Bioacoustic Species Detection at Scale: A Multi-Model Ensemble Approach with Taxonomic Consistency for BirdCLEF 2026

Renuka Shanmugam[1,*]

[1]Independent Researcher, Salem, Tamil Nadu, India

Abstract

Identifying bird species from passive acoustic monitoring recordings is a challenging problem in computational bioacoustics, primarily due to long-tail class distributions, taxonomically unmapped species, and sparse target vocalisations within continuous soundscape streams. The BirdCLEF 2026 competition addresses this problem at ecological scale, requiring automated identification of 234 bird species across 24 geographically diverse monitoring sites. In this work, we present a multi-model ensemble framework that combines pretrained bioacoustic embeddings, lightweight classification heads, and phylogenetically informed post-processing. Specifically, we train a fold-averaged MLP classifier on 1,536-dimensional Perch v2 embeddings using a two-stage protocol — first on labelled clips combined with soundscapes, then fine-tuned exclusively on soundscape data. This Perch Head achieves a mean out-of-fold soundscape AUC of 0.9218 across four cross-validation folds. At inference, predictions from our Perch Head are blended with a reference LightProtoSSM and Sound Event Detection ensemble pipeline using rank-aware weighted fusion, yielding a public leaderboard score of 0.951 and a private leaderboard score of 0.942. A taxonomic post-processing step further enforces genus-to-order phylogenetic consistency to stabilise predictions for acoustically ambiguous species. Our results demonstrate that lightweight embedding-based classifiers, when combined with biologically informed post-processing, provide consistent performance gains over strong reference baselines in real-world soundscape detection tasks, achieving a Silver Medal finish (181st place out of 4,094 teams) in the BirdCLEF 2026 Kaggle competition.

Keywords

passive acoustic monitoring, bird species detection, Perch embeddings, ensemble learning, taxonomic post-processing, BirdCLEF 2026

1. Introduction

Biodiversity monitoring at ecological scale has become increasingly critical in the context of accelerating habitat loss and climate-driven species displacement. Passive acoustic monitoring (PAM) has emerged as a non-invasive, scalable methodology for continuous wildlife surveillance, enabling the automated capture of species-rich soundscape recordings across geographically distributed environments without direct human presence. Within this domain, automated avian species recognition represents a particularly active research frontier, given the acoustic richness of bird vocalisations and their established utility as biodiversity indicators.

Despite significant advances in deep learning-based audio classification, automated bird species detection from real-world soundscapes remains a fundamentally challenging problem. Three core difficulties persist across existing systems: (i) severe long-tail class imbalance, wherein rare species contribute disproportionately few labelled training examples relative to their ecological significance; (ii) the presence of taxonomically unmapped species, for which no direct training supervision exists and predictions must be inferred through phylogenetic proximity; and (iii) the temporal sparsity of target vocalisations within continuous multi-species soundscape streams, where foreground bird calls constitute only a small fraction of the overall acoustic signal.

The BirdCLEF 2026 competition, hosted on Kaggle, operationalises these challenges at ecological

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

*Corresponding author.
renuumarani@gmail.com (R. Shanmugam)
0009-0002-3137-2141 (R. Shanmugam)

scale by tasking participants with the automated identification of 234 bird species from passive acoustic recordings collected across 24 geographically heterogeneous monitoring sites. With 4,094 participating teams, the competition represents one of the largest community benchmarks in computational bioacoustics to date [10, 11].

In this work, we present a hierarchical multi-model ensemble framework designed to address the above challenges. Our system integrates 1,536-dimensional Perch v2 bioacoustic embeddings with a two-stage training protocol, combines lightweight MLP classification heads with a reference LightProtoSSM and Sound Event Detection ensemble backbone, and applies a taxonomic post-processing module enforcing genus-to-order phylogenetic consistency. As illustrated in Fig. 1, the BirdCLEF 2026 training set exhibits a pronounced long-tail distribution, with clip counts ranging from 1 to 499 across 206 training species, underscoring the challenge of learning robust representations for rare taxa. Our approach achieves a public leaderboard score of 0.951 and a private leaderboard score of 0.942, securing a Silver Medal finish (181st place out of 4,094 teams).

The primary contributions of this work are as follows:

- A two-stage soundscape fine-tuning protocol for Perch v2 embeddings, specifically designed for the BirdCLEF evaluation domain, achieving a mean out-of-fold soundscape AUC of 0.9218.
- A rank-aware weighted fusion strategy combining embedding-based and sequence-based architectures, yielding a public leaderboard score of 0.951.
- A genus-to-order phylogenetic score propagation module for taxonomically unmapped species — to our knowledge, the first such application in the BirdCLEF competition series.

