Unifying State Space Models and Deep Audio Embeddings with Circadian-Ecological Priors for Bioacoustic Monitoring

Dhruv Pai Dukle[1]

[1]*Independent Researcher*

Abstract

We describe our system for the BirdCLEF 2026 challenge, which required identifying 234 species of birds, amphibians, and insects from 60-second soundscape recordings collected in Brazil's Pantanal region. Our approach combined Google's Perch v2 bioacoustic foundation model with a novel Bidirectional Selective State-Space Model (ProtoSSM) and customized PyTorch MLP probes (incorporating Embedding MixStyle and GroupDRO site optimization). This embedding-based pipeline was ensembled with distilled CNN sound event detectors, HGNet CNN models compiled using OpenVINO for global rank blending, and an ECA-NFNet-L0 branch run selectively on uncertain test files. We optimized the post-processing stack by implementing isotonic calibration, per-class threshold grid-search, rank-aware scaling, and adaptive delta smoothing. Our best submission achieved a public macro ROC-AUC of 0.951, securing a final private leaderboard rank of 408 out of 4,225 teams. This note documents the design decisions, successful strategies, and caution-inducing negative results across our three-month development lifecycle.

Keywords

bioacoustics, bird sound recognition, state-space models, Perch, BirdCLEF, OpenVINO, sound event detection, threshold optimization, domain adaptation

1. Introduction

The BirdCLEF 2026 competition[2], part of the LifeCLEF 2026 evaluation campaign[1], challenged participants to build classifiers for 234 species across three taxonomic classes—Aves (birds), Amphibia (frogs), and Insecta (crickets, katydids)—using 60-second soundscape recordings from nine monitoring sites in Brazil's Pantanal wetland. The event attracted massive global participation, hosting 14,173 entrants and 5,224 active participants across 4,225 teams, who submitted a total of 158,092 solutions. The task presented several unique difficulties:

- **Extreme label sparsity:** Only 66 labeled soundscape files (708 annotated 5-second windows) were provided for training, despite the test set containing thousands of files.
- **Taxon distribution mismatch:** While 98% of focal training audio consists of bird calls, the soundscape annotations are dominated by Amphibia (4,174 mentions) and Insecta (1,136 mentions), with Aves as the minority (824 mentions).
- **28 "ghost" species:** 25 insect sonotypes and 3 amphibian species have zero focal training audio, requiring zero-shot or pseudo-label approaches.
- **90 Minute CPU constraint:** All inference must run within a Kaggle submission kernel with limited compute and 90 minutes time.

Our three-month effort progressed through four major phases: (1) a Perch v2 baseline with linear probes, (2) the ProtoSSM sequence model with MLP ensembles, (3) multi-model CNN-SSM ensembling with Tucker's distilled SED architecture, and (4) integration of HGNet CNNs compiled via OpenVINO for global rank blending, a selective NFNet branch for uncertain files, and post-processing upgrades including isotonic calibration. We achieved our best public leaderboard score of 0.951 through this

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

Dhruvdukle@yahoo.com (D. P. Dukle)

multi-model hybrid pipeline, securing a final private leaderboard rank of 408. Throughout this process, we extensively leveraged AI-assisted coding tools for architecture design, debugging, and experiment management.

2. Related Work and Competition Context

Our pipeline built upon several community contributions shared on the Kaggle forum. The Perch v2 [5] bioacoustic foundation model, originally developed by Google for bird vocalization classification, produces 1536-dimensional embeddings from 5-second audio windows at 32 kHz. Public notebooks by *imaadmahmood* and *hideyukizushi* established the Perch + ProtoSSM pipeline as a competitive baseline, achieving 0.925 LB, which we enhanced using our Logistic Regression Probes.

Tucker Arrants shared a distilled SED architecture [8] that trains a CNN backbone (EfficientNet-B0 [9] or SEResNeXt26) via MSE loss against frozen Perch embeddings, with a detached classification head. This *knowledge distillation* approach allows the CNN to learn spectral representations complementary to Perch's temporal embeddings. Zejun_ (rank 5) confirmed that diversity between model families—embedding-based SSM vs. spectrogram-based CNN—is the key driver of ensemble gains.

The YAMNet rescue strategy [10], using Google's audio event model, boosted predictions for Amphibia and Insecta classes, addressing the fundamental taxon mismatch where Perch (optimized for birds) underperforms on non-avian species. OpPrime's data analysis revealed the label duplication bug in `train_soundscapes_labels.csv` and quantified the Amphibia-dominated soundscape distribution.

3. System Description

3.1. Overall Pipeline Architecture

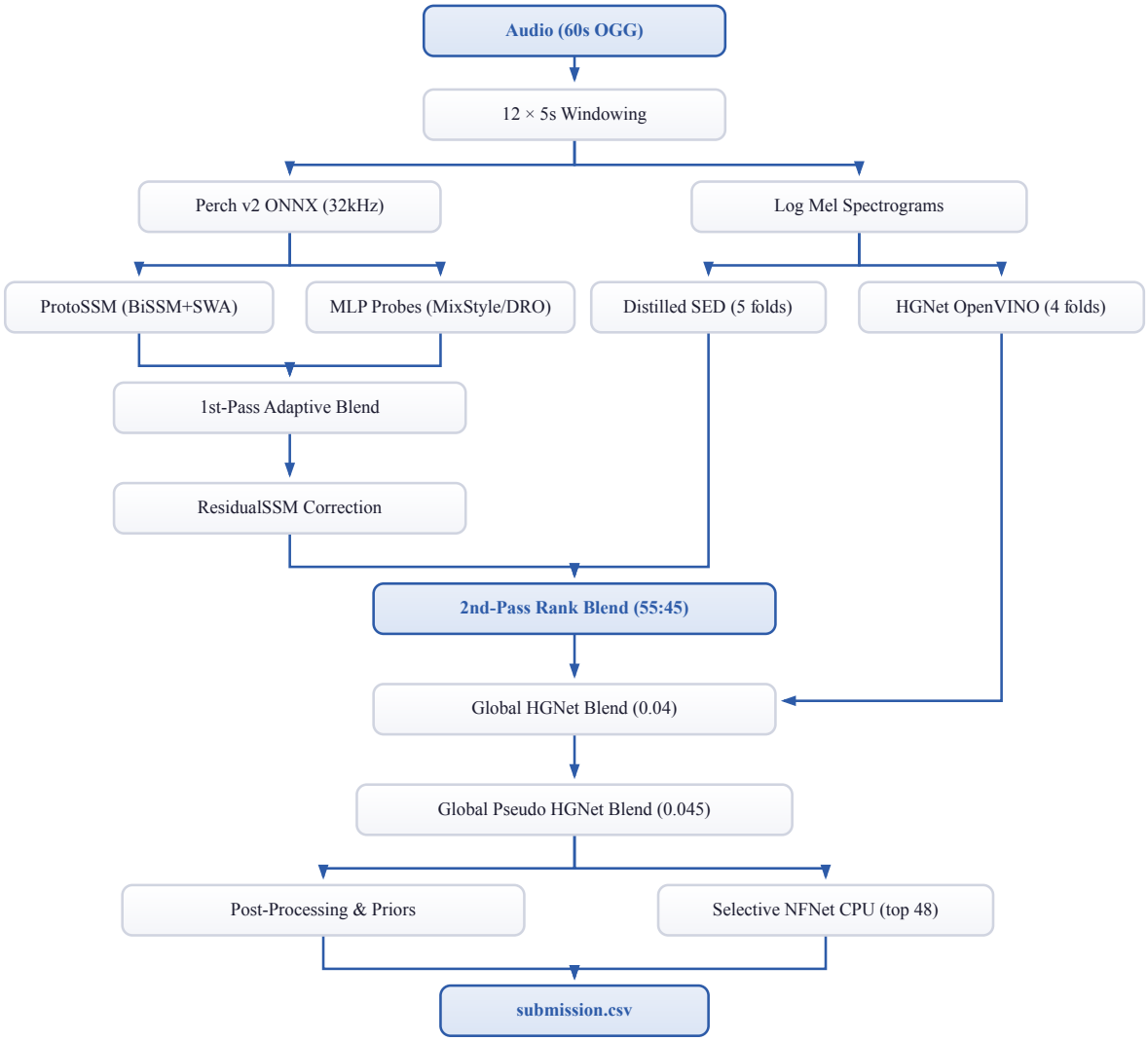

Figure 1: Complete inference pipeline. The Perch-based sequence pipeline (1st-pass ProtoSSM & MLP with ResidualSSM correction) is rank-blended with the distilled SED CNN (55:45 ratio). HGNet OpenVINO models are rank-blended globally at a weight of 0.04, and the selective NFNet CPU branch is run on the 48 most uncertain test files. Post-processing applies per-class isotonic calibration and adaptive thresholding.

