Cross-Mutated Semi-Supervised Learning and Selective State-Space Fusion for BirdCLEF+ 2026 Soundscape Recognition

Musa Tur Farazi[1], Nufayer Jahan Reza[1]

[1]Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

Abstract

We present our solution to the BirdCLEF+ 2026 soundscape recognition challenge, which achieved approximately 96.0% on the Public Leaderboard and 95.4% on the Private Leaderboard. The task required detecting 234 animal sound classes in consecutive 5-second segments of one-minute passive-monitoring recordings from the Brazilian Pantanal. Our system is based on five key components: a strong SED ensemble with ECA-NFNet-L0 and EfficientNetV2-S backbones, in-domain adaptation with soundscape pseudo-labels, noisy-student training with power-sharpened soft targets and waveform mixup, cross-family distillation to improve model diversity after self-training saturated, and a Perch-based temporal branch using a selective state-space sequence model and per-class MLP probes. At inference time, we combine the SED and Perch branches by per-class rank blending and apply lightweight taxonomy-aware post-processing for selected insect groups. We conduct ablation analyses to evaluate the contribution of each training stage, compare same-family self-training with cross-mutation distillation, and analyze the effect of adding the Perch branch. Our results show that soundscape-domain pseudo supervision provides the largest single improvement, while backbone diversity and representation-based temporal modeling further improve the robustness of the final ensemble.

Keywords

BirdCLEF, bioacoustics, pseudo-labeling, noisy student, state-space models, ensemble learning

1. Introduction

BirdCLEF+ 2026 [1, 2] addressed multi-label acoustic event recognition in one-minute passive-monitoring soundscapes. For each recording, participants were required to estimate the presence of 234 target classes in consecutive 5-second segments by their approaches. The taxonomy in the competition dataset comprised bird species together with non-avian acoustic targets, including insects and amphibians whose vocalizations often exhibited texture-like temporal structure. The competition and dataset were challenging for two main reasons. First, the supervision was weak and sparse, since many target events occupied only a short portion of a 60-second recording. Second, several non-bird classes were better characterized as persistent or background acoustic textures than as isolated call events, which limited the effectiveness of methods designed only for short, discrete bird vocalizations.

Our approach followed an iterative semi-supervised distillation strategy. We first trained supervised sound event detection models and used them to annotate unlabeled soundscape segments with soft pseudo-labels. These pseudo-labels were subsequently sharpened, combined with labeled recordings through waveform-level augmentation, and used to train stronger student models that were promoted to new teachers. After same-family self-training started to saturate and occasionally degraded leaderboard performance, we switched to a cross-mutation scheme in which each model family was supervised by teachers from different backbone families. In addition to this SED branch, we trained ProtoSSM, a Perch-embedding sequence model, and included it as an artifact-based inference component. The final submission was produced by rank-blending the ProtoSSM/Perch predictions with a strong ensemble of SED models.

Our principal contributions are:

- a cross-mutation distillation scheme that breaks same-family self-training saturation by using pseudo-label teachers from a different backbone lineage;

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

- power-sharpened waveform mixup as a robust method for consuming noisy soft pseudo-labels without amplifying teacher errors;
- ProtoSSM, a lightweight Mamba selective state-space model with per-class prototype readouts that models the twelve-window temporal structure of each soundscape using Perch embeddings;
- an empirical diagnosis of same-family self-training saturation, showing that repeating the pseudo-label cycle within one backbone lineage can degrade performance even when the teacher ensemble is strong.

2. Related Work

The BirdCLEF competition series has consistently pushed the state of the art in passive acoustic monitoring. Kahl et al. introduced endangered-species identification from soundscape recordings in 2022 through LifeCLEF [3] and expanded geographic scope and taxonomic breadth in subsequent editions [4, 5]. BirdCLEF+ 2025 [6] extended the target set beyond birds to include amphibians, insects, and mammals in the Magdalena Valley, Colombia. Our approach addresses the BirdCLEF+ 2026 task [1, 2], which adopts the same multi-taxonomic recognition format as its predecessor but is evaluated on a distinct 234-class taxonomy recorded in the Brazilian Pantanal. As the competition has grown in scale, top-performing systems have converged on a shared recipe: CNN backbones applied to log-mel spectrograms, followed by attention-based pooling to aggregate frame-level evidence into clip-level predictions. Kong et al. [7] formalized this design at scale with PANNs, showing that class-wise attention weights naturally align with weakly labeled audio. BirdNET [8] applied the same spectrogram-CNN paradigm specifically to avian species, setting a strong bioacoustic baseline across hundreds of classes.

CNN backbone choice has a measurable impact on downstream performance. NFNet [9] replaces batch normalization with adaptive gradient clipping, achieving competitive accuracy at large batch sizes without the instability that normalization-free training can introduce at scale. ECA-Net [10] augments any CNN with a lightweight channel-attention module that uses a single 1-D convolution over neighboring channels in place of the fully connected squeeze-and-excitation gate. EfficientNetV2 [11] revisits model scaling with fused-MBConv blocks and progressive image resizing to improve training speed without sacrificing accuracy. These three lines of work directly motivated the choice of ECA-NFNet-L0 and EfficientNetV2-S as the SED backbone families in our system.

The severe class imbalance and domain gap between focal training recordings and passive-monitoring test soundscapes made semi-supervised learning essential. Lee [12] established pseudo-labeling as a simple and effective baseline for semi-supervised learning, and Xie et al. [13] scaled it into the noisy student framework, where a student trained with augmentation and stochastic noise on teacher soft labels was iteratively promoted to teacher. Rather than hard-thresholding pseudo-labels, our pipeline applied a power transformation to sharpen soft targets while retaining calibration. Waveform-level mixup [14] served as the primary augmentation, creating interpolated multi-label targets that better reflected the overlapping species structure of soundscape files. When direct self-training saturated, knowledge distillation [15] with cross-family teachers—where one backbone architecture supervised a student from a different family—provided additional diversity and prevented reinforcing the same model errors.

Complementing the SED branch, Ghani et al. [16] utilized Perch, a global birdsong embedding model covering thousands of species, and showed that its representations transfer strongly across geographies and low-resource taxa. In our system, Perch embeddings were consumed by ProtoSSM, a sequence model that processed the twelve 5-second windows of each soundscape. ProtoSSM was built on the Mamba selective state-space architecture [17], which extended structured state spaces [18] by making the recurrence parameters input-dependent, enabling linear-time sequence modeling with content-aware gating. Per-class prototype embeddings, which followed the cosine-similarity readout of prototypical networks [19], provided a few-shot-style inductive bias that was especially useful for rare species where supervised training data was limited.

3. Solution Overview and Overall Pipeline

Figure 1 summarizes our end-to-end pipeline. We organized our system into two branches: a multi-stage SED ensemble that we trained through supervised learning, pseudo-concatenation, noisy-student refinement, and cross-mutation distillation; and an artifact-only Perch–ProtoSSM branch that we used to model temporal context across the twelve evaluation windows of each file. For the inference pipeline, we combined the two branches by per-class rank blending and then applied lightweight taxonomy-aware post-processing.

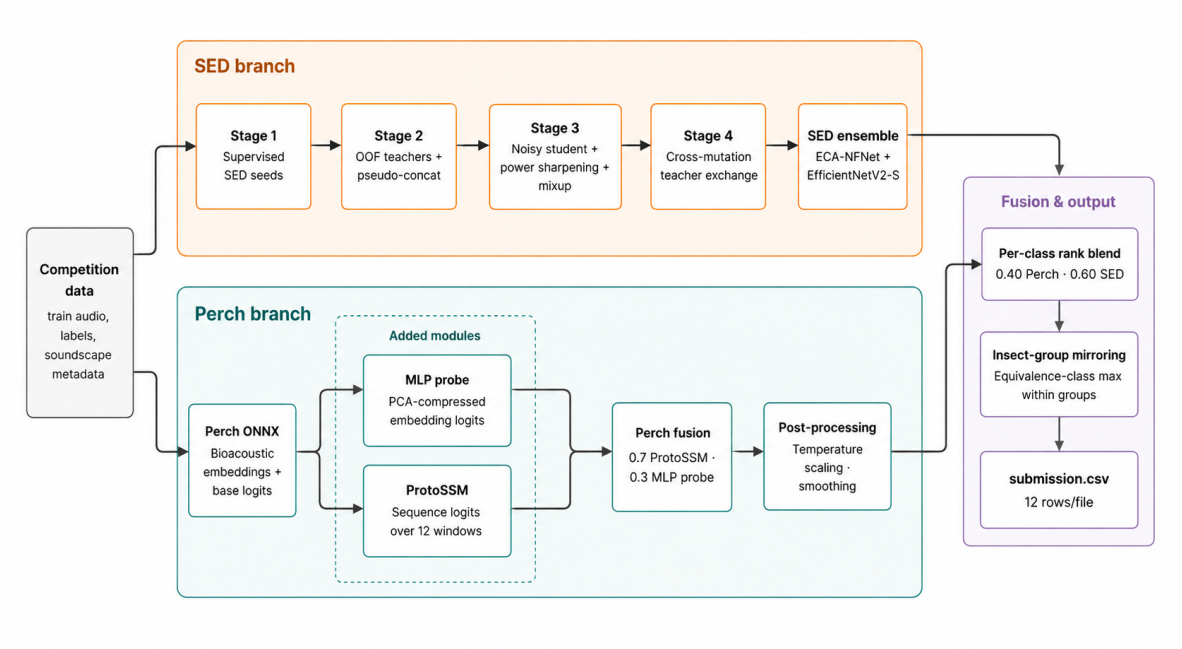

Figure 1: Overall solution workflow: multi-stage SED training, Perch–ProtoSSM inference, rank blending, and taxonomy-aware post-processing.

4. Task Definition and Notation

We denoted a 60-second soundscape by

$$x_f \in \mathbb{R}^{60 \cdot 32000},$$

where f indexed the file and the audio was sampled at 32 kHz. We divided each file into

$$K = 12$$

non-overlapping 5-second evaluation windows. We used the target label dimension

$$C = 234.$$

For each file, our model produced a prediction matrix, which can be written as,

$$P_f \in [0,1]^{K \times C},$$

where $P_{f,t,c}$ represented our predicted probability that class c was active in window t of file f.

