Rank Before Fusion: A Calibration-Robust Three-Branch Bioacoustic Ensemble for BirdCLEF+ 2026

Jingwei Cheng[1,*,†], Jinhao Zhao[1,†], Runlai Ran[1], Xiaojie Yuan[1,‡] and Yihang Wang[1,‡]

[1]*Chongqing University of Posts and Telecommunications, Chongqing 400065, China*

Abstract

This working note describes our 18th place solution for BirdCLEF+ 2026. The final submission used a three-branch ensemble composed of two sound event detection (SED) models with different mel-spectrogram configurations and a Perch/ProtoSSM branch that combined pretrained bioacoustic embeddings, metadata priors, and sequence-level correction. The main engineering choice was to avoid direct averaging of heterogeneous probabilities. Instead, we aligned all submissions by `row_id`, converted each branch output into per-class percentile ranks, averaged the ranks, and applied a final file-level top-window post-processing step. We also used a restricted pseudo-labeling procedure: pseudo labels were generated by a single distilled model and were used only for Perch single-model MixUp training. Our best public/private leaderboard scores were 0.954/0.957. Ablation logs indicate that incremental changes such as weighted sampling, broader audio usage, class-type dependent temporal smoothing, baseline bug correction, and rank-based fusion were more important than any single large model change. We additionally use an executed Kaggle notebook to report dataset diagnostics and provide paper-ready supporting figures.

Keywords

BirdCLEF+, bioacoustics, sound event detection, rank ensemble, Perch, pseudo labeling

1. Introduction

Automated identification of bird and animal vocalizations is an important component of large-scale biodiversity monitoring. The BirdCLEF+ 2026 task was part of the LifeCLEF 2026 evaluation campaign [1, 2]. It continues this line of work by asking participants to recognize target species from soundscape recordings under strong domain shift, background noise, and imperfect temporal localization. Each test soundscape is a 60 second audio file, and systems must predict species presence for each 5 second window.

Our solution was built around a simple observation from the competition: no single branch was reliable across all species, sites, and acoustic contexts. Different mel resolutions, pretrained bioacoustic embeddings, temporal models, and metadata priors captured different failure modes. The final solution therefore emphasized diversity and calibration-robust fusion rather than a single large model. We finished 18th on the private leaderboard with a three-branch rank ensemble. We are grateful to Kaggle and the BirdCLEF+ 2026 organizers for providing the competition platform and data.

The main contributions of this working note are:

- a three-branch architecture combining two SED branches and a Perch/ProtoSSM sequence branch;
- a calibration-robust rank-before-fusion strategy for combining heterogeneous bioacoustic branches whose probability scales are not directly comparable;
- a restricted pseudo-labeling procedure that uses one distilled model to train only a Perch single-model branch with MixUp;

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

*Corresponding author.

†These authors contributed equally.

‡Only Jingwei Cheng, Jinhao Zhao, and Runlai Ran participated in the competition. Xiaojie Yuan and Yihang Wang did not participate in the competition and contributed only to manuscript preparation.

cjwing2024@outlook.com (J. Cheng); 2593457390@qq.com (J. Zhao); rrrl615@outlook.com (R. Ran); 19123303019@163.com (X. Yuan); liyuu148103275@gmail.com (Y. Wang)

0009-0005-3085-8934 (J. Cheng)

- practical ablation evidence from competition-side experiment logs;
- a reproducible Kaggle EDA notebook for class imbalance, label-density, and audio-domain diagnostics.

2. Related Work

BirdCLEF systems commonly combine spectrogram frontends with image CNN backbones, often borrowing architectures from large-scale vision classification. EfficientNet-style models [3] and NFNet-style models [4] were therefore natural candidates for our SED branches. Weakly supervised SED models also motivate our use of both clip-level and frame-level objectives: attention over frame logits can convert weak clip labels into temporally localized evidence [5].

Transfer learning has become especially important in bioacoustics. BirdNET [6] and global birdsong embeddings [7] show that large pretrained acoustic models can transfer useful species-level and habitat-level information to downstream recognition tasks. Our Perch branch follows this direction by using pretrained embeddings both as direct features and as a distillation target for a smaller SED model.

The remaining components are standard but important engineering tools for noisy multilabel audio. MixUp [8], SpecAugment [9], focal loss [10], and pseudo labels [11] were all used conservatively. In our final system, pseudo labels were not a broad semi-supervised training loop; they were restricted to Perch single-model MixUp because the ablation signal was split-sensitive.

3. Task and Data

Each soundscape is split into 12 non-overlapping 5 second windows. For every window, the system predicts 234 target classes in the canonical order given by `sample_submission.csv`. We used two supervised data sources: `train_audio`, containing focal recordings with primary and secondary labels, and `train_soundscapes_labels.csv`, containing labeled soundscape windows. The focal recordings provide broad class coverage but are weakly localized in time. The labeled soundscapes are closer to the evaluation format because their labels are attached to 5 second windows, but they are much smaller and site-dependent.

