Overview of BirdCLEF+ 2026: Acoustic Species Identification in the Pantanal, South America

Stefan Kahl[1,2], Tom Denton[3], Larissa Sayuri Moreira Sugai[2], Liliana Piatti[4], Wener Hugo Arruda Moreno[5], Mariana Motti Barbosa[4], Maximilian Eibl[1], Carolline Zatta Fieker[6], Carolina Martins Garcia[9], Daiene Louveira Hokama Sousa[4], João Emílio de Almeida Júnior[4], Ryan Christopher Kridler[4], Mario Lasseck[2], Alyson Vieira de Melo[4], Henning Müller[10], Matheus de Oliveira Neves[4], Matheus Gonçalves dos Reis[6], Lucas Korzune Sampaio Teles[6,7], Karl-L. Schuchmann[6,8], José Luiz Massao Moreira Sugai[4], Kirk Thiago Pedroso Azevedo[6], Priscila do Nascimento Lopes[4], Marinez Isaac Marques[6,8], Holger Klinck[2], Hervé Glotin[11], Hervé Goëau[12], Willem-Pier Vellinga[13], Robert Planqué[13] and Alexis Joly[14]

[1]*Chemnitz University of Technology, Chemnitz, Germany*
[2]*Cornell K. Lisa Yang Center for Conservation Bioacoustics, Cornell University, Ithaca, NY, USA*
[3]*Google DeepMind, San Francisco, USA*
[4]*Universidade Federal de Mato Grosso do Sul (UFMS), Instituto de Biociências Campo Grande, Brazil*
[5]*Instituto Homem Pantaneiro, Corumbá, Brazil*
[6]*Instituto Nacional de Pesquisa do Pantanal (INPP), Cuiabá, Brazil*
[7]*Universidade Federal de Mato Grosso (UFMT), Instituto de Computação, Cuiabá, MT, Brazil*
[8]*Universidade Federal de Mato Grosso, (UFMT) Instituto de Biociências, Cuiabá, MT, Brasil.*
[9]*Sauá Consultoria Ambiental, Campo Grande, Brazil*
[10]*University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland*
[11]*LIS, Université de Toulon, CNRS, Toulon, France*
[12]*CIRAD, UMR AMAP, Montpellier, France*
[13]*Xeno-canto Foundation, Groningen, Netherlands*
[14]*Inria, LIRMM, University of Montpellier, CNRS, Montpellier, France*

Abstract

The BirdCLEF+ 2026 challenge evaluated automated acoustic species identification in the Pantanal, South America—the world's largest continental wetland. This edition tasked participants with developing machine learning models to classify multi-taxonomic vocalizations from birds, amphibians, insects, mammals, and reptiles in continuous, complex neotropical soundscapes. Participants addressed training data constraints, including a 28-species focal training data gap in which target species had no direct focal recordings but were represented in training soundscapes, as well as strict platform constraints, specifically a 90-minute CPU-only inference limit. To address these challenges, participants designed strategies using bioacoustic foundation models. Distilled and ensembled models were developed to align student architectures with foundation model embeddings, helping prevent public leaderboard overfitting and respecting CPU runtime limits. Other techniques included iterative pseudo-labeling of unlabeled local soundscapes, specialized taxon-specific backbones for acoustically distinct groups, rank-gating composite blending, and post-processing informed by circadian and seasonal ecological priors. The competition attracted 5,026 participants, organized into 4,094 teams, who submitted 152,756 runs. The highest-performing system achieved a private leaderboard ROC-AUC of 0.965. The accompanying working notes describe various model designs and parameterizations, offering insights into neotropical passive acoustic monitoring.

Keywords

LifeCLEF, insect, amphibian, mammal, bird, song, call, species, retrieval, audio, collection, identification, fine-grained classification, evaluation, benchmark, bioacoustics, passive acoustic monitoring, PAM, Pantanal

CLEF 2026 Working Notes, September 21–24, 2026, Jena, Germany
stefan.kahl@cornell.edu (S. Kahl); tmd@google.com (T. Denton); lsugai@cornell.edu (L. S. M. Sugai); liliana.piatti@ufms.br (L. Piatti)
0000-0002-2411-8877 (S. Kahl); 0000-0003-1078-7268 (H. Klinck); 0000-0003-3296-3795 (H. Goëau); 0000-0003-3886-5088 (W. Vellinga); 0000-0002-0489-5425 (R. Planqué); 0000-0002-2161-9940 (A. Joly)

1. Introduction

Mobile and habitat-diverse species are valuable biodiversity indicators, as changes in their populations and community composition reflect the success of ecological restoration efforts. These species often respond rapidly to environmental changes, making them useful for detecting early signs of ecological improvement or degradation. However, traditional observer-based surveys across large areas are both costly and logistically demanding, often requiring extensive fieldwork, expertise, and repeated visits to remote locations, which frequently limit the spatio-temporal scale and resolution of monitoring efforts. In contrast, passive acoustic monitoring (PAM), combined with AI, offers a scalable and non-invasive solution that enables conservationists to collect and analyze vast amounts of ecological data [1, 2, 3, 4]. PAM systems can operate continuously for extended periods in challenging environments, capturing vocal activity from a wide range of taxa, including birds, amphibians, mammals, and insects [5]. Automated species identification has advanced rapidly due to deep learning architectures designed for audio analysis [6]. When paired with PAM, automated species identification enables researchers to monitor biodiversity across ecologically relevant scales, allowing more timely and data-driven reviews to inform decision-making.

The Pantanal is the largest continental wetland on the planet (150,000-195,000 km^2), home to exuberant biodiversity embedded in a mosaic of landscapes naturally shaped by flood-pulse dynamics. Cattle ranching is the main economic activity, with a presence in the area for more than 300 years, and 95% of the biome is privately owned. Surprisingly, 85% of the lowland still holds natural habitats (information primarily derived from satellite imagery), suggesting a balance between traditional, non-intensive ranching methods and conservation of natural landscapes. However, very little is known about the trends and status of animal biodiversity—specifically birds, mammals, amphibians, reptiles, and invertebrates. Recently, the combination of hydrological changes due to habitat loss and conversion in the headwaters and floodplain and climate change has culminated in massive megafires, with one-third of the biome burned in 2020, killing an estimated 17 million vertebrates [7]. Given that biome-wide changes are expected to be exacerbated by future projections of climate and habitat change, assessing and monitoring changes in animal biodiversity is a key missing piece of information for comprehensive landscape planning and prioritization of conservation interventions [7, 8].

To address these environmental challenges and advance automated biodiversity monitoring tools, the BirdCLEF+ challenge was established as a standard benchmark for bioacoustic classification. BirdCLEF+ is hosted annually as part of the LifeCLEF research initiative, a major international evaluation campaign focused on machine learning challenges for biodiversity understanding and ecosystem management [9]. Over the years, the BirdCLEF series has evolved to evaluate species identification in complex soundscapes across diverse geographic regions, including the Western Ghats of India and the tropical forests of Colombia [10, 11, 12, 13]. The 2026 edition of the competition marked a critical expansion, moving beyond birds to target a multi-taxonomic cohort in the South American Pantanal. As in previous years, the competition was hosted on the Kaggle platform (https://www.kaggle.com/c/birdclef-2026).