For supervised training, we represented the observed labels by

$$y_{i,c} \in \{0,1\},$$

where i indexed a training crop or labeled window. For pseudo-labeled soundscapes, our teacher models produced soft labels

$$q_{i,c} \in [0,1].$$

5. Dataset Exploration

We performed a focused exploration of the competition metadata and audio to understand the structure of the task before choosing model families or training schedules. We examined the target taxonomy, per-class recording counts, source quality, geographic metadata, labeled soundscape coverage, and representative spectrogram morphology. Figures 2–8 summarized this analysis and showed how our later design choices were grounded in the observed data rather than selected only from generic audio-classification practice.

Our EDA highlighted three properties that shaped the final solution. First, the taxonomy was not purely avian: although birds dominated the label space, the target set also contained non-bird animal classes and texture-like acoustic targets, especially insects and amphibians. We therefore treated the task as general soundscape recognition instead of bird-only call detection. Second, the isolated train_audio recordings were strongly imbalanced across species and recording sources. This made it unlikely that a supervised-only model would learn reliable decision boundaries for rare or noisy classes, motivating our later use of balancing, pseudo-labeling, and auxiliary soundscape supervision. Third, the labeled soundscape segments were particularly valuable because they were closer to the test distribution than isolated focal recordings: they contained overlapping species, background noise, variable distance to the microphone, and long-duration acoustic textures. For this reason, we used the EDA not only as a descriptive step, but also as a guide for pseudo-label filtering, waveform-level mixing, backbone diversity, and the final fusion between SED models and Perch-based sequence models.

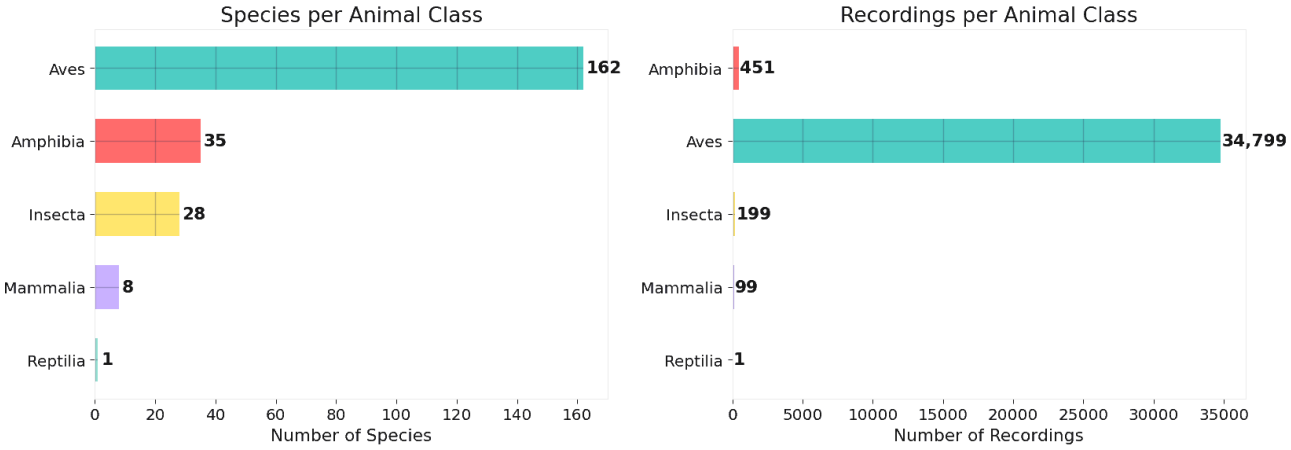

Figure 2: Taxonomic breakdown: 234 target classes dominated by birds, with non-bird acoustic textures (insects, amphibians) requiring class-specific handling.

Figure 2 shows that birds dominated the 234-class label space, but the task also included amphibians, insects, and other animal classes. This matters because these groups did not share the same acoustic structure. Bird detections were often short, harmonic, and event-like, whereas insect and amphibian labels could correspond to sustained broadband or rhythmic background textures. Consequently, through our approach, we treated the task as general soundscape recognition rather than bird-only classification, and later post-processing kept special handling for texture-like insect groups.

The class-level taxonomy also guided the construction of balanced training batches. Although non-bird groups account for a smaller fraction of the label space, they can be acoustically salient in passive-monitoring recordings and may dominate the full-minute background texture. Treating these classes simply as rare labels would therefore underestimate their contribution to the soundscape. For this reason, in later training stages we retained multi-taxonomic sampling and avoided bird-specific heuristics when selecting pseudo-labeled soundscape windows.

Figure 3 further shows a pronounced long-tailed distribution in the number of available recordings per class. A small subset of classes has many isolated focal recordings, whereas many target species have limited direct supervision. This imbalance makes a purely supervised objective vulnerable to biased

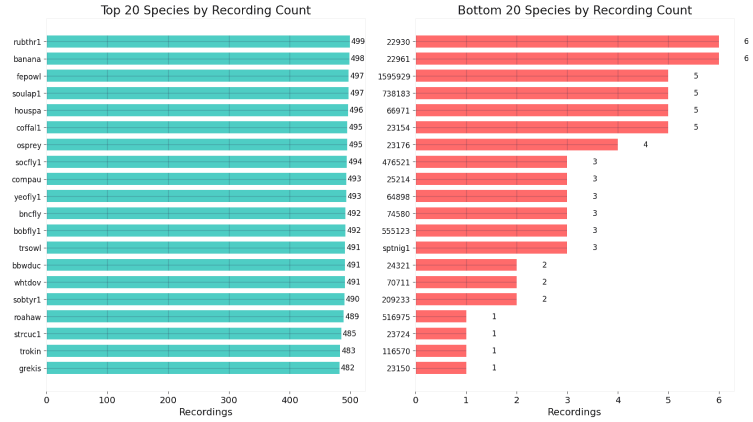

Figure 3: Long-tailed species recording distribution across target classes.

optimization: common species appear disproportionately often in random minibatches, rare species provide few positive examples, and class-wise threshold selection becomes less stable. Accordingly, the pseudo-labeling stages, repeated soundscape sampling, focal loss, and mixup were used not only to improve domain matching, but also to increase the exposure of tail classes to positive or weakly positive training windows.

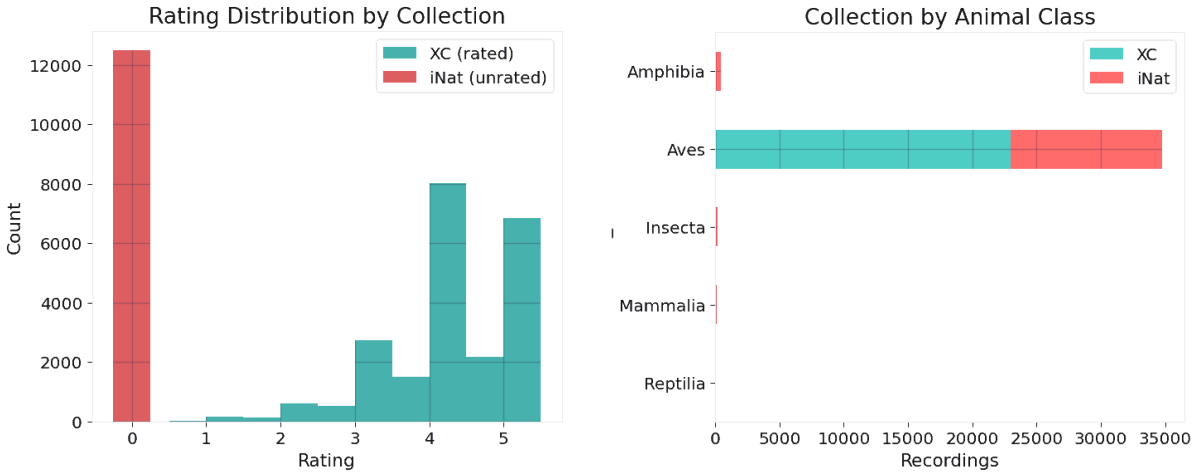

Figure 4: Recording quality and collection source. The panels show the Xeno-canto quality-rating distribution and source composition by animal class.

Figure 4 shows the comparison between recording source and quality metadata. We found that the focal training data combined community-contributed recordings made with heterogeneous microphones, background noise levels, caller distances, and annotation conventions. Xeno-canto files included explicit quality ratings, whereas iNaturalist samples did not provide the same quality scale. This heterogeneity led us to use conservative regularization: we favored label smoothing, stochastic augmentation, BCE–focal loss, and soft pseudo-labels over hard filtering rules that might discard useful but noisy rare-class evidence.

Our source-quality analysis also showed that metadata could not be applied uniformly across the full corpus. A high-quality Xeno-canto clip and an unrated iNaturalist sample may both contain useful target evidence, but their noise profiles and annotation reliability differ. We therefore relied more on augmentation and teacher confidence than on rigid source-specific exclusion rules.

The geographic map in Figure 5 highlights one of the most important sources of domain shift in the task. The isolated training recordings spanned a broad geographic range and were collected under many different ecological and recording conditions, whereas the evaluation soundscapes were concentrated

in the Brazilian Pantanal. Consequently, a model that performed well on focal recordings could still be poorly calibrated for the deployment domain. It could learn species-level vocal signatures, but misestimate the likelihood of those species under Pantanal-specific background noise, microphone placement, seasonal acoustic activity, and local co-occurrence patterns. The gap was therefore not only a geographic mismatch, but also a mismatch in acoustic context: clean or near-focal clips differed substantially from one-minute passive-monitoring recordings in which target calls were distant, overlapping, intermittent, or embedded in persistent insect and amphibian choruses.

All these observations shaped both validation strategy and training design. Local out-of-fold scores were useful for detecting gross model failures and comparing supervised baselines, but they were treated as an incomplete proxy for leaderboard performance because the local folds could not fully reproduce the Pantanal soundscape distribution. We therefore used pseudo-labeling on soundscape-like audio as the main adaptation mechanism. Teacher models transferred taxonomic knowledge from the diverse focal recordings into unlabeled or weakly labeled soundscape windows, while later student stages learned from those soft labels under augmentations that better matched the evaluation setting. In this sense, pseudo-labeling served as the practical bridge between broad supervised taxonomic coverage and the narrower, context-dependent acoustic conditions of the test soundscapes.

