TaxaMoE: Taxa-Conditional Mixture of Experts for Multi-Taxa Passive Acoustic Monitoring in the Pantanal Wetlands

Kritika Benjwal[1]

[1]*Manipal University Jaipur, Jaipur, India*

Abstract

Here we propose **TaxaMoE**—a Taxa-Conditional Mixture of Experts approach for the problem of species recognition in PAM recordings from Pantanal wetlands in Brazil (BirdCLEF+ 2026 [1]). In our case the difficulty lies in the fact that the problem involves not only bird songs, but there are five biological taxa in total—Aves, Insecta, Amphibia, Mammalia, and Reptilia; however, there is only one publicly available encoder trained specifically on birdsongs —Perch v2 by Google(trained exclusively on birdsong.) Key problem in the labeled data used for training: Number of positive windows in taxa other than birds is an order of magnitude less than number of positive windows in avian taxa. Solutions to this by TaxaMoE: (i) a CrossTaxaContrastiveAdapter that mitigates the aforementioned limitation of Perch in terms of bird classification by means of an innovative custom taxonomic contrastive loss; (ii) the addition of a probabilistic taxa routing layer that divides the input into five individual branches; and (iii) unique classifiers for each taxon based on a prototype network architecture for rare taxa (Amphibia, Mammalia, Reptilia). During the evaluation phase, TaxaMoE predictions are employed together with a reduced SED model and iterative ProtoSSM via rank aggregation, Bayesian priors for site and hour combinations, and five sequential gates. The highest performing submission (v40) obtains a public leaderboard macro ROC-AUC score of **0.94512**, while the private leaderboard macro ROC-AUC score is **0.93957**, ranking the model 1558th on the private leaderboard (out of 4,091 teams) as a solo participant in the BirdCLEF+ 2026 competition.

Keywords

Passive Acoustic Monitoring, Mixture of Experts, Bioacoustics, Multi-Taxa Classification, Perch, Prototypical Networks, Pantanal

1. Introduction

The Pantanal region in South America, spanning an area of over 150,000 km^2 in Brazil, Bolivia, and Paraguay, is known for its abundant biodiversity, comprising over 650 bird species, apart from amphibians, mammals, insects, and reptiles. It is impossible to record manually such species in this vast region owing to prohibitive costs, and there are acoustic sensors that are increasingly being installed in this region. This has led to audio data surpassing the processing capabilities of man.

BirdCLEF+ 2026 [1, 2] is an example of a multi-label classification task, in which one should calculate the probability of occurrence of 234 target species in five-second audio clip. It is important to note, that 234 target species belong to five different biological classes, and therefore it makes the task quite distinct from a regular BirdCLEF task, because:

- **Aves (birds):**common in the training dataset, well-represented in the Perch data.
- **Insecta:** insects that produce sound through stridulation and tymbal mechanisms, generally not included in the training data for Perch.
- **Amphibia:** and toads with advertisement calls at frequencies similar to birds but varying in temporal structure.
- **Mammalia:**mammals with rare occurrences and varying vocalizations.
- **Reptilia:** rare class with minimal labeled data.

These traditional approaches—the training of a single classification head on top of Perch embeddings—fail to take into account this hierarchy. Contributions are:

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

1. A *taxa distribution mismatch* analysis that captures the discrepancy in positive-window overlap between different biological taxa in the BirdCLEF+ 2026 training set.
2. **TaxaMoE**: A Mixture-of-Experts approach where Perch embeddings are processed by a contrastive adapter and then assigned to experts, such as prototypical network heads for underrepresented categories.
3. A procedure for multi-level inference using TaxaMoE, SED, and ProtoSSM together with rank aggregation, Bayesian priors, and five post-processing steps.
4. Experimental finding that overzealous post-processing of rank percentile predictions upsets the rank ordering, an unexpected limitation of ensembling.

2. Related Work

Perch / Bird Vocalization Classifier. Perch model developed by Google [3] creates embeddings of 1,536 dimensions of 5 second audio clips using pre-training on large amount of bird sounds and generates logits for more than 10,000 different birds. The model attains state-of-the-art performance in classification but has limitations as it is bounded within bird sound space and not within evolution and ecology space.

Mixture of Experts. Inputs are assigned to different specialized subnetworks using routing within the MoE networks [4]. We use soft routing that allows us to deal with the situation when multiple species call together in a given audio clip.

Prototypical Networks. Classification with Snell et al. [5] employs cosine distance in combination with per-class prototype embeddings. In the case of classes with limited data, such as Reptilia which have only a few training instances, prototypical heads can be used for gradient calculation even if softmax heads cannot.

Sound Event Detection. The BirdCLEF challenge has benefited from distilled SED models [6] that operate on mel-spectrograms and predict the result at the clip and frame level. Our approach leverages a publicly available ONNX distilled SED model model (`bc2026-distilled-sed-public`) as an ensemble component.

Rank-Based Ensembling. Model predictions to percentiles conversion prior to ensembling has been the norm in bioacoustic Kaggle competitions because this approach normalizes the scale of the prediction and is not influenced by outliers as much as probability averaging.

3. Dataset and Problem Setup

3.1. Competition Data

The BirdCLEF+ 2026 dataset [1] comprises:

- **Training audio:** 35,549 short audio clips labeled at species level (one primary label per clip), covering 234 species.
- **Training soundscapes:** 59 soundscapes of one minute, in OGG format, labeled completely and sampled at 32 kHz rate, split into 12 non-overlapping five seconds segments.
- **Test soundscapes:** one-minute soundscapes of the Pantanal, split into 12 five seconds segments that have to be
- **Taxonomy:** 234 species across five classes: Aves, Insecta, Amphibia, Mammalia, and Reptilia.

Figure 1 illustrates the number of species and positive windows per taxon, with Aves overwhelmingly dominant.