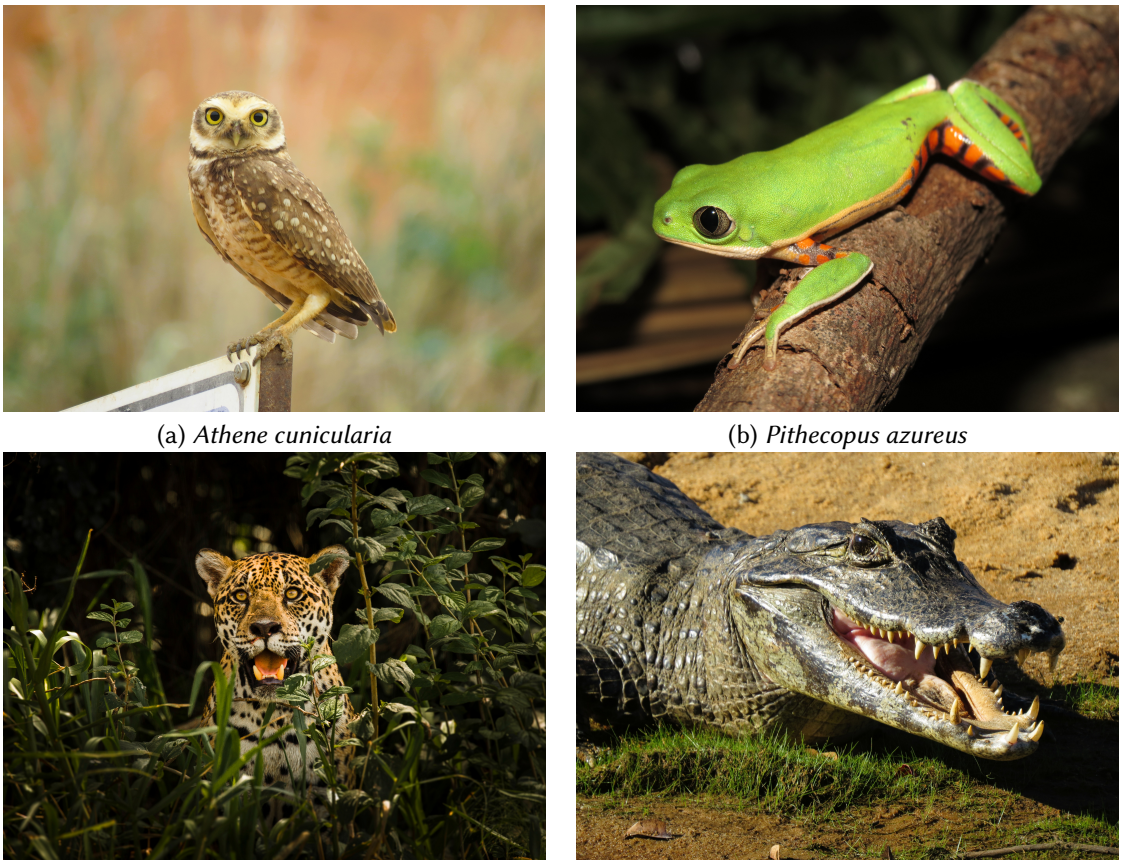
This competition aimed to advance automated, multi-taxonomic species identification in soundscape data from the Brazilian part of the Pantanal (State of Mato Grosso do Sul). Key objectives included detecting species across diverse taxonomic groups, developing machine learning models that recognize rare and endangered species with limited training data, and leveraging unlabeled data to improve detection and classification performance.

2. The Pantanal Multi-Taxonomic Dataset

The competition dataset was organized using a mix of crowd-sourced recordings (for training), local soundscapes labeled by domain experts (for validation and evaluation), and unlabeled soundscapes (for semi-supervised training). Neotropical soundscapes pose unique acoustic challenges, including high species density, overlapping calls (acoustic choruses), and significant background noise from wind, rain, and insects, all of which require robust feature representations [14].

2.1. Taxonomy and Target Classes

The competition targeted **234 species** in total, spanning five distinct biological classes. The distribution of species across taxonomic groups and their representation across dataset splits is detailed in Table 1. Figure 1 illustrates representative species for four of these taxonomic classes. While bioacoustic models are traditionally optimized for avian identification, recent work has shown that large deep learning models can be adapted to monitor other vocalizing classes, such as amphibians [15, 16].



(a) *Athene cunicularia* (b) *Pithecopus azureus*

(c) *Panthera onca* (d) *Caiman yacare*

Figure 1: Examples of target taxonomic groups in the Pantanal. The dataset includes vocalizations from diverse taxonomic classes: (a) Burrowing Owl (*Athene cunicularia*), representing Aves (Photo: Matheus O. Neves); (b) Southern Orange-legged Leaf Frog (*Pithecopus azureus*), representing Amphibia (Photo: Matheus O. Neves); (c) Jaguar (*Panthera onca*), representing Mammalia (Photo: Leonardo Ramos); and (d) Southern Spectacled Caiman (*Caiman yacare*), representing Reptilia (Photo: Matheus O. Neves).

2.2. Focal Training Data and Species Representation

The primary training set consisted of **35,549 focal recordings** from iNaturalist [17] and Xeno-canto [18], representing **206 species**. For the remaining **28 target species**, no direct focal training recordings were included, so models had to generalize from external source domains, leverage unlabeled local

Table 1
Taxonomic distribution and dataset split statistics across the 234 target species in the BirdCLEF+ 2026 dataset.

Class	Target Species	Train Species	No Focal Data	Train Clips	Soundscape Species (Train / Total)	Labeled Segments (Train / Total)
Aves	162	162	0	34,799	28 / 102	824 / 14,072
Amphibia	35	32	3	451	17 / 30	4,174 / 24,822
Insecta	28	3	25	199	25 / 25	1,136 / 5,096
Mammalia	8	8	0	99	4 / 8	84 / 1,278
Reptilia	1	1	0	1	1 / 1	26 / 206
Total	**234**	**206**	**28**	**35,549**	**75 / 166**	**6,244 / 45,474**

soundscapes, or use labeled segments in the soundscape training partition. For the insect class, these 25 target categories were represented as acoustically distinct "sonotypes" (i.e., defined by their spectral-temporal signatures in local soundscapes rather than standard taxonomic species names) because standard focal templates were unavailable. Although these 28 species lacked clean focal recordings, labeled segments of these classes in soundscape format were provided in the training soundscapes split (75 species were labeled in the training soundscapes overall, including all 25 insect sonotypes, 17 amphibian species, 28 avian species, 4 mammal species, and 1 reptile species), offering participants in-context training examples.

2.3. Passive Monitoring Deployments

Validation and test soundscapes were collected from a grid of **35 passive acoustic monitoring stations** deployed across **18 distinct locations** in Mato Grosso do Sul, Brazil (including SESC Baía das Pedras, Base de Estudos do Pantanal - UFMS, municipality of Porto Murtinho, Fazenda Caiman, and Serra do Amolar). The recording stations spanned latitudes from $-21.60°$ to $-16.49°$ and longitudes from $-57.78°$ to $-55.67°$. The deployment hardware comprised two microphone systems, consisting of 23 SwiftOne (Cornell K. Lisa Yang Center for Conservation Bioacoustics, Ithaca, NY, USA) and 12 Song Meter SM4 (Wildlife Acoustics Inc., Maynard, MA, USA) recorders. Recorders were configured to capture audio continuously (24/7) to ensure that the temporal dynamics of the soundscapes were fully represented. This extensive spatial and temporal coverage is critical for estimating species richness and understanding community composition from large-scale PAM surveys [19, 20].

2.4. Expert Annotations

Soundscape labeling was conducted by a collaborative network of **12 Brazilian biological experts** affiliated with institutions including the *Universidade Federal de Mato Grosso do Sul (UFMS)*, the *Instituto Nacional de Pesquisa do Pantanal (INPP)*, the *Instituto Homem Pantaneiro (IHP)*, and *Sauá Consultoria Ambiental*. A total of **1,305 soundscape files** were processed: **657 annotated** and **648 left unannotated** for semi-supervised training. The annotated files contained **1,478 labeled 5-second intervals** (covering 75 species), enabling participants to align their models directly with the validation soundscape's acoustic distribution.

3. BirdCLEF+ 2026 Competition Setup

The challenge was hosted on Kaggle in a code competition format with hidden test data. Participants identified species from short 5-second audio clips extracted from soundscape recordings.