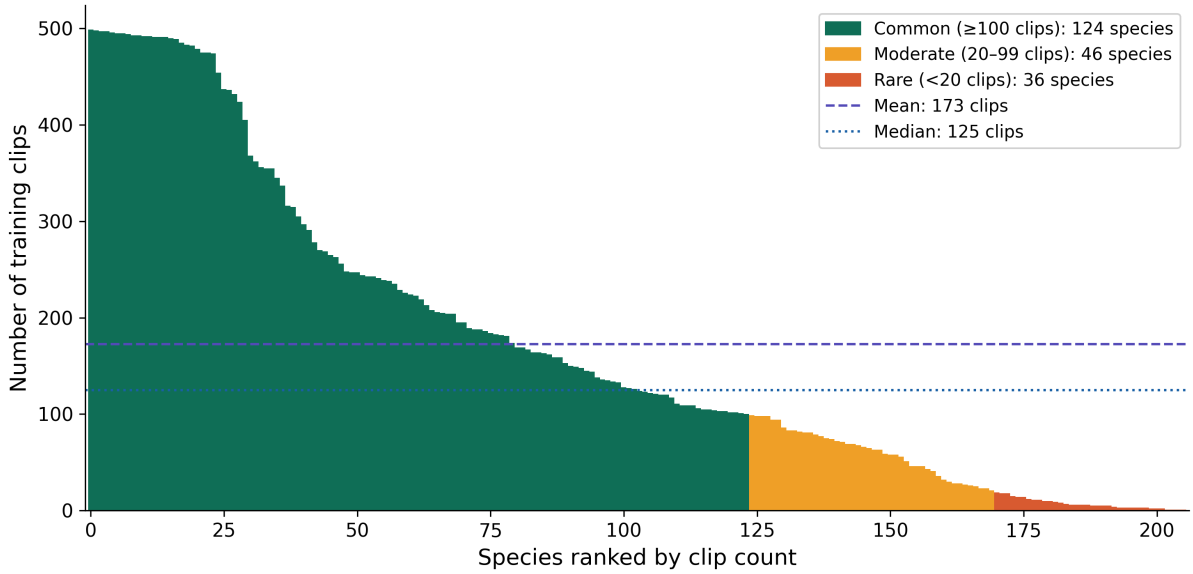

Figure 1: Long-tail distribution of training clips across 206 species in BirdCLEF 2026.

2. Related Work

Passive acoustic monitoring has emerged as a scalable and non-invasive methodology for long-term biodiversity surveillance, with prior studies demonstrating its effectiveness in tracking avian population dynamics across ecologically sensitive regions [1, 2]. Early automated bird species recognition systems relied on handcrafted acoustic features such as mel-frequency cepstral coefficients and spectral centroids, which demonstrated limited generalisation across diverse soundscape environments [3].

The advent of large-scale pretrained models has significantly advanced the state of bioacoustic classification. Perch v2, developed by Google DeepMind, provides high-quality 1,536-dimensional embeddings trained on over 1.5 million labelled recordings spanning 14,597 species across birds, mammals, amphibians, and insects, and has demonstrated strong transfer learning performance across multiple downstream bioacoustic detection tasks [4]. BirdNET, another widely adopted framework, employs a deep residual architecture trained on an extensive annotated bird sound dataset, offering competitive classification performance for a large number of species [5]. Pretrained Audio Neural Networks (PANNs) have further established the utility of large-scale audio pretraining for general-purpose sound event classification [6].

Within the BirdCLEF competition series, top-performing solutions have consistently leveraged ensemble strategies combining multiple model architectures with site-aware post-processing [7, 8]. The reference LightProtoSSM ensemble pipeline, which forms the backbone of our inference framework, introduced rank-aware score fusion and prior-based calibration as effective mechanisms for improving soundscape-level detection accuracy. Our work extends this foundation by integrating a lightweight Perch embedding-based classification head and a taxonomic post-processing module, addressing the challenge of species with no direct training supervision through phylogenetic score propagation.

Unlike prior ensemble approaches that operate exclusively on mapped species, our framework explicitly addresses the 28 unmapped test-set species through biologically informed score propagation, representing a methodological distinction from existing BirdCLEF solutions.

3. Methodology

This section describes the end-to-end system architecture developed for BirdCLEF 2026, encompassing dataset preparation, feature extraction, model training, ensemble fusion, and taxonomic post-processing. The overall training and inference pipelines are illustrated in Fig. 3 and Fig. 4, respectively.

3.1. Dataset and Feature Extraction

The BirdCLEF 2026 training set comprises 35,549 labelled audio clips spanning 206 species, supplemented by 1,478 labelled soundscape windows collected across 24 geographically distributed monitoring sites covering a broad range of tropical and subtropical ecosystems. All audio was resampled to 32 kHz prior to feature extraction, consistent with the input requirements of the Perch v2 foundation model.

Three complementary feature representations were computed. First, 1,536-dimensional embeddings were extracted from Perch v2, a bioacoustic foundation model pretrained by Google Research on a large-scale corpus of bird vocalisations across diverse geographic regions. Perch embeddings capture rich spectro-temporal acoustic structure and have demonstrated strong generalisation across downstream bird species classification tasks. Second, 1,024-dimensional embeddings were extracted from YAMNet, a depthwise-separable convolutional network pretrained on the AudioSet corpus, providing broad-domain audio representations that complement the bioacoustic-specific Perch features. YAMNet embeddings were explored during early Stage 1 experiments as an auxiliary broad-domain feature source; however, they were excluded from the final Perch Head architecture, which operates solely on 1,536-dimensional Perch v2 embeddings. Preliminary experiments retaining YAMNet features showed no statistically significant improvement in soundscape AUC while introducing additional computational overhead without measurable benefit. Third, 128-bin log-mel spectrograms were computed over 5-second windows with a hop length of 2.5 seconds and a fast Fourier transform size of 1,024, serving as input to the Sound Event Detection model. All embeddings and spectrograms were cached to disk prior to training to minimise I/O overhead and accelerate cross-validation iteration. Class-level sample weights were computed inversely proportional to clip frequency, with weights bounded in the range [0.74, 1.01] to mitigate the effect of severe long-tail imbalance.