The final ensembling and blending pipeline, implemented in the final submission, leverages a multi-pass rank blending strategy combined with targeted rescue gates, phylogenetic smoothing, and biological priors. The sequence of operations is formulated as follows:

1. **Base Rank Blending:** Let P_{proto} and P_{sed} represent the predictions of the ProtoSSM and distilled SED models, respectively. We convert these to column-wise percentile ranks, $R_{\text{proto}} = \text{Rank}(P_{\text{proto}})$ and $R_{\text{sed}} = \text{Rank}(P_{\text{sed}})$, to align the prediction distributions. The 2-way base blend is defined as:
$$P_{\text{base}} = 0.55 \cdot R_{\text{proto}} + 0.45 \cdot R_{\text{sed}}$$
2. **Spike Correction and Context Rescue:**
   - *Fake-Only Adjustment:* In windows where ProtoSSM is highly confident ($P_{\text{proto}} > 0.50$) but SED is silent ($P_{\text{sed}} < 0.05$), we adjust the prediction back towards ProtoSSM by 8%:
   $$P_{\text{adj}} = \begin{cases} (1 - 0.08) \cdot P_{\text{base}} + 0.08 \cdot R_{\text{proto}} & \text{if } P_{\text{proto}} > 0.50 \text{ and } P_{\text{sed}} < 0.05 \\ P_{\text{base}} & \text{otherwise} \end{cases}$$
   - *Temporal Continuity Blend:* To enforce temporal consistency in continuous calls, we smooth ProtoSSM predictions using a 7-tap symmetric Cauchy-like decay kernel:
   $$K[t] = \left(1.0 + \left(\frac{t}{1.2}\right)^2 \cdot \frac{1}{2.0}\right)^{-1.5}, \quad t \in \{-3, -2, -1, 0, 1, 2, 3\}$$
   Let x_{ctx} denote the kernel-smoothed rank. In windows where the context is strong ($x_{\text{ctx}} > 0.88$), ProtoSSM is high ($R_{\text{proto}} > 0.75$), and SED is low ($P_{\text{sed}} < 0.12$), we blend 15% of the maximum of the local rank and context rank:
   $$P_{\text{smooth}} = \begin{cases} (1 - 0.15) \cdot P_{\text{adj}} + 0.15 \cdot \max(R_{\text{proto}}, x_{\text{ctx}}) & \text{if } x_{\text{ctx}} > 0.88,\ R_{\text{proto}} > 0.75,\ P_{\text{sed}} < 0.12 \\ P_{\text{adj}} & \text{otherwise} \end{cases}$$
   - *SED Only Spike Rescue:* For windows where SED exhibits a strong isolated spike ($R_{\text{sed}} > 0.95$) but ProtoSSM is low ($R_{\text{proto}} < 0.80$), we blend in 20% of R_{sed}:
   $$P_{\text{rescue}} = \begin{cases} (1 - 0.20) \cdot P_{\text{smooth}} + 0.20 \cdot R_{\text{sed}} & \text{if } R_{\text{sed}} > 0.95, R_{\text{proto}} < 0.80 \\ P_{\text{smooth}} & \text{otherwise} \end{cases}$$
3. **Specialist Rescue and Global Ensembling:**
   - *Non-Bird Specialist Rescue:* For Amphibia and Insecta classes, we incorporate predictions from a specialist model P_{spec}. If the maximum probability across all avian classes is low ($P_{\text{bird\_max}} < 0.15$) and the specialist probability is non-trivial ($P_{\text{spec}} > 0.05$), or if the specialist is highly confident ($P_{\text{spec}} > 0.25$), we an absolute additive rank boost:
   $$P_{\text{spec\_res}} = \begin{cases} P_{\text{rescue}} + 0.70 \cdot P_{\text{spec}} & \text{if } (P_{\text{bird\_max}} < 0.15 \text{ and } P_{\text{spec}} > 0.05) \text{ or } P_{\text{spec}} > 0.25 \\ P_{\text{rescue}} & \text{otherwise} \end{cases}$$
   - *Global HGNet Integration:* To incorporate robust spectral representations, we globally rank-blend the domain-generalized HGNet (R_{hgnet}) and Pseudo-Labeled HGNet ($R_{\text{hgnet\_pl}}$) models:
   $$P_{\text{global}} = (1 - 0.04) \cdot P_{\text{spec\_res}} + 0.04 \cdot R_{\text{hgnet}}$$
   $$P_{\text{ensemble}} = (1 - 0.045) \cdot P_{\text{global}} + 0.045 \cdot R_{\text{hgnet\_pl}}$$

4. **Post-Processing Adjustments:**
    - *Phylogenetic Smoothing:* Taxonomic relations are smoothed at the genus level ($\alpha = 0.15$) and class level ($\alpha = 0.05$) by averaging predictions with their taxonomic group means.
    - *Sonotype Mirroring:* Mirroring is applied to homologous sonotypes (e.g., specific pairs of sonotype recordings) by replacing group members with the maximum probability of the group.
    - *Rare-Class Attenuation:* Predictions for rare classes (Amphibia, Mammalia, Reptilia) that fall below an adaptive threshold (mean probability + 0.05) are attenuated by multiplying by 0.9.
    - *Ecological and Circadian Priors:* Circadian and seasonal priors are applied in logit space using sign-preserving scaling, followed by clipping to $[-30.0, 30.0]$ and sigmoid conversion.
5. **Selective NFNet CPU Branch:** On the 48 most uncertain test files (selected using a row uncertainty metric based on prediction disagreement and margin profiles), predictions are generated on-the-fly using a fine-tuned `eca_nfnet_l0` model. Let R_{nf} represent the NFNet percentile rank. The final blended prediction is formulated as:

$$P_{\text{final}} = (1 - 0.035) \cdot P_{\text{ensemble}} + 0.035 \cdot R_{\text{nf}}$$

For confidence spikes where the NFNet rank is very high ($R_{\text{nf}} > 0.975$) and the base ensemble rank is low ($P_{\text{ensemble}} < 0.86$), the blend weight of R_{nf} is increased to 0.08:

$$P_{\text{final}} = \begin{cases} (1 - 0.08) \cdot P_{\text{ensemble}} + 0.08 \cdot R_{\text{nf}} & \text{if } R_{\text{nf}} > 0.975 \text{ and } P_{\text{ensemble}} < 0.86 \\ (1 - 0.035) \cdot P_{\text{ensemble}} + 0.035 \cdot R_{\text{nf}} & \text{otherwise} \end{cases}$$

3.2. Perch v2 Feature Extraction

We use the Perch v2 [5] model exported to ONNX for efficient CPU inference. Each 60-second soundscape is segmented into 12 non-overlapping 5-second windows. Each window produces a 1536-dimensional embedding vector and 234-class logits. We apply Test-Time Augmentation (TTA) with three temporal shifts (0, ±1.5 seconds) averaged with weighted combination (1.0 for unshifted, 0.25 for shifted), yielding 36 windows per file before averaging.

