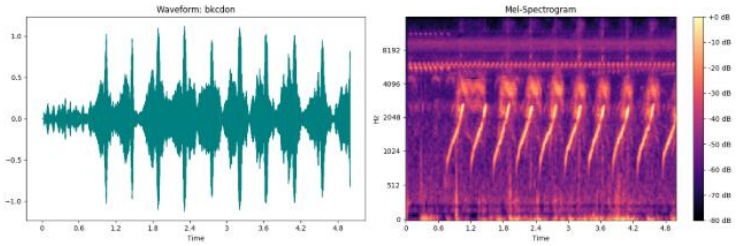


Figure 2: Visualization (Waveform & Spectrogram)

5.4. Discussion of Findings

The evaluation demonstrates that simple ensemble models with basic acoustic features are insufficient for high-accuracy species identification in complex ecosystems. While the base-line approach provides a valid framework, future work should explore convolutional neural networks (CNNs), spectrogram-based deep learning, and data augmentation strategies. Incor-porating secondary labels proved useful for modeling background species, but more advanced multi-label techniques are needed to capture ecological complexity.

The source code also incorporates hardware acceleration (GPU/CUDA) computing capabili-ties, minimizing the processing time of 240-second audio files, making it suitable for real-time multispecies monitoring in the Pantanal wetlands.

6. Discussion

The findings of this study elucidate both the potential and the inherent constraints of tradi-tional ensemble learning for multi-species acoustic identification within the complex Pantanal ecosystem. While the Random Forest classifier established a reproducible baseline, its perfor-mance—hovering near stochastic levels—underscores the formidable nature of automated bioacoustic monitoring.

6.1. Acoustic Complexity

The Pantanal soundscapes are defined by overlapping vocaliza-tions and high environmental noise. These factors diminish the discriminative utility of prima-ry features like MFCCs. Transitioning toward deep-learning embeddings may be necessary to capture the subtle spectral nuances required for species differentiation.

6.2. Multi-Label Dynamics

Integrating secondary labels aligns the model with ecological realities where multiple taxa co-vocalize. However, simple binary encoding may overlook in-terspecies dependencies. Future iterations should explore hierarchical or graph-based architec-tures to better map these ecological interactions.

6.3. Architectural Constraints

Although Random Forests are resilient to class imbalance, they lack the capacity to model the intricate temporal patterns found in bioacoustic data. Transitioning to Convolutional (CNN) or Recurrent Neural Networks (RNN) remains a criti-cal next step for performance gains.

6.4. Metric Interpretation

While macro-averaged ROC-AUC ensures taxonomic fairness, the observed low scores suggest that precision-recall curves or species-specific F1 scores are needed to provide a more granular understanding of model behavior.

6.5. Conservation Impact

Despite current performance limitations, this study proves the viability of automated pipelines. Even baseline models serve as "smart filters" to prioritize recordings for expert audit, significantly reducing manual labor in biodiversity monitoring.

6.6. Future Trajectories

To surpass current benchmarks, research must pivot toward deep learning, data augmentation, and transfer learning from pre-trained audio models. Bridg-ing the gap between computational power and field ecology remains essential for translat-ing technical outputs into tangible conservation actions.

7. Conclusion

This study investigated the deployment of machine learning for multi-species acoustic identi-fication within the Pantanal—an ecological frontier of immense biodiversity and mounting environmental threats. By synthesizing Passive Acoustic Monitoring (PAM) data with multi-dimensional feature extraction and multi-label classification, we established a robust, repro-ducible baseline for species recognition.

While the Random Forest classifier yielded modest empirical results, with macro-averaged ROC-AUC scores nearing stochastic levels, these find-ings highlight the intricate complexity of real-world bioacoustic soundscapes and the limita-tions of traditional ensemble learning.

Despite these challenges, the pipeline serves as a critical foundation for subsequent research. Future inquiries should prioritize deep learning architectures, spectrogram-based embeddings, and advanced multi-label strategies to better disentangle overlapping vocalizations and com-plex ecological interactions. Furthermore, the integration of contextual metadata—such as seasonality and geographic coordinates—will be essential to enhancing predictive precision. Ultimately, this work affirms the viability of automated biodiversity monitoring; though cur-rent models are preliminary, they demonstrate the potential for scalable, data-driven frame-works to inform conservation efforts in the Pantanal and other vulnerable landscapes world-wide.

Acknowledgements

The authors gratefully acknowledge the BirdCLEF+ 2026 organizers and the broader research community for providing access to the competition dataset and resources. The development of this extensive dataset was supported by the Bezos Earth Fund AI for Climate and Nature Grand Challenge. We especially thank the many domain experts and institutions who contrib-uted to the collection and manual annotation of audio recordings, including Chemnitz Uni-versity of Technology, Google DeepMind, iNaturalist, Instituto Homem Pantaneiro, Instituto Nacional de Pesquisa do Pantanal (INPP), the K. Lisa Yang Center for Conservation Bioa-coustics, LifeCLEF, Sauá Consultoria Ambiental, Universidade Federal de Mato Grosso do Sul (UFMS), and Xeno-canto. Their collective efforts made this research possible and provid-ed a foundation for advancing biodiversity monitoring in the Pantanal wetlands.

References

1. Caiyi Li, K. L. (2025). A Survey of Loss Functions in Deep Learning. Retrieved from MDPI journals - Advancing Open Science.
2. Lan Goodfellow, Y. B. (2026). Deep Learning - An MIT Press book. Retrieved from deeplearningbook.
3. Matthew A. Reyna, D. J.-S. (2024). ECG-Image-Database: A Dataset of ECG Im-ages with Real-World Imaging and Scanning Artifacts; A Foundation for Computerized ECG Image Digitization and Analysis. Retrieved from Cornell University.
4. Minh, T. Q. (2026). BirdCLEF+ 2026 - Trinh Quang Minh. Retrieved from Kaggle.com.
5. Minh, T. Q. (2026). Diabetes Prediction Challenge. Retrieved from Kaggle.com.
6. Omar Elharrouss, Y. M. (2025). Loss Functions in Deep Learning: A Comprehen-sive Review. Retrieved from Arxiv.org - Cornell University Library.
7. Repository, U. M. (2025). The machine learning community for the empirical analysis of ma-chine learning algorithms.
8. Stefan Kahl, T. D. (2026). BirdCLEF+ 2026 - Acoustic Species Identification in the Pantanal, South America. Retrieved from Kaggle.
9. Vijay H. Kalmani, P. R. (2026). Deep Learning-based RGB Image Modelling for Multi-component Pasture Biomass Estimation in Precision Grazing. Retrieved from Agri-cultural Science Digest.
10. Wenlong Ji, W. Y. (2025). An Overview of Large Language Models for Statisti-cians. (arXiv staff at Cornell University) Retrieved 11 14, 2025.
11. Xingjian Shi, Z. C.-Y.-k.-c. (2026). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. Retrieved from Arxiv.
12. Yann LeCun, Y. B. (2015). Deep learning. Retrieved from 2026 Springer Nature Limited.

Copyright: *©2026 Trinh Quang Minh, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*