3.1. Evaluation Protocol

The evaluation metric was a variant of macro-averaged ROC-AUC computed on 5-second soundscape clips, excluding classes with no true positive labels in the private test set. This approach allowed us to

assess model performance without tuning confidence thresholds and emphasized species-level rather than segment-level accuracy [21]. Because raw classifier confidence scores are often unitless and show species-specific relationships to prediction accuracy, using rank-based metrics like ROC-AUC helps standardize performance assessments without requiring species-specific threshold calibration [22]. The hidden test set consisted of **approximately 600 soundscape files**, each exactly 1 minute long (sampled at 32 kHz), yielding a total of **7,200 evaluation segments** of 5 seconds each.

3.2. Submission Constraints

To respect real-world deployment constraints and encourage computationally efficient solutions, submissions were run in a strict **CPU-only notebook environment with a 90-minute runtime limit** (restricted to 2 CPU cores, 16 GB of RAM, and 30 GB of disk space). This restriction simulated the computational constraints of edge-device deployment for real-world bioacoustic monitoring. Under these hardware constraints, a single unoptimized out-of-the-box inference pass of a large deep learning model (such as Google's Perch 2.0 foundation model) on the 600-file test set required approximately 30 to 45 minutes, so large ensembles of raw foundation models would easily exceed the 90-minute limit. Consequently, this constraint prevented the deployment of massive, unoptimized ensembles of heavy deep learning models and prioritized architectures optimized for speed and efficiency (such as ONNX-quantized networks or lightweight student backbones).

4. Results of BirdCLEF+ 2026

The BirdCLEF+ 2026 competition attracted **5,026 participants**, organized in 4,094 teams, who submitted 152,756 runs.

4.1. Leaderboard Results

The final ranking (Figure 2) was determined by the private leaderboard score, computed on 65% of the test set and kept hidden until the deadline. During the competition, participants had access to a public leaderboard based on 35% of the test set.

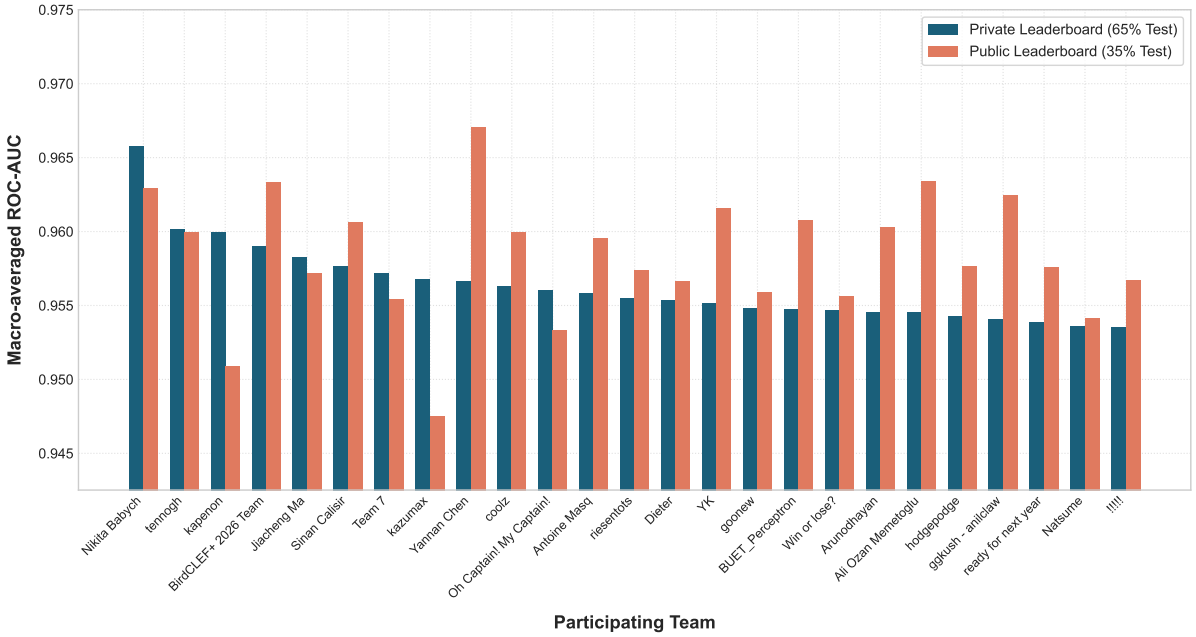


Figure 2: BirdCLEF+ 2026 leaderboard results. Public and private scores for the top-performing submissions. The top submissions exhibit close scoring alignment, with the first-place submission achieving a private leaderboard ROC-AUC of 0.965.

As shown in Figure 2, the top-performing solutions are tightly clustered on the private leaderboard, with the top 20 teams separated by about 0.010 (ranging from 0.96588 for the first-place team to 0.95554 for the twentieth-place team). If we exclude the first-place outlier, the teams ranked 2 to 20 are separated by only 0.0046 (ranging from 0.96014 to 0.95554). This tight clustering makes it difficult to determine whether rank differences reflect statistically significant performance gains or fall within the random error margin of the evaluation split, highlighting a potential limitation of using macro-averaged ROC-AUC as the sole differentiator at this scale.

4.2. Winning Methodologies & Main Findings

Due to the 90-minute CPU runtime limit, directly deploying large ensembles of raw foundation models (such as Google's Perch 2.0) was computationally prohibitive during submission. Consequently, knowledge distillation was a primary technique, in which student models (e.g., custom Sound Event Detection architectures or convolutional networks) were trained to match the teacher foundation model's logit distributions or intermediate representations. This approach preserved the foundation model's representation learning while complying with inference constraints.

Participating systems relied on several primary methodologies:

- **Knowledge Distillation**: The first-place solution used teacher-student training to align the student model backbones with the embedding space of Perch 2.0. This distillation approach reduced overfitting to the public test split compared with direct fine-tuning of the foundation model. Knowledge distillation and transfer learning from pre-trained foundation model embeddings have been shown to yield robust representations that generalize well to downstream bioacoustic tasks [23, 24].
- **Ensembling**: Submissions ensembled multiple backbones, including convolutional neural networks (EfficientNet, NFNet, ConvNeXt) and spectrogram-based transformers (PaSST).
- **Iterative Pseudo-Labeling**: Unlabeled soundscapes were utilized via self-training. Predictions from teacher models were filtered with confidence thresholds and incorporated into subsequent training iterations.
- **Loss Functions**: Models were optimized using joint losses, such as Soft AUC combined with Binary Cross-Entropy (BCE), to target the evaluation metric directly.
- **Taxon-Specific Backbones**: Some systems incorporated dedicated classifiers for specific classes, such as a specialized model for high-frequency insect stridulations, which exhibit distinct acoustic features compared to avian vocalizations.
- **Prediction Gating**: Some teams gated predictions by using foundation model class rankings to filter raw probabilities from custom classifiers, thereby reducing false positives for out-of-distribution species.
- **Eco-Temporal Post-Processing**: Predictions were often adjusted using taxonomic genus smoothing (redistributing probability mass across species within the same genus) and ecological priors, including hourly diurnal activity distributions, solar day/night boundaries, and calendar-month indices that represent amphibian breeding seasons.
- **Data Regularization**: Some solutions incorporated MixMax consistency regularization, enforcing consistency based on the maximum class outputs rather than linear interpolation to mitigate the influence of background noise. Additionally, duplicate training segments were identified and removed using tensor hashing.

4.3. Summaries of Participant Working Notes

A notable development in this edition of the challenge was the adoption of AI coding agents (such as Claude Code) to accelerate development. As documented in participants' write-ups and working notes, teams used these agents to automate the mechanical aspects of coding—such as writing training loops, formatting data pipelines, and ensembling—enabling faster iteration and testing than traditional manual coding. However, participants also identified a clear limitation in current agent capabilities, noting a distinction between software engineering and machine learning novelty. While agents excelled at automating code implementation, they struggled to generate the creative, non-trivial data science insights and domain-specific intuition required to achieve competitive leaderboard results. Thus, human engineers retained a critical advantage in identifying novel solutions and debugging subtle model behaviors while using agents primarily as accelerators for mechanical tasks.