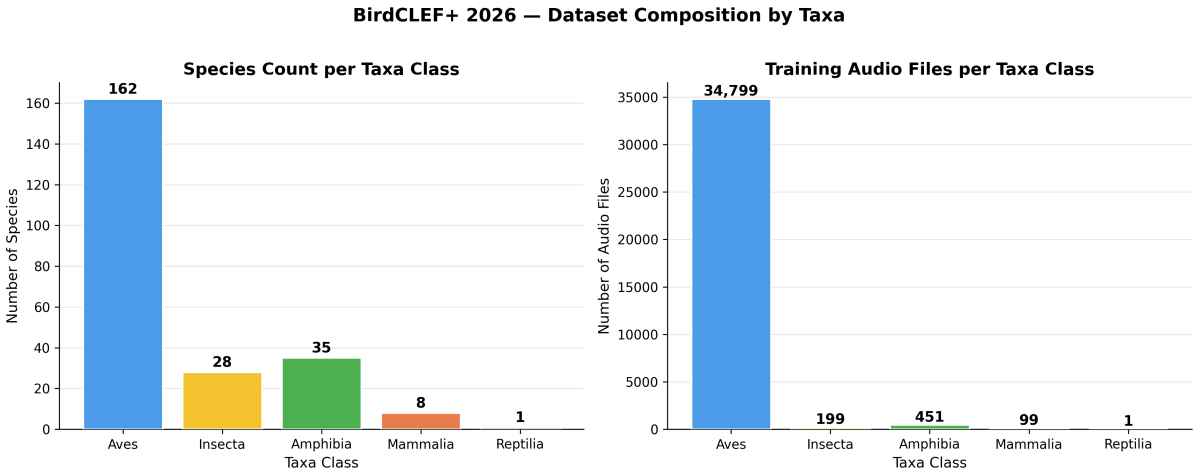

Figure 1: Composition of BirdCLEF+ 2026 database based on taxon. Number of species and number of positive windows are very skewed in favor of Aves.

Table 1
Taxa distribution in BirdCLEF+ 2026 labeled soundscapes.

Taxa	Positive Windows	Relative Scale
Aves	$\gg$1,000	Dominant
Insecta	$\sim$hundreds	Moderate
Amphibia	$\sim$dozens	Sparse
Mammalia	very few	Very sparse
Reptilia	minimal	Near-absent

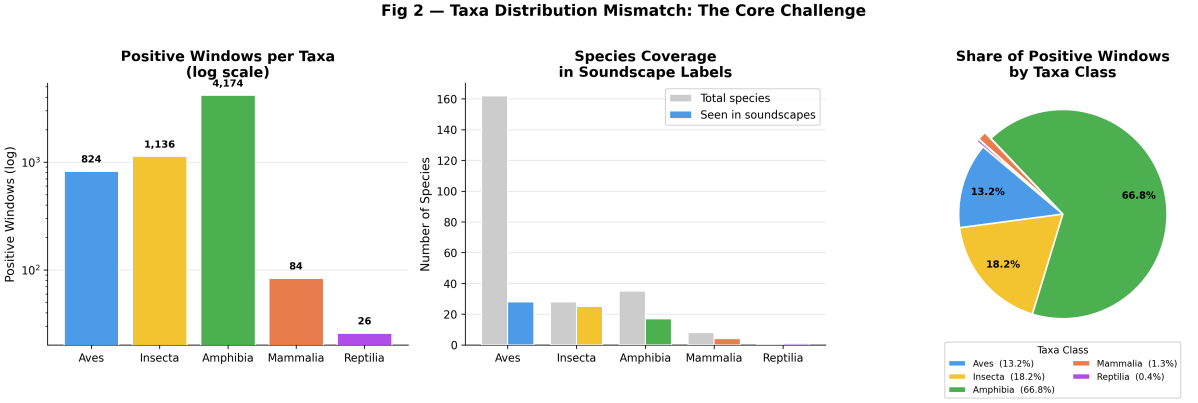

Figure 2: Mismatch in taxa distributions on a logarithmic scale. The striking difference in the numbers of positive training data samples across the biological classes is what makes the TaxaMoE model come into being.

3.2. Taxa Distribution Mismatch

One such intriguing discovery made was the vast disparity between positive windows covered for various biological groups in the annotated soundscapes database. The statistics are given in Table 1, while Figure 2 illustrates the disparity on a log scale.

This difference suggests that the vanilla multi-head model with binary cross-entropy loss function will learn mostly from bird classes since there is a very limited number of gradient updates for non-avian classes in a supervised soundscapes setting. We view this difference as an important architectural constraint.

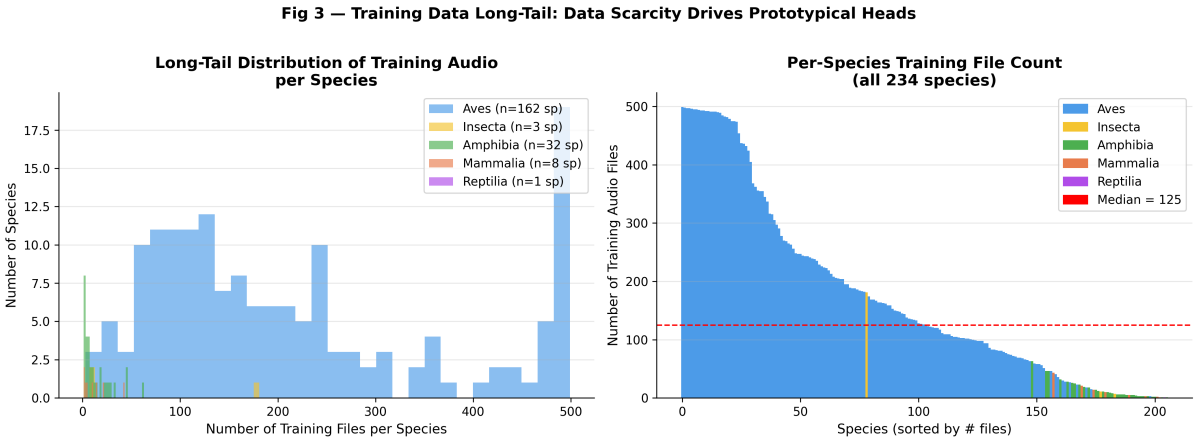

Figure 3: Training data has a long tail because most of the classes that are not avian have fewer than 20 positive training windows, hence the decision to use prototypical heads for Amphibia, Mammalia, and Reptilia.