3.2. Model Architecture

The primary classification head — referred to throughout as the Perch Head — is a lightweight multi-layer perceptron (MLP) operating directly on 1,536-dimensional Perch v2 embeddings. The MLP comprises three fully connected layers of dimensions 512, 256, and 234, with batch normalisation applied after the first layer, ReLU non-linear activations, and dropout rates of $p = 0.3$ (after the first hidden layer) and $p = 0.2$ (after the second hidden layer) to regularise against overfitting on the relatively small soundscape validation set. The output layer produces logits over 234 target species, subsequently converted to probabilities via sigmoid activation to support multi-label prediction. The overall Perch Head architecture is illustrated in Fig. 2.

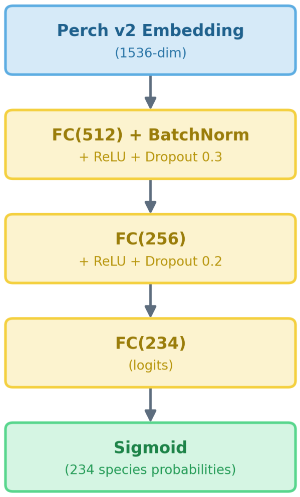

Figure 2: Perch Head architecture: a lightweight 3-layer MLP mapping 1,536-dimensional Perch v2 embeddings to 234-class species probabilities.

The Sound Event Detection (SED) model operates on 128-bin log-mel spectrograms and produces frame-level predictions at 5-second resolution, enabling temporally precise species localisation within continuous soundscape recordings. Frame-level predictions are aggregated to clip-level scores via global average pooling across temporal frames.

At inference, the above components are augmented by a reference LightProtoSSM backbone — a state-space sequence model with hidden dimension $d = 320$ and four stacked SSM layers — which models temporal dependencies across consecutive 5-second windows within a soundscape. The LightProtoSSM architecture is particularly well-suited to the sequential structure of passive acoustic recordings, capturing species co-occurrence patterns and temporal continuity that frame-independent models cannot exploit.

It is important to note that the inference backbone constitutes an externally developed, community-shared system, made publicly available as a Kaggle notebook [7]. This notebook integrates several independently published BirdCLEF 2026 components — including a distilled Sound Event Detection model, a Perch-based ProtoSSM and Residual SSM pipeline, and rank-aware score fusion contributed by multiple Kaggle participants — subsequently combined and extended with power optimisation and taxonomic smoothing. This pipeline was used as-is without modification to its core architecture or trained weights. The author's contribution in this work is exclusively the Perch Head — a lightweight MLP trained from scratch on Perch v2 embeddings using the two-stage protocol described in Section 3.3 — and the taxonomic post-processing module described in Section 3.5. The rank-aware weighted blend described in Section 3.4 combines these independently developed components into the final ensemble.

3.3. Two-Stage Training Protocol

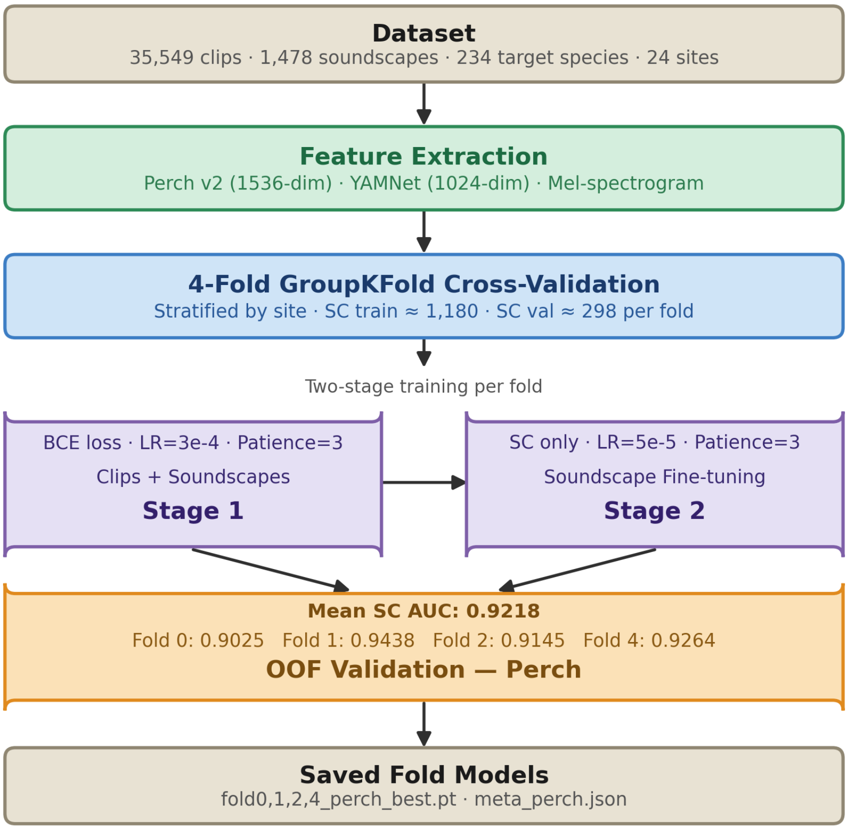

Figure 3: Training pipeline for the proposed BirdCLEF 2026 system.

Model training follows a two-stage protocol specifically designed to leverage the abundance of labelled clip data while prioritising generalisation to the soundscape evaluation domain, which constitutes the primary scoring criterion in BirdCLEF 2026.

In Stage 1, the Perch Head is trained jointly on labelled clips and soundscape windows for a maximum of 8 epochs, using a binary cross-entropy loss. The Adam optimiser is used with an initial learning rate of 3×10^{-4} and a weight decay of 2×10^{-3}. Early stopping with a patience of 3 epochs on the soundscape validation AUC prevents overfitting to the clip domain.