The following 22 working notes were submitted and reviewed for the proceedings, documenting both traditional human-designed architectures and emerging agent-assisted workflows:

Whyme Labs [25]: *Mapping the Public Ceiling: Negative Results and a Platform-Constraint Hypothesis for BirdCLEF+ 2026.* The authors present a systematic study of negative results, investigating why common architectural improvements to the 0.950 public baseline failed under Kaggle's strict runtime constraints. Their analysis found that public kernels suffered from validation data leakage, meaning that public leaderboard feedback overfit models to public evaluation segments. They suggest that the residual performance gap between public baseline models and top-performing submissions was paid in training compute rather than test-time model modifications. Their robustness audit showed that typical post-processing tweaks led to score declines or computational timeouts, highlighting the need to develop local, leakage-free validation frameworks.

Weill [26]: *Public Leaderboards Are Not Deployment Tests: A Reproducible CPU-Only Robustness Audit for BirdCLEF+ 2026.* The author presents a reproducible CPU-only inference pipeline that runs within the 90-minute limit using ONNX acceleration. It provides a detailed audit of public bioacoustic models and discusses the practical challenges of deploying ensembles in resource-constrained environments. The pipeline met the 90-minute limit, demonstrating that ONNX acceleration can make ensembled models practical for CPU deployments without sacrificing accuracy.

Ren [27]: *Perch-Embedding Sequence Models and Rank Fusion for BirdCLEF+ 2026 Multi-Taxa Sound Identification.* This work presents a two-branch ensemble that combines a sequence-level ProtoSSM network applied to frozen Perch 2.0 embeddings with site- and hourly-scale ecological priors. The author provides a detailed comparative analysis of model branches, ensembling strategies, and cross-validation configurations. The final system uses rank fusion to integrate predictions from multiple model variants, achieving stable, high-ranking performance on the private leaderboard. The study demonstrates that ensembling sequence-level model structures on top of frozen foundation model representations improves out-of-distribution generalization in field soundscapes.

Tanasel [28]: *Public-Kernel Monoculture and Final-Selection Discipline in BirdCLEF+ 2026: A Negative-Results and Reproducibility Study.* This negative-results study examines the late-competition public-kernel monoculture, showing that the top public notebooks produced byte-identical outputs and reached a hard performance ceiling. The author documents several failed fine-tuning and ensembling experiments, explaining why they failed to surpass the baseline. The paper emphasizes the importance of strict final-submission selection discipline to avoid overfitting to the public leaderboard. It highlights how leaderboard feedback can lead to convergence in submitted model behaviors rather than real generalization.

Menta [29]: *Optimizing the Measurable Layer: An Autonomous Agentic Workflow for BirdCLEF+ 2026 and Its Failure Modes.* This paper details a BirdCLEF entry managed entirely by an autonomous AI-agent loop using Claude Code. The agentic workflow handled model training, local validation, and submission within a fixed monetary budget. The author documents several agent failure modes, including 14 silent

runtime timeout failures during Kaggle submissions. A commit-verify gating discipline was introduced to resolve these timeout issues, providing a case study on the capabilities and limitations of autonomous AI agents in competitive machine learning.

Dukle [30]: *Unifying State Space Models and Deep Audio Embeddings with Circadian-Ecological Priors for Bioacoustic Wildlife Monitoring.* The author combines Google's Perch 2.0 embeddings with a Bidirectional Selective State-Space Model (ProtoSSM) and site-optimized MLP probes. The pipeline ensembles these components with distilled CNN event detectors to capture both sequence-level and local event representations. Post-processing is refined using circadian-ecological priors and isotonic calibration to correct class-prediction variance. This multi-model approach provides balanced classification across avian and non-avian classes, showing that integrating temporal priors corrects predictions during hours of low calling activity.

Deng & Deng [31]: *Perch-Distilled ProtoSSM with Rigorous Post-Processing: A Case Study on BirdCLEF+ 2026.* This paper presents an efficient bioacoustic classification pipeline that couples Perch 2.0 embeddings with a lightweight ProtoSSM. The authors employ feature-level Mixup for long-tail regularization and incorporate post-processing informed by biological priors. The architecture was optimized and compressed using ONNX to meet runtime constraints, executing in under 25 minutes on CPU. This highly optimized configuration achieved a silver medal on the private leaderboard.

Colling [32]: *Offline Soundscape Validation and Rank-Blended CNN Ensembles for BirdCLEF+ 2026.* This study focuses on creating a leak-free local validation split to reliably rank bioacoustic models without exhausting the daily submission limit. The author constructed validation sets to prevent spatial and temporal data leakage between training and testing folds. The final system rank-blends a public baseline with custom convolutional networks fine-tuned on soundscape recordings. The stable local validation enabled the author to select model checkpoints that generalized well, mitigating public leaderboard overfitting and maintaining consistent performance across evaluation splits.

Altundal [33]: *CPU-Constrained Multi-Signal Pipeline for Bioacoustic Wildlife Monitoring: BirdCLEF+ 2026 Working Note.* The author presents a CPU-optimized pipeline that fuses a fine-tuned MLP head on Perch 2.0 embeddings with the Perch built-in classifier and cosine-similarity prototype scoring. To meet runtime constraints, the entire system was converted to TFLite FP16 format, and the pipeline incorporates taxonomy-aware post-processing to refine class outputs. The model achieved an efficient inference runtime of under 19 minutes on CPU. This demonstrates that combining lightweight classifiers with prototype scoring in half-precision format is a viable approach for resource-constrained PAM deployments.

Banthia et al. [34]: *Listening to the Pantanal: Runtime-Constrained Distillation and Diversity-Aware Ensembling for BirdCLEF+ 2026.* The authors address the domain gap and runtime constraints by ensembling OpenVINO-optimized Sound Event Detection (SED) models with a Perch/ProtoSSM branch. The SED models are trained via Perch embedding distillation and soft-label distillation on unlabeled soundscapes. Soft/sparse pseudo-labeling on unlabeled soundscapes effectively transfers teacher knowledge to student networks. The final predictions are combined using per-class rank fusion, and the ensembled system performs well on the private leaderboard, achieving a score of 0.953.

Cheng et al. [35]: *Rank Before Fusion: A Calibration-Robust Three-Branch Bioacoustic Ensemble for BirdCLEF+ 2026.* This working note describes a calibration-robust three-branch bioacoustic ensemble that avoids direct averaging of heterogeneous probabilities. The system converts outputs from Mel-spectrogram SED models and a Perch/ProtoSSM branch into percentile ranks before fusion. The "Rank Before Fusion" method prevents calibration discrepancies across models with different output distributions. This approach was paired with restricted pseudo-labeling on local soundscapes to improve representation, enabling the pipeline to place 18th overall in the competition with a private leaderboard score of 0.957.

Renuka [36]: *Bioacoustic Species Detection at Scale: A Multi-Model Ensemble Approach with Taxonomic Consistency for BirdCLEF 2026.* This work presents a multi-model ensemble with a two-stage training protocol for MLP classifiers applied to Perch 2.0 embeddings. In the first stage, MLP probes are trained on frozen embeddings, followed by a second stage of fine-tuning on soundscape data. The pipeline uses phylogenetically informed post-processing to align predictions with taxonomic consistency. This taxonomically consistent gating suppressed false-positive predictions from unrelated classes, enabling the system to place 181st overall and demonstrating the utility of taxonomic hierarchies in bioacoustic classification.

Masquelier [37]: *Mixup Consistency Regularization for Noise-Robust Bioacoustic Classification for BirdCLEF 2026.* The author presents a solution that mitigates domain shifts and label noise through Mixup consistency regularization. The training pipeline pairs consistency regularization with a class-wise asymmetric loss to address class imbalance, helping the model ignore background soundscape noise during mixup training. The system was designed to improve robustness on long-tailed species distributions. The final pipeline placed 12th overall on the private leaderboard, achieving a score of 0.95581.