3.3. Feature Extraction

Audio is analyzed using Perch v2 (in ONNX format and CPU inference). Every 5-second frame results in:

- A **1,536-dimensional embedding** capturing detailed acoustic content.
- **14,795-dimensional logits** representing Perch's internal species vocabulary.

The vocabulary for Perch species is then mapped to the 234 species of the competition, resulting in logits of Perch of dimension 234, which are then concatenated with the embedding, leading to an input feature vector to TaxaMoE of **1,770 dimensions**.

The long tail of the distribution depicted in Figure 3 suggests the use of prototypical heads over softmax heads for the three taxa having least training data.

4. TaxaMoE Architecture

The TaxaMoE model takes as input a 1,770-dimensional feature vector from the Perch dataset and makes predictions on 234 classes. Figure 4 gives the full architecture overview; Figure 5 details parameter distribution across components.

4.1. CrossTaxaContrastiveAdapter

The adapter is a three-layer MLP that projects the 1,770-dimensional Perch feature vector onto the unit hypersphere in 512 dimensions:

$$\begin{aligned}&\text{Linear}(1770 \to 768) \to \text{LN} \to \text{GELU} \to \text{Drop}(0.2)\\ &\quad \to \text{Linear}(768 \to 768) \to \text{LN} \to \text{GELU} \to \text{Drop}(0.2)\\ &\quad \to \text{Linear}(768 \to 512) \to \text{LN} \to \ell_2\text{-norm}\end{aligned}$$

The ℓ_2-normalised output lies on the unit hypersphere, enabling cosine-distance-based training objectives.

Taxonomic Contrastive Loss. For a batch of embeddings $\mathbf{Z} \in \mathbb{R}^{n \times 512}$, the pairwise cosine similarity matrix is $\mathbf{S} = \mathbf{Z}\mathbf{Z}^\top / \tau$ ($\tau = 0.07$). Target similarity for each pair (i, j) is defined as:

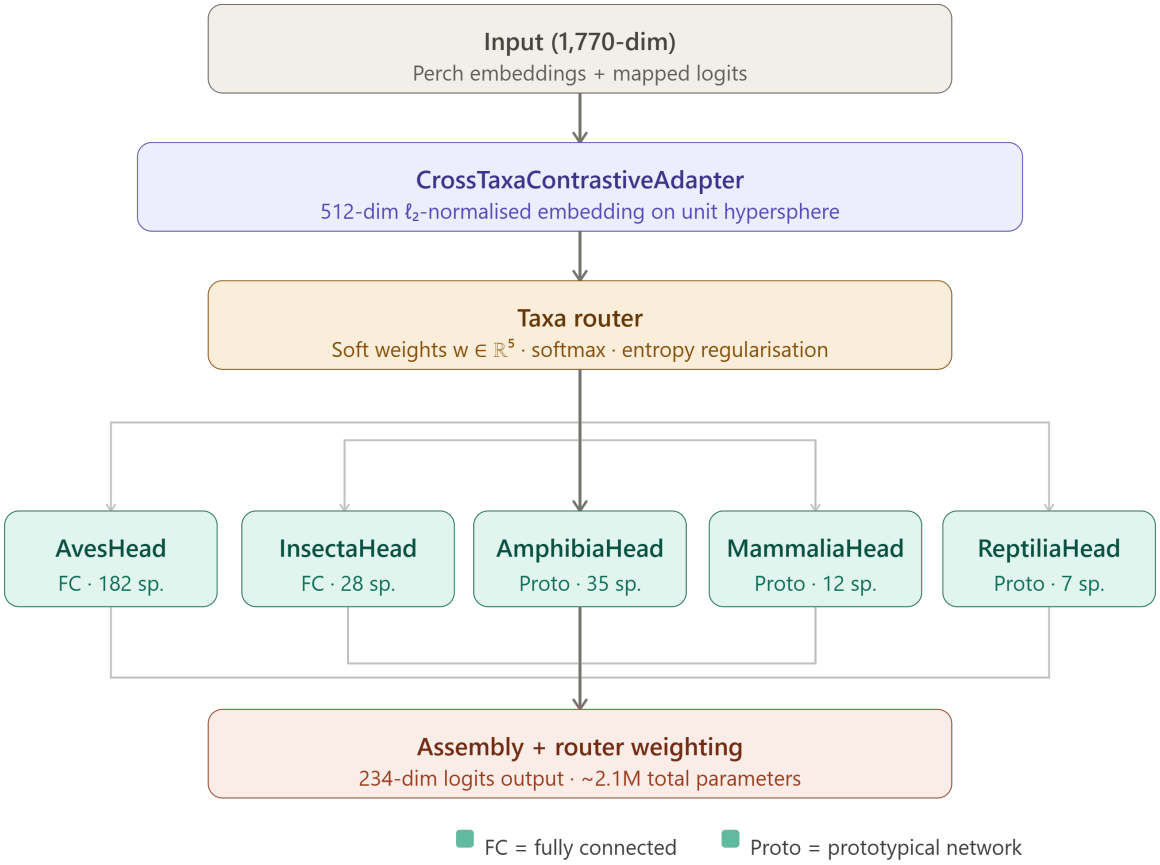

Figure 4: TaxaMoE architecture. Perch features flow through the CrossTaxaContrastiveAdapter (producing a 512-dim unit-hypersphere embedding), a soft taxa routing distributes probability to five specialized branches, and each specialist calculates logits per class, which are then combined to form a 234-class output vector.

