BIRDCLEF+ 2026 : Acoustic species identification in the Pantanal

Wildlife species identification from passive acoustic monitoring

TEAM 2

Donggeun · Haiwon · Seunghoon · Woohyuk

Programming for Data Science

Project overview

ORGANIZER

Kaggle × CLEF

Supported by Google DeepMind

PARTICIPANTS

4,955

across 4,001 teams

PRIZE POOL

$45,000

1st place — $15,000

DURATION

Mar 11 – Jun 3

2026 (ongoing)

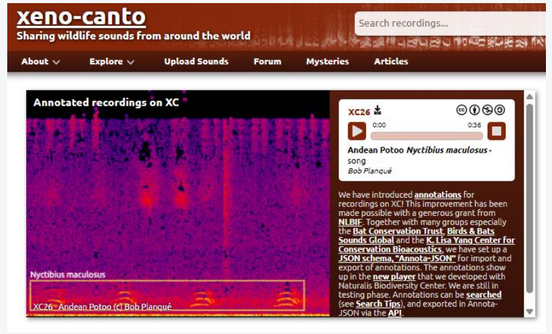
What is BirdCLEF+?

The Pantanal wetlands span 150,000+ km² across Brazil — home to over 650 bird species and countless other animals. A growing network of acoustic recorders runs continuously to capture wildlife sounds across habitats and seasons.

The challenge

The volume of audio is too large to review manually, and labeled species data is scarce. The goal: develop ML models that identify 234 wildlife species from 5-second windows of passive acoustic monitoring data.

- Training data comes from two citizen science platforms, ‘Xeno-Canto’ and ‘iNaturalist’
  - People upload wildlife sound recordings from around the world there





Competition



SIGNAL

1,000+ acoustic recorders deployed across the Pantanal wetlands, running continuously.

METHOD

ML species identification

234 species, 5-second windows, macro-averaged ROC-AUC.

IMPACT

Scalable biodiversity monitoring for climate research, conservation, ecosystem management.

Conventional monitoring across vast, remote regions is expensive and logistically demanding.
This is a real, unsolved scientific bottleneck — not a toy problem.

Task definition

One 1-minute test soundscape = 12 consecutive 5-second prediction windows

W01 0–5s	W02 5–10	W03 10–15	W04 15–20	W05 20–25	W06 25–30	W07 30–35	W08 35–40	W09 40–45	W10 45–50	W11 50–55	W12 55–60

↓ **for each window: output P(species present) for all 234 species**

INPUT	5-second audio windows from 1-minute field recordings · 32 kHz, .ogg file format
OUTPUT	Probability of presence for each of **234** species · Aves, Amphibia, Mammalia, Insecta, Reptilia
METRIC	Macro-averaged ROC-AUC · skips classes with no true positives in test set
CONSTRAINT	CPU notebook ≤ 90 min · GPU disabled, no internet, pre-trained models allowed

Class imbalance is severe — some insect sonotypes and rare reptiles have fewer than 10 positive training examples, and 50%+ of 5-second windows contain no vocalization at all.

EDA · dataset structure & class imbalance

TRAIN CLIPS
35,549
Xeno-Canto
+ iNaturalist

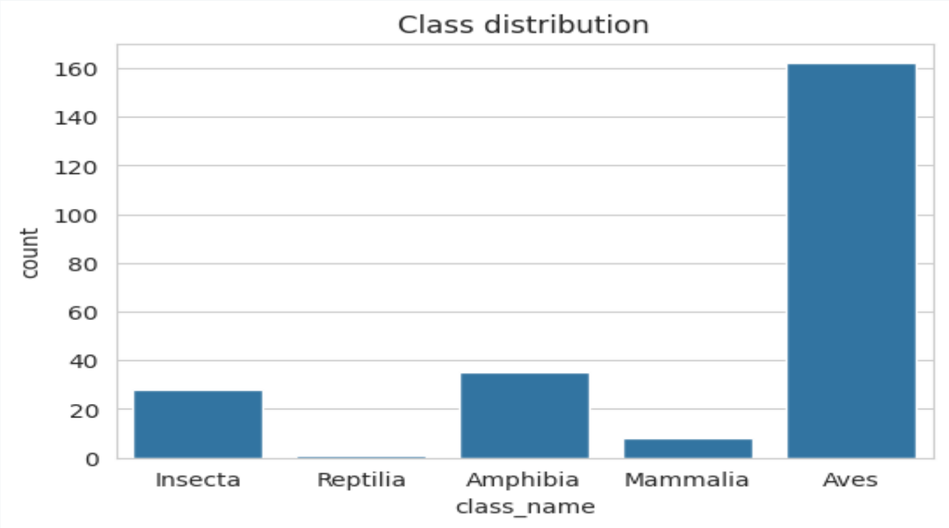
SPECIES (TRAIN / TARGET)
206 / 234
28 species absent