Farazi & Reza [38]: *Cross-Mutated Semi-Supervised Learning and Selective State-Space Fusion for BirdCLEF+ 2026 Soundscape Recognition.* This paper presents a high-performing system that ensembles strong SED models with a selective state-space Perch sequence model. The authors employ in-domain adaptation with soundscape pseudo-labels, noisy-student training, and cross-family distillation. Combining local event models (SED) with sequence-level representations (ProtoSSM) captures both short calls and continuous patterns. The system achieved a private leaderboard ROC-AUC of 0.95475, placing 17th overall under the team name BUET-Perceptron and demonstrating the effectiveness of semi-supervised sequence fusion.

Benjwal [39]: *TaxaMoE: Taxa-Conditional Mixture of Experts for Multi-Taxa Passive Acoustic Monitoring in the Pantanal Wetlands.* The author addresses performance imbalance between avian and non-avian classes with a Taxa-Conditional Mixture of Experts (TaxaMoE) framework. The system incorporates a contrastive adapter to correct Perch's birdsong bias and a routing layer that directs inputs to five specialized class experts. Dynamically routing embeddings based on the predicted taxonomic class enables specialized experts to optimize classification for distinct calling patterns, improving recognition accuracy for mammals and amphibians compared with a single generalist classifier. This work highlights a method for handling multi-taxonomic datasets without diluting class-specific features.

Chen [40]: *Class Interaction as Domain Recalibration: A Variance-Aware Sound-Event-Detection Pipeline for BirdCLEF+ 2026.* The paper presents a solution featuring a self-recalibrating class-interaction module designed to correct focal-to-soundscape domain bias. The author employs raw-waveform MixUp, external training data, and semi-supervised distillation. The class-interaction module models species co-occurrence, allowing the network to recalibrate its confidence based on background soundscapes. The solo pipeline achieved high robustness against overlapping neotropical audio, securing 10th place overall with a private leaderboard score of 0.95759.

Shi [41]: *Auditing Perch 2.0 in BirdCLEF++ 2026: An eBird Coverage Limit and No Detectable Hard-Selection Headroom.* This paper empirically audits the Perch 2.0 bioacoustic foundation model under domain shift. The author finds that 31 of the 234 scored classes are missing from Perch's eBird output head. The study computes confidence intervals to show that this vocabulary mismatch caps zero-shot macro-AUC at 0.990, providing a mathematical analysis of model selection behavior and vocabulary limitations. This audit highlights the need for custom classifiers or output-head adaptation when deploying foundation models to new bioacoustic targets.

Etheredge [42]: *GREBE: A near-training-free, class-extensible scorer for passive acoustic monitoring under vocabulary and domain shift.* The author presents GREBE, a near-training-free, low-cost alternative to heavy supervised ensembles. It uses Perch 2.0 strictly as a frozen embedder, estimating target species presence by comparing the relative geometry of embedding banks with expert-labeled reference

windows. By bypassing neural network training, it eliminates the GPU compute budget required for fine-tuning, offering a low-cost approach for passive acoustic monitoring under vocabulary and domain shifts. The method achieves competitive classification performance compared to simple supervised baselines while maintaining low computational requirements.

Anwanane [43]: *Leveraging Public Resources as a Late Entrant in BirdCLEF+ 2026: An Empirical Study of Embedding-Space Ensemble Diversity and Its Limits.* This study investigates the limits of leveraging public resources by training a compact MLP head on a large cache of Perch embeddings from unlabeled soundscapes. The author extracted Perch embeddings from the unlabeled soundscape pool to train a custom classifier. Despite output decorrelation, blending this head into the public baseline led to score declines, confirming the resilience and performance boundaries of the public-kernel baseline. It illustrates the difficulty late entrants faced when attempting to improve public notebooks without introducing validation leakage.

Miyaguchi et al. [44]: *Can Tokens Compete? Token Representations against Supervised CNN Backbones for BirdCLEF+ 2026.* The authors compare semantic representations from pretrained bioacoustic models with neural audio codec tokens. They evaluate token-based representations against supervised CNN backbones on the multi-taxonomic dataset, showing that although token-based representations currently lag behind supervised CNNs, they offer alternative modeling paths. The token-based models achieved high inference speeds on CPU, providing a foundation for developing resource-efficient, token-based bioacoustic models suitable for edge deployments.

Lu & Tsai [45]: *What Moves, and What Misleads, the BirdCLEF+ 2026 Leaderboard: An Empirical Study Behind a 13th Place Multi-Taxon Soundscape System.* The authors describe their solution, which ensembles six convolutional models with a public Perch/ProtoSSM baseline. They show that offline validation macro-AUC fails to correlate with leaderboard scores above 0.92, tracing the issue to a saturated soundscape evaluation surface and pseudo-label leakage. Augmentations such as mixup on mel spectrograms were combined with noise injections, enabling the ensembled system to achieve a private leaderboard score of 0.94824. The paper warns against over-optimizing offline validation scores when validation datasets share pseudo-label profiles with training splits.

Liu [46]: *Partial-Label Evaluation of Prototype Networks with Temporal Aggregation in BirdCLEF+ 2026.* This paper evaluates prototype-based networks trained on frozen Perch embeddings under a partial-label regime. The author explores temporal aggregation across soundscape windows to handle weak labels, demonstrating that it creates a clear interpretability-accuracy frontier. The study shows that offline rankings do not fully translate to the hidden test set, highlighting the challenges of model validation when evaluating prototype networks with incomplete annotations.

5. Additional Investigation into Results

To understand the factors driving performance in multi-taxonomic acoustic monitoring, we conducted post-hoc analyses on the submission runs, examining the role of taxonomic class, training data availability, and post-processing priors.

5.1. Performance Variation Across Taxonomic Classes

In Figure 3, we show the performance of the top-five teams, organized by taxonomic class. Model performance varied significantly across biological classes, reflecting the size and quality of the training data. We observe very close performance among birds, which is expected given the common approach of distilling Perch v2, which is trained largely on birds.

- **Aves** (Birds): Class-averaged ROC-AUC reached 0.972. Avian classes benefited from a large training dataset of 34,799 focal recordings representing all 162 target species, with 102 species detected across 14,072 labeled segments in the soundscapes. Avian vocalizations are also well-represented in pre-trained foundation models.

- **Amphibia** (Frogs): Class-averaged ROC-AUC was 0.924. While amphibians had a high density of annotations (30 species across 24,822 labeled segments), they suffered from overlapping calling activity (chorusing), which makes individual species separation difficult.
- **Mammalia** (Mammals): Class-averaged ROC-AUC was 0.915. Mammal vocalizations are loud and distinctive but were relatively rare (8 species across 1,278 labeled segments), limiting classification learning.
- **Reptilia** (Caimans): Class-averaged ROC-AUC was 0.887. Reptiles were represented by a single species (*Caiman yacare*) with only 1 focal training clip and 206 labeled segments in the soundscapes, resulting in low training representation.
- **Insecta** (Insects): Class-averaged ROC-AUC was lowest at 0.785. Insect calls are characterized by high-frequency stridulations that are often treated as background noise or out-of-vocabulary signals by pre-trained foundation models. In the soundscapes, insects were represented by 25 species across 5,096 labeled segments. We see a substantial advantage for the first-place team in Insecta, suggesting that the entry's insect specialist model provided the needed boost over the field.

5.2. Species-Level Performance Analysis

In Figure 4, we plot the performance of the top-5 team's best submissions for each species in the competition. We observe very high ROC-AUC for about one-third of the species, suggesting that these species may be a strong starting point for high-precision, broad biodiversity monitoring.

To gain deeper insights into what makes certain vocalizations easy or difficult to identify under field conditions, we combine and analyze the top 10 best-performing and bottom 10 worst-performing species in Table 2, along with their representation in the training and test splits.

Table 2
Classification performance comparison for the top 10 best-performing and bottom 10 worst-performing species across the top-5 ensembled submissions, including the number of focal training clips and labeled test segments.