3.1. Validation

Reliable validation was one of the main practical difficulties, as is common in BirdCLEF competitions. We used stratified folds for focal recordings based on the primary label. For labeled soundscape windows, we grouped by filename, so that windows from the same 60 second recording never crossed train and validation folds. We also monitored non-S22 macro AUC during development because the S22 subset appeared noisier in the labeled soundscapes.

Despite these precautions, we did not treat local validation as a precise estimate of private leaderboard performance. It was most useful for detecting large regressions, leakage risks, and row-order bugs. Near the final leaderboard range, small changes were often dominated by public/private split variance, which is why the ablations in Section 5 are reported as directional evidence rather than as a fully controlled study.

3.2. Data Diagnostics Notebook

An executed Kaggle notebook provides the dataset diagnostics used in this note. The released metadata contain 35,549 focal training recordings, 234 target classes, and 1,478 labeled soundscape windows. The focal recordings are highly imbalanced: the maximum-to-minimum primary label count ratio is 499, the median class has 125 recordings, and 25 classes have fewer than 10 recordings. In contrast, labeled soundscape windows contain an average of 4.22 positive labels per 5 second window. The class-frequency correlation between focal recordings and labeled soundscape windows is negative

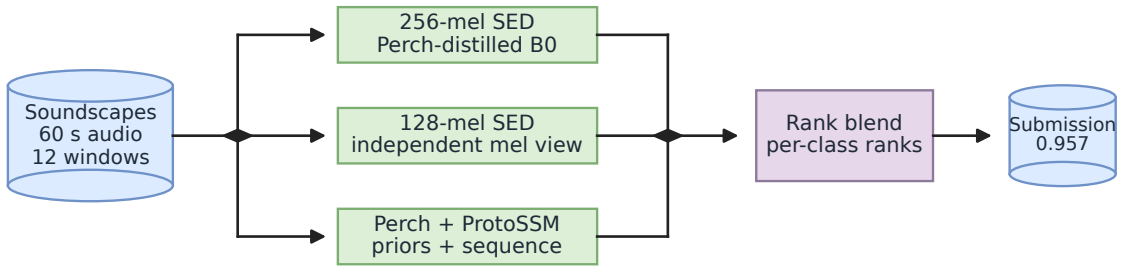

Figure 1: Overview of the private-best three-branch rank ensemble.

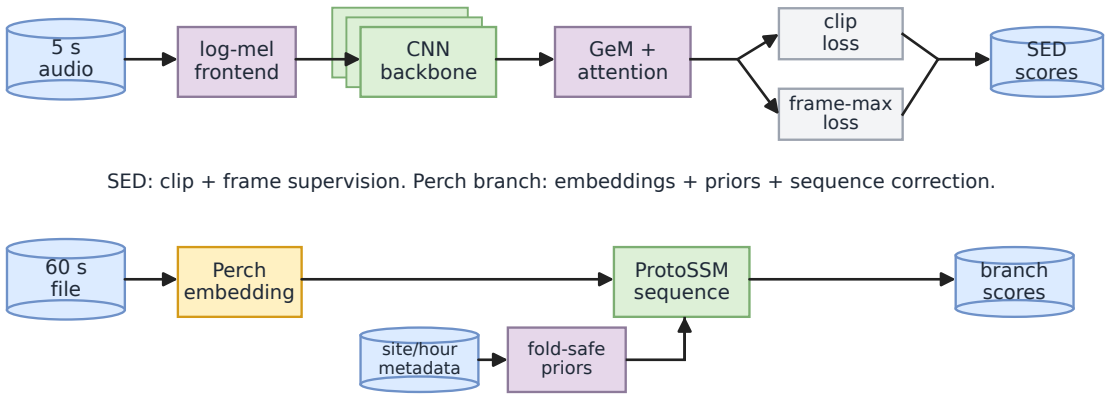

Figure 2: Modeling details for the SED and Perch/ProtoSSM branches.

(Spearman $r = -0.427$), which supports our emphasis on domain-robust fusion and soundscape-aware validation. Full diagnostic figures are shown in Appendix A.

4. Method

Figure 1 summarizes the final private-best system. The best submission used three explicit branches: a 256-mel SED model, a 128-mel SED model, and a Perch/ProtoSSM branch. A later four-model experiment added an NFNet-style branch but did not improve the private leaderboard score.

Figure 2 shows the two main modeling views used by the system. The SED branches make local 5 second decisions from log-mel spectrograms, while the Perch branch operates at the 60 second file level with pretrained embeddings, metadata priors, and sequence correction.