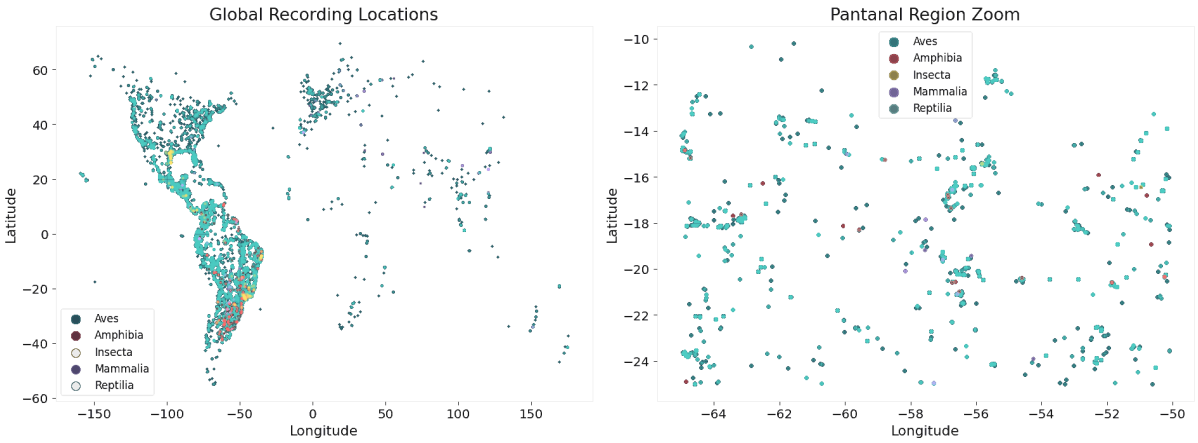

Figure 5: Geographic spread of training recordings. Global collection sites contrast with the Pantanal-focused test soundscapes.

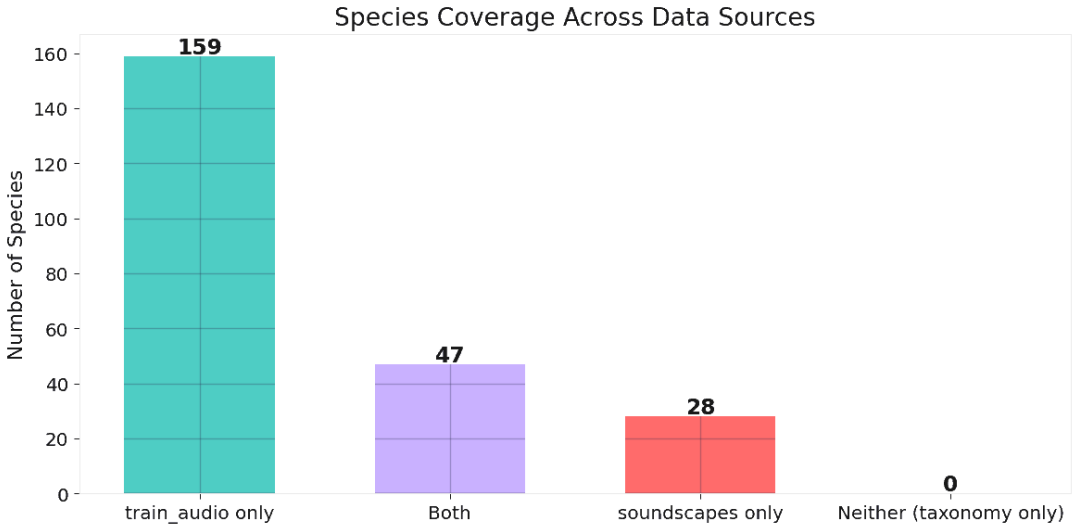

Figure 6: Overlap between taxonomy, isolated audio, and labeled soundscapes.

Figure 6 shows that the taxonomy, isolated recordings, and labeled soundscape annotations provided

complementary, rather than interchangeable, sources of supervision. Isolated audio offered broad taxonomic coverage, whereas labeled soundscapes provided examples that most closely matched the evaluation domain. Because some targets were better represented in one source than the other, our training recipe combined focal recordings with soundscape examples instead of treating either source as sufficient on its own.

This overlap structure also informed the pseudo-labeling protocol. Classes observed only in isolated recordings required careful transfer into soundscape conditions, while classes present in labeled soundscapes offered especially valuable examples of background noise, overlapping vocalizations, and weak temporal supervision. We therefore used out-of-fold pseudo-label generation to expand soundscape-style supervision without feeding a model's own training labels back into itself. The same reasoning guided later stages, where focal clips and soundscape windows were mixed within a unified training stream rather than optimized as separate problems.

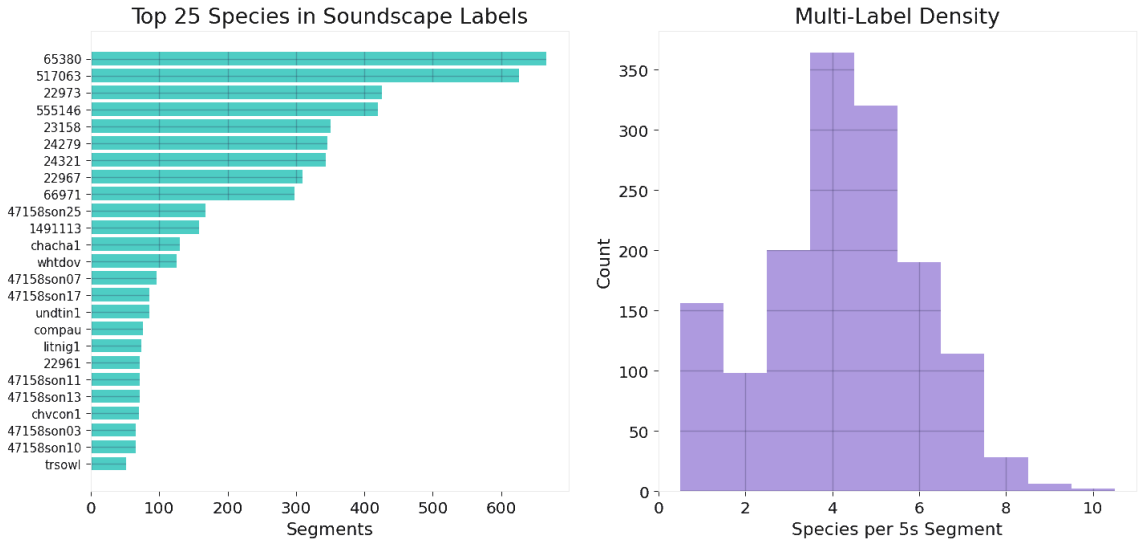

Figure 7: Labeled soundscape segment distribution. The panels show the most frequent soundscape labels and the number of active species per 5-second segment.

The labeled-soundscape analysis in Figure 7 further shows that the evaluation unit was not a full recording but a sequence of 5-second windows. Positive evidence was sparse in time, and different classes had different segment densities across the labeled soundscapes. This supported the choice of SED models that emitted window-level probabilities, and it also justified using longer context for selected EfficientNetV2-S models: surrounding windows could contain habitat, time-of-day, or repeated-call context even when the target call was brief.

The two panels in Figure 7 were especially useful for choosing sampling strategy. The top-species plot identified labels that repeatedly occur in soundscape annotations, while the multi-label-density plot showed that a single 5-second segment could contain several simultaneous targets. These observations led to confidence-weighted soundscape sampling and raw-waveform mixup, because training examples often needed to represent mixtures rather than isolated single-species calls.

Finally, Figure 8 provided qualitative waveform and spectrogram examples across animal groups. These examples confirmed that the label space mixed several acoustic regimes: narrow-band bird vocalizations, pulse trains, noisy broadband calls, and nearly stationary insect textures. The visual differences motivated log-mel spectrogram inputs, SpecAugment-style masking, raw-waveform mixup, and the final ensemble design in which a task-specific SED branch was complemented by a Perch embedding branch with temporal sequence modeling.