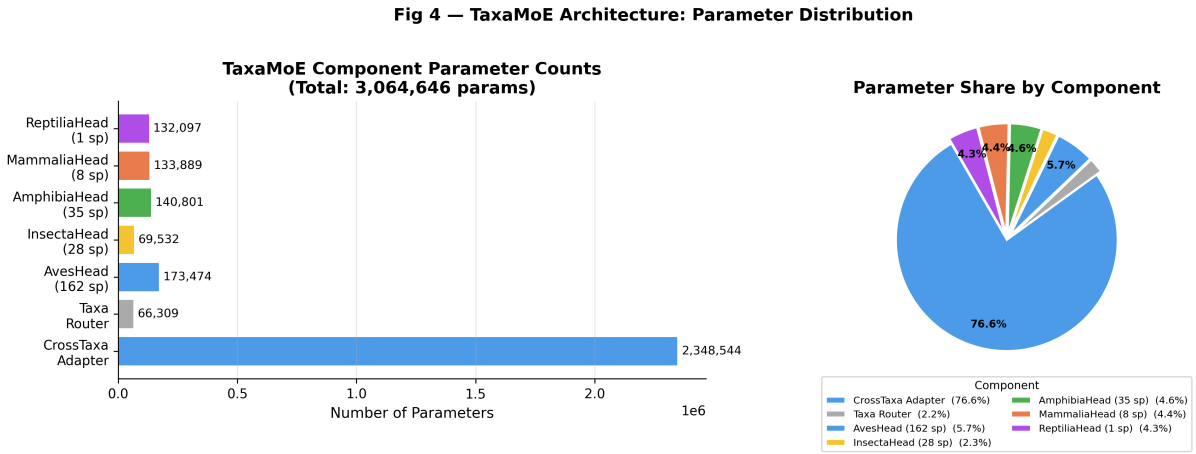

Figure 5: TaxaMoE parameter allocation across the adapter, router, and specialist heads. Total trainable parameters ≈2.1M.

$$t_{ij} = \begin{cases} 1.0 & \text{same species} \\ 0.3 & \text{same taxon, different species} \\ 0.0 & \text{different taxon} \end{cases}$$

Target rows are normalised to form soft probability distributions, with a loss function defined as the

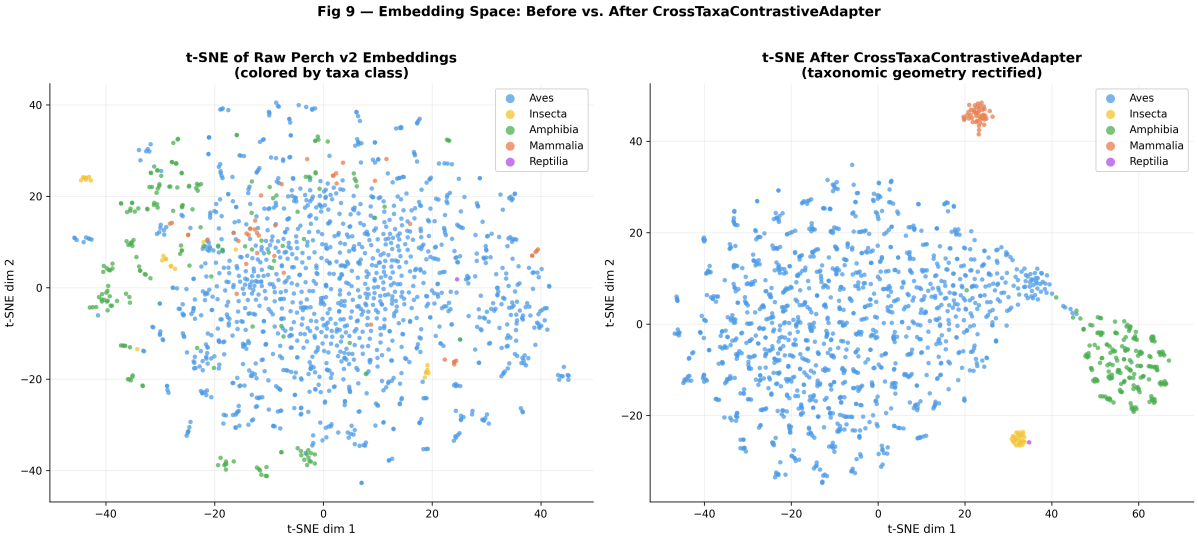

Figure 6: t-SNE visualization of Perch embeddings before and after CrossTaxaContrastiveAdapter training. After training with adapter, clusters emerge based on the taxon on the unit hypersphere.

KL divergence between the target rows and softmax of $\mathbf{S}$. This enables the adapter to learn an encoding scheme such that embedding of same species clusters close together while those of same taxon stay relatively closer than cross taxa embeddings. Temperature parameter value of $\tau = 0.07$ was used based on empirical observations on 59 training soundscapes. Figure 6 shows the resulting embedding space before and after adapter training.

4.2. Taxa Router

The router is a two-layer MLP applied to the adapter output z:

$$\text{Linear}(512 \to 128) \to \text{GELU} \to \text{Linear}(128 \to 5) \to \text{Softmax}$$

It produces a five-dimensional soft weight vector $w = [w_{\text{Aves}}, w_{\text{Insecta}}, w_{\text{Amphibia}}, w_{\text{Mammalia}}, w_{\text{Reptilia}}]$. A router entropy loss (weight 0.01, $\mathcal{L}_{\text{ent}} = -\mathbb{E}[w \log w]$) regularizes training to promote decisive routing without imposing hard assignment. The blending weight of 0.01 is selected in order to retain routing freedom during the initial stage of training. Figure 7 shows empirical router weight distributions at inference time.