SOUNDSCAPE SEGS
1,478
all 5-sec, expert labeled

X-C SHARE
64.8%
23,043 clips

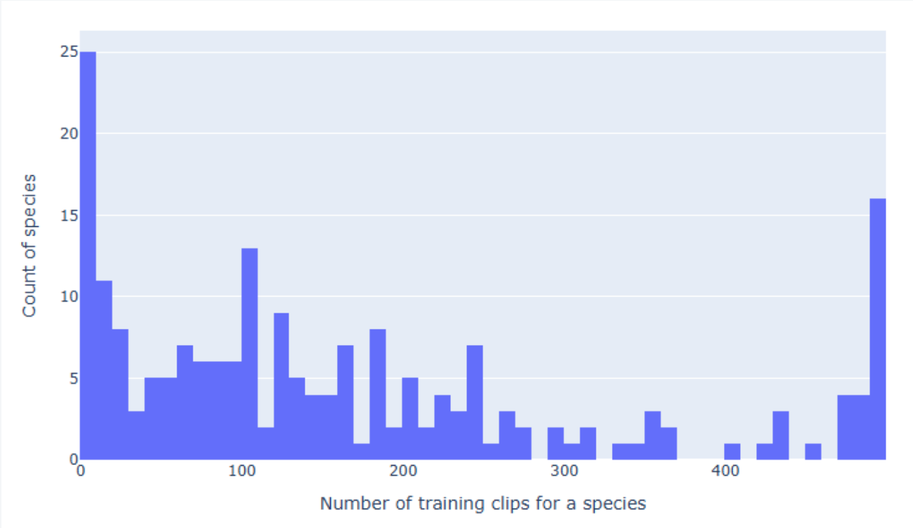
iNatural SHARE
35.2%
12,506 clips

SPECIES BY TAXONOMIC CLASS



Not a bird-only challenge. Five taxa, fundamentally different acoustic profiles — the model must generalize across call types.

CLIPS PER SPECIES — LONG-TAIL



Heavy class imbalance. Rare species are easy to underfit and poorly calibrated — **focal loss function** tuning is essential, not optional.

EDA · the domain gap

TRAIN_AUDIO — WHAT WE TRAIN ON

- 35,549 individual species clips
- Single species per clip, edited
- XC 64.8% + iNat 35.2%
- Curated, rated quality
- Recorded **globally**

Validation AUC 0.9588

→

DOMAIN GAP

Δ = 0.1718

the model learned recording artifacts, not species acoustics.

TEST SOUNDSCAPES — WHAT WE PREDICT ON

- ~600 files, 1 minute each
- Multiple species vocalizing at once
- Wind, rain, background noise
- Field-collected, raw outdoor audio
- **Pantanal, Brazil only**

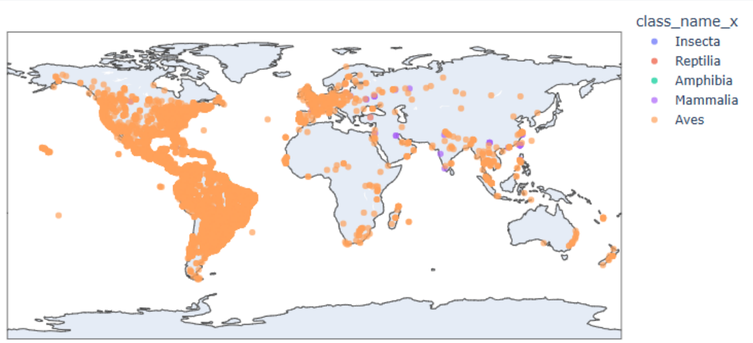
Actual LB score 0.787

SPECIES COVERAGE MISMATCH

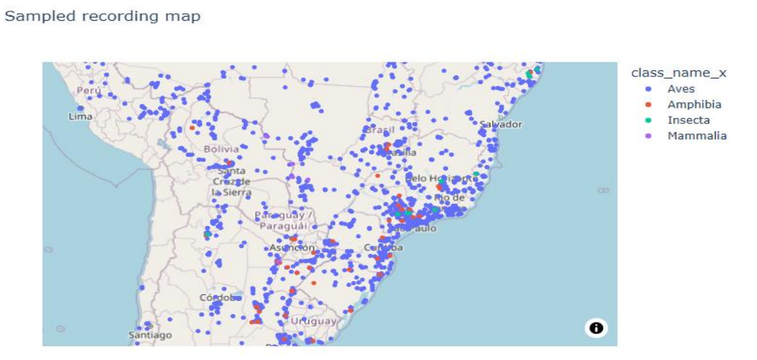
Category	Count
Species with individual clips	206
Species in labeled soundscapes	75
In both	47
Only in soundscapes — NOT in 'train_audio' ★	**28**
Only in individual clips	159

28 target species have zero clips in 'train_audio':
learnable only from soundscape label files

TRAIN LOCATIONS VS. TEST REGION



TRAINED GLOBALLY → TESTED LOCALLY



Base model · Perch v2

Why domain matters more than model size.