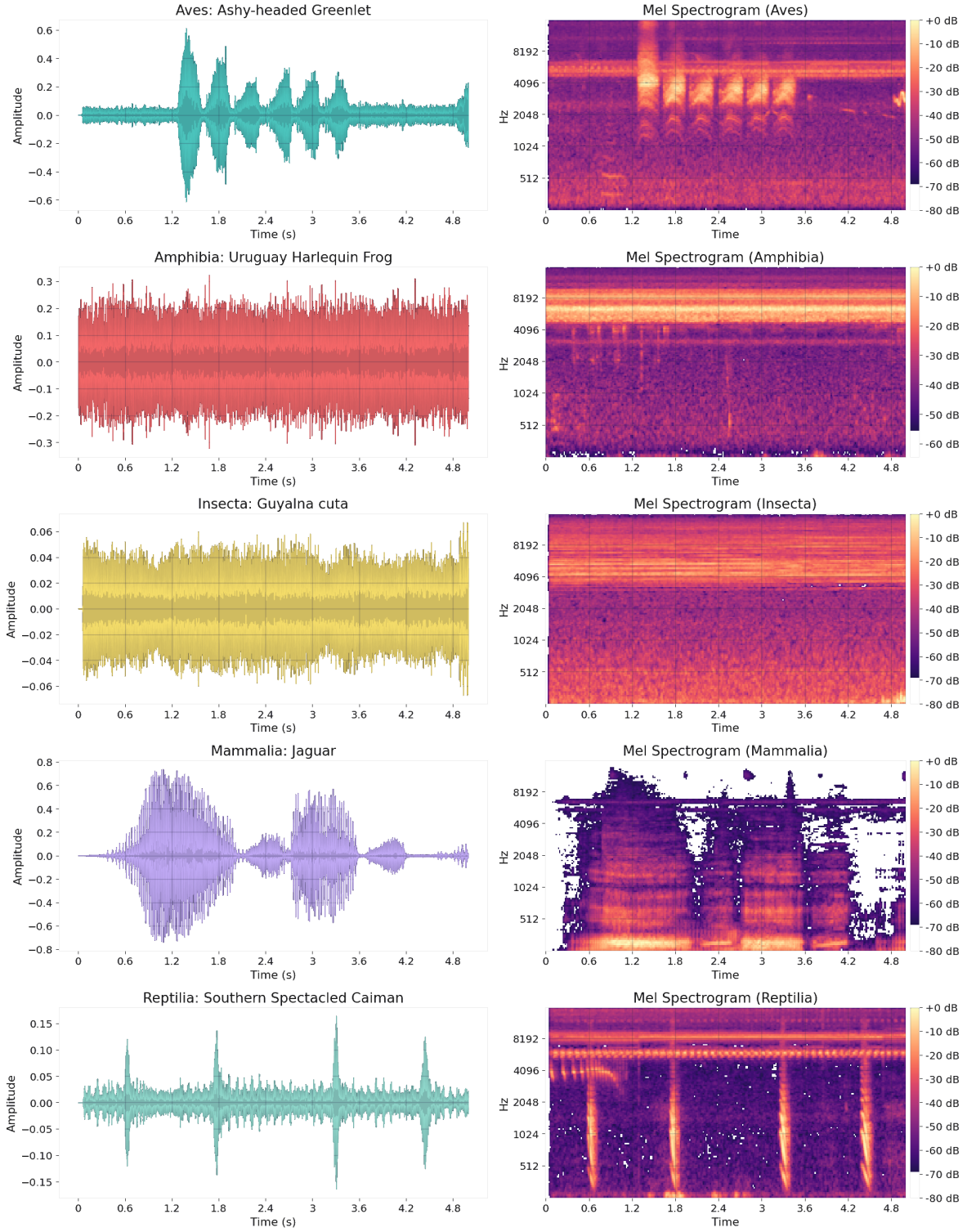

Figure 8: Example waveforms and spectrograms across animal classes. Distinct acoustic signatures (bird events, amphibian rhythms, insect textures, mammal calls) motivated class-type-specific post-processing.

6. Validation Strategy

We used the Kaggle public leaderboard (public LB) as our primary evaluation signal during development. The public LB evaluated each submission on a held-out subset of test soundscapes using ROC-AUC averaged over the 234 target classes. Unless stated otherwise, all stage-level scores reported in the following sections refer to this public leaderboard metric.

For pseudo-label generation, we adopted an out-of-fold (OOF) procedure to reduce in-sample leakage. In each fold, we trained the supervised models on four-fifths of the labeled data and generated predictions for the held-out fifth. We then aggregated the held-out predictions across folds to obtain full coverage of the training set. The resulting OOF pseudo-label table was used as the pseudo-label source for Stage 2. This design reduced the risk that a teacher model would simply reproduce labels for examples it had already seen during supervised training.

Across the checkpoints reported in this paper, our public and private leaderboard scores were typically within $\pm 0.3\%$ of each other. We therefore treated the public LB as a useful, although imperfect, proxy for final performance. Local OOF scores were less reliable because the labeled focal recordings in train_audio were collected under diverse conditions, whereas the test soundscapes came from a fixed set of Pantanal monitoring sites. Consequently, we interpreted small public-LB gains cautiously and promoted architectural or training changes only when they were supported by repeated runs or by improvements in complementary model families.

7. Training of Supervised Seed Models

In the first stage, we trained three SED models using only the labeled competition data. We used these models both as supervised baselines and as architecturally diverse teachers for the later pseudo-labeling stages. For clarity throughout the paper, we referred to the 2024 pre-trained tf_efficientnetv2_s_in21k variant as EfficientNetV2-S (In21k-2024), and to the complementary BigXCV2Ext checkpoint lineage as EfficientNetV2-S (BigXCV2Ext).

We initialized all three backbones from publicly released BirdCLEF-pretrained checkpoints. For EfficientNetV2-S (In21k-2024), we used an EfficientNetV2-S checkpoint pretrained on large-scale birdcall data from previous BirdCLEF editions and initialized from ImageNet-21k weights. For EfficientNetV2-S (BigXCV2Ext), we used a complementary checkpoint lineage trained on an extended Xeno-Canto-based pretraining corpus. For ECA-NFNet-L0, we used a matching BirdCLEF-pretrained checkpoint initialized from a similar large-scale birdcall pretraining setup. In each case, we transferred only the backbone weights and re-initialized the attention classification head for the 234-class BirdCLEF+ 2026 taxonomy.

Unless explicitly stated otherwise, all stage scores in the following training tables reported public and private leaderboard values.

Table 1
Initial labeled-only supervised models. Leaderboard scores are approximate single-model scores.

ID	Backbone	Public LB	Private LB
S1-A	ECA-NFNet-L0	90.6%	92.0%
S1-B	EfficientNetV2-S (In21k-2024)	90.5%	91.7%
S1-C	EfficientNetV2-S (BigXCV2Ext)	90.1%	90.6%

7.1. SED Input Representation

We fed each SED model a waveform crop of duration T seconds, choosing $T \in \{5, 10, 20\}$ according to the model family and training stage. We used 5-second crops to match the official evaluation window exactly, and we used longer 10- and 20-second crops when additional acoustic context was beneficial. These longer crops helped us capture calls that began near segment boundaries, repeated phrases that spanned adjacent windows, and texture-like insect or amphibian classes whose identity was more reliably inferred from persistence over time than from a single short impulse.

For each crop, we either read the raw waveform at 32 kHz or resampled it to 32 kHz, then converted it into a power spectrogram using the short-time Fourier transform. We then projected the power spectrum onto a 128-bin mel filterbank and applied logarithmic compression:

$$M = \log\left(\epsilon + \mathrm{Mel}\left(|\,\mathrm{STFT}(x)|^2\right)\right),$$

where ϵ prevented numerical instability in silent or near-silent time–frequency cells. We used the mel representation to reduce the frequency dimension while preserving perceptually and biologically meaningful frequency structure, and we used the logarithm to compress the large dynamic range between nearby loud vocalizations and distant background calls. This representation let us reuse image-style CNN backbones while retaining the time–frequency patterns needed for sound event detection.

We normalized the resulting spectrogram independently for each crop:

$$\hat{M} = \frac{M - \mu(M)}{\sigma(M) + \epsilon},$$

and then scaled it to the range $[0, 1]$. This per-crop normalization reduced our sensitivity to recording gain, microphone differences, and background loudness, all of which varied substantially across focal recordings and passive-monitoring soundscapes. The final min–max scaling gave our SED backbones a consistent input range and prevented unusually loud recordings from dominating early convolutional activations. During inference on a full 60-second soundscape, we applied the same preprocessing window by window so that each model produced predictions aligned with the twelve required 5-second segments.

7.2. Attention SED Head

For a backbone feature sequence $h_\tau \in \mathbb{R}^D$, we used the SED attention head to compute frame logits $z_{\tau,c}$ and class-specific attention scores $u_{\tau,c}$:

$$z_{\tau,c} = w_c^\top h_\tau + b_c,$$

$$a_{\tau,c} = \frac{\exp(\tanh(u_{\tau,c}))}{\sum_{\tau'} \exp(\tanh(u_{\tau',c}))}.$$

We then obtained the clip-level logit and probability through attention pooling:

$$\ell_c = \sum_\tau a_{\tau,c} z_{\tau,c}, \qquad p_c = \sigma(\ell_c).$$

For models that exposed frame logits, we optionally used frame-max blending during inference:

$$p_c^{\text{blend}} = (1 - \beta) p_c^{\text{clip}} + \beta \max_\tau \sigma(z_{\tau,c}).$$

In our final inference stack, we set $\beta = 0.40$ for ECA-NFNet and $\beta = 0.25$ for the EfficientNetV2-S family. The higher value for ECA-NFNet reflects its more reliably localized frame predictions on 5-second windows; the lower value for EfficientNetV2-S reduces noise from frame logits in the longer-context (10s/20s) models where adjacent non-target frames dominate.

7.3. Training Objective and Augmentation

We trained all models with a combined Binary Cross-Entropy and Focal loss:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}}(\ell_c, \tilde{y}_c) + \mathcal{L}_{\text{Focal}}(\ell_c, \tilde{y}_c;\ \alpha{=}0.25,\ \gamma{=}2),$$

where $\tilde{y}_c = y_c(1 - \epsilon) + \epsilon/C$ applied label smoothing with $\epsilon = 0.05$. During training, we used three waveform and spectrogram augmentations. First, we applied MixUp with probability 0.5, averaged two raw waveforms at $\lambda = 0.5$, and combined their label vectors with an element-wise clamped sum. Second, we used RandomFiltering to perturb the frequency response by interpolating random gain values in the range $[-20, 0]$ dB across four spectral bands. Third, we applied SpecAugment with probability 0.3, using up to three time masks of width ≤ 20 frames and up to three frequency masks of width ≤ 10 mel bins. We also repeated soundscape windows eight times per epoch within the training sampler to compensate for their small count relative to isolated recordings.

8. Pseudo-Concatenation

In the second stage, we generated OOF pseudo-labels from an ensemble of the supervised EfficientNetV2-S (In21k-2024/BigXCV2Ext) models. We then concatenated these labels with the labeled competition data. Unlike the later noisy-student setting, we used a direct pseudo-concat approach in this stage: we added both labeled and pseudo-labeled examples to the training table and optimized them with binary cross-entropy.

For labeled examples, we used

$$\mathcal{L}_{\text{sup}} = -\frac{1}{C}\sum_{c=1}^{C}\left[y_c \log p_c + (1-y_c)\log(1-p_c)\right].$$

For pseudo-labeled examples, we used

$$\mathcal{L}_{\text{pseudo}} = -\frac{1}{C}\sum_{c=1}^{C}\left[q_c \log p_c + (1-q_c)\log(1-p_c)\right].$$

We defined the total training objective as

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{pseudo}}\mathcal{L}_{\text{pseudo}},$$

with λ_{pseudo} absorbed into the sampling frequency and batch construction.