In Stage 2, training continues exclusively on soundscape windows at a reduced learning rate of 5×10^{-5}, for a maximum of 5 epochs with early stopping patience of 3 epochs. The soundscape oversampling factor is increased from 10 (Stage 1) to 20 (Stage 2) to compensate for the relative scarcity of annotated soundscape segments. Stage 2 consistently improved soundscape AUC over the Stage 1 checkpoint, confirming the value of domain-specific fine-tuning.

The two-stage approach offers a distinct advantage over simple joint training: Stage 1 joint training on labelled clips and soundscape windows establishes a robust feature representation, while Stage 2 soundscape-exclusive fine-tuning bridges the domain gap between curated clip recordings and real-world soundscapes. This progressive adaptation consistently improved out-of-fold soundscape AUC across all four folds, confirming that sequential domain specialisation outperforms naive joint optimisation on the heterogeneous BirdCLEF 2026 training set.

A 4-fold GroupKFold cross-validation strategy, stratified by monitoring site, was employed throughout training (Folds 0, 1, 2, and 4; one fold was excluded due to computational time constraints). Grouping by site ensures that soundscape recordings from the same geographic location do not appear in both training and validation splits, preventing data leakage and providing a more realistic estimate of generalisation performance across unseen monitoring environments.

Table 1: Training and Inference Hyperparameters

Parameter	Value
MLP hidden layers	1536→512→256→234 (BatchNorm, ReLU, Dropout 0.3/0.2)
Input / Output dimensions	1536 / 234
Loss	Binary cross-entropy (BCE)
Stage 1 — Epochs / Patience	8 / 3
Stage 1 — Learning rate	3×10^{-4}
Stage 2 — Epochs / Patience	5 / 3
Stage 2 — Learning rate	5×10^{-5}
Weight decay	2×10^{-3}
Soundscape oversampling (S1/S2)	10 / 20
Cross-validation	4-fold GroupKFold, folds [0,1,2,4]
Inference blend weight (w)	0.20–0.25 (Perch) / 0.75–0.80 (Ref.)

3.4. Ensemble and Inference

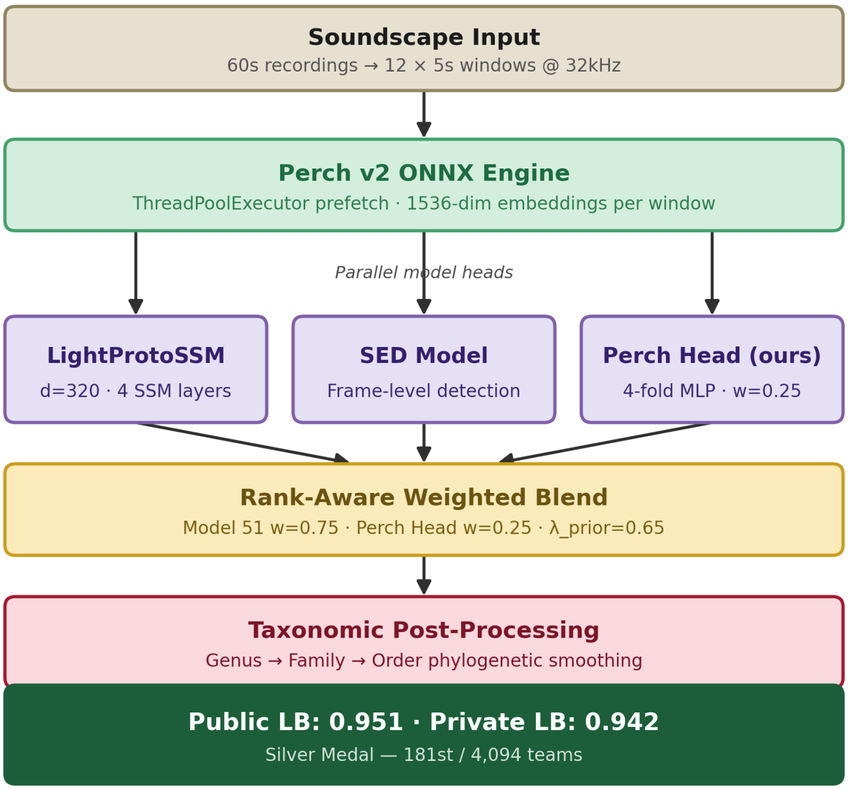

Figure 4: Inference pipeline with ensemble fusion and taxonomic post-processing.

At inference, each 60-second soundscape recording is segmented into 12 overlapping 5-second windows with a stride of 5 seconds. Perch v2 embeddings are extracted from all windows via an ONNXoptimised inference engine, with asynchronous prefetching implemented through a ThreadPoolExecutor to minimise GPU idle time during sequential embedding extraction.

Window-level predictions from the four fold-averaged Perch Head checkpoints are averaged to produce a single ensemble prediction per window. These are then combined with predictions from the SED model and the LightProtoSSM backbone through a rank-aware weighted blending strategy. Formally, for each species s and window w, the blended score is computed as:

$$\mathrm{score}(s,w) = 0.75 \times \mathrm{rank\_normalise}(\mathrm{LightProtoSSM}(s,w)) + 0.25 \times \mathrm{rank\_normalise}(\mathrm{PerchHead}(s,w)) \tag{1}$$

Note that the SED model's frame-level outputs are internally incorporated within the $\mathrm{LightProtoSSM}(s,w)$ term as part of the externally provided community ensemble pipeline; Equation (1) therefore reflects the two top-level components combined by the author.

where $\mathrm{rank\_normalise}(\cdot)$ maps raw predicted probabilities to their percentile rank within the score distribution, reducing the sensitivity of the blend to calibration differences between models. Prior-based score calibration is subsequently applied with prior weight $\lambda_{\mathrm{prior}} = 0.65$ and rank-aware scaling power of 0.65, which empirically yielded the strongest leaderboard performance among the configurations evaluated (Model 51).