Table 1
Main SED branch configurations.

Branch	Mel bins	Hop	Backbone
SED-1	256	512	EfficientNet-B0
SED-2	128	512	EfficientNetV2-S / EfficientNet-B0 style
SED-4	192	768	NFNet-L0

4.1. SED Branches

The CNN branches use a sound event detection formulation over 5 second audio chunks. Audio is resampled to 32 kHz mono and converted to log-mel spectrograms. The main SED settings are shown in Table 1. The backbone family was based on EfficientNet-style models [3] and an NFNet variant [4].

The SED head uses a one-channel image backbone, GeM pooling over the frequency axis, a 512-dimensional bottleneck, attention logits, and class logits over the time axis. The clip-level prediction is computed from attention-weighted frame logits. Training supervises both the clip prediction and the strongest frame prediction:

$$\mathscr{L} = 0.5\,\mathrm{BCE}(z_{\mathrm{clip}}, y) + 0.5\,\mathrm{BCE}(z_{\mathrm{max-frame}}, y). \tag{1}$$

We used variants of focal BCE [10], label smoothing, class-frequency weighted sampling, background/no-call noise, SpecAugment [9], and MixUp [8] between focal and soundscape examples.

4.2. Perch Distillation and ProtoSSM

The 256-mel EfficientNet-B0 branch additionally used Perch v2 as a teacher [12]. A frozen Perch ONNX model produced a 1536-dimensional embedding for each 5 second chunk. This embedding was projected to 512 dimensions with a fixed random orthogonal projector, and the student model predicted it through an MSE distillation head. The goal was to pull the student representation toward a more general bioacoustic embedding space while still optimizing the competition labels.

The third branch was intentionally different from the mel-CNN branches. Perch v2 provided raw class scores and embeddings for every 5 second window. We then added fold-safe metadata priors based on site, hour, and site-hour tables. These priors were smoothed in a taxonomy-aware way, with different behavior for event-like species and texture-like classes such as insects and amphibians.

The temporal model was ProtoSSM. It treated each 60 second soundscape as a sequence of 12 windows and learned corrections over Perch scores and embeddings. The branch also trained class-wise MLP probes on PCA-compressed Perch embeddings plus raw, prior, and base scores. ProtoSSM outputs, MLP/probe scores, and priors were then blended, with a residual SSM correction available as a second pass.

4.3. Pseudo Labels

Pseudo labels were used more narrowly than the final ensemble might suggest. They were generated by the second distilled single model, not by the final rank ensemble. The unlabeled soundscapes were processed in five shards, merged back into canonical `row_id` order, and used as soft labels only for Perch single-model MixUp training. This follows the general idea of pseudo-labeling [11], but keeps the teacher and student scope intentionally narrow. Figure 3 illustrates this restricted pseudo-labeling pipeline.

For pseudo MixUp, focal recordings and pseudo-labeled soundscape windows were mixed with equal weights:

$$x_{\mathrm{mix}} = 0.5\,x_{\mathrm{focal}} + 0.5\,x_{\mathrm{pseudo}}, \tag{2}$$

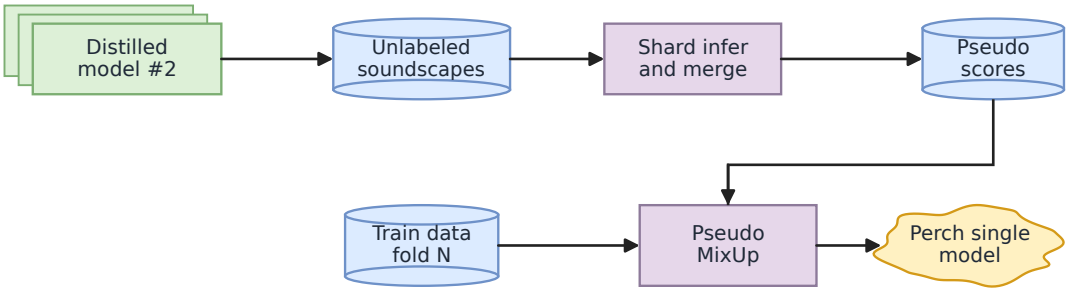

Figure 3: Pseudo labels were produced by one distilled model and used only for Perch single-model MixUp training.

$$y_{\text{mix}} = 0.5\, y_{\text{focal}} + 0.5\, y_{\text{pseudo}}. \tag{3}$$

Before mixing, pseudo scores were sharpened with $y_{\text{pseudo}} = s^{1.6}$. Validation files were excluded from the pseudo pool within each fold to reduce leakage risk.