Table 2
Pseudo-concat retraining stage using the shared EfficientNetV2-S OOF pseudo-label file. RegNetY-016 and HGNetV2-B5 were introduced as exploratory backbones at this stage with no Stage-1 baseline; both were excluded from the final inference ensemble after failing to improve the combined ECA-NFNet and EfficientNetV2-S family ensemble.

Model	Public LB	Private LB
ECA-NFNet-L0	93.6%	93.9%
EfficientNetV2-S (In21k-2024)	93.4%	93.3%
RegNetY-016	93.0%	92.6%
HGNetV2-B5	92.8%	93.0%

9. Noisy Student With Power-Transformed Pseudo Labels

In the third stage, we changed how we consumed pseudo-labels. Instead of directly concatenating pseudo-labeled examples, we formed a teacher ensemble from the second-stage ECA-NFNet and EfficientNetV2-S (In21k-2024) models. This teacher ensemble reached approximately 94% public LB, and we used it to train a new generation of student models.

9.1. Power Transformation

Our main pseudo-label preprocessing step was a probability power transformation. Given a teacher probability $q_{i,c}$, we defined the sharpened pseudo-label as

$$\tilde{q}_{i,c} = q_{i,c}^{\gamma},$$

where $\gamma > 1$. We set the exponent separately for each backbone family, depending on how aggressively we wanted to sharpen that teacher's soft labels:

$$\gamma = \begin{cases} 1/0.65 \approx 1.54, & \text{EfficientNetV2-S family}, \\ 1/0.60 \approx 1.67, & \text{ECA-NFNet-L0}. \end{cases}$$

This suppresses weak low-confidence noise while retaining high-confidence signals, since $q < 1 \Rightarrow q^\gamma < q$ with stronger suppression at larger γ.

Before power transformation, soundscape windows with maximum class probability below $\tau_{\min} = 0.03$ were discarded entirely. After transformation, very small residual values were zeroed:

$$\tilde{q}_{i,c} = \begin{cases} q_{i,c}^\gamma, & q_{i,c}^\gamma \geq \tau_{\text{trim}}, \\ 0, & \text{otherwise}, \end{cases}$$

with $\tau_{\text{trim}} = 0.005$.

9.2. Weighted Pseudo Sampling

After applying the per-window filter ($\tau_{\min} = 0.03$) and the post-power trim ($\tau_{\text{trim}} = 0.005$), we sampled the retained pseudo-labeled soundscapes according to their aggregate confidence. For a file f with windows $t = 1, \ldots, K$, we defined the file confidence score as

$$s_f = \sum_{c=1}^{C} \max_t \tilde{q}_{f,t,c}.$$

We sampled files with larger s_f more frequently through a weighted random sampler. This reduced how often we trained on almost-unlabeled or highly uncertain soundscapes, while still keeping confident pseudo-labeled examples in the training stream.

9.3. Raw-Waveform Mixup

We found raw-waveform mixup to be the most effective way to use pseudo labels. We mixed labeled competition audio with pseudo-labeled soundscape crops so that each student saw both reliable foreground labels and target-domain acoustic backgrounds. Given a labeled sample (x, y) and a pseudo-labeled sample $(u, \tilde{q})$, we constructed the mixed input and target as

$$\bar{x} = \lambda x + (1 - \lambda)u,$$

$$\bar{y} = \lambda y + (1 - \lambda)\tilde{q}.$$

Our best setting used a constant

$$\lambda = 0.5.$$

This choice balanced clear labeled foregrounds with target-domain soundscape backgrounds. We avoided very small beta-distribution mixup coefficients because they often suppressed the meaningful source signal.

Table 3
Noisy-student models we trained with the stage-2 ECA-NFNet and EfficientNetV2-S (In21k-2024) teacher ensemble.

Student	Init.	Public LB	Private LB
EfficientNetV2-S (BigXCV2Ext)	Stage-1 seed	93.7%	94.1%
EfficientNetV2-S (In21k-2024)	Stage-2 model	93.8%	93.9%
RegNetY-016	Stage-2 model	93.6%	93.8%
ECA-NFNet-L0	Stage-2 model	94.2%	94.3%

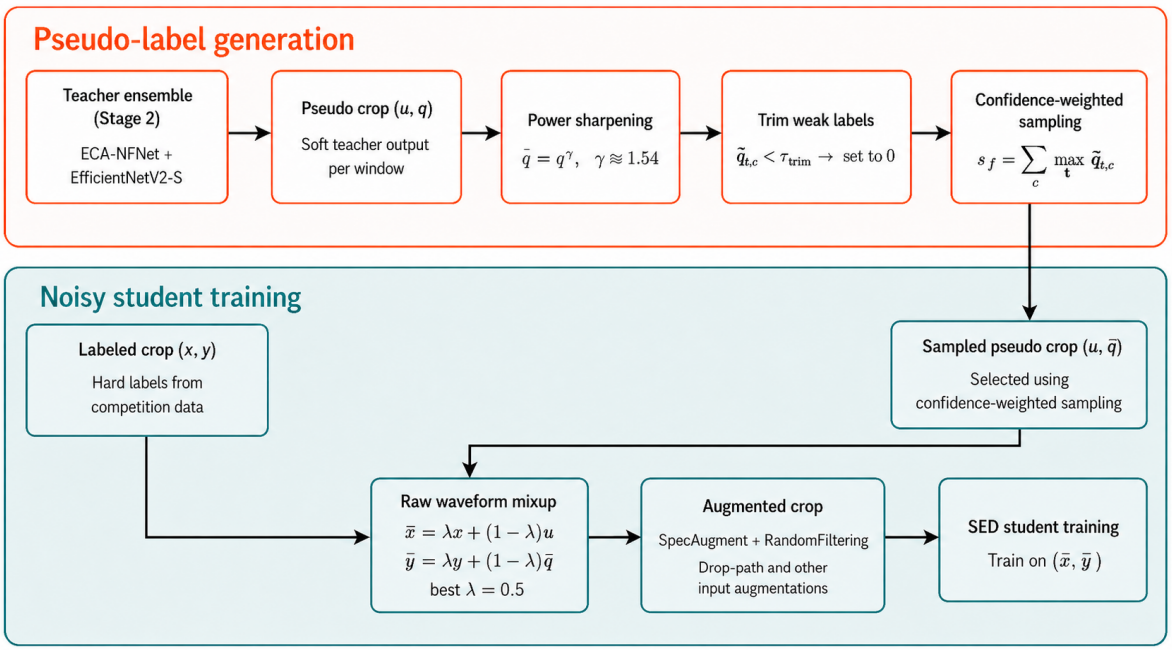

Figure 9: Our noisy-student training loop. We power-sharpened and trimmed teacher soft labels, confidence-sampled pseudo soundscapes, mixed them with labeled crops at λ=0.5, and trained students with SpecAugment and RandomFiltering.

Table 4
Direct self-training results from stage 3 to stage 4.

Model	Previous stage public LB	Previous stage private LB	Self-trained public LB	Self-trained private LB
EfficientNetV2-S (BigXCV2Ext)	93.7%	94.1%	93.5%	93.4%
EfficientNetV2-S (In21k-2024)	93.8%	93.9%	93.5%	93.5%
RegNetY-016	93.6%	93.8%	93.0%	92.9%
ECA-NFNet-L0	94.2%	94.3%	94.1%	94.0%

10. Self-Training Saturation and Cross-Mutation

We then attempted one more self-training iteration by using each third-stage model as its own teacher. As shown in Table 4, this same-family update degraded most models.

These results suggested that same-family self-training was beginning to reinforce each teacher's mistakes rather than add new information. We therefore switched to cross-mutation distillation, where each student was supervised by teacher predictions from different backbone lineages rather than by its own family. Because the ECA-NFNet and EfficientNetV2-S families differed in architecture, pretraining history, and effective temporal context, this policy exposed each student to different errors and class affinities. We used cross-mutation mainly to preserve model diversity: even when standalone gains were modest, the resulting students supplied less-correlated decision boundaries for the final SED ensemble.

A notable result from cross-mutation was that V2S-BigXCV2Ext gained 1.1 pp on the public leaderboard (93.7% $\rightarrow$ 94.8%) but *lost* 0.3 pp on the private set (94.1% $\rightarrow$ 93.8%). We attribute this to the ECA-NFNet teacher providing strong signal on the specific public subset, which the student absorbed without equally improving private-set robustness. This outcome illustrates the risk of optimizing against a single public-leaderboard signal and motivated our later use of ensemble diversity rather than further self-training iterations. RegNetY-016 and HGNetV2-B5, although trained through Stages 2–4, were excluded from the final inference ensemble because they added no measurable improvement to the

ECA-NFNet and EfficientNetV2-S family combination.

Table 5
Our cross-mutation teacher policy. V2S denotes EfficientNetV2-S.

Student	Cross-family teacher(s)	Public LB	Private LB
ECA-NFNet-L0	Stage-3 V2S-BigXCV2Ext + V2S-In21k	94.3%	94.5%
V2S-BigXCV2Ext	Stage-3/4 ECA-NFNet	94.8%	93.8%
RegNetY-016	Stage-3/4 V2S-BigXCV2Ext + ECA-NFNet	93.9%	93.9%

11. Perch and ProtoSSM Stage

Before introducing the Perch branch, our strongest final-inference system was an SED-only ensemble that combined the ECA-NFNet family with the EfficientNetV2-S (In21k-2024/BigXCV2Ext) models. This ensemble already provided a competitive task-specific baseline, achieving approximately 95.5% public LB and 95.3% private LB. However, this branch was trained entirely within the BirdCLEF+ 2026 label space and therefore remained exposed to the same weaknesses as the supervised and pseudo-labeled SED models: limited rare-class evidence, imperfect calibration on passive-monitoring soundscapes, and sensitivity to local teacher errors. To add a complementary source of information, we introduced a separate Perch-based sequence branch. We used Perch as a generic bioacoustic representation model that supplied both a high-dimensional embedding and a class-logit vector for each 5-second window. For file f and window t, we denoted these outputs as

$$e_{f,t} \in \mathbb{R}^{1536}, \qquad r_{f,t} \in \mathbb{R}^{234},$$

where $e_{f,t}$ was the Perch embedding and $r_{f,t}$ was the raw Perch logit vector. We then used these features as the input to a lightweight temporal stacker designed specifically for the twelve-window structure of the evaluation soundscapes.