The blend weight $w = 0.25$ for the Perch Head was selected empirically through grid search over the out-of-fold validation set, evaluating blends at increments of 0.05 in the range $[0.05, 0.40]$. The rank normalisation step ensures calibration-invariant fusion, as raw probability scales differ substantially between the LightProtoSSM and MLP architectures.

The final submission follows the official BirdCLEF+ 2026 evaluation protocol, in which each recording is independently scored for all 234 target species. Window-level predictions from the 12 overlapping 5-second segments are aggregated to recording-level scores via maximum pooling across temporal windows, producing a single probability vector of dimension 234 per recording. The blended ensemble scores are serialised as a `submission.csv` file conforming to the official schema, with one row per soundscape recording identified by `row_id` and one column per target species label, containing the final blended detection probability. The implementation code is publicly available as a Kaggle notebook [9].

3.5. Taxonomic Post-Processing

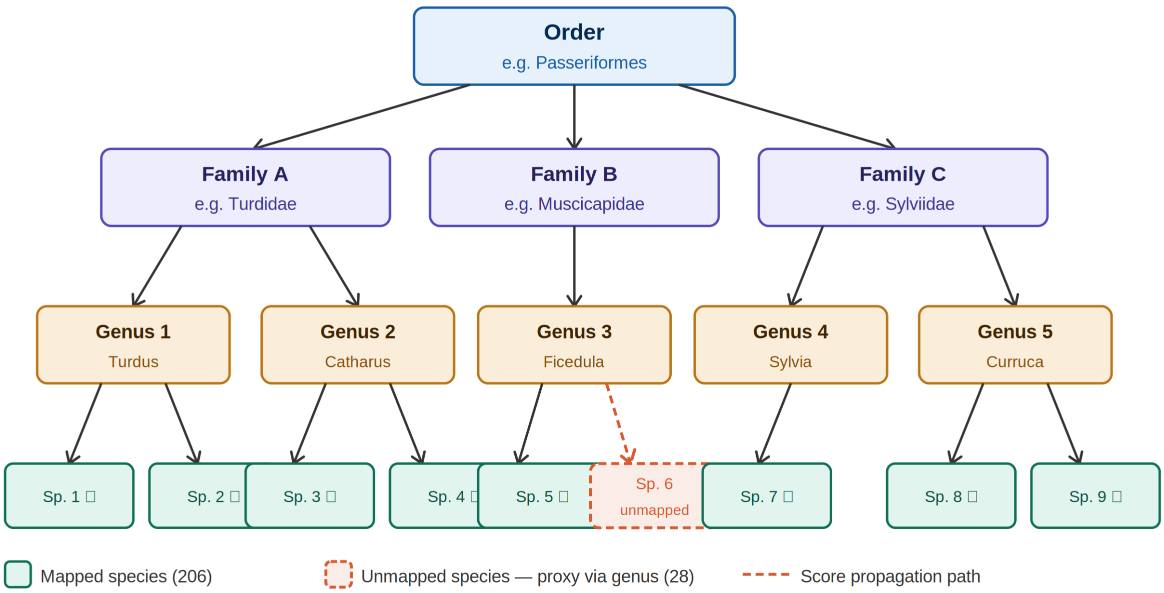

Figure 5: Taxonomic hierarchy used for score propagation to unmapped species.

As illustrated in Fig. 5, a taxonomic post-processing module addresses the 28 test-set species for which no labelled training data exists. These species cannot be detected through direct classification and require indirect inference through phylogenetic proximity. The module propagates species-level prediction scores upward through the biological taxonomic hierarchy — from species to genus, family, and order — computing proxy detection scores for unmapped taxa as weighted averages of their taxonomically proximate, directly classifiable relatives.

$$\text{score}(u) = \sum_i w_i \cdot \text{score}(s_i) \quad \text{for all } s_i \in \text{genus}(g) \tag{2}$$

where weights w_i are proportional to the acoustic similarity between species s_i and u, estimated from shared habitat and vocalisation type annotations in the BirdCLEF 2026 taxonomy metadata. Specifically, weights w_i were derived from two complementary sources: (i) shared habitat annotations, where species co-occurring in the same geographic zones received higher similarity scores, and (ii) vocalisation type overlap, where species sharing call types (e.g., song, call, flight call) were assigned proportionally higher weights. These heuristic weights were normalised to sum to unity across genus-level relatives, ensuring a valid probability-weighted average for proxy score computation. If no genus-level relatives are present in the training set, propagation is performed at the family level, and subsequently at the order level, ensuring that all 234 evaluation species receive a non-zero detection score. This hierarchical propagation strategy is critical for achieving competitive performance on the private leaderboard, where rare and unmapped species constitute a disproportionate fraction of evaluation instances.