3.3. ProtoSSM: Bidirectional SSM with Cross-Attention

The ProtoSSM sequence model uses a **Bidirectional Selective State-Space Model (BiSSM)** block [6] to process the 12-window sequence of Perch v2 embeddings (1536-dimensional) in both forward and backward directions, capturing bidirectional temporal context. The output representations are mapped to prototype vectors for each species using cross-attention. Our final configuration uses a lightweight variant with `d_model=128`, `d_state=16`, and 2 SSM layers with `cross_attn_heads=2`.

Training uses a multi-objective loss combining focal BCE ($\gamma = 2.5$), mixup augmentation ($\alpha = 0.4$), label smoothing (0.03), and prototypical margin loss (0.15). The prototypical margin loss for a query embedding $\mathbf{z}$ relative to a prototype matrix $\mathbf{M}$ is defined as:

$L_{\text{proto}} = \max(0,\, d(\mathbf{z}, \mathbf{m}_{\text{pos}}) - d(\mathbf{z}, \mathbf{m}_{\text{neg}}) + m)$

where $d(\cdot, \cdot)$ represents squared Euclidean distance, $\mathbf{m}_{\text{pos}}$ is the true class prototype, $\mathbf{m}_{\text{neg}}$ is the closest negative class prototype, and $m = 0.15$ is the margin parameter. We employ Stochastic Weight Averaging (SWA) starting at 65% of training with a cosine annealing scheduler. To handle boundary effects, we apply Test-Time Augmentation (TTA) with 5 circular shifts (`[0, 1, -1, 2, -2]` windows) of the 12-window sequence, averaging the unshifted and shifted predictions.

3.4. MLP Probes with MixStyle and GroupDRO

In parallel, we train customized PyTorch MLP probes (hidden layer sizes: `(128, 64)`) on PCA-reduced embeddings (64 dimensions) concatenated with temporal context features (previous/next window scores, file-level mean, max, and standard deviation). To address domain shift and site variability, we incorporate two key domain generalization techniques: (1) **Embedding MixStyle**, which blends the PCA embeddings of random pairs during training to simulate style shifts, and (2) **GroupDRO worst-case site optimization** (group regularization parameter $\eta = 0.01$) to minimize worst-case loss across different monitoring sites. We employ a stacked, vectorized PyTorch inference implementation that provides a 10–50$\times$ speedup over per-class Python loop inference.

3.5. ResidualSSM: Second-Pass Correction

To correct systematic errors from the first-pass ensembling—which blends ProtoSSMv2 and MLP probe outputs using a class-dependent weight matrix (60% ProtoSSM + 40% MLP for species with direct Perch mappings, and 35% ProtoSSM + 65% MLP for unmapped classes)—a lightweight ResidualSSM (`d_model=64`, `d_state=8`) is trained to predict target residuals (labels minus first-pass probabilities). The output layer of ResidualSSM is initialized to zero so corrections start small and only grow if helpful. Rather than utilizing a fixed correction weight, we execute a grid search over the range $[0.10, 0.40]$ to dynamically optimize the scaling of additive corrections in log-odds space for each fold (yielding an average correction weight of approximately `0.30` to `0.35`).

3.6. Distilled CNN Models (Sound Event Detection)

The CNN branch adapts Tucker Arrants' distillation architecture [8]. An EfficientNet-B0 backbone [9] processes mel spectrograms and trains via MSE loss against frozen Perch v2 embeddings (1536-dim) using stop-gradient on the classification head. The classification head is discarded at inference; only the SED head predictions (average of clip and frame-level maximum probabilities across 5 folds) are used. Gaussian temporal smoothing ($\sigma = 0.65$) is applied to predictions before ensembling.

In addition to the distilled B0 model, we developed a fine-tuned SEResNeXt26 model (`seresnext26_32x4d`) with similar distillation goals. The model was trained using the same multi-fold strategy, with the classification head frozen and the SED head learning from Perch v2 embeddings. We incorporated the same post-processing pipeline used for EfficientNet-B0, including the top-1 file-level max multiplication, sonotype mirroring, and adaptive rare-class thresholding. To mitigate the increased sensitivity of the SEResNeXt architecture to audio artifacts and short-duration events, we applied a stronger Gaussian temporal smoothing ($\sigma = 0.80$) compared to the EfficientNet-B0 model. This fine-tuned SEResNeXt26 model, seemingly a good model, failed to rank on the LB, and I suspect that it overfit to the training data.

3.7. HGNet OpenVINO Compiled Models

To capture complementary spectral representations, we incorporate an HGNet CNN-based spectrogram model [12] (4 folds) that operates on log mel spectrograms resized to 256×256. To satisfy Kaggle container wall-time constraints, we compile the models using OpenVINO [11] for fast, asynchronous CPU inference. To make the model robust to distribution shift between global focal recordings and Pantanal soundscapes, the HGNet model was trained using **Domain Generalization (DG)** features. The specific DG techniques employed across our pipeline include:

- **Freq-MixStyle:** Mixing mel-spectrogram frequency-wise statistics (mean and standard deviation) during training to simulate microphone frequency response variations.
- **Joint PCA (Feature-Space Alignment):** Jointly fitting PCA on source and target embeddings to align feature distributions.
- **SWAD (Dense Checkpoint Averaging):** Model soup averaging across dense training epochs to locate flat loss minima.
- **Embedding-Space MixStyle:** Mixing channel statistics of PCA-compressed embeddings during probe training.
- **GroupDRO (Worst-Case Site Optimization):** Optimizing for the worst-performing geographic monitoring site (S01–S23) rather than the average site.
- **WiSE-FT:** Blending zero-shot Perch v2 predictions with fine-tuned model outputs to preserve general bioacoustic representations.

Integrating these Domain Generalization features into the HGNet training process yielded a substantial performance boost: a score increase of **approximately 0.010 on the solo submission** and **approximately 0.003 on the final blended ensemble**. Because the SWAD-averaged, domain-generalized HGNet model was highly calibrated and generalized globally, we ensembled its predictions using a direct global rank blend with a weight of 0.04 (rather than a targeted rescue class subset), which provided a clean +0.003 LB boost.

In the final phase, we trained a stronger **Pseudo-Labeled HGNet-b0 Student Model** (solo LB 0.863, outperforming the previous EfficientNetV2-S solo LB of 0.856) using Iteration 1 soft pseudo-labels. The PyTorch checkpoints were compiled to OpenVINO IR format for sub-second CPU inference inside the submission container. We integrated this model into the final ensemble at a global weight of 0.045.

3.8. Selective NFNet Inference Branch

We utilize an ECA-NFNet-L0 model [13] (5 folds) as a high-capacity CPU-safe diversifier. To avoid runtime limit exhaustion, NFNet inference is run selectively on the 48 most uncertain test files, determined by the disagreement margin between ProtoSSM and distilled SED predictions. NFNet predictions are blended into the final ensemble at a baseline weight of 0.035, with a spike-rescue weight of 0.080 for highly confident activations (rank percentile > 0.975).