11.1. Prior-Augmented Base Scores

We first augmented the raw Perch logits with coarse ecological context. Let $b_{f,t,c}$ denote the prior-fused Perch base logit for class c. From the labeled soundscapes, we estimated empirical site-level and hour-level class frequencies, which we combined into a site–hour prior

$$\pi_{s,h,c} = \Pr(y_c = 1 \mid \text{site} = s, \text{hour} = h),$$

estimated as the empirical mean over labeled soundscape windows from that site and hour. We then formed the base score as

$$b_{f,t,c} = r_{f,t,c} + \lambda_{\tau(c)} \cdot \text{logit}(\pi_{s_f,h_f,c}),$$

where $\tau(c) \in \{\text{event}, \text{texture}\}$ indicated the acoustic type of class c. We set $\lambda_{\text{event}} = 0.45$ for event-like bird classes and $\lambda_{\text{texture}} = 1.10$ for texture-like insect and amphibian classes. These values were selected by grid search over the candidate set $\{0.25, 0.30, \ldots, 0.70\}$ using the public leaderboard as the objective; the higher texture weight reflected the observation that site and time-of-day context provided stronger evidence for persistent acoustic classes than for short isolated events.

11.2. Prototype-Augmented Selective State-Space Model Architecture

We designed ProtoSSM as a prototype-augmented selective state-space model that represented each soundscape as a short temporal sequence of 12 evaluation windows. At each time step, we concatenated the Perch embedding, Perch logits, site embedding, hour embedding, and positional information:

$$v_t = [e_t, r_t, m_{\text{site}}, m_{\text{hour}}, p_t].$$

We projected this vector into the model dimension and normalized it before temporal processing:

$$h_t^{(0)} = \mathrm{LN}(W_v v_t).$$

We then passed the sequence through three selective state-space model (SSM) layers, which allowed the model to propagate evidence across adjacent and non-adjacent windows with linear-time sequence processing:

$$h_{1:K}^{(\ell+1)} = \mathrm{SSM}_\ell(h_{1:K}^{(\ell)}), \qquad \ell = 0, 1, 2.$$

Because some species evidence appeared as repeated calls or persistent acoustic texture, we also added a cross-attention block so that windows could exchange global context:

$$\tilde{H} = \mathrm{MHA}(H, H, H),$$

where our final implementation used 4 cross-attention heads.

To improve rare-class behavior, we added a prototype readout in addition to the standard classifier head. For each class c, we learned a prototype vector μ_c and computed a cosine-similarity logit

$$s_{t,c} = \frac{\langle W_p h_t, \mu_c \rangle}{\|W_p h_t\|_2 \|\mu_c\|_2}.$$

We combined the standard classifier score, the prototype similarity, and a class bias to obtain the ProtoSSM logit:

$$z_{t,c}^P = w_c^\top h_t + \eta s_{t,c} + b_c.$$

Our final artifact-only ProtoSSM configuration used

$$d_{\text{model}} = 192, \quad d_{\text{state}} = 12, \quad L = 3, \quad \text{dropout} = 0.22, \quad n_{\text{sites}} \leq 20.$$

This configuration produced a model with 1,805,402 parameters. We trained ProtoSSM on the full labeled soundscape set (59 files, 708 evaluation windows) using focal BCE loss ($\gamma = 2.0$), Adam optimizer ($\text{lr} = 5 \times 10^{-4}$, no weight decay), and cosine annealing with warm restarts over 30 epochs. Both the ProtoSSM weights and the per-class MLP probe artifacts were pre-computed offline and loaded by the inference pipeline at submission time. Full training parameters are listed in Table 8 (Appendix A). The submitted Kaggle inference notebook, trained SED checkpoints, ProtoSSM weights, and MLP probe artifacts are available upon request from the authors.

11.3. MLP Probe Branch

In parallel with ProtoSSM, we trained a second Perch branch based on PCA-compressed embeddings and handcrafted temporal features. We standardized each Perch embedding and projected it to a 128-dimensional PCA space:

$$z_t = \mathrm{PCA}_{128}(\mathrm{standardize}(e_t)).$$

For each class, we trained an MLP probe only when that class had at least 5 positive windows in the full-file labeled soundscape annotations (59 files $\times$ 12 windows). Classes below this threshold were excluded from probe training; at inference time their scores relied solely on the prior-fused base logit $b_{t,c}$. This constraint avoided fitting unstable per-class probes for extremely sparse targets. The probe feature vector combined the compressed embedding, raw Perch evidence, prior-fused logits, class priors, and local temporal summaries:

$$\phi_{t,c} = [z_t,\ r_{t,c},\ b_{t,c},\ \pi_{t,c},\ \mathrm{prev}(b),\ \mathrm{next}(b),\ \mathrm{mean}(b),\ \max(b),\ \mathrm{std}(b)].$$

We mixed the probe output with the prior-fused base logit rather than replacing the base score entirely:

$$z_{t,c}^M = (1 - \alpha) b_{t,c} + \alpha f_c(\phi_{t,c}),$$

with

$$\alpha = 0.45.$$

This interpolation preserved the robustness of the Perch prior score while allowing the MLP probe to correct class-specific temporal and contextual patterns.

11.4. Perch Branch Fusion

Finally, we fused the ProtoSSM and MLP-probe branches at the logit level. This allowed the sequence model and the simpler per-class probes to contribute complementary evidence before probability calibration and downstream rank blending:

$$z_{t,c}^{\text{Perch}} = w_P z_{t,c}^{P} + (1 - w_P) z_{t,c}^{M},$$

where our final artifact-only inference setup used

$$w_P = 0.7.$$

We gave ProtoSSM the larger weight because it modeled full-recording temporal structure, while the MLP branch served as a lower-capacity correction based on compressed embeddings and handcrafted temporal summaries; w_P was chosen by a coarse public-leaderboard-guided sweep over fixed Perch artifacts.

12. Final Inference Ensemble

At submission time, we used the system strictly in inference mode. All SED checkpoints, ProtoSSM weights, and MLP-probe artifacts were pre-trained before submission, and we did not perform pseudo-label generation, ProtoSSM fitting, or MLP-probe training during the final run. This separation was important for reproducibility: the submitted pipeline only loaded fixed artifacts, generated branch-level predictions, applied deterministic post-processing, and wrote the final ranked submission file.

Figure 10 summarized the inference-time data flow. We kept the SED and Perch-based branches independent until the final fusion step so that each branch could preserve its own ranking behavior before calibration-sensitive operations were applied.

12.1. SED Family Ensemble

Our final SED branch combined model families that differed in backbone, pretraining source, and temporal context length. This branch supplied the task-specific acoustic evidence that was later combined with the Perch-based temporal branch. The branch consisted of:

- an ECA-NFNet-L0 group;
- EfficientNetV2-S (In21k-2024), context 5s;
- EfficientNetV2-S (In21k-2024), context 20s;
- EfficientNetV2-S (BigXCV2Ext), context 20s;
- EfficientNetV2-S (BigXCV2Ext), second iteration, context 10s.

We used the 5-second models to preserve exact alignment with the evaluation windows and the longer-context models to capture calls crossing window boundaries, repeated phrases, and persistent acoustic textures. Within the EfficientNetV2-S family, we normalized the raw tuning constants

$$(0.30,\ 0.20,\ 1.00,\ 0.65).$$

Thus, the EfficientNetV2-S ensemble prediction was

$$p^{E2S} = \frac{0.30 p^{\text{In21k24},5s} + 0.20 p^{\text{In21k24},20s} + 1.00 p^{\text{BigXCV2Ext},20s} + 0.65 p^{\text{BigXCV2Ext},10s}}{0.30 + 0.20 + 1.00 + 0.65}.$$

We then combined the ECA-NFNet and EfficientNetV2-S families with equal weight:

$$p^{SED} = 0.5 p^{ECA} + 0.5 p^{E2S}.$$

This equal-weight family fusion gave the final SED branch a balance between the high single-model strength of the ECA-NFNet lineage and the context diversity of the EfficientNetV2-S models. The EfficientNetV2-S internal weights and equal family blend were chosen by simple public-leaderboard-guided trial blends over fixed checkpoints.

12.2. Post-Processing of Perch Branch

We post-processed the Perch branch before combining it with the SED ensemble because the Perch-derived logits and SED probabilities were not calibrated on the same scale. First, we converted the Perch branch logits to probabilities with class-type temperature scaling:

$$p_{t,c} = \sigma\left(\frac{z_{t,c}^{\text{Perch}}}{T_c}\right),$$

where

$$T_c = \begin{cases} 0.95, & c \in \{\text{Insecta}, \text{Amphibia}\}, \\ 1.10, & \text{otherwise.} \end{cases}$$

We used a lower temperature for insect and amphibian classes to preserve sharper responses for texture-like signals, and a slightly higher temperature for the remaining classes to reduce over-confident isolated-window predictions.

We then applied a sequence of lightweight temporal and confidence-based transformations. File-level confidence scaling used the top two windows as an estimate of whether class c had reliable support anywhere in file f. We defined

$$\text{Top2}_c(f) = \text{indices of the two largest values in } \{p_{f,j,c}\}_{j=1}^{K}.$$

The scaled probability was then

$$p'_{f,t,c} = p_{f,t,c} \cdot \frac{1}{2} \sum_{j \in \text{Top2}_c(f)} p_{f,j,c}.$$

Rank-aware scaling further emphasized files with at least one strong class-specific response:

$$p''_{f,t,c} = p'_{f,t,c} \cdot \left(\max_j p'_{f,j,c}\right)^{0.4}.$$

For temporal smoothing, we mixed each middle window with its immediate neighbors using an adaptive coefficient that decreased when the current window already had high confidence:

$$\hat{p}_{f,t,c} = (1 - a_{f,t}) p''_{f,t,c} + a_{f,t} \frac{p''_{f,t-1,c} + p''_{f,t+1,c}}{2},$$

where

$$a_{f,t} = 0.20 \left(1 - \max_c p''_{f,t,c}\right).$$

Finally, we applied per-class threshold sharpening to remap probabilities through class-specific thresholds τ_c:

$$g(p; \tau_c) = \begin{cases} \frac{1}{2} \frac{p}{\tau_c + \epsilon}, & p \le \tau_c, \\ \frac{1}{2} + \frac{1}{2} \frac{p - \tau_c}{1 - \tau_c + \epsilon}, & p > \tau_c. \end{cases}$$

Together, these operations made the Perch branch more consistent with the temporal structure of one-minute soundscapes before fusion with the SED ensemble.