ARCHITECTURE	EfficientNet-B3 · ~12M parameters
TRAINING DATA	Xeno-Canto · iNaturalist · Animal Sound Archive
COVERAGE	~15,000 species globally
OUTPUT	5-sec audio → 1,536-dimensional feature vector
FORMAT	ONNX format · optimized for CPU inference

CPU strategy — embedding caching. Extract embeddings once, save to disk, load at submission. No recomputation during the 90-minute window. This single decision makes the entire pipeline feasible on CPU.

① DOMAIN-MATCHED PRE-TRAINING

- Trained on millions of wildlife vocalizations — the same acoustic domain as our test soundscapes.
- EfficientNet (ImageNet) learns visual textures; Perch learns species-specific call patterns.

② COMPLETE SPECIES COVERAGE

- ~15,000 species globally.
- All 234 BirdCLEF+ classes included, no zero-shot unknown-species problem.

③ STRONG OUT-OF-THE-BOX QUALITY

- *"A simple linear probe on Perch embeddings achieves competitive performance out of the box."*
  - Perch documentation, Google Research

EfficientNet-B0 (ImageNet pre-training) → LB 0.787

Perch v2 raw embedding (no head tuning) → LB 0.737

Perch v2 + ProtoSSM + Bayesian Prior → LB 0.927

* Perch raw: high-quality embeddings but MLP head only — the head and post-processing drive the remaining gain.

Full pipeline

60-sec soundscape → 12 × 5-sec windows

Perch v2 ONNX → 1,536-dim embeddings

[cached to disk — extracted once, loaded at submission]

SSM branch

ProtoSSM + ResidualSSM

LB 0.928

Rank-Power branch

$0.6 \cdot R(p_{proto}) + 0.4 \cdot R(p_{sed}) \rightarrow G()$

LB 0.949

Weighted blend

$p_{final} = \alpha \times p_{ssm} + (1 - \alpha) \times p_{rank}$

Third model

Applied only to recordings the first two models disagreed on

SUBMISSION

G() — FIVE-STEP POST-PROCESSING

1. Temporal smoothing
2. File confidence scaling
3. Rank-aware scaling power = 0.40
4. Adaptive smoothing
5. File confidence gate t = 0.25

R() = rank transform ; aligns the two models' output scales before blending.

Three models — what they are?

SSM

State Space Model

A model that processes a sequence step by step, carrying memory of what came before. Each window informs the next.

Like reading a sentence word by word — you don't forget what came before.

Prototype

Each species has one learned representative sound pattern. A new recording is classified by measuring how similar it sounds to each prototype.

Like recognizing a friend's voice — you match it against a mental template, not a rulebook.

ProtoSSM combines both – it sequences 12 audio windows in order (SSM) while classifying each window by comparing it to species prototypes.

ResidualSSM

A second, smaller model that looks at what the first model got wrong and corrects it. It does not retrain the main model — it only adjusts the output.

Like a proofreader who fixes recurring mistakes without rewriting the whole document.

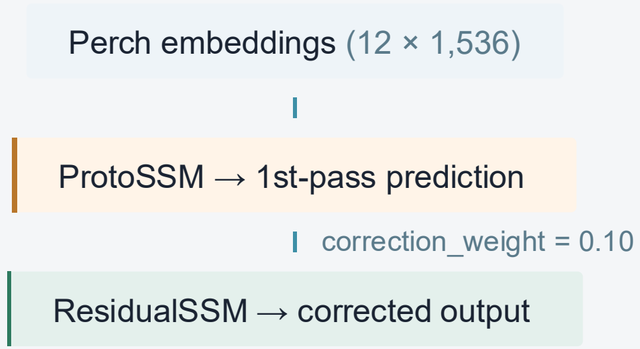
These three components make up the SSM branch.

01 · ProtoSSM

This branch treats a 60-second soundscape as a 12-step temporal sequence and classifies species using learned prototypes.