4.4. Inference and Rank Fusion

The SED models were exported to OpenVINO XML and run in batches of 128 soundscape files. A shared audio cache and mel cache reduced repeated decoding and frontend computation. For each SED fold, we combined clip and frame predictions as

$$p = 0.5\, \sigma(z_{\text{clip}}) + 0.5\, \sigma(z_{\text{max-frame}}). \tag{4}$$

Fold predictions were averaged and then smoothed with class-type dependent 3-point kernels. We used $[0.20, 0.60, 0.20]$ for event-like classes and $[0.35, 0.30, 0.35]$ for texture-like classes. This step was split-sensitive: in a later removal check, disabling it improved our private-side score by about $+0.002$, but reduced public LB by about 0.0005. We therefore interpret it as a stability and calibration choice, not as a uniformly positive leaderboard trick. The Perch/ProtoSSM branch kept its own post-processing, including taxonomy temperature scaling, file-level confidence scaling, rank-aware scaling, delta smoothing, and per-class threshold sharpening.

The final fusion stage is shown in Figure 4. Instead of averaging raw probabilities, we first aligned all branch outputs by `row_id`. Then, for each branch and class, scores were converted to global percentile ranks:

$$r_{m,c}(i) = \text{percentile\_rank}\big(p_{m,c}(i)\big), \tag{5}$$

where m is the branch, c is the class, and i indexes the row. The final score was the equal-weighted average of the ranks. This preserves ordering while reducing calibration mismatch among branches. We preferred this percentile-rank fusion over standard probability averaging because the three branches used different frontends, objectives, calibration behavior, and post-processing assumptions. In competition logs, direct probability averaging was sensitive to score-scale mismatch: an over-confident branch could dominate classes for which its absolute probabilities were poorly calibrated. Percentile ranks preserved within-class ordering while reducing dependence on absolute score scale, which made the fusion more stable across heterogeneous branches. After rank fusion, we applied a file-level top-window

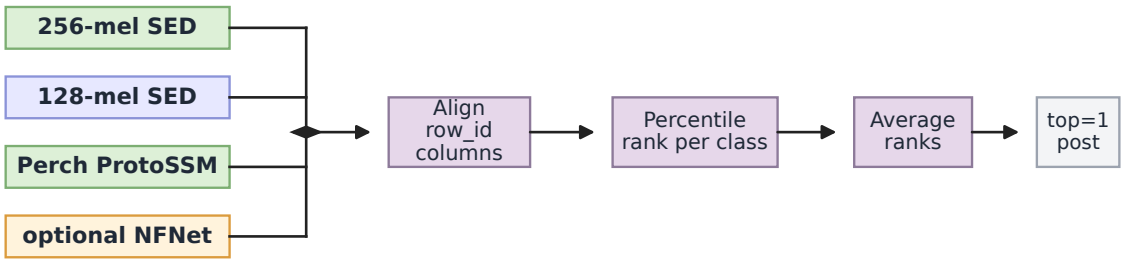

Figure 4: Rank fusion converts heterogeneous branch outputs into per-class percentile ranks before averaging.

Table 2
Main leaderboard submissions.

Submission	Public LB	Private LB
3-model rank ensemble with TTA	0.954	0.957
4-model rank ensemble	0.954	0.955

Table 3
V2S branch ablation log.

Experiment	LB	Note
initial V2S baseline	0.918	initial reference
clean-audio subset	0.918	private +0.001
moderate weighted sampling	0.924	WRS = 0.5; public gain
alternative weighted sampling	0.919	WRS = -1; private +0.004
file-level top-1 post-process	0.926	strongest-window file scaling
full-audio training	0.929	broader audio usage
class-type temporal smoothing	0.933	event/texture kernels
corrected V2S stack	0.942	baseline bug fix; cumulative changes

post-process: for each 60 second file and class, the strongest window scaled the predictions for all 12 windows.

5. Experiments and Ablations

Table 2 shows the two main late-stage submissions. The best private result came from the three-model TTA ensemble. The four-model ensemble was competitive but did not improve private LB, suggesting that additional diversity can also introduce private-set noise.

The ablation evidence in Table 3 and Table 4 comes from competition-side experiment logs rather than a single perfectly controlled offline protocol. We therefore interpret these numbers as directional evidence about which engineering choices were useful.

Class-type temporal smoothing used different 3-point kernels for event-like and texture-like classes. A later removal check moved public and private scores in opposite directions: private-side score increased by about +0.002, while public LB decreased by about 0.0005. For the corrected V2S stack row, we later found a bug in the earlier baseline path; the jump to 0.942 is therefore better read as the cumulative

Table 4
Perch/distillation and pseudo-label ablation log.

Experiment	LB	Note
row-order alignment sensitivity	0.917 / 0.914	alignment check changed LB by -0.003
512-dim Perch projection	0.917	private -0.004
Perch single without pseudo labels	0.926	stronger public LB
Perch single with pseudo-label MixUp	0.923	lower public, better private log