12.3. Rank Blending With SED Ensemble

We fused the final SED branch with the post-processed Perch branch by per-class rank blending rather than by direct probability averaging. This choice avoided assuming that the SED and Perch branches had comparable probability calibration, while still preserving their relative ordering of likely positives. For each class c and all submission rows n, we computed

$$R^P_{n,c} = \text{rank}(p^{Perch}_{n,c}), \qquad R^S_{n,c} = \text{rank}(p^{SED}_{n,c}).$$

We then formed the final rank score as

$$p_{n,c}^{final} = \frac{(1 - w_S)R_{n,c}^{P} + w_S R_{n,c}^{S}}{N},$$

where N was the total number of submission rows and

$$w_S = 0.60.$$

We assigned the SED branch the larger weight because it was directly optimized for the BirdCLEF+ 2026 label space, while the Perch branch contributed complementary representation and temporal evidence. We chose $w_S = 0.60$ from the same coarse public-leaderboard-guided sweep over rank-blend ratios.

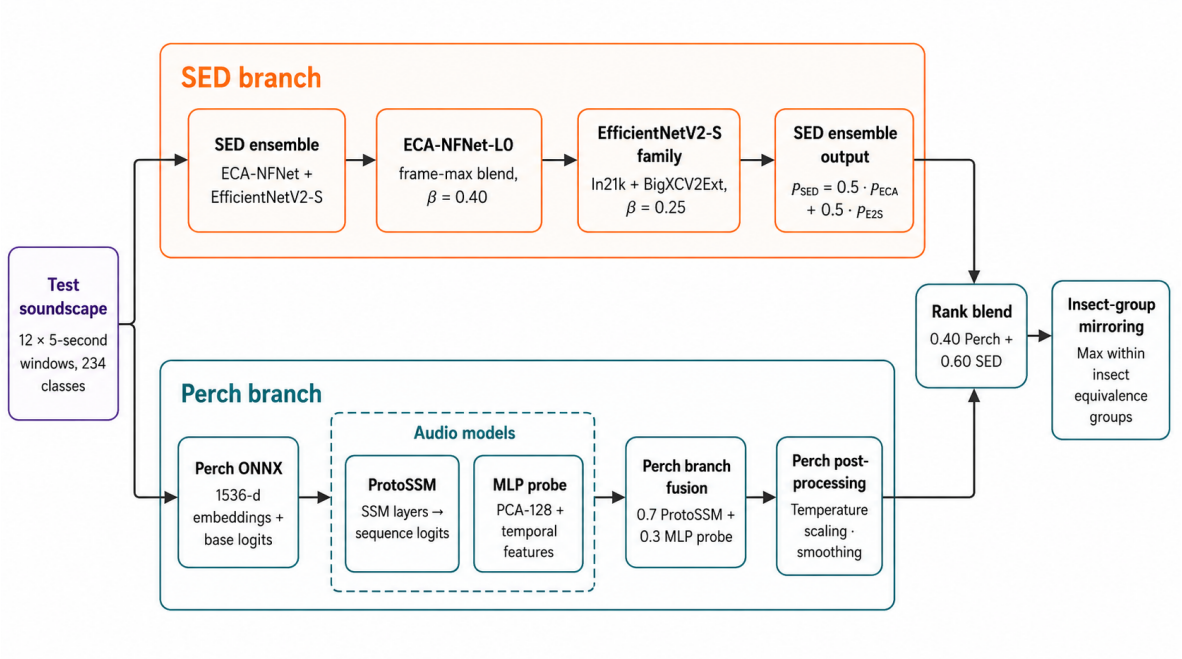

Figure 10: Final inference stack. We ran the SED ensemble and the Perch–ProtoSSM artifact branch independently, post-processed the Perch probabilities, combined both branches by per-class rank blending, mirrored selected insect equivalence groups, and submitted the resulting scores.

After rank blending, we applied a final taxonomy-aware correction for selected insect-like classes. We defined four hand-curated equivalence groups by cross-referencing BirdCLEF+ 2026 taxonomy metadata with spectral inspection of labeled soundscape segments whose sonotypes were near-identical. The resulting groups, identified by Xeno-canto sonotype IDs, were: {47158son15,47158son16}; {47158son09,47158son12}; {47158son02,47158son14}; and {47158son13,47158son21,47158son22,47158son23}. For each group G, we mirrored scores by taking the maximum within the group:

$$p_{n,c} \leftarrow \max_{j \in G} p_{n,j}, \qquad c \in G.$$

This rule reflected the observation that these insect species shared nearly identical persistent texture patterns in the Pantanal soundscapes and were acoustically indistinguishable at the window level.

13. Results

13.1. Stepwise Ablation

We report the main development trajectory in Table 6. Following the convention used in CLEF working notes, we present both public and private leaderboard scores and interpret them primarily as comparative

evidence between successive system variants rather than as exact estimates of generalization. Each row corresponds to the best representative model or ensemble available after the corresponding stage and before it was used as a teacher or component in the next stage.

Table 6
Stepwise ablation of our BirdCLEF+ 2026 system. Scores are approximate leaderboard values for the best representative model or ensemble at each stage.

Stage	Best public LB	Best private LB
Supervised (labeled data only)	90.6%	92.0%
Pseudo-concatenation	93.6%	93.9%
Noisy student (power sharpening + waveform mixup)	94.2%	94.3%
Cross-mutation distillation	94.3%	94.5%
SED-only ensemble (Stages 1–4 combined)	95.5%	95.3%
Our Full Pipeline	**96.0%**	**95.4%**

The ablation showed that the largest single improvement came from moving beyond labeled-only training. Our supervised seed models reached approximately 90.6% public LB, while Stage 2 pseudo-concatenation improved the best public score to approximately 93.6%. This +3.0% absolute gain indicated that the additional soundscape-style supervision was more important than small backbone-level changes at the beginning of the pipeline. In particular, pseudo-labeling helped bridge the mismatch between isolated focal recordings and passive-monitoring evaluation soundscapes.

The noisy-student stage gave a smaller but consistent improvement, increasing the best public score from approximately 93.6% to 94.2%. We attribute this gain to three changes that made pseudo-label consumption more robust: power sharpening suppressed low-confidence teacher noise, confidence-weighted sampling reduced the influence of nearly empty soundscapes, and raw-waveform mixup exposed the student to target-domain backgrounds while preserving reliable labeled foregrounds. Cross-mutation distillation added only a marginal single-model gain (94.2% $\rightarrow$ 94.3% public LB) but increased model diversity. We caution that individual cross-mutation gains on the public set did not always transfer to the private set: V2S-BigXCV2Ext improved 1.1 pp publicly but lost 0.3 pp privately, a divergence we discuss further in Section 14. This diversity was nevertheless important when we assembled the SED-only ensemble, which reached approximately 95.5% public LB and 95.3% private LB.

Finally, adding the Perch branch through rank blending improved the public leaderboard score from approximately 95.5% to 96.0%, while the private score remained close to the SED-only result. We interpreted this as evidence that the Perch branch supplied complementary information, but that most of the robust private-set performance already came from the strong SED ensemble. The small public–private gap in the final two rows also suggested that the rank-blended system did not substantially overfit the public subset.

13.2. Self-Training Saturation

Our Stage 4 experiments showed that additional self-training was not automatically beneficial. As reported in Table 4, direct same-family self-training was neutral or slightly harmful for three out of four model families, with degradations between 0.1% and 0.6%. This pattern was important because it indicated that simply repeating the pseudo-labeling cycle within the same backbone lineage tended to reinforce teacher-specific errors rather than introduce new information.

We therefore treated same-family self-training saturation as a signal to prioritize teacher diversity. Cross-mutation distillation used predictions from different backbone families to supervise each student, which helped preserve complementary decision boundaries for the later ensemble. Although the per-model improvement from cross-mutation was modest, its value became visible at the ensemble level, where diverse SED families produced a larger gain than any individual Stage 4 model.

13.3. Perch Branch Contribution

Before introducing Perch, our SED-only ensemble reached approximately 95.5% public LB and 95.3% private LB. Rank-blending with the Perch–ProtoSSM branch raised the public score to approximately 96.0% (+0.5 pp) and the private score to approximately 95.4% (+0.1 pp). We acknowledge that a +0.5%/+0.1% improvement is a weak aggregate signal — the gain appears modest in overall ROC-AUC terms but we believe it is concentrated in three settings that the aggregate number does not highlight.

First, ProtoSSM had access to the full 60-second temporal context while each SED window operated on an isolated 5–20-second clip; for species with repeated calls or for persistent acoustic textures (insect and amphibian classes), cross-window evidence propagation through the SSM layers should provide information not available to any single SED window. Second, Perch embeddings were pre-trained on a much broader taxonomic base than our BirdCLEF-specific SED models; for rare classes with very few positive training windows, embedding-space prototype similarity can substitute for insufficient direct supervision. Third, the rank-blending strategy preserved heterogeneous ranking information without requiring both branches to share the same calibration, which reduced the risk that Perch miscalibration would harm the already-strong SED rankings.

We did not collect per-taxonomy AUC breakdowns (Aves vs Insecta vs Amphibia) during development. A class-level analysis would be needed to confirm or challenge the per-group hypothesis above and is a clear direction for future work.