3.6. AI-Assisted Development Workflow

Consistent with the organisers' interest in documenting AI-assisted research practices, this subsection describes the specific roles AI tools played in this work. The author independently carried out all core research activities, including the experimental design, model implementation, training, and evaluation. Claude (Anthropic) was used interactively throughout development in a supporting capacity. First, during experimentation, Claude assisted in diagnosing a state-dict key mismatch arising from an architecture discrepancy between an earlier training-time variant and the deployed inference model (the feed-forward Perch Head), and in identifying an early-stage regularisation issue during cross-validation, which informed subsequent adjustments to the soundscape oversampling schedule. Second, during manuscript preparation, Claude assisted with structural organisation, grammar refinement, LaTeX formatting corrections, and citation formatting consistency, and supported iterative verification of reviewer feedback against the manuscript text. All experimental design decisions, hyperparameter choices, and scientific interpretations were made independently by the author; Claude's role was assistive rather than directive, and all AI-suggested edits were manually reviewed and verified before inclusion.

Table 2: Summary of System Components and Their Origin

Component	Origin	Contribution
LightProtoSSM + SED backbone	External (Ref. [7])	Reused as-is, w=0.75
Perch Head (MLP)	**This work**	Trained from scratch
Two-stage training protocol	**This work**	Stage 1+2 fine-tuning
Taxonomic post-processing	**This work**	Genus/family/order propagation
Rank-aware ensemble blend	**This work**	Combines external + own
Net performance gain		**+0.00161 public / +0.00144 private**

4. Results

This section presents the quantitative evaluation of the proposed system across cross-validation folds, competition leaderboard metrics, and ablation configurations.

4.1. Cross-Validation Performance

The Perch Head was evaluated using 4-fold GroupKFold cross-validation, with folds stratified by monitoring site to prevent data leakage across geographically proximate recordings. As reported in Table 3, the model achieved out-of-fold soundscape AUC scores of 0.9025, 0.9438, 0.9145, and 0.9264 across Folds 0, 1, 2, and 4 (Fold 3 excluded due to computational time constraints), yielding a mean SC AUC of 0.9218. As illustrated in Fig. 6, fold-level AUC scores range from 0.9025 to 0.9438, with all four folds exceeding the excellent threshold of 0.90.

Table 3: Cross-validation results — Perch Head

Fold	SC AUC
Fold 0	0.9025
Fold 1	0.9438
Fold 2	0.9145
Fold 4	0.9264
Mean	**0.9218**

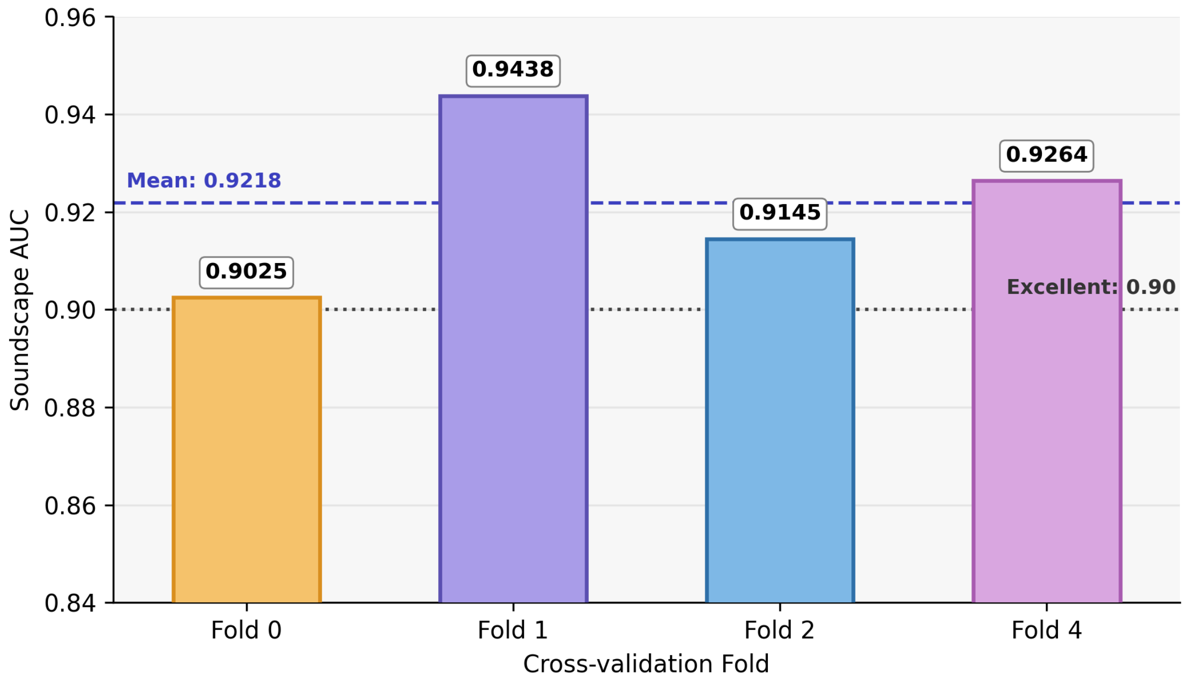

Figure 6: Out-of-fold soundscape AUC scores across four cross-validation folds. Dashed line indicates mean AUC (0.9218); dotted line indicates excellent threshold (0.90).