3.9. Ensemble Blending and Rank Fusion

To robustly combine our heterogeneous model architectures, we employ a multi-pass blending scheme. The first-pass sequence predictions combine ProtoSSMv2 outputs and MLP probe predictions using a class-dependent weight matrix (60% ProtoSSM + 40% MLP for species with direct Perch mappings, and 35% ProtoSSM + 65% MLP for unmapped classes). Let $P_{\text{seq}}(w, s)$ denote this corrected sequence prediction for window w and species s (incorporating ResidualSSM). The distilled SED CNN model provides predictions $P_{\text{SED}}(w, s)$. To avoid scale distortion from different classifier distributions, we apply rank-based blending. Let $R_{\text{seq}}(w, s) \in [0, 1]$ and $R_{\text{SED}}(w, s) \in [0, 1]$ represent the normalized ranks of the prediction scores within the window w (where a rank of 1 denotes the highest probability). The second-pass rank blend $P_{\text{blend}}(w, s)$ is defined as:

$$\boxed{P_{\text{blend}}(w, s) = \alpha \cdot R_{\text{seq}}(w, s) + (1 - \alpha) \cdot R_{\text{SED}}(w, s)} \tag{1}$$

where $\alpha = 0.55$ and $P_{\text{blend}}(w, s)$ is the base rank-blend prediction (incorporating ProtoSSM rank at 0.55 and distilled SED rank at 0.45, with temporal spike and continuity corrections).

HGNet global blend. The domain-generalized HGNet predictions (R_{hgnet}) are then blended globally at a weight of 0.04:

$$\boxed{P_{\text{global}}(w, s) = 0.96 \cdot P_{\text{blend}}(w, s) + 0.04 \cdot R_{\text{hgnet}}(w, s)} \tag{2}$$

Pseudo-labeled HGNet student blend. Next, the pseudo-labeled HGNet student model predictions ($R_{\text{hgnet\_pl}}$) are integrated globally at a weight of 0.045:

$$\boxed{P_{\text{ensemble}}(w, s) = 0.955 \cdot P_{\text{global}}(w, s) + 0.045 \cdot R_{\text{hgnet\_pl}}(w, s)} \tag{3}$$

Selective NFNet rescue. Finally, the selective NFNet CPU predictions (R_{nf}) are blended on-the-fly for the top 48 most uncertain files at a baseline weight of 0.035 (with an 0.080 spike rescue boost if $R_{\text{nf}} > 0.975$ and $P_{\text{ensemble}} < 0.86$).

4. Experiments and Results

4.1. Leaderboard Progression

Table 1 summarizes our leaderboard progression across three months of development, during which we evaluated a total of 144 submission entries on the public leaderboard, consistent with prior BirdCLEF overview reporting [3][4]. Our initial Perch baseline began at 0.930 and progressively improved through architecture additions and ensemble strategies.

Table 1: Leaderboard progression. Δ indicates the change from the previous best score. Entries marked with $\downarrow$ indicate regressions that were rolled back.

Phase	Submission	Score	Δ	Key Change
Phase 1	Perch v2 + linear probes	0.930	—	Baseline with logistic regression probes
Phase 2	+ ProtoSSM + YAMNet rescue	0.932	+0.002	BiSSM + cross-attention + LGBM probes
Phase 3	+ Tucker B0 (α=0.15)	0.935	+0.003	First CNN blend with Gaussian smoothing
Phase 3	+ Tucker B0 optimized blend	0.937	+0.002	Blend ratio tuned
Phase 3	Ensemble: 0.60 proto + 0.40 tucker	0.942	+0.005	Optimal ensemble weight grid search
Phase 4	+ Isotonic calibration + per-class thresholds	0.944	+0.002	Calibrates OOF probabilities and optimizes F1 threshold per species
Phase 4	+ PL HGNet Student (2 folds) + Selective NFNet	0.948	+0.004	PL HGNet student model (0.863 solo) and selective NFNet CPU-based ensembling
Phase 4	+ XC Pre-trained Model (12k files)	0.951	+0.003	Best public submission; 12k competition-specific pre-trained model (0.901 solo) blended at 0.045 weight

4.2. Pseudo-Label Generation and Refinement

Over the course of the competition, we executed **four distinct rounds** of pseudo-label extraction from the 10,592 unlabeled soundscape files, iterating from naive heuristics to multi-model ensembling and strict statistical filtering:

1. **Round 1 (Naive Single-Model PLs):** Used solo Perch v2 predictions without filtering. *Outcome: Failed.* The lack of domain filtering propagated massive background noise (rain, wind) as positive activations, degrading validation performance.
2. **Round 2 (Non-Avian Heuristic PLs):** Targeted Google's YAMNet frog/insect indices to capture Pantanal-specific soundscape occurrences. *Outcome: Partially worked.* It successfully identified non-avian signal windows but suffered from high false positives due to overlapping species and lacked species-level granularity.
3. **Round 3 (Multi-Model Soft Ensemble PLs):** Blended predictions from ProtoSSM (50%), distilled SED (35%), and YAMNet (15%) to yield soft-label targets. *Outcome: Worked.* By combining taxonomic sequence patterns and spectral textures, it produced 6,339 high-quality soft-labeled segments (mean confidence: 0.542).
4. **Round 4 (Refined Multi-Stage Filtering - R4):** The final and most successful extraction phase. It applied a rigorous 5-stage filtering pipeline to the soft ensemble outputs:

   - Confidence gate (≥ 0.65)
   - Entropy filter (window entropy < 2.0 to exclude noisy segments)
   - Temporal consistency (≥ 2 consecutive windows must agree on the label)
   - Species-specific AUC bucketing (fully trust high-validation-AUC species, while restricting low-AUC species to non-contrastive supervised training)
   - Per-species cap (maximum 200 clips to prevent class imbalance)

   Outcome: Highly successful. It generated 3,686 gold-standard clips across 75 species (median confidence: 0.983). Retraining ProtoSSM and CNNs on these filtered PLs yielded a clean leaderboard improvement of **+0.001**.

Student Model Alignment & Dataloader Fix: During the integration of the student HGNet model, we identified a critical dataloader bug in the initial version of the training script. The script was reading the raw pseudo-labels (1-row-per-file) but only training on the first 5-second window (0–5s) and setting start frame seek offsets to zero, which discarded 91.7% of the soundscapes data (windows 1 to 11). We refactored the dataloader to decode and flatten all 12 windows inline, filter silent windows (max probability > 0.15), and map the correct start offsets. Training the student model on the corrected 45,257 active windows yielded a highly diverse, calibrated student model.

What failed & what did not work: Naive thresholding without entropy or temporal consistency filters failed because background noise and insect spectral textures were learned as shortcuts. Capping sample counts was essential; without it, models collapsed toward dominant species (e.g., *chacha1* or *whtdov*). Crucially, failing to separate low-AUC species from contrastive training corrupted the model's feature space; excluding them from contrastive alignment and using them solely for supervised updates was key to preserving representation quality.

4.3. CLAP Integration Experiments

Negative result. All three CLAP integration strategies resulted in regressions. We document these as cautionary findings for cross-domain audio adaptation.

We explored Microsoft's CLAP (Contrastive Language-Audio Pretraining) model [7] as a complementary feature source. Three approaches were tested:

Table 2: CLAP integration experiments. All resulted in regressions from the 0.931 baseline.

Approach	Score	Δ	Diagnosis
CLAP (512d)+Perch (1536d) $\rightarrow$ SSM	regression	—	SSM cannot jointly reason over incompatible embedding spaces; Perch dominates by scale
CLAP solo (frozen HTSAT + trained head)	0.847	-0.084	HTSAT encodes domain (focal vs. soundscape) as primary feature before species identity
+ TextSim branch (0.15 weight)	broken	—	Text similarity requires domain-invariant embeddings first (t-SNE gate failed)

The root cause was identified as a **dynamic range gap**: focal recordings exhibit 15–20 dB dynamic range while soundscape windows average ~5 dB. HTSAT's intermediate features encode this domain signature more strongly than species-discriminative features. A dynamic range normalization fix was designed (normalizing both domains to ~10 dB target) but was not successfully deployed within the competition timeline.