14. Discussion

Our experiments showed that the effectiveness of pseudo-labeling depended not only on the strength of the teacher model, but also on the way in which the pseudo labels were filtered, sharpened, sampled, and mixed with labeled data. The largest improvement came from introducing soundscape-domain pseudo supervision, which reduced the mismatch between isolated focal recordings and the passive-monitoring evaluation data. However, later iterations also showed that naive self-training could saturate quickly. When we trained a student from the same model family as its teacher, the student often inherited the teacher's errors and produced little or no leaderboard improvement. This motivated our use of power-transformed labels, confidence-weighted pseudo sampling, raw-waveform mixup, stochastic regularization, and finally cross-family teacher mutation.

We also found that model diversity was more useful than repeated refinement of a single backbone lineage. Cross-mutation distillation produced only modest single-model gains, but it helped maintain complementary decision boundaries across the SED ensemble. This was consistent with the multi-label and multi-taxonomic nature of the task: short bird vocalizations, amphibian rhythms, insect textures, and other animal sounds were not equally well represented by a single model family or context length. In our final system, the ensemble benefited from combining 5-second models aligned with the evaluation windows and longer-context models that could capture boundary-crossing events and persistent acoustic patterns.

The Perch branch contributed a different form of evidence from the SED models. Whereas our SED branch was optimized directly for the BirdCLEF+ 2026 taxonomy, Perch provided a general bioacoustic embedding learned from a broader representation space. ProtoSSM then used this embedding to model the temporal arrangement of the twelve 5-second windows in a soundscape. This was useful because many target events were not independent across windows: calls could repeat, habitats and site priors could persist, and insect or amphibian classes could appear as continuous acoustic textures. Thus, the Perch–ProtoSSM branch complemented the CNN ensemble even though its standalone contribution was smaller than the contribution of pseudo-labeling.

Our final fusion strategy reflected the fact that the SED and Perch branches were not calibrated on the same probability scale. Instead of averaging probabilities directly, we used per-class rank blending:

$$p_{n,c}^{final} = \frac{0.40 \cdot \mathrm{rank}(p_{n,c}^{Perch}) + 0.60 \cdot \mathrm{rank}(p_{n,c}^{SED})}{N}.$$

This preserved the ordering information from both branches and was more stable than arithmetic averaging of heterogeneous probabilities. We assigned the larger weight to the SED branch because it was trained directly on the target label space, while the Perch branch supplied complementary embedding and temporal evidence.

14.1. Limitations

Our system had several practical limitations. First, the Kaggle public leaderboard was the primary validation signal throughout development. We count approximately 20 scalar hyperparameters selected this way: γ_{V2S}, γ_{ECA}, $\tau_{\min}$, τ_{trim}, λ_{mix}, β_{ECA}, β_{V2S}, the four EfficientNetV2-S internal blend weights, λ_{event}, $\lambda_{\mathrm{texture}}$, T_{event}, T_{texture}, the two temporal-smoothing alphas, α_{probe}, w_P, and w_S. In addition, the 234 per-class probability thresholds τ_c applied during Perch post-processing were each optimized individually against the public subset, yielding a total of over 250 free parameters selected without a held-out validation set. The V2S-BigXCV2Ext cross-mutation result (94.8% public, 93.8% private — 0.3 pp worse than its Stage-3 private score of 94.1%) is a concrete demonstration of the resulting overfit risk. Small stage-to-stage differences in our ablation table should therefore be interpreted cautiously rather than as reliable generalization estimates. Second, our cross-mutation teachers were still drawn from the same competition data and from a limited set of backbone families, so the diversity gain was bounded compared with using fully independent model lineages or external bioacoustic classifiers.

Additional limitations came from hand-tuned components. The power-sharpening exponents ($\gamma = 1/0.65$ and $\gamma = 1/0.60$) were selected from a small empirical search and were not re-optimized separately for every pseudo-labeling stage. Similarly, the insect-group mirroring rule was designed for the BirdCLEF+ 2026 taxonomy and would need to be revalidated for a different label set. Finally, our approach prioritized leaderboard performance and inference-time robustness over interpretability; future work could analyze class-wise error patterns, site-specific behavior, and the contribution of individual temporal windows in more detail.

15. Conclusion

We presented our final BirdCLEF+ 2026 soundscape recognition system, which combined multi-stage SED training with a Perch-based temporal inference branch. Starting from supervised ECA-NFNet and EfficientNetV2-S models near 90% public LB, we improved performance through pseudo-concatenation, noisy-student training with power-sharpened pseudo labels, raw-waveform mixup, and cross-mutation distillation. The strongest SED ensemble reached approximately 95.5% public LB and 95.3% private LB.

The final submission combined this SED ensemble with an artifact-only Perch–ProtoSSM branch, post-processing, taxonomy-aware insect mirroring, and per-class rank blending. This complete pipeline reached approximately 96.0% public LB and 95.4% private LB. Overall, our results indicated that the most important ingredients were soundscape-domain pseudo supervision, robust pseudo-label consumption, backbone diversity, and a complementary representation-based temporal branch.

Acknowledgments

We thank the BirdCLEF+ 2026 organizers and the Cornell Lab of Ornithology for providing the competition data, evaluation infrastructure, and Perch embedding model. We also acknowledge the Xeno-canto community for the focal bird recordings that form the backbone of the training data.

Declaration on Generative AI

Large language model tools were used to assist with drafting and editing portions of this manuscript. All technical content, experimental results, numerical values, and scientific claims were verified by the authors.

References

[1] 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. d. A. Júnior, R. C. Kridler, M. Lasseck, A. V. d. Melo, H. Müller, M. G. d. Neves, M. d. O. Reis, L. K. Sampaio Teles, K.-L. Schuchmann, L. Sugai, K. T. P. Azevedo, P. d. N. Lopes, M. I. Marques, H. Klinck, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, A. Joly, Overview of BirdCLEF+ 2026: Acoustic species identification in the Pantanal, South America, 2026. Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum.

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

[3] S. Kahl, A. Navine, T. Denton, H. Klinck, P. Hart, H. Müller, H. Goëau, A. Joly, Overview of BirdCLEF 2022: Endangered bird species recognition in soundscape recordings, 2022. CLEF 2022 Working Notes, CEUR Workshop Proceedings, volume 3180.

[4] S. Kahl, T. Denton, H. Klinck, A. Joly, H. Müller, F.-R. Stöter, T. Lorieul, Overview of BirdCLEF 2023: Automated bird species identification in eastern africa, 2023. CLEF 2023 Working Notes, CEUR Workshop Proceedings, volume 3497.

[5] S. Kahl, T. Denton, H. Klinck, A. Joly, H. Müller, F.-R. Stöter, T. Lorieul, Overview of BirdCLEF 2024: Acoustic identification of under-studied bird species in the western ghats, 2024. CLEF 2024 Working Notes, CEUR Workshop Proceedings, volume 3740.

[6] S. Kahl, T. Denton, H. Klinck, A. Joly, H. Müller, F.-R. Stöter, T. Lorieul, Overview of BirdCLEF+ 2025: Multi-taxonomic sound identification in the middle magdalena, colombia, 2025. CLEF 2025 Working Notes, CEUR Workshop Proceedings.

[7] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, PANNs: Large-scale pretrained audio neural networks for audio pattern recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020) 2880–2894.

[8] S. Kahl, C. M. Wood, M. Eibl, H. Klinck, BirdNET: A deep learning solution for avian diversity monitoring, Ecological Informatics 61 (2021) 101236.

[9] A. Brock, S. De, S. L. Smith, K. Simonyan, High-performance large-scale image recognition without normalization, in: Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, 2021, pp. 1059–1071.

[10] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-net: Efficient channel attention for deep convolutional neural networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11534–11542.

[11] M. Tan, Q. V. Le, EfficientNetV2: Smaller models and faster training, in: Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, 2021, pp. 10096–10106.

[12] D.-H. Lee, Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, 2013. ICML 2013 Workshop on Challenges in Representation Learning.

[13] Q. Xie, M.-T. Luong, E. Hovy, Q. V. Le, Self-training with noisy student improves ImageNet classification, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10687–10698.

[14] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, 2018. International Conference on Learning Representations.

[15] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, 2015. NeurIPS 2014 Deep Learning Workshop, arXiv:1503.02531.

[16] B. Ghani, T. Denton, S. Kahl, H. Klinck, Global birdsong embeddings enable superior transfer learning for bioacoustic classification, Scientific Reports 13 (2023) 22876. doi:10.1038/s41598-023-49989-z.

[17] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, 2023.

ArXiv:2312.00752.
[18] A. Gu, K. Goel, C. Ré, Efficiently modeling long sequences with structured state spaces, 2022. International Conference on Learning Representations.
[19] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, 2017. Advances in Neural Information Processing Systems, volume 30.

A. Training and Spectrogram Parameters

Table 7 lists the core spectrogram and training constants shared across the SED models. Table 8 lists the ProtoSSM and MLP probe training configuration.

Table 7
Core SED preprocessing and training constants.

Parameter	Value
Sample rate	32 kHz
Mel filterbanks	128
Frequency range	20 Hz – 16 kHz
FFT size / hop length	2048 / 512
Mel scale / clipping	Slaney / 80 dB
Per-crop normalization	Mean/std, then min–max to $[0, 1]$
Training loss	BCE + Focal ($\alpha = 0.25$, $\gamma = 2.0$)
Label smoothing ϵ	0.05
Pseudo-label sharpening	$1/0.65$ (V2S), $1/0.60$ (ECA)
Pseudo filtering	$\tau_{\min} = 0.03$, $\tau_{\text{trim}} = 0.005$
Waveform mixup	Probability 0.50, $\lambda = 0.5$
SpecAugment	Prob. 0.30; 3 time / 3 frequency masks
RandomFiltering	$[-20, 0]$ dB
Soundscape repeat per epoch	$8\times$

Table 8
ProtoSSM and MLP probe training configuration.

Parameter	Value
ProtoSSM	
Training set	Full labeled soundscape set (59 files, 708 windows)
Loss	Focal BCE ($\gamma = 2.0$)
Optimizer	Adam
Learning rate	5×10^{-4}
LR schedule	Cosine annealing with warm restarts
Epochs	30
Weight decay	0
MLP probe (per-class)	
PCA dim	128
Minimum positive windows	5
Learning rate	5×10^{-4}
Early stopping	Enabled
Probe blend weight α	0.45