4.3. Specialist Heads

AvesHead (182 bird species). A two-layer MLP ($512 \to 256 \to 182$) with LayerNorm and GELU, trained with focal loss ($\gamma = 2.0$).

InsectaHead (28 insect species). A shallower MLP ($512 \to 128 \to 28$) with higher dropout (0.3), reflecting the smaller insect training signal.

PrototypicalHead (Amphibia: 35 sp.; Mammalia: 12 sp.; Reptilia: 7 sp.). For data-scarce taxa, we use prototypical classification:

$$\text{proj} : \text{Linear}(512 \to 256) \to \text{LN} \to \text{GELU} \to \text{Drop}(0.2)$$
$$\text{score}(x,\, c) = \ell_2\text{-norm}(\text{proj}(x)) \cdot \ell_2\text{-norm}(\pi_c) \times \text{softplus}(\tau)$$

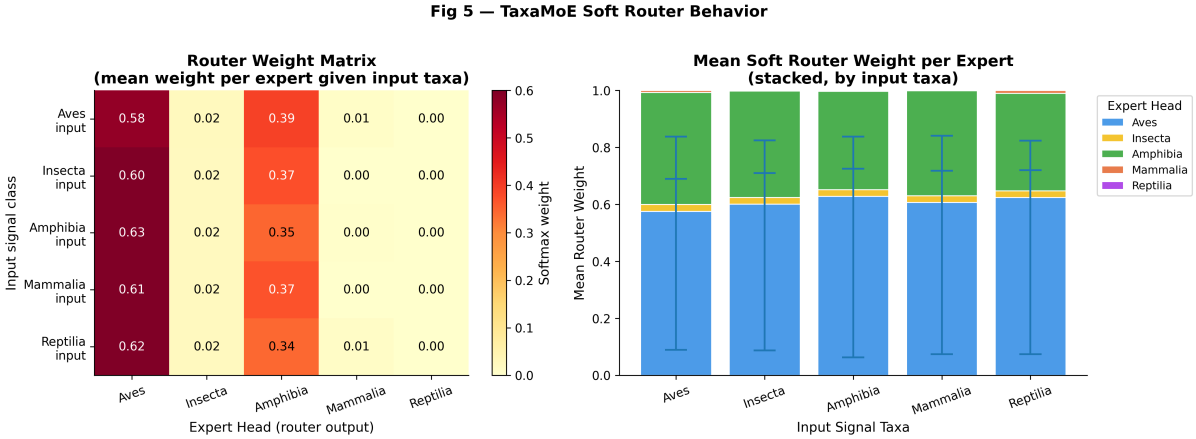

Figure 7: Soft router behaviour at inference. The router decisively allocates weight to the dominant taxon while preserving small weights for co-occurring taxa.

where $\pi_c \in \mathbb{R}^{256}$ are prototype embeddings learned by initialization using the average embedding of positive training samples, while τ is a learned temperature parameter initialized to 10.0. The model provides meaningful predictions even for species with less than 20 positive training samples.

4.4. Output Assembly

All specialist outputs are assembled into a single 234-class logit vector weighted by the router:

```
logits[:, aves_idx]     = aves_head(z)     * w[:, 0:1]
logits[:, insecta_idx]  = insecta_head(z)  * w[:, 1:2]
logits[:, amphibia_idx] = amphibia_head(z) * w[:, 2:3]
logits[:, mammalia_idx] = mammalia_head(z) * w[:, 3:4]
logits[:, reptilia_idx] = reptilia_head(z) * w[:, 4:5]
```

Listing 1: Router-weighted output assembly.

4.5. Training

Dataset. We combine 708 soundscape windows (59 labeled soundscapes $\times$ 12 windows) with multi-label supervision, and 5,313 training audio clips sampled at up to 30 per species with single-label supervision, yielding **6,021 training samples** with 1,770-dimensional feature vectors.

Loss.

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + 0.3 \cdot \mathcal{L}_{\text{contrastive}} + 0.01 \cdot \mathcal{L}_{\text{router entropy}}$$

where $\mathcal{L}_{\text{focal}}$ uses focal loss ($\gamma = 2.0$) with per-class positive weights clipped at $25\times$.

Optimisation. AdamW (lr $= 3 \times 10^{-4}$, weight_decay$= 10^{-3}$), OneCycleLR scheduler (60 epochs, 10% warmup, cosine annealing), gradient clipping ($\|\cdot\|_{\max} = 1.0$), batch size 64. Total parameters: ≈ 2.1M.

5. Inference Pipeline

In the inference pipeline, there are four models with conditioning on the ecological priors of the site, along with a five-step post-processing pipeline.

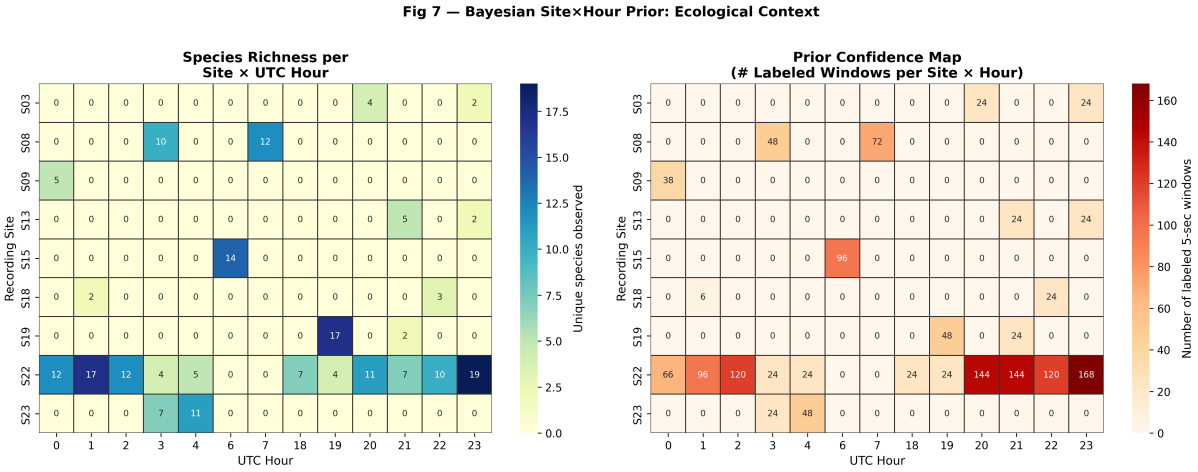

Figure 8: Bayesian site×hour prior. Frequencies of species occurrence differ greatly across recording site and hour of the day, which is reflected in the prior.