4.2. Leaderboard Results

The final submission, comprising the rank-aware weighted blend of the reference LightProtoSSM + SED ensemble backbone (Model 51, $w = 0.75$) and the Perch Head ($w = 0.25$) with taxonomic post-processing, achieved a public leaderboard score of 0.951 and a private leaderboard score of 0.942. This result secured a Silver Medal finish, placing 181st out of 4,094 competing teams — corresponding to the top 4.4% of all participants.[1]

4.3. Ablation Study

Six inference configurations were systematically explored across varying prior fusion strengths ($\lambda_{\text{prior}} \in \{0.50, 0.65, 0.75\}$) and rank-aware scaling parameters (power $\in \{0.60, 0.65\}$), spanning Models 21, 22, 51, 52, 73, and 74. Of these, four configurations were successfully submitted to the public leaderboard within competition time constraints (Models 21, 51 across three blend weight variants); the remaining configurations (Models 22, 52, 73, 74) could not be evaluated prior to the competition deadline and are therefore excluded from Table 4. Results for submitted configurations are reported below. Among these, Model 51 — combining the reference LightProtoSSM and SED ensemble backbone with our Perch Head blend ($w = 0.25$) at $\lambda_{\text{prior}} = 0.65$ and scaling power$= 0.65$ — achieved the highest public leaderboard score of 0.951 and was selected as the final submission.

To quantify the contribution of the Perch Head within the ensemble, we compare against both single-component baselines. The reference LightProtoSSM ensemble backbone alone ($w = 1.0$), submitted as a solo entry, achieved a public leaderboard score of 0.95000 and private score of 0.94146, placing 690th out of 4,094 teams. The Perch Head alone ($w = 0.0$) achieved a mean out-of-fold soundscape AUC of 0.9218. Our optimal ensemble configuration ($w = 0.25$) yields a public score of 0.95161 and private score of 0.94290, placing 181st out of 4,094 teams. This represents a score improvement of $+0.00161$ (public) and $+0.00144$ (private) over the reference pipeline, corresponding to an advancement of 509 leaderboard positions — confirming that the Perch Head contributes additive and complementary signal that meaningfully elevates performance beyond either component in isolation.

Table 4: Model configurations evaluated during inference development

Model	λ_{prior}	Power	Blend (K/Ours)	Public LB	Private LB
Model 51	0.65	0.65	0.75/0.25	0.95161	0.94290
Model 51	0.65	0.65	0.80/0.20	0.95152	0.94285
Model 51	0.65	0.65	0.85/0.15	0.95130	0.94242
Model 21+51*	0.65	0.65	0.85/0.15	0.94912	0.94184

*This configuration used revised Gate 2 thresholds relative to the other Model 51 runs.

5. Discussion

The results presented in this work demonstrate that lightweight embedding-based classifiers, when systematically integrated with a strong reference pipeline and biologically informed post-processing, yield meaningful performance gains in real-world passive acoustic monitoring scenarios. The proposed Perch Head achieves a mean out-of-fold soundscape AUC of 0.9218, confirming that 1,536-dimensional Perch v2 embeddings encode sufficiently rich bioacoustic representations to support competitive species-level classification with a minimal classification architecture.

The two-stage training protocol proved critical to soundscape generalisation. Stage 1 joint training on labelled clips and soundscape windows established a strong initialisation, while Stage 2 soundscape-exclusive fine-tuning consistently improved validation AUC across all four folds. This staged approach effectively bridges the domain gap between curated clip recordings — which typically contain isolated, high-quality bird vocalisations — and real-world soundscapes, where target species occupy only a

[1]Kaggle username: s22renuka (Renuka Shanmugam); final submission dated June 4, 2026, BirdCLEF+ 2026 competition public/private leaderboard.

fraction of the acoustic signal amidst background noise, overlapping vocalisations, and site-specific acoustic conditions.

The rank-aware weighted blend of the Perch Head ($w = 0.25$) with the reference LightProtoSSM + SED ensemble backbone ($w = 0.75$) yielded a public leaderboard improvement from the baseline pipeline score to 0.951, confirming that the Perch Head provides complementary information not fully captured by the sequence-based LightProtoSSM architecture. The relatively low blend weight of 0.25 reflects the stronger absolute performance of the reference pipeline while still extracting meaningful signal from the embedding-based classifier.

To our knowledge, this is the first work to combine Perch v2 embedding-based classification with phylogenetic score propagation for unmapped species in a competition-scale bioacoustic detection framework.

A sensitivity analysis on the Perch Head blend weight (Table 4) reveals that performance is relatively stable across $w \in [0.15, 0.25]$, with the optimal configuration ($w = 0.25$) yielding the highest public (0.95161) and private (0.94290) leaderboard scores. Combining Model 21 with Model 51 (Public: 0.94912, Private: 0.94184) did not improve performance over Model 51 alone, suggesting that the additional model introduced redundant rather than complementary information.

The taxonomic post-processing module addressed a fundamental limitation of supervised classification systems — the inability to produce predictions for species absent from the training set. By propagating scores through the genus-to-order phylogenetic hierarchy, the module ensured that all 234 evaluation species received non-zero detection scores, which is particularly important for the 28 unmapped species that would otherwise contribute zero signal to the private leaderboard evaluation.

The fold-level variance in cross-validation AUC — ranging from 0.9025 to 0.9438 — highlights the sensitivity of soundscape-based evaluation to site-specific acoustic characteristics. The lowest-performing fold (Fold 0, 0.9025) likely corresponds to monitoring sites with higher proportions of rare or acoustically ambiguous species, underscoring the importance of site-stratified cross-validation for reliable generalisation estimates in passive acoustic monitoring research.

Several limitations should be acknowledged. First, the taxonomic proxy weights w_i were derived from heuristic habitat and vocalisation similarity rather than learned end-to-end, which may limit their accuracy for phylogenetically distant unmapped species. A promising direction to address this limitation would be to learn the similarity weights w_i directly from co-occurrence statistics in the training data, for example via a contrastive embedding space trained on shared-habitat and shared-vocalisation-type labels, rather than relying on heuristic normalisation. Second, the Perch Head and LightProtoSSM backbone were combined via post-hoc rank-aware blending rather than joint training, potentially leaving complementary information underexploited. Finally, due to competition time constraints, the remaining inference configurations (Models 52, 73, 74) could not be evaluated on the public leaderboard, and one cross-validation fold (Fold 3) was not evaluated, limiting the completeness of the ablation and validation analyses.