ARCHITECTURE

INPUT	12 × 1,536-dim Perch embeddings pre-cached
SSM CORE	Sequence model : processes 12 windows in order
HEAD	One learned prototype per species · cosine distance
RESIDUAL	ResidualSSM · corrects systematic under-predictions from the first pass



WHY SSM?

A 60-sec soundscape is a temporal sequence — the 12 windows are not independent. If a species calls in window 3, it is likely still present in window 4. The SSM processes chunks in order and models this continuity explicitly.

WHY PROTOTYPES?

Each species has one learned representative embedding. Classification by cosine distance generalizes better to rare species with very few positive examples.

WHY RESIDUALSSM?

ProtoSSM systematically under-predicts certain species. A lightweight second pass (d_model=128) corrects these bias patterns without retraining the main branch.

RESULT → **LB 0.928**

02 · Hyperparameter tuning

ARCHITECTURE IMPROVEMENTS

① Cosine annealing restart

The optimizer resets every 20 epochs and avoids sharp local minima that flat schedules cannot escape.

② MLP head expansion

(128,) → (256, 128). A two-layer head captures richer interactions inside the 1,536-dim Perch space.

③ ResidualSSM strengthening

Doubles the correction capacity for systematic under-prediction in the main branch.

PRIOR BIRDCLEF 2025 LITERATURE APPLIED

File-Level Aggregation	Private LB +0.014 in 2025 solutions. Reduces overconfident negatives.
‘focal_gamma’ tuning	Concentrates loss on rare, hard species; down-weights easy majority examples.
MixUp augmentation	Especially strong on private LB ; blended clips simulate multi-species soundscapes.
MLP width 128 → 192	Empirical optimum confirmed via controlled sweep.

Every technique was OOF(Out-of-fold)-validated before being kept; nothing was added on faith.

RESULT → **LB 0.942 → 0.943**

03 · Rank-power fusion

ProtoSSM + SED ensemble — same input, different error profiles.

$$p_{rank} = G(0.6 \cdot R(p_{proto}) + 0.4 \cdot R(p_{sed}))$$

R() — class-wise rank transform, removes calibration mismatch between ProtoSSM logits and SED probabilities.

WHY SED ENSEMBLE?

ProtoSSM captures temporal continuity well but smooths over short, localized call events. SED specializes in exactly those sharp, brief activations. The two models make systematically different errors — rank-blending reduces total error.

BirdCLEF 2025 2nd-place reported ensemble contribution of +0.007 on private LB.

WHY ONNX?

SED weights are pre-saved before submission — at inference, load and forward only. ONNX runtime is ~150× faster than TensorFlow SavedModel on CPU, which keeps us inside the 90-minute limit.

ECOLOGICAL PRIOR INJECTION

$$p_{adjusted} = (1 - \lambda) \cdot p_{model} + \lambda \cdot p_{prior}$$

$\lambda = 0.75$, by experiment.

Three-tier prior:

We built separate priors for recording site, time of day, and their combination!

- Nocturnal species — up-weighted at night hours
- Habitat-specific species — up-weighted at matching sites

Reduces false positives for ecologically implausible species without requiring additional labeled data.

RESULT → **LB 0.945 → 0.946**

04 · Residual diversity branch

RANK-POWER

main predictor · LB 0.949

SSM

diversity path · LB 0.928

WHY MODELS HELP EACH OTHERS?

Two copies of the same model repeat the same errors. A useful ensemble requires *genuinely different* error profiles.

SSM	misses short, sharp call events
Rank-Power (SED + rank)	over-smooths rare species

> *When the two branches agree, prediction confidence is high. When they disagree, the minority branch corrects main-branch mistakes.*

THIRD MODEL — APPLIED SELECTIVELY

After the two main models make their predictions, we compare their outputs file by file. Where they agree, we trust the result. Where they disagree significantly, we run a third model on just those files as a tiebreaker.

Applying it to all 600 test files would exceed the 90-minute CPU limit. So we ran it only on the files where the two models disagreed most.