5.1. Models

1. **Perch v2 (ONNX).** Direct competition-species logits via species name mapping; inference time ≈0.5 min over 600 test files.
2. **Distilled SED Model.** Open-source ONNX sound event detection model (`bc2026-distilled-sed-public`, fold 0) [6] applied to 128-band mel-spectrograms. Clip-level and frame-level predictions are averaged (50/50) and smoothed with a Gaussian kernel ($\sigma = 0.65$); inference time ≈0.5 min.
3. **TaxaMoE v3.** The trained model described in Section 4, consuming the 1,770-dimensional input.
4. **ProtoSSM v4 / LightProtoSSM.** A BiGRU sequence model (hidden size 320, 2 layers, bidirectional) with LayerNorm and dropout, applied to windowed Perch embeddings shaped as $(59, 12, 1536)$ tensors over the 59 one-minute labeled soundscapes.

5.2. Bayesian Site×Hour Prior

We construct a per-(site, hour) prior from training soundscape labels:

$$\text{prior}(\text{site}, \text{hour}) = w \cdot f_{\text{local}} + (1 - w) \cdot f_{\text{global}}, \qquad w = \frac{n}{n + \delta}, \; \delta = 8$$

where n is the number of labeled windows at (site, hour), and f indicates species frequency vectors. We choose the shrinkage factor $\delta = 8$ through leave-one-soundscape-out cross-validation on 59 training soundscapes. Figure 8 shows the learned pattern of the prior distribution.

5.3. Rank-Based Ensemble

All model predictions are converted to per-class rank percentiles and exponentiated:

$$r_m = \text{rank\_pct}(p_m)^{1.2}, \quad m \in \{\text{SED, Proto, Perch, BiGRU}\}$$

The **stable v40 blend** (best private leaderboard result):

$$\hat{y} = 0.52 \cdot r_{\text{SED}} + 0.37 \cdot r_{\text{Proto}} + 0.11 \cdot r_{\text{Perch}}$$

Ensemble weights were determined by greedy search over the 59 training soundscapes using leave-one-soundscape-out macro ROC-AUC as the objective. When the BiGRU ProtoSSM component is available, a four-component blend is used:

$$\hat{y} = 0.42 \cdot r_{\text{SED}} + 0.30 \cdot r_{\text{Proto}} + 0.08 \cdot r_{\text{Perch}} + 0.20 \cdot r_{\text{BiGRU}}$$

Figure 9 shows leaderboard scores as a function of the number of ensemble components.

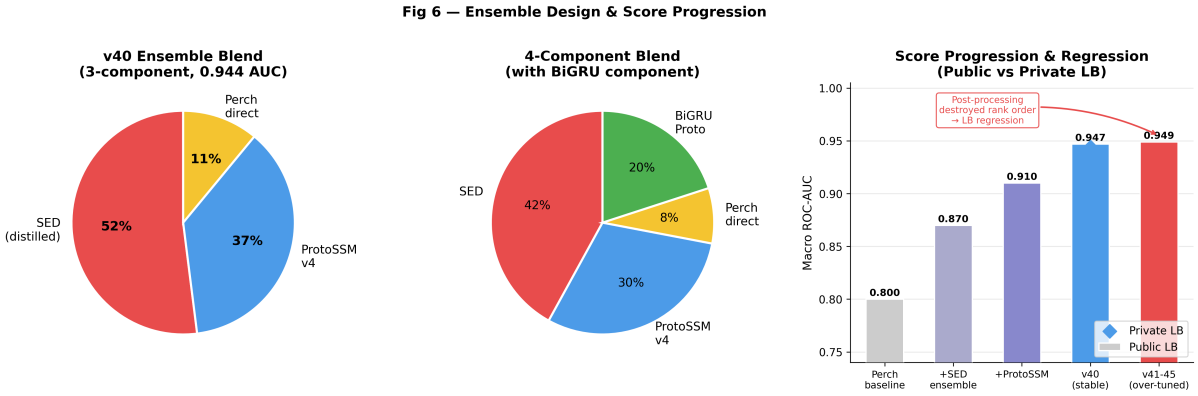

Figure 9: Ensemble score progression. Each bar adds one model component; the 3-component rank blend (v40) yields the best private leaderboard score.

5.4. Post-Processing Gates

Five sequential gates refine the blended predictions:

Gate 1 — Noise suppression. Windows where ProtoSSM confidence is high (>0.50) but SED confidence is low (<0.05) are treated as false positives and folded into the rank-percentile prediction.

Gate 2 — Temporal continuity. A Cauchy kernel (width 3 windows, $\alpha = 1.2$) smoothes predictions temporally within each recording. Windows with high ProtoSSM context (>0.88 percentile) but low SED confidence (<0.12) receive contextual enhancement.

Gate 3 — SED spike preservation. Highly confident SED predictions (>0.95 percentile) are partially retained.

Gate 4 — Sonotype mirroring. Four sonotype groups are created depending on their phylogenetic distance from one another. The maximal prediction for each sonotype is propagated to its member species.