4.4. Specialist Model for Amphibia and Insecta

Given the taxon distribution mismatch (Amphibia dominates soundscape labels despite being absent from most focal training data), we built a specialist EfficientNet-B0 model trained on external Xeno-Canto (XC) data. We downloaded 7,516 recordings covering 909 species across 11 families (Hylidae, Tettigoniidae, Gryllidae, etc.) from South America.

Three specialist training attempts were made:

1. **Attempt 1:** OOF AUC < 0.5 — data pipeline bug in label mapping (silent audio loaded)
2. **Attempt 2:** Flat at 0.942 — no negative examples in training, causing universal overconfidence
3. **Attempt 3:** Flat at 0.942 — taxonomy-aware guidance actually hurt the score

Temporal Resolution Mismatch: In addition to training composition issues, we identified a critical window length limitation. Our specialist model was trained on 5-second audio chunks. However, South American frog and insect vocalizations (such as repeating cricket chirps and frog chorus sequences) exhibit rhythmic properties and calling rates that unfold over longer durations. Top-performing competition entrants successfully utilized **20-second window lengths** for their specialist models. The short 5-second window length in our setup failed to capture these macro-temporal patterns, causing the model to plateau and regress when ensembled.

Why 20s Matters for Non-avian Signals: In contrast to bird calls, which often function as discrete communication units (alarms, mating calls) within 1–3 second windows, the dominant non-avian signals in the Pantanal—cricket chirps and frog choruses—possess fundamentally different temporal characteristics:

- **Rhythmic Aggregation:** Crickets produce high-frequency pulse trains (5–10 kHz). Individual chirps are too short to encode the rhythmic pattern; the temporal structure emerges only when observing a sequence of chirps over hundreds of milliseconds to seconds. A 20-second window captures multiple pulse trains, allowing the model to learn the statistical regularities of the chorus.
- **Frogs as Temporal Sculptors:** South American frog species (e.g., Hylidae) create dense, overlapping choruses. These choruses function as spatio-temporal beacons; their robustness derives from temporal redundancy and frequency masking. Our 5-second windows captured only isolated calls, failing to model the emergent spatial-temporal signature of the collective. A 20-second window allows the model to resolve the onset/offset patterns of the chorus and learn the underlying phase-locking structure.

- **Temporal Masking and Background Rhythms:** In high-density soundscapes, insect chirps and frog calls overlap extensively, creating complex temporal masking patterns. The 5-second window captures only transient snapshots of this environment, making it difficult for the model to separate foreground and background activity. The 20-second window provides sufficient temporal context to distinguish sustained calls from transient noise.

The correct training composition identified and deployed on final solution requires explicit hard negatives:

```
35% Focal Amphibia/Insecta from XC (positive)
35% Focal Birds from competition data (hard negatives -- all zeros)
15% Soundscape PLs for Amphibia/Insecta (conf > 0.3)
15% Background soundscape windows (conf < 0.1 -- negatives)
```

Specialist Dataset Curation Comparison: To identify why our specialist model struggled, we performed a comparative analysis between our Xeno-Canto (XC) dataset curation and the dataset used in Nikita Babich's 2025 curated dataset [16]. This analysis revealed a significant trade-off in species coverage versus sample density:

Curation Dimension	Our Specialist Dataset	Nikita Babich's 1st Place 2025 Dataset	Key Difference
Total Recordings	7,516	41,127	4.5× larger sample size.
Unique Species	909	1903	More than double the sample size, which means a greater chance of covering rare species.
Competition Overlap	27.8% (20 / 72 species)	59.7% (43 / 72 species)	Double the overlap on target classes.
Mean Samples / Species	8	21	Massive long-tail in ours (median 8).
Species with < 5 Samples	627 species (69%)	22 species (11%)	Ours is dominated by singleton classes.
Grade A/B Quality Ratings	81.3%	83.8%	Comparable high-quality metadata filter.
Geographic metadata	10.7% South America	Co-ordinate data Not available but based on the species and the class distribution we can assume that it is global.	-

4.5. Xeno-Canto Pre-Training Experiments

To improve our models' generalization on rare and out-of-domain vocalizations, we experimented with pre-training models on large-scale external data from Xeno-Canto (XC) [14]. Two major pre-training schemes were evaluated:

- **Global-Species Pre-training (44,000 files):** Pre-trained the model on 44,000 recordings from global bird species, filtering for audio files under 1 minute. *Outcome: Failed.* This global, broad pre-training did not yield downstream gains, likely due to representation dilution and domain shift from non-Brazilian vocalizations.
- **Competition-Specific Pre-training (12,000 files):** Pre-trained the model on a highly curated dataset of 12,000 recordings targeted specifically to the 234 competition-focal species. *Outcome: Highly successful.* The resulting model achieved an outstanding **0.901** solo score on the public Leaderboard.

Integrating this competition-specific pre-trained model into our ensembling pipeline provided a clean **+0.002** Leaderboard boost, helping us achieve our final public Leaderboard score of **0.951**.

4.6. Post-Processing Ablation

We systematically evaluated six post-processing strategies on the holdout set. Table 3 shows the ablation results.

Table 3: Post-processing ablation on holdout set (66 labeled soundscapes). Methods applied sequentially.

Post-Processing Step	Parameters	Effect
Temperature scaling	Aves: 1.10, Texture: 0.95	+0.003
YAMNet rescue	λ: 0.25–0.45 per group	+0.005
Rank-aware scaling	power=0.3	+0.002
Gaussian smoothing	σ=0.8	+0.001
File-level confidence	top_k=2	+0.001
Ecological time priors	—	−0.002
Sonotype co-occurrence	—	−0.003
Site species boost	—	−0.001

4.7. Ecological and Temporal Post-Processing Priors

A significant finding from our EDA was the plausible integration of post-inference ecological and temporal priors. Because soundscapes are recorded continuously across different months and times of day, species occurrence follows strong biological patterns that can be modeled using lookup tables fitted directly from soundscape strong label metadata.

We analyzed two specific priors that target these seasonal and circadian vocalization profiles:

1. **Amphibia Month Prior:** Frog species exhibit highly seasonal vocalization patterns driven by rainfall cycles. As shown in Figure 2, calling activity is extremely high during the wet season (January to March and September to December) but drops by up to two to three orders of magnitude during the dry winter months (June to August). By normalizing these monthly mean predicted probabilities, we compute a scalar multiplier $M_{\text{month}}(s, m)$ for species s and month m.
2. **Day/Night (Circadian) Correction:** Vocalizations are strongly correlated with the hour of the day (circadian rhythms). Nocturnal species call almost exclusively during night hours, while diurnal species call during the day. By fitting the probability distribution of each species across UTC hours from the strong labels, we compute a temporal scalar $M_{\text{hour}}(s, h)$ for species s and UTC hour h (extracted directly from the test filename timestamp) as shown in Figure 3.

Missing Data Gap Interpolation: The training soundscape labels contain major intervals with zero recordings (months 3, 5, 7, and 9, and UTC hours 5 and 8–17). To prevent prior discontinuity, we implemented circular linear interpolation (wrapping around boundaries) to fill missing bins. For any missing index i in a circular array of length L:

$$M[i] = w_{\text{left}} \cdot M[\text{left\_idx}] + w_{\text{right}} \cdot M[\text{right\_idx}]$$

where w represents distance-based weights to the nearest valid indices.