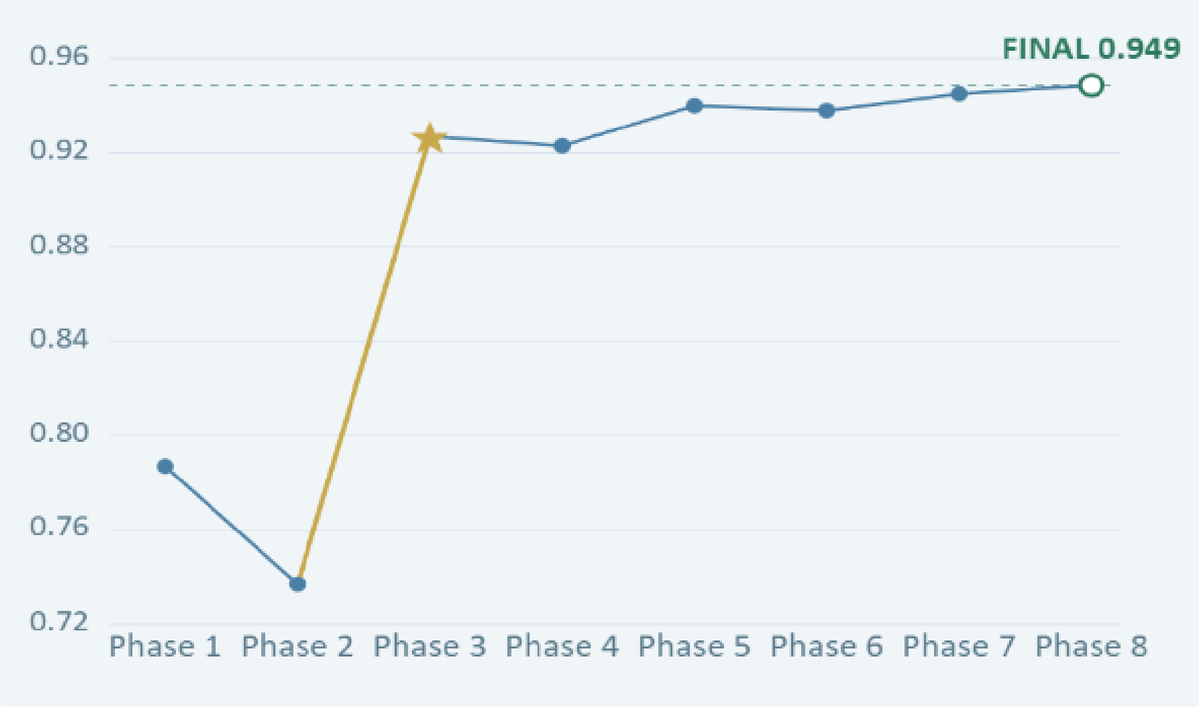
ProtoSSM: species A SED: species A → agree → done

ProtoSSM: species A SED: species B → disagree → run third model

FINAL → **LB 0.949**

Score progression

Submissions across 8 phases.



#	Approach	LB
1	EfficientNet-B0 baseline	0.787
2	Perch v2 raw embeddings	0.737
3 ★	**ProtoSSM + Bayesian prior**	**0.927**
4	Hyperparameter tuning	0.923
5	Post-processing rebuilt	0.940
6	rank_aware_power tuning	0.938
7	3-way fusion complete	0.945
8 ☆	**Parameter freeze**	**0.949**

Final result

#	Team	Members	Score
153	KIM DGDG		0.949

Public leaderboard, as of May 28 2026 — competition closes Jun 3, 2026.

COMPETITION SCALE

14,173 entrants · 4,955 participants · 4,001 teams

138,212 total submissions · top ~3.8% among all teams (ongoing).

RANK
153rd
of 4,001 teams

SCORE
0.949
macro-avg ROC-AUC

WHAT WE CONTRIBUTED

- Perch ONNX + ProtoSSM + SED rank-fusion + prior
- A CPU-safe pipeline with the 90-min constraint as a primary design criterion
- Evidence that domain-matched pre-training outweighs model scale in bioacoustic transfer
- Ecological priors (site × hour, λ = 0.75) that reduce false positives without new labels

What we learned

01

Pre-training domain alignment > model size

Replacing EfficientNet (ImageNet) with Perch (bioacoustics) added **+0.190** with no change to downstream architecture. When the pre-training domain matches the deployment domain, a smaller model consistently outperforms a larger mismatched one.

02

Hardware limits are design constraints, not obstacles

The 90-minute CPU restriction forced disciplined choices: embedding caching, ONNX conversion, pre-saved SED weights, and selective NFNet on 48 files only. Each constraint produced a leaner, more reproducible decision.

03

Calibration is part of the model

Bayesian prior injection (λ = 0.75, site × hour), the rank transform R(), and the five-step G() pipeline together contributed more than any single architectural change. Getting the prediction distribution right matters as much as getting the architecture right.

04

Ensemble diversity requires different error sources

Blending two outputs from the same model family repeats errors. The 2.64% weight on yukiZ works because its error profile (SSM path) genuinely differs from Rank-Power (SED + rank fusion) — different training path, different failure modes, different corrections.

Thank you / Questions?

TEAM 2

Donggeun · Haiwon · Seunghoon · Woohyuk

Programming for Data Science