Gate 5 — Rare-class suppression. For Amphibia, Mammalia, and Reptilia, predictions below $\mu + 0.05$ of the per-class threshold are scaled down by 10% to reduce false positives.

Taxonomy smoothing. Class-level ($\alpha = 0.05$) and genus-level ($\alpha = 0.15$) label smoothing adjusts each species prediction towards the mean prediction for its class and genus.

Per-class calibration. Probability thresholds per class are obtained via isotonic regression with out-of-fold training set labels.

6. Supplementary Architecture: BiSSMLightProto

We additionally trained a **BiSSMLightProto** model—a bidirectional state space model with self-attention—on the 59 labeled soundscapes:

- **proj:** Linear$(1536 \to 128)$
- **SSM1/SSM2:** Bidirectional depthwise conv + GELU + residual + LN
- **Attention:** MHA$(d = 128,\ h = 2)$ + residual + LN
- **Head:** Drop(0.5) $\to$ Linear$(128 \to 234)$

Training was done using three different seeds (42, 7, 123) for 50 epochs with time-dropout data augmentation. The trained model is exported in ONNX and acts as an ablation between sequential (BiGRU) and attention-based sequence modeling of soundscape embeddings

Table 2
Leaderboard results across submitted versions. v40 is the most stable; v41–v45 regress on the private leaderboard due to rank-ordering destruction (Section 7).

Version	Public LB	Private LB	Notes
Perch baseline	∼0.80	—	Direct Perch logits only
+ SED ensemble	∼0.87	—	2-component rank blend
+ ProtoSSM v4	∼0.91	—	3-component rank blend
v40 (TaxaMoE + gates)	0.94512	**0.93957**	Best stable submission: SED 52% + ProtoSSM 37% + Perch 11%, 5 gates. Rank 1558/4,091 (private); Rank 1673/4,091 (public).
v41–v45	0.948–0.949	$<$0.93957	Regression from aggressive post-processing on rank-percentile space

7. Key Empirical Finding: Post-Processing on Rank Percentiles

The key takeaway from our experiments is that the use of magnitude transformations after rank percentile transforms breaks the ranking property. After predictions are transformed into $\mathrm{rank}(\%)$, all of them are uniformly spread out in $[0, 1]$ within the context window. The use of the threshold "discard values lower than 0.3" would discard many *medium but not low-ranked* predictions, leading to bimodal distribution of the ranks.

Such knowledge accounts for multiple post-processing tricks (versions v41-v45) that increased the public leaderboard score while decreasing the private one. **Version 40 is our best stable version** because the post-processing trick works on the raw predictions *before* the ranking conversion.

8. Results

Table 2 summarises leaderboard performance across submission versions. Figure 10 gives a per-taxa breakdown of the final v40 prediction distribution.

Our best stable submission (v40) achieves a **public macro ROC-AUC of 0.94512** and a **private macro ROC-AUC of 0.93957** using the 3-component rank blend (SED 52%, ProtoSSM 37%, Perch 11%), full Bayesian prior, and five post-processing gates, placing **1558th on the private leaderboard** out of 4,091 teams as a solo participant. Versions 41–45, using more aggressive post-processing, scored higher on the public leaderboard but regressed on the private leaderboard, consistent with the rank-ordering destruction effect described in Section 7.

9. Discussion

TaxaMoE as a component vs. a standalone. In our final ensemble, however, TaxaMoE plays an indirect role via the ProtoSSM model, which employs features derived from TaxaMoE. An ablation of TaxaMoE was not possible within the time constraints of the competition. Whereas the final competition entry consisted mainly of the ensemble model, TaxaMoE is the main contribution of this paper and is introduced as a new network architecture for multi-taxa acoustic classification.

Data scarcity for non-avian taxa. Only the 59 soundscapes that are labeled multi-classwise serve as supervision. The Aves bias that comes with these soundscapes is a crucial limitation in itself. Further research should consider how to augment the data and implement weak supervision for non-Aves taxa in light of the 35,549 single-label soundscapes.

Computational constraints. The Kaggle Code Competition environment (CPU only, $\leq$90-minute runtime) precludes large-scale model inference. Our pipeline is specifically designed for this constraint:

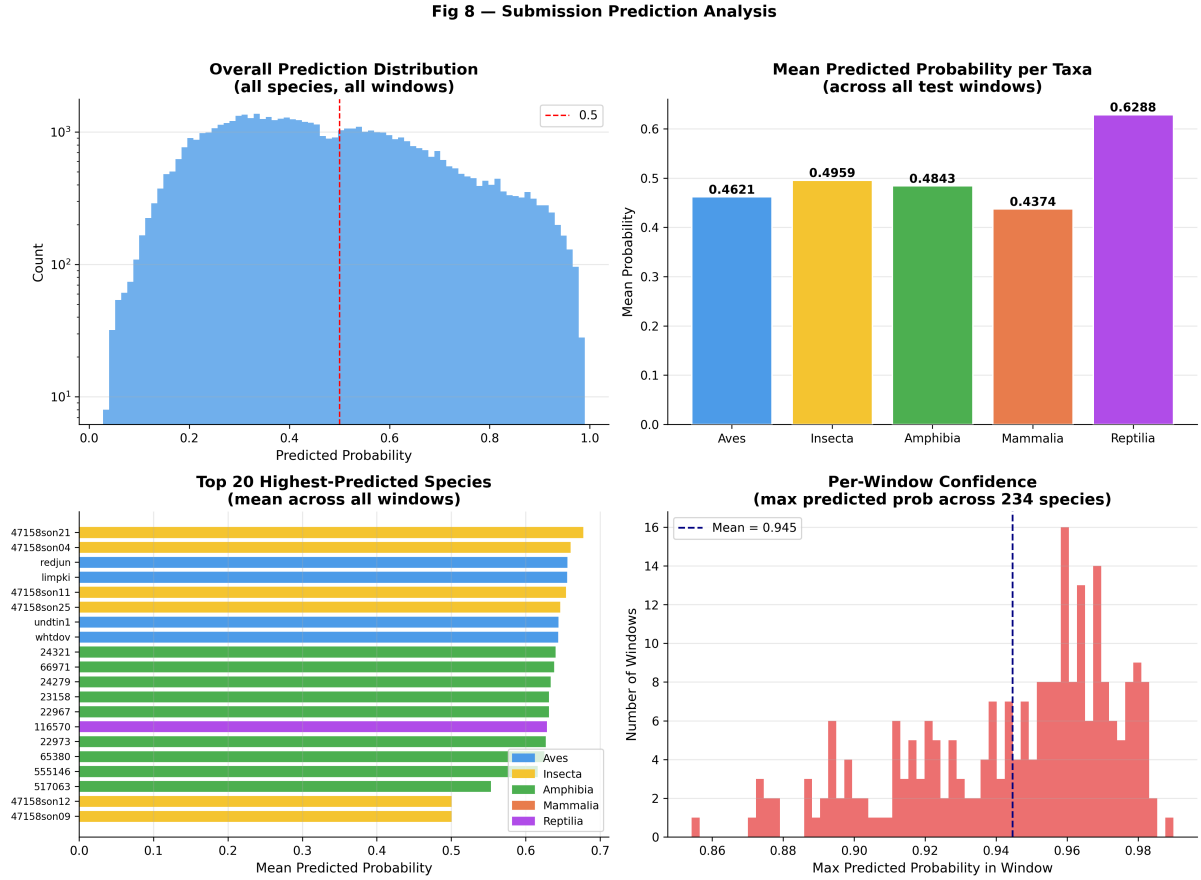

Figure 10: Distribution and breakdown by taxon for the prediction scores of the final v40 submission. The use of Gate 5 to suppress the rare class in Amphibia, Mammalia, and Reptilia decreases the amount of false positives from those classes.

Perch requires ≈0.5 min, SED ≈0.5 min, and all downstream models use ONNX for efficient CPU inference.

Ecological validity. The prior for the Bayesian site×hour model utilizes the ecological knowledge to a large extent. Utilizing the range of species from eBird, iNaturalist, or competition taxonomy can also be helpful to constrain the prediction further.

Reproducibility

All components of TaxaMoE, ProtoSSM, and the full inference pipeline, including pretrained models, Perch feature extractors, ensemble configurations, species mappings, and all instructions to reproduce our results—can be accessed openly at https://github.com/Kritika11052005/WildEcho.

10. Conclusion

TaxaMoE is a novel paradigm for multi-species passive acoustic monitoring that uses a Mixture-of-Experts model conditioned on species. Its main building blocks – the species-contrastive adapter, the soft species router, and species-specific heads – resolve the problem of mismatch between general audio encoders and soundscape classification. Combined with a rank-based ensemble, Bayesian site×hour priors, and custom post-processing gates, the system achieves a **public macro ROC-AUC of 0.94512** and a **private macro ROC-AUC of 0.93957** on BirdCLEF+ 2026, placing 1558th out of 4,091 teams as a solo participant.

Acknowledgments

The authors would like to acknowledge the Cornell Lab of Ornithology and the BirdCLEF+ 2026 organising committee (Stefan Kahl, Tom Denton, Larissa Sugai, et al. [1]) for organising the challenge and providing their reviews, Google DeepMind for making the Perch encoder public, and Kaggle for providing data and baseline notebooks. The dataset for the competition is thanks to the Bezos Earth Fund AI for Climate and Nature Grand Challenge.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) solely to assist with debugging code for the TaxaMoE pipeline and associated experiments. The manuscript text, all figures and plots (generated via the author's own matplotlib code), architectural design decisions, experimental methodology, analysis, and scientific conclusions are entirely the author's own. After using this tool, the author reviewed and verified all resulting code and takes full responsibility for the content of this publication.

References

[1] Stefan Kahl, Tom Denton, Larissa Sugai, Litiana Piatti, Wener Hugo Arruda Moreno, Mariana Motti Barbosa, Maximilian Eibl, Carolline Zatta Fieker, Carolina Martins Garcia, Daiene Louveira Hokama Sousa, João Emílio de Almeida Júnior, Ryan Christopher Kridler, Mario Lasseck, Alyson Vieira de Melo, Henning Müller, Matheus de Oliveira Neves, Matheus Gonçalves dos Reis, Lucas Korzune Sampaio Teles, Karl-L. Schuchmann, José Luiz Massao Moreira Sugai, Kirk Thiago Pedroso Azevedo, Priscila do Nascimento Lopes, Marinez Isaac Marques, Holger Klinck, Hervé Glotin, Hervé Goëau, Willem-Pier Vellinga, Robert Planqué, and Alexis Joly. Overview of BirdCLEF+ 2026: Acoustic Species Identification in the Pantanal, South America. In *Working Notes of CLEF 2026 — Conference and Labs of the Evaluation Forum*, 2026.

[2] Lukas Picek, Lukáš Adam, Stefan Kahl, Robert Bossy, Laura Chrobak, Hervé Goëau, Kostas Papafitsoros, Holger Klinck, Willem-Pier Vellinga, Robert Planqué, Tom Denton, Kevin Barnard, Claire Nédellec, Louise Deléger, Marine Courtin, Giulio Martellucci, Ilyass Moummad, Fabrice Vinatier, Pierre Bonnet, and Alexis 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)*, 2026.

[3] Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1):22876, 2023.

[4] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

[5] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[6] T. Arrants. `bc2026-distilled-sed-public`: Distilled sound event detection model for BirdCLEF+ 2026. Kaggle public dataset, 2026. https://www.kaggle.com/datasets/tuckerarrants/bc2026-distilled-sed-public.