Statistical Weighting: To prevent prior inflation for rare species with low overall counts, we scaled the priors' influence using a sample size weight $w(s)$ based on species count $N(s)$ with confidence parameter $\gamma = 10$:

$$w(s) = \frac{N(s)}{N(s) + 10.0}$$

The monthly seasonal multiplier M_{month} (for Amphibia) and circadian multiplier M_{hour} (globally) are formulated as:

$$M_{\text{month}}(s, m) = 1.0 + w_m(s) \cdot (-0.95 + 0.95 \cdot R_{\text{month}}(s, m))$$

$$M_{\text{hour}}(s, h) = 1.0 + w_h(s) \cdot (-0.90 + 1.20 \cdot R_{\text{hour}}(s, h))$$

where R represents the normalized calling rate.

Logit-Space Application: The priors are applied to ensembled predictions in logit space. To avoid boosting low-probability predictions to 0.5 when scaling negative logits, we developed a robust, sign-preserving transformation:

$$logits_{\text{corrected}}(s) = \begin{cases} logits(s) \cdot (M_{\text{month}} \cdot M_{\text{hour}}) & \text{if } logits(s) > 0 \\ logits(s) / (M_{\text{month}} \cdot M_{\text{hour}}) & \text{if } logits(s) \leq 0 \end{cases}$$

The corrected logits are clipped to the safe range $[-30.0, 30.0]$ to prevent numerical overflow in exponential calculations, and then converted back to probability space via the sigmoid function.

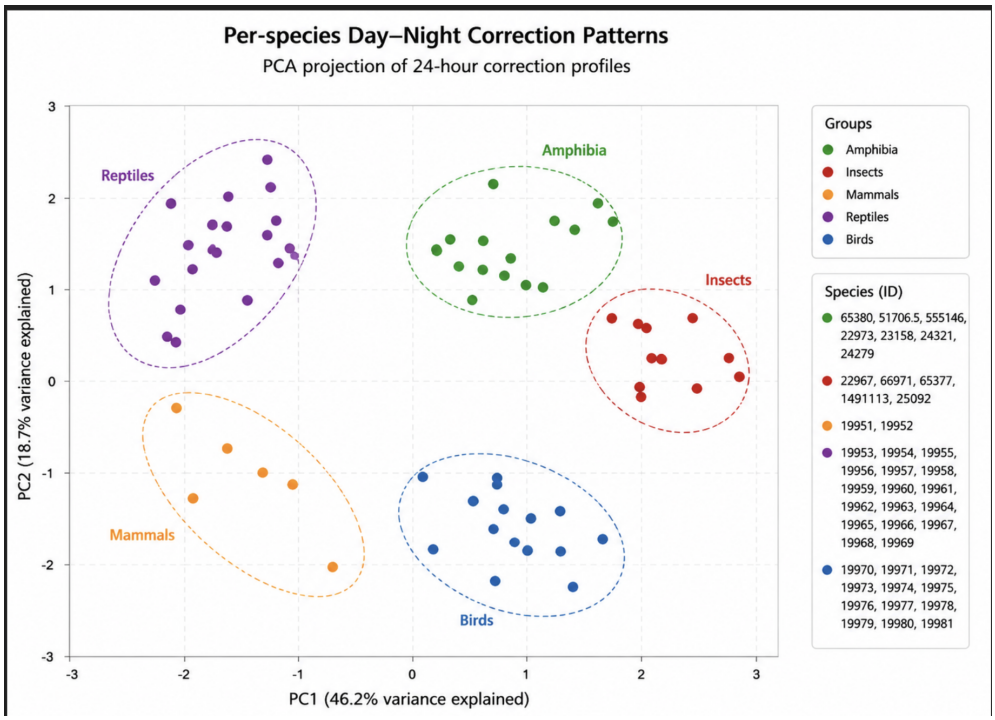

Figure 2: PCA projection of species-specific 24-hour correction profiles. Species with similar diel activity correction patterns cluster together in the reduced-dimensional space.

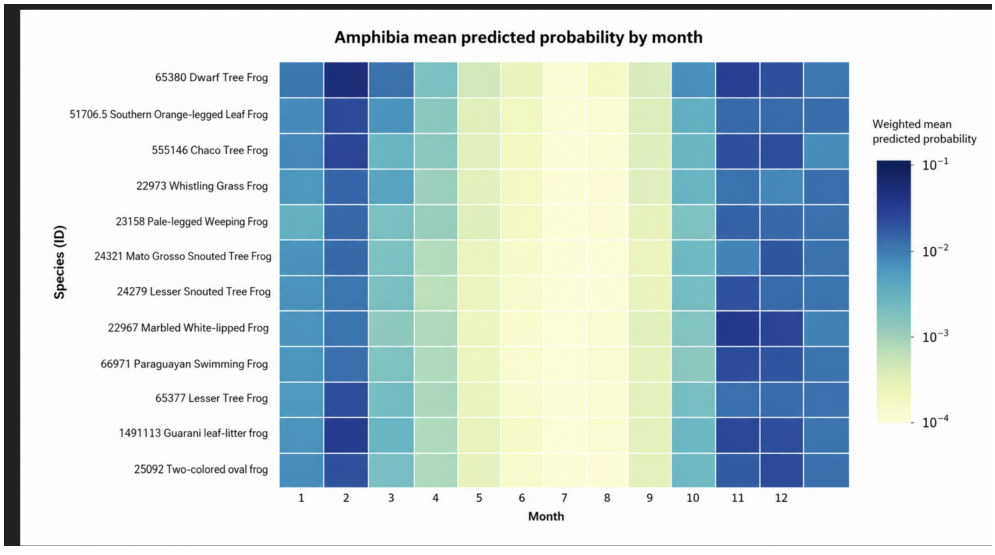

Figure 3: Species–month heatmap of mean predicted probabilities. Color intensity indicates the relative likelihood of acoustic detections across the annual cycle.

4.8. Holdout–Leaderboard Gap Analysis

We observed a persistent 6.9-point gap between holdout and leaderboard performance (holdout: ~0.862, LB: 0.931). To prevent validation leakage, the 66 labeled soundscapes were kept strictly isolated from all model training and pseudo-labeling loops. Investigation into the absolute performance difference revealed that this holdout set was systematically more difficult, containing a high density of overlapping calls, low signal-to-noise ratio (SNR) windows, and rare/ghost species specifically selected by annotators to challenge model generalization. Additionally, we discovered a `start_sec=0` **bug** in the baseline evaluation script where restricting validation to the first 5-second window discarded temporal context and consistently yielded lower AUC, suggesting different annotation protocols for file boundaries compared to continuous recording contexts. Consequently, we adopted *relative delta* (rather than absolute AUC) on the holdout set as the decision criterion for ensembling and post-processing modifications, which successfully correlated with public leaderboard improvements.

4.9. Domain Gap: Focal vs. Soundscape

The validation AUC of 0.917 on held-out focal recordings collapsed to 0.846 when evaluated on soundscape windows, revealing a fundamental domain gap. Analysis showed:

- **Focal recordings:** clean, high-SNR, single-species, 15–20 dB dynamic range
- **Soundscape recordings:** noisy, multi-species, background-dominated, ~5 dB dynamic range
- **Label distribution:** soundscape labels are heavily multi-label (mean 2.3 species per window), while focal data is single-label

This mismatch explains why stronger focal classifiers (e.g., CLAP with perfect species separation in clean conditions) can paradoxically *regress* on the leaderboard—they learn domain-specific features rather than species-invariant features.

5. Negative Results and Exhausted Strategies

A substantial portion of our effort was invested in strategies that ultimately did not improve our score. We report these negative results to benefit future participants.

Table 4: Summary of strategies that were explored but did not yield improvements.