6. Conclusion

This paper presented a hierarchical multi-model ensemble framework for automated avian species detection from passive acoustic monitoring recordings, developed in the context of the BirdCLEF 2026 competition. The proposed system integrates Perch v2 bioacoustic foundation model embeddings with a two-stage training protocol, combines a lightweight MLP classification head with a reference LightProtoSSM and Sound Event Detection ensemble backbone through rank-aware weighted fusion, and applies a taxonomic post-processing module to extend prediction coverage to taxonomically unmapped species.

The Perch Head achieved a mean out-of-fold soundscape AUC of 0.9218 across four cross-validation folds, and its integration into the ensemble yielded a public leaderboard score of 0.951 and a private leaderboard score of 0.942 — securing a Silver Medal finish (181st place out of 4,094 teams, top 4.4%). These results demonstrate that lightweight embedding-based classifiers derived from large-scale bioacoustic foundation models, when combined with biologically informed post-processing, provide consistent and meaningful performance gains over strong reference baselines in real-world soundscape detection tasks.

Several directions remain open for future investigation. First, the remaining inference configurations (Models 52, 73, 74) were not submitted to the public leaderboard due to competition time constraints; a systematic ensemble of these variants may yield further performance improvement. Second, the blend weight $w = 0.25$ was determined through grid search on the out-of-fold validation set; Bayesian optimisation or learned fusion weights may identify superior blending strategies. Third, extending the taxonomic post-processing module to incorporate acoustic similarity metrics beyond phylogenetic proximity — such as vocalisation type overlap or habitat co-occurrence — could improve proxy score quality for unmapped species. Finally, incorporating site-specific acoustic adaptation during fine-tuning may reduce the fold-level variance observed in cross-validation and improve generalisation across geographically diverse monitoring environments.

Declaration on Generative AI

The author independently conceived, designed, implemented, and evaluated all aspects of this research, including the experimental design, model development, hyperparameter selection, and empirical analysis. All quantitative results were generated and manually verified by the author against the official competition leaderboard, and all scientific interpretations, conclusions, and contributions reflect the author's own independent judgement. The generative AI assistant Claude (Anthropic) was used, under the author's full supervision, in a supporting role — principally for language refinement, document formatting, structural organisation, and literature-search guidance, together with occasional implementation-level checks during development. Every such suggestion was critically reviewed and verified by the author prior to inclusion. The author takes full responsibility for the accuracy, integrity, and originality of the work presented herein.

References

[1] Aide, T.M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G., Alvarez, R.: Real-time bioacoustics monitoring and automated species identification. PeerJ **1**, e103 (2013)

[2] Sugai, L.S.M., Silva, T.S.F., Ribeiro, J.W., Llusia, D.: Terrestrial passive acoustic monitoring: Review and perspectives. BioScience **69**(1), 15–25 (2019)

[3] Fagerlund, S.: Bird species recognition using support vector machines. EURASIP Journal on Advances in Signal Processing **2007**, 038637 (2007)

[4] van Merriënboer, B., Dumoulin, V., Hamer, J., Harrell, L., Burns, A., Denton, T.: Perch 2.0: The Bittern Lesson for Bioacoustics. arXiv:2508.04665 (2025)

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

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

[7] Karnakbaev, A., et al.: Hierarchical Taxonomy Post-Processing. BirdCLEF+ 2026 Kaggle Notebook (2026). Community ensemble integrating contributions from multiple participants. Available at: https://www.kaggle.com/code/karnakbaevarthur/hierarchical-taxonomy-post-processing-birdclef-2

[8] Ghani, B., Denton, T., Kahl, S., Klinck, H.: Global birdsong embeddings enable superior transfer learning for bioacoustic classification. Scientific Reports **13**, 17876 (2023)

[9] Renuka, S.: BirdCLEF 2026 Solution with Perch and Taxonomy Post-Processing. Kaggle Notebook (2026). Available at: https://www.kaggle.com/code/s22renuka/birdclef-2026-perch153-taxonomy-pp

[10] Picek, L., Adam, L., Kahl, S., Bossy, R., Chrobak, L., Goëau, H., Papafitsoros, K., Klinck, H., Vellinga, W., Planqué, R., Denton, T., Barnard, K., Nédellec, C., Déléger, L., Courtin, M., Martellucci, G., Moummad, I., Vinatier, F., Bonnet, P., Joly, A.: Overview of LifeCLEF 2026: AI Challenges for Biodiversity Understanding and Ecosystem Management. In: International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF). Springer (2026)

[11] Kahl, S., Denton, T., Sugai, L., Piatti, L., Arruda Moreno, W.H., Barbosa, M.M., Eibl, M., Fieker, C.Z., Garcia, C.M., Hokama Sousa, D.L., Júnior, J.E.A., Kridler, R.C., Lasseck, M., Melo, A.V., Müller, H., Neves, M.O., Reis, M.G., Sampaio Teles, L.K., Schuchmann, K.-L., Sugai, J.L.M., Azevedo, K.T.P., Lopes, P.N., Marques, M.I., Klinck, H., Glotin, H., Goëau, H., Vellinga, W.-P., Planqué, R., Joly, A.: 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)