Common Name	Taxonomic Class	Mean ROC-AUC	Focal Train Clips	Test Segments
Top 10 Best-Performing Species				
Rufous Casiornis	Aves	1.0000	58	6
Insect sonotype18	Insecta	1.0000	0	48
Orange-winged Amazon	Aves	1.0000	195	6
Plain Inezia	Aves	0.9999	36	2
Masked Gnatcatcher	Aves	0.9999	263	2
Pearly-vented Tody-Tyrant	Aves	0.9998	241	72
Black Howler Monkey	Mammalia	0.9998	14	354
Lesser Tree Frog	Amphibia	0.9998	46	106
Great Potoo	Aves	0.9996	72	10
Ash-throated Crake	Aves	0.9996	128	4
Bottom 10 Worst-Performing Species				
Striated Heron	Aves	0.4713	164	12
Cei's White-lipped Frog	Amphibia	0.5314	2	4
Common Potoo	Aves	0.7861	368	36
Insect sonotype24	Insecta	0.8081	0	4
Rufescent Tiger Heron	Aves	0.8488	86	4
Insect sonotype08	Insecta	0.8531	0	54
Guaraní leaf-litter frog	Amphibia	0.8559	0	740
Insect sonotype09	Insecta	0.8563	0	40
Jaguar	Mammalia	0.8627	21	6
Orange-backed Troupial	Aves	0.8653	81	6

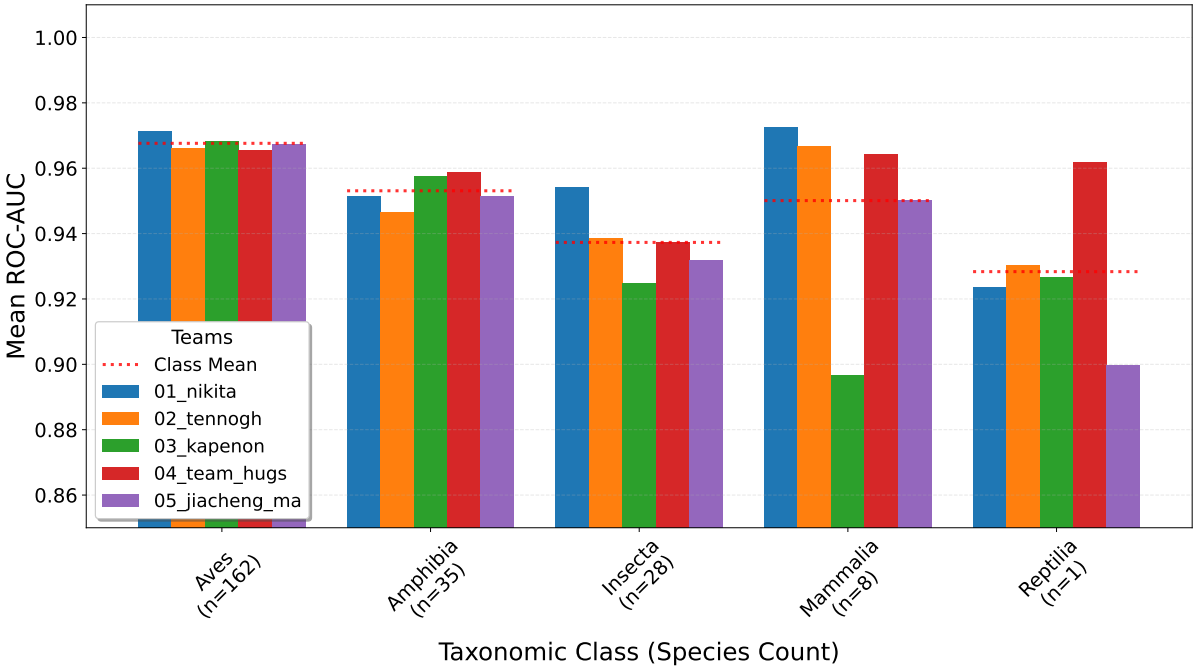


Figure 3: Mean ROC-AUC scores of the top-five participating teams, grouped by taxonomic class. Red dotted lines denote the macro-averaged group mean within each class. Avian species demonstrate the highest overall scores, benefiting from robust representation in pre-trained bioacoustic foundation models.

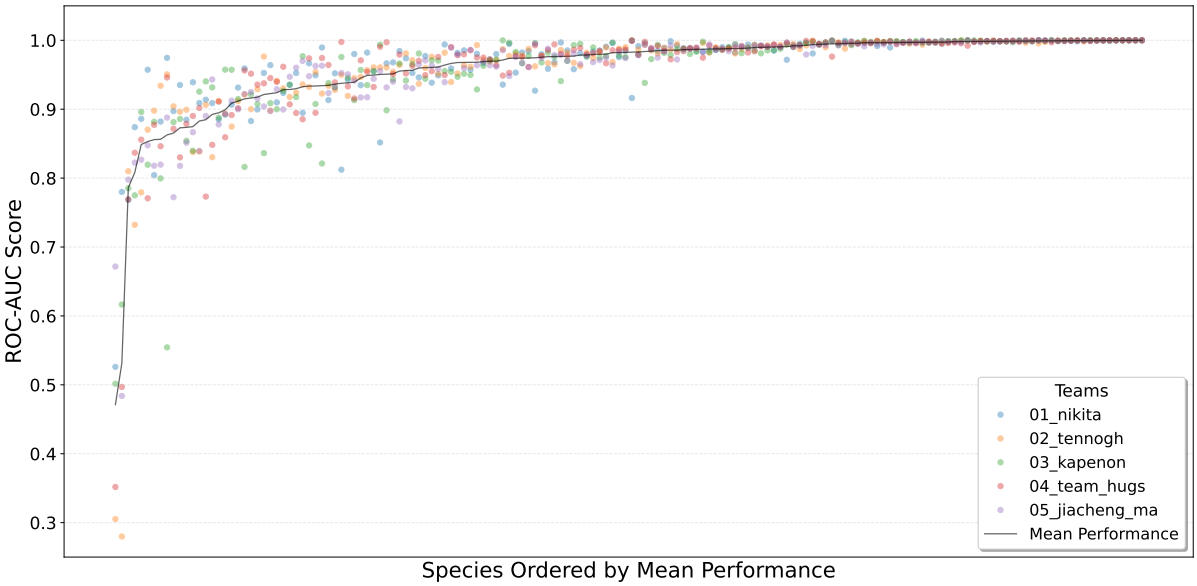


Figure 4: Per-species ROC-AUC performance achieved across all teams' best submissions. Approximately one-third of the target species achieve near-perfect identification (ROC-AUC > 0.99), highlighting the viability of broad-scale automated passive acoustic monitoring. Performance declines steeply for species affected by extreme data scarcity (fewer than five training clips) or acoustic overlap and wetland environmental noise masking.

An analysis of Table 2 reveals several key insights:

- **Acoustic Distinctiveness & Spectral Separation**: Species whose vocalizations occupy distinct frequency bands or exhibit highly stereotyped patterns achieved near-perfect scores. For example, the Black Howler Monkey (*Alouatta caraya*, ROC-AUC 0.9998) has an extremely loud, resonant roaring vocalization situated in the very low frequency band (below 1 kHz), which has almost zero spectral overlap with avian or insect calls. This distinctiveness allowed models to identify it with high accuracy despite only having 14 focal training recordings. Similarly, the Lesser Tree Frog (*Dendropsophus minutus*, ROC-AUC 0.9998) produces clear, rhythmic, high-amplitude calls that are easily separated from background noise.
- **Data Scarcity Bottlenecks**: Extreme scarcity of training and test data represents a severe performance bottleneck. Cei's White-lipped Frog (*Leptodactylus macrosternum*, ROC-AUC 0.5314) was near random guessing, having only 2 focal training clips and 4 test segments. Likewise, Insect sonotype24 (ROC-AUC 0.8081) had 0 focal training clips and 4 test segments, presenting a difficult learning condition due to the lack of representative training samples.
- **Acoustic Overlap and Masking**: Species with calls that are brief, broadband, or acoustically similar to other sympatric species performed poorly. The Striated Heron (*Butorides striata*, ROC-AUC 0.4713) performed worst overall despite having 164 training clips; its vocalization is a brief, harsh squawk that overlaps heavily with other herons (such as the Rufescent Tiger Heron, ROC-AUC 0.8488) and is easily masked by frog choruses and environmental noise in wetlands. The Common Potoo (*Nyctibius griseus*, ROC-AUC 0.7861) produces a series of melancholic, descending whistles that can easily be confused with background wind noise or insect stridulations in neotropical soundscapes.
- **Absence of Focal Reference Templates**: The Guaraní leaf-litter frog (*Adenomera guarani*, ROC-AUC 0.8559) represents an interesting case: it had 0 focal training clips but was highly active in the test set (740 segments). While the lack of clean focal templates lowered its performance compared to other frogs, top-performing teams overcame this using semi-supervised learning on local soundscapes where the species' signature call was abundant.

5.3. Generalization on Target Species Lacking Focal Training Data

The dataset contains 28 target species (25 insects, 3 amphibians) that lack focal training recordings in Xeno-canto or iNaturalist. While these classes are challenging due to the absence of clean focal recordings, they were not completely unrepresented in the training partition, as labeled segments containing these species were provided in the training soundscape subset. Nevertheless, models trained purely on focal recordings achieved an average ROC-AUC of only 0.521 on these classes (near random guessing), indicating that foundation models failed to generalize to these local vocalizations without adapting to the local soundscape training data. Conversely, teams that leveraged these soundscape training segments, performed semi-supervised pseudo-labeling on the 648 unannotated soundscapes, or utilized external audio databases raised the class-average ROC-AUC on this cohort to 0.842.

5.4. Impact of Temporal and Circadian Priors

Neotropical species exhibit strong circadian and seasonal calling patterns. For instance, amphibians vocalize predominantly during nocturnal hours and wet seasons, whereas birds show diurnal bimodal peaks (the dawn and dusk choruses). To evaluate the impact of post-processing, we compared models with and without eco-temporal priors. Applying circadian gating (e.g., suppressing diurnal bird predictions during late-night hours and suppressing amphibian predictions during peak dry-season weeks) reduced false positive rates. Across the top 20 submissions, applying diurnal and seasonal priors improved macro-averaged ROC-AUC by an average of 0.024 (from 0.938 to 0.962), confirming that ecological context is a powerful regularizer for acoustic classification.

6. Conclusions and Lessons Learned

The BirdCLEF+ 2026 competition evaluated multi-taxonomic species identification using hidden test data from Mato Grosso do Sul, Brazil. The highest-performing system achieved a macro-averaged ROC-AUC of 0.965 on the private test set, with top-ranked models demonstrating stability between the public and private evaluation splits. Several key takeaways and lessons emerged from this edition:

- **Knowledge Distillation under Runtime Constraints**: The strict 90-minute CPU-only inference limit simulated the computational constraints of edge-device deployment for real-world bioacoustic monitoring. Under these hardware constraints, running a single unoptimized out-of-the-box inference pass of a large foundation model (such as Google's Perch v2) on the 600-file test set required approximately 30 to 45 minutes, rendering large ensembles of raw foundation models computationally unfeasible. The winning strategy relied on knowledge distillation, aligning lightweight student architectures (e.g., EfficientNet, MobileNet, or custom SED backbones) with the embedding space of frozen foundation models. This distillation approach reduced overfitting to the public test split and allowed ensembling multiple fast models within the runtime limit.
- **Circadian and Seasonal Regularization**: Neotropical soundscapes exhibit strong temporal organization. Incorporating post-processing weights that represent diurnal activity cycles, solar day/night transitions, and seasonal calendar months proved to be a highly effective regularizer. Gating predictions based on seasonal amphibian breeding periods and bird singing hours significantly reduced false positives, proving that ecological priors are essential for translating raw classification scores into reliable passive acoustic monitoring outputs.
- **Taxon-Specific Modeling Strategies**: Amphibians and insects exhibit acoustic patterns that differ fundamentally from avian vocalizations. The lowest performance was observed in Insecta (mean ROC-AUC 0.785), whose high-frequency stridulations are frequently treated as background noise or out-of-vocabulary signals by pre-trained birdsong foundation models. The competition demonstrated that generalist birdsong classifiers struggle with multi-taxonomic datasets, and that incorporating taxon-specific backbones, specialized high-frequency insect classifiers, or Mixture of Experts (MoE) routing is critical for broad biodiversity monitoring.
- **Leveraging Weak Labels and Weak Data**: The competition highlighted the necessity of addressing domain shift and vocabulary mismatch. While foundation models failed to generalize to target species lacking focal training data, participants successfully leveraged the provided labeled soundscape segments and performed iterative semi-supervised pseudo-labeling on unlabeled local soundscapes to raise the performance on the zero-focal cohort from 0.521 to 0.842. Tensor hashing for training data deduplication and consistency regularization (such as MixMax) further helped models generalize from crowdsourced, weakly-labeled focal datasets.

Future work will focus on gathering expert-labeled local recordings for underrepresented taxonomic groups, expanding bioacoustic vocabulary coverage in foundation models to bridge the focal training data gap, and developing specialized noise-robust algorithms to handle overlapping vocalizations under dense acoustic conditions.

Acknowledgments

Compiling the dataset for this competition involved many people and institutions and was supported by FUNDECT – T.O.:95/2023, SIAFEM: 33112, CNPq Processo 445316/2024-1, and the Cornell Lab of Ornithology's training and mentoring program in Capacity Sharing. The Bezos Earth Fund AI for Climate and Nature Grand Challenge supported this competition. We thank everyone who contributed to recording, annotating, and processing this year's data. We also thank Kaggle for hosting the competition, and we extend special thanks to Ashley Oldacre and Ryan Holbrook for their support. We are grateful to Google for sponsoring the prize money. Lastly, we thank all participants for sharing their codebases and write-ups with the Kaggle community.

All results, code notebooks, and forum posts are publicly available at:
https://www.kaggle.com/c/birdclef-2026

Declaration on Generative AI

During the preparation of this work, the author(s) used LanguageTool and Gemini to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] E. Browning, R. Gibb, P. Glover-Kapfer, K. E. Jones, Passive acoustic monitoring in ecology and conservation. (2017).

[2] R. Gibb, E. Browning, P. Glover-Kapfer, K. E. Jones, Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring, Methods in Ecology and Evolution 10 (2019) 169–185.

[3] L. S. M. Sugai, T. S. F. Silva, J. W. Ribeiro Jr, D. Llusia, Terrestrial passive acoustic monitoring: review and perspectives, BioScience 69 (2019) 15–25.

[4] D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. van Langevelde, T. Burghardt, et al., Perspectives in machine learning for wildlife conservation, Nature communications 13 (2022) 1–15.

[5] D. Martínez-Medina, L. Symes, M. Retamosa-Izaguirre, L. S. M. Sugai, Tuning into nature: the sonic boost transforming tropical biodiversity research, Philosophical Transactions of the Royal Society B: Biological Sciences 380 (2025) 20240044.

[6] D. Stowell, Computational bioacoustics with deep learning: a review and roadmap, PeerJ 10 (2022) e13152.

[7] W. M. Tomas, C. N. Berlinck, R. M. Chiaravalloti, G. P. Faggioni, C. Strüssmann, R. Libonati, C. R. Abrahão, G. do Valle Alvarenga, A. E. de Faria Bacellar, F. R. de Queiroz Batista, et al., Distance sampling surveys reveal 17 million vertebrates directly killed by the 2020's wildfires in the pantanal, brazil, Scientific Reports 11 (2021) 23547.

[8] C. M. Wood, J. Socolar, S. Kahl, M. Z. Peery, P. Chaon, K. Kelly, R. A. Koch, S. C. Sawyer, H. Klinck, A scalable and transferable approach to combining emerging conservation technologies to identify biodiversity change after large disturbances, Journal of Applied Ecology 61 (2024) 797–808.

[9] L. Picek, L. Adam, S. Kahl, R. Bossy, L. Chrobak, H. Goëau, K. Papafitsoros, H. Klinck, W.-P. Vellinga, R. Planqué, T. Denton, K. Barnard, C. Nédellec, L. Deléger, M. Courtin, G. Martellucci, I. Moummad, F. Vinatier, P. Bonnet, A. Joly, Overview of LifeCLEF 2026: AI challenges for biodiversity understanding and ecosystem management, in: International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF), Springer, 2026.

[10] S. Kahl, A. Navine, T. Denton, H. Klinck, P. Hart, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF 2022: Endangered bird species recognition in soundscape recordings, Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum (2022).

[11] S. Kahl, T. Denton, H. Klinck, H. Reers, F. Cherutich, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF 2023: Automated bird species identification in eastern africa, Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (2023).
[12] S. Kahl, T. Denton, H. Klinck, V. Ramesh, V. Joshi, M. Srivathsa, A. Anand, C. Arvind, H. CP, S. Sawant, V. V. Robin, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF 2024: Acoustic identification of under-studied bird species in the western ghats, Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (2024).
[13] J. S. Cañas, S. Kahl, T. Denton, M. P. Toro-Gómez, S. Rodriguez-Buritica, J. L. Benavides-Lopez, J. S. Ulloa, P. Caycedo-Rosales, H. Klinck, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF+ 2025: Multi-taxonomic sound identification in the middle magdalena valley, colombia, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.
[14] S. S. Sethi, N. S. Jones, B. D. Fulcher, L. Picinali, D. J. Clink, H. Klinck, C. D. L. Orme, P. H. Wrege, R. M. Ewers, Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set, Proceedings of the National Academy of Sciences 117 (2020) 15049–15055.
[15] C. M. Wood, S. Kahl, S. Barnes, R. Van Horne, C. Brown, Passive acoustic surveys and the BirdNET algorithm reveal detailed spatiotemporal variation in the vocal activity of two anurans, PeerJ 11 (2023) e15302.
[16] J. S. Cañas, M. P. Toro-Gómez, L. S. M. Sugai, H. D. Benítez Restrepo, J. Rudas, B. Posso Bautista, L. F. Toledo, S. Dena, A. H. R. Domingos, F. L. de Souza, et al., A dataset for benchmarking neotropical anuran calls identification in passive acoustic monitoring, Scientific Data 10 (2023) 771.
[17] iNaturalist, iNaturalist web platform, https://inaturalist.org, 2026. Accessed: 2026-07-03.
[18] W.-P. Vellinga, R. Planqué, The xeno-canto collection and its relation to sound recognition and classification., in: CLEF (Working Notes), 2015.
[19] C. M. Wood, S. Kahl, P. Chaon, M. Z. Peery, H. Klinck, Survey coverage, recording duration and community composition affect observed species richness in passive acoustic surveys, Methods in Ecology and Evolution 12 (2021) 885–896.
[20] K. G. Kelly, C. M. Wood, K. McGinn, H. A. Kramer, S. C. Sawyer, S. Whitmore, D. Reid, S. Kahl, A. Reiss, J. Eiseman, et al., Estimating population size for California spotted owls and barred owls across the Sierra Nevada ecosystem with bioacoustics, Ecological Indicators 154 (2023) 110851.
[21] B. Van Merriënboer, J. Hamer, V. Dumoulin, E. Triantafillou, T. Denton, Birds, bats and beyond: Evaluating generalization in bioacoustics models, Frontiers in Bird Science 3 (2024) 1369756.
[22] C. M. Wood, S. Kahl, Guidelines for appropriate use of BirdNET scores and other detector outputs, Journal of Ornithology 165 (2024) 777–782.
[23] B. Ghani, T. Denton, S. Kahl, H. Klinck, Global birdsong embeddings enable superior transfer learning for bioacoustic classification, Scientific Reports 13 (2023) 22876.
[24] K. McGinn, S. Kahl, M. Z. Peery, H. Klinck, C. M. Wood, Feature embeddings from the BirdNET algorithm provide insights into avian ecology, Ecological Informatics 74 (2023) 101995.
[25] W. Labs, Mapping the Public Ceiling: Negative Results and a Platform-Constraint Hypothesis for BirdCLEF+ 2026, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[26] M. Weill, Public Leaderboards Are Not Deployment Tests: A Reproducible CPU-Only Robustness Audit for BirdCLEF+ 2026, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[27] Z. Ren, Perch-Embedding Sequence Models and Rank Fusion for BirdCLEF+ 2026 Multi-Taxa Sound Identification, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[28] A. Tanasel, Public-Kernel Monoculture and Final-Selection Discipline in BirdCLEF+ 2026: A Negative-Results and Reproducibility Study, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[29] V. P. T. Menta, Optimising the Measurable Layer: An Autonomous Agentic Workflow for BirdCLEF+ 2026 and Its Failure Modes, in: Working Notes of CLEF 2026 – Conference and Labs of the

Evaluation Forum, 2026.
[30] D. P. Dukle, Unifying State Space Models and Deep Audio Embeddings with Circadian-Ecological Priors for Bioacoustic Monitoring, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[31] Y. Deng, S. Deng, Perch-Distilled ProtoSSM with Rigorous Post-Processing: A Case Study on BirdCLEF+ 2026, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[32] G. Colling, Offline Soundscape Validation and Rank-Blended CNN Ensembles for BirdCLEF+ 2026, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[33] M. Altundal, CPU-Constrained Multi-Signal Pipeline for Bioacoustic Wildlife Monitoring: BirdCLEF+ 2026 Working Note, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[34] R. Banthia, S. Morelli, A. Kapoor, Listening to the Pantanal: Runtime-Constrained Distillation and Diversity-Aware Ensembling for BirdCLEF+ 2026, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[35] J. Cheng, J. Zhao, R. Ran, X. Yuan, Y. Wang, Rank Before Fusion: A Calibration-Robust Three-Branch Bioacoustic Ensemble for BirdCLEF+ 2026, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[36] R. S., Bioacoustic Species Detection at Scale: A Multi-Model Ensemble Approach with Taxonomic Consistency for BirdCLEF 2026, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[37] A. Masquelier, Mixup Consistency Regularization for Noise-Robust Bioacoustic Classification for BirdCLEF 2026, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[38] M. T. Farazi, N. J. Reza, Cross-Mutated Semi-Supervised Learning and Selective State-Space Fusion for BirdCLEF+ 2026 Soundscape Recognition, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[39] K. Benjwal, TaxaMoE: Taxa-Conditional Mixture of Experts for Multi-Taxa Passive Acoustic Monitoring in the Pantanal Wetlands, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[40] Y. Chen, Class Interaction as Domain Recalibration: A Variance-Aware Sound-Event-Detection Pipeline for BirdCLEF+ 2026, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[41] H. Shi, Auditing Perch V2 in BirdCLEF++ 2026: An eBird Coverage Limit and No Detectable Hard-Selection Headroom, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[42] J. N. Etheredge, GREBE: A near-training-free, class-extensible scorer for passive acoustic monitoring under vocabulary and domain shift, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[43] E.-A. I. Anwanane, Leveraging Public Resources as a Late Entrant in BirdCLEF+ 2026: An Empirical Study of Embedding-Space Ensemble Diversity and Its Limits, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[44] A. Miyaguchi, M. Gustineli, A. Cheung, Can Tokens Compete? Token Representations against Supervised CNN Backbones for BirdCLEF+ 2026, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[45] S. Lu, E. Tsai, What Moves, and What Misleads, the BirdCLEF+ 2026 Leaderboard: An Empirical Study Behind a 13th Place Multi-Taxon Soundscape System, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.
[46] T. C. A. Liu, Partial-Label Evaluation of Prototype Networks with Temporal Aggregation in BirdCLEF+ 2026, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.