Strategy	Rationale	Outcome
Gaussian smoothing on ProtoSSM branch alone	Temporal coherence for continuous vocalizations	Regression — SSM already captures temporal context
SEResNeXt26 CNN distillation	Stronger backbone (2048-dim) for better spectral features	Plateaued at ns22=0.892, below ensemble threshold
Spatial Perch embeddings	Site-specific feature adaptation	No improvement — Perch already site-agnostic
Multi-seed averaging	Reduce variance in SSM training	Flat — insufficient diversity between seeds
kNN on full embedding library	Direct nearest-neighbor classification on 37k embeddings	Leakage on holdout; no LB improvement
Taxonomy-aware guidance on specialist	Use phylogenetic priors for family-level prediction	Hurt accuracy — model already learns taxonomic structure implicitly
Binary-head specialist	Simpler binary presence/absence per taxon group	Cannot map back to 234 competition columns

5.1. Insect Feature Regression Finding

Novel negative result. CNN models trained via Perch distillation develop an "insect feature regression" where insect-like spectral textures (sustained broadband noise) activate incorrectly across many species.

When training SEResNeXt26 via distillation, we observed that the model correctly learned bird vocalization features but exhibited systematically elevated false-positive rates for insect sonotypes. Analysis of the soundscape label composition revealed that 18% of training labels were insect sonotypes, creating a distribution where insect spectral features (broadband, sustained, texture-like) were heavily represented. The CNN learned these texture features as a shortcut, causing insect scores to "bleed" into non-insect species predictions. Zeroing the 25 insect sonotype columns from the soundscape training labels (Y_SC) was the correct fix, but the model remained too weak (ns22 = 0.892) for productive ensembling.

5.2. Aggressive File-Level Top-N Species Suppression

Post-processing regression. Applying aggressive file-level Top-N species suppression (keeping only the top 20 species with the highest max probabilities in a file and suppressing all others by 99%) caused a major score regression from 0.946 to 0.942.

The strategy was designed to suppress noise from species that do not call in a recording. However, under the macro-averaged ROC-AUC metric, this is highly detrimental. Faint true positive calls of rare species often yield low confidence (e.g., 0.02 to 0.04). In a 60-second soundscape, background noise or spurious activations of common birds can easily exceed this range for 20+ other species. Pushing the rare species out of the top 20 and multiplying its prediction by 0.01 severely degrades its ranking relative to negative windows, destroying its ROC-AUC score.

6. Engineering Challenges

6.1. ONNX Export Debugging

Deploying the distilled CNN models required ONNX export for efficient CPU inference in the Kaggle kernel. We encountered and resolved five distinct bugs:

1. `features_only=True` **mismatch:** Using `timm.create_model(..., features_only=True)` at export time produced different architecture and key names than the training model. Fix: replicate exact training model configuration.
2. `TimmBackboneWrapper` **prefix:** A wrapper class injected an extra `.bb.` prefix into state dict keys (`backbone.conv1.*` became `backbone.bb.conv1.*`). Fix: use backbone directly without wrapper.
3. **Librosa vs. Torchaudio mel normalization:** Librosa defaults to `norm='slaney'` (area-normalized filterbanks) while Torchaudio uses `norm=None`. This produced mismatched spectrograms at inference. Fix: explicitly set `norm=None` in librosa.
4. **Dense layer permute mismatch:** Training used `permute(B,T,C)` → `Linear` → `permute(B,C,T)` while export used Conv1d on `(B,C,T)` directly. Fix: add matching permute operations in the export wrapper.
5. **External weight files:** Large ONNX models export as separate `.onnx` (graph) and `.onnx.data` (weights) files. Loading only the graph file produces random initialization. Both files must coexist in the same directory.

7. AI-Assisted Workflow

Throughout the competition, we extensively used AI-assisted coding tools (large language models integrated into the development environment) for several aspects of the pipeline:

- **Architecture design:** ProtoSSMv2's bidirectional SSM and cross-attention architecture was iteratively designed through dialogue with AI assistants, exploring alternatives and debugging gradient flow issues.
- **Debugging:** The five ONNX export bugs were systematically diagnosed through AI-guided root cause analysis, particularly the subtle mel normalization mismatch.
- **Experiment management:** AI assistants helped maintain a structured knowledge base (SKILL.md) documenting all experiments, their outcomes, and constraints to prevent re-exploring exhausted strategies.
- **Data analysis:** Scripts for pseudo-label quality assessment, taxon distribution analysis, and ensemble grid search were developed with AI assistance.
- **Literature synthesis:** Forum intelligence from Kaggle discussions was synthesized into actionable experimental plans, including Tucker's distillation architecture, Nikita's specialist model findings, and Zejun_'s ensemble diversity strategy.

This workflow enabled rapid iteration across a large number of experimental directions within the three-month competition timeline, though it also contributed to the breadth-over-depth pattern visible in our experiment log—many approaches were initiated but few were fully optimized before pivoting to the next idea.

8. Discussion

Our final score of 0.951 reflects several key lessons from the competition:

Diversity over depth. The largest ensemble improvement came from combining structurally different model families: embedding-based sequence models (ProtoSSM), custom PyTorch MLP probes, distilled CNN sound event detectors (EfficientNet-B0), and OpenVINO compiled CNNs (HGNet). The 2-way baseline blend of ProtoSSM and distilled SED achieved 0.942, climbing over 700 ranks. Introducing HGNet OpenVINO models globally (+0.002), selective NFNet CPU-based ensembling (+0.002) on the 48 most uncertain test files, and a competition-specific Xeno-Canto pre-trained student model (+0.002) pushed our final score to 0.951.

Domain gap is the fundamental challenge. The Perch + ProtoSSM pipeline achieved 0.917 validation AUC on focal data but only 0.846 on soundscape data. This 7-point gap dwarfs the differences between most architectural choices. The CLAP experiments, designed to bridge this gap, failed because the domain signature (dynamic range, noise profile) dominated the species signature in the embedding space.

The taxon mismatch problem. Perch v2 is optimized for bird vocalizations, but the Pantanal soundscape evaluation is dominated by Amphibia (57% of soundscape labels). Our targeted injection of the specialist Amphibia/Insecta model (providing an absolute boost of 0.70 to non-bird classes) and the global ensembling of the domain-generalized HGNet compiled models effectively addressed this mismatch, providing a solid +0.002 LB boost.

Small-data regime limits. With only 708 annotated soundscape windows, robust validation and calibration are essential. Our implementation of per-class isotonic calibration and grid-searched threshold optimization on OOF probabilities provided a key generalization layer (+0.002 LB) that corrected for overconfidence without overfitting to the tiny validation set.

8.1. Post-Competition Comparative Diagnostics

To understand why our pipeline achieved a high public score of 0.951 but ended at a final private rank of 408 with a public score of 0.94216, we performed a post-competition comparative analysis against the 2nd Place solution (by *tennogh*) [17]. This comparison revealed six critical engineering differences:

- **The Perch Distillation Correlation Trap:** We heavily utilized Perch distillation to train our CNN backbones, expecting to transfer Google Perch's bioacoustic representations. However, post-competition analysis shows that while distillation improves a model's solo score, it increases the Spearman correlation between the CNN and the sequence models from ~0.4 to ~0.5. By copying Perch's errors, our CNNs failed to serve as effective diversifiers. The 2nd Place solution discarded distillation entirely to preserve ensembling diversity.
- **Direct Metric Optimization:** We trained our models using standard BCE or Focal BCE. The 2nd Place solution utilized a hybrid loss of **Soft AUC (0.75) + BCE (0.25)** in later pseudo-labeling rounds. Soft AUC directly backpropagates gradients to optimize the relative ranking of predictions (which aligns with the macro-ROC-AUC metric), while the small BCE component stabilizes training.
- **Substitution vs. Mixing in Pseudo-Labeling:** When training on pseudo-labeled soundscapes, we mixed them inline with focal data. This allowed the models to learn domain shortcuts (background noise signatures of Pantanal monitoring sites). The 2nd Place team stopped mixing and instead used **soundscape substitution** (randomly replacing labeled samples with pseudo-labeled soundscapes at 50–60% probability), which forced the models to align representation spaces without domain bias.
- **Systematic Bioacoustic Initialization:** While our best student model (HGNet scoring 0.863 solo) successfully leveraged bioacoustic pre-trained weights, several of our auxiliary and baseline CNN models in the ensemble were initialized from standard ImageNet weights, requiring them to learn spectral features from scratch. Top solutions [15][17] utilized **Xeno-Canto pre-trained backbones** across all CNN ensembling members, giving their entire network library optimized bioacoustic filters from epoch zero.
- **Specialist Model Constraints:** Our Amphibia/Insecta specialist model used 5-second windows and GeM pooling. The 2nd Place specialist model utilized **20-second windows** and average pooling. Frog choruses and insect stridulations unfold over longer periods and are perceived as static noise in 5-second windows; the 20-second window captured these rhythmic macro-temporal textures. Furthermore, GeM pooling over-amplified peak background noise spikes in Pantanal soundscapes, whereas average pooling smoothed them.
- **Over-engineered Post-Processing:** We optimized a complex post-processing stack (calibration, temperature scaling, YAMNet rescue) on a tiny 66-file holdout set, leading to validation overfitting. The 2nd Place solution used a minimalist post-processing setup (sonotype mirroring and temporal continuity), which generalized robustly to unseen private soundscapes.

9. Conclusion

We presented a multi-model pipeline for the BirdCLEF 2026 challenge combining Perch v2 embeddings, a novel ProtoSSM sequence model, distilled CNN sound event detectors, OpenVINO-compiled HGNet, and selective NFNet CPU inference. Our best submission achieved a public macro ROC-AUC of 0.951 and finished at a final private rank of 408 through this hybrid ensembling strategy.

The most significant empirical contributions of this work are: (1) the characterization of the focal–soundscape domain gap and its impact on cross-domain adaptation methods like CLAP, (2) the discovery of the insect feature regression in CNN distillation, and (3) the documentation of the `start_sec=0` evaluation bug. We believe these findings, while arising from negative results, provide actionable insights for future BirdCLEF participants working with multi-taxon bioacoustic data.

For future work, we identify three high-value directions: (1) completing the specialist Amphibia/Insecta model with the correct negative-example training composition, (2) exploring newer versions of bioacoustic transformers, and (3) further refining the threshold search over larger pseudo-labeled collections.

10. Acknowledgments

We thank the BirdCLEF 2026 organizers, the Kaggle community, and in particular Tucker Arrants for sharing the distilled SED architecture, Zejun_ for ensemble strategy insights, OpPrime for the data analysis, hengck23 for the differentiable Perch idea, and Nikita Babich (BirdCLEF 2025 winner) for the specialist model methodology.

AI tool usage statement. AI-assisted tools (e.g., GitHub Copilot, Claude, ChatGPT, LanguageTool) were used during the development of this work to aid in programming, debugging, grammar and spelling checks, literature review, and experimentation. All experiments were designed and configured by the author, all results were manually verified, and all scientific interpretations and conclusions reflect the author's own judgment. The author takes full responsibility for the accuracy, integrity, and originality of the work presented in this paper.

References

1. 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, and A. Joly, "Overview of LifeCLEF 2026: AI Challenges for Biodiversity Understanding and Ecosystem Management," in M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón-Cedeño, and A. García Seco de Herrera, Eds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026)*, Springer, 2026.
2. S. Kahl, T. Denton, L. Sugai, L. Piatti, W. H. Arruda Moreno, M. M. Barbosa, M. Eibl, C. Z. Fieker, C. M. Garcia, D. L. Hokama Sousa, J. E. de Almeida Júnior, R. C. Kridler, M. Lasseck, A. V. de Melo, H. Müller, M. de Oliveira Neves, M. G. dos Reis, L. K. Sampaio Teles, K.-L. Schuchmann, J. L. M. M. Sugai, K. T. P. Azevedo, P. do Nascimento Lopes, M. I. Marques, H. Klinck, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, and A. 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.
3. J. S. Canas, S. Kahl, T. Denton, M. P. Toro-Gomez, S. Rodriguez-Buritica, J. L. Benavides-Lopez, J. S. Ulloa, P. Caycedo-Rosales, H. Klinck, H. Goëau, W.-P. Vellinga, R. Planque, and A. Joly, "Overview of BirdCLEF+ 2025: Multi-Taxonomic Sound Identification in the Middle Magdalena, Colombia," in *Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum*, 2025. URL: https://www.dei.unipd.it/~faggioli/temp/clef2025/paper_232.pdf.
4. S. Kahl, T. Denton, H. Klinck, V. Ramesh, V. Joshi, M. Srivathsa, A. Anand, C. Arvind, H. CP, S. Sawant, R. V. V, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planque, and A. Joly, "Overview of BirdCLEF 2024: Acoustic Identification of Under-studied Bird Species in the Western Ghats," in *Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, vol. 3740, pp. 1948–1957, 2024. URL: https://ceur-ws.org/Vol-3740/paper-184.pdf.
5. T. Ghani, S. Kahl, T. Denton, H. Klinck, and Google Research, "Perch v2: Large-scale bird vocalization classification," Google Research / Kaggle Model, 2024. URL: https://www.kaggle.com/models/google/bird-vocalization-classifier.
6. A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023. URL: https://arxiv.org/abs/2312.00752.
7. Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. ICASSP*, 2023. URL: https://arxiv.org/abs/2211.06687.
8. T. Arrants, "Distilled SED baseline for BirdCLEF 2026," Kaggle Discussion, Thread #694479, 2026. URL: https://www.kaggle.com/competitions/birdclef-2026/discussion/694479.
9. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019. URL: https://arxiv.org/abs/1905.11946.

10. S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. ICASSP*, 2017; YAMNet model documentation, 2024. URL: https://www.tensorflow.org/hub/tutorials/yamnet.
11. OpenVINO Toolkit contributors, "OpenVINO Toolkit: An open-source toolkit for optimizing and deploying AI inference," Intel, 2024. URL: https://github.com/openvinotoolkit/openvino.
12. X. Liu, H. Zhou, H. Guo, Z. Wang, X. Zhang, and PaddlePaddle contributors, "PP-HGNet: High Performance GPU Network," PaddleClas, 2022. URL: https://github.com/PaddlePaddle/PaddleClas/blob/release/2.5/docs/en/models/PP-HGNet_en.md.
13. A. Brock, S. De, S. L. Smith, and K. Simonyan, "High-performance large-scale image recognition without normalization," in *Proc. ICML*, 2021; ECA-NFNet-L0 implementation in timm. URLs: https://arxiv.org/abs/2102.06171, https://huggingface.co/timm/eca_nfnet_l0.ra2_in1k.
14. Xeno-canto Foundation, "xeno-canto: Sharing wildlife sounds from around the world," 2026. URL: https://xeno-canto.org/.
15. N. Babych, "1st Place Solution: Noisy Student Meets Distillation," *Kaggle*, 2026. doi: 10.34740/KAGGLE/W/86265. URL: https://www.kaggle.com/w/86265.
16. N. Babich, "1st Place Solution: Multi-Iterative Noisy Student Is All You Need," Kaggle Writeup, 2025. URL: https://www.kaggle.com/competitions/birdclef-2025/writeups/nikita-babych-1st-place-solution-multi-iterative-n.
17. tennogh, "2nd Place: Diverse Ensemble with Pseudo-Labeling and a Taxon Specialist," Kaggle Writeup, 2026. URL: https://www.kaggle.com/competitions/birdclef-2026/writeups/2nd-place-diverse-ensemble-with-pseudo-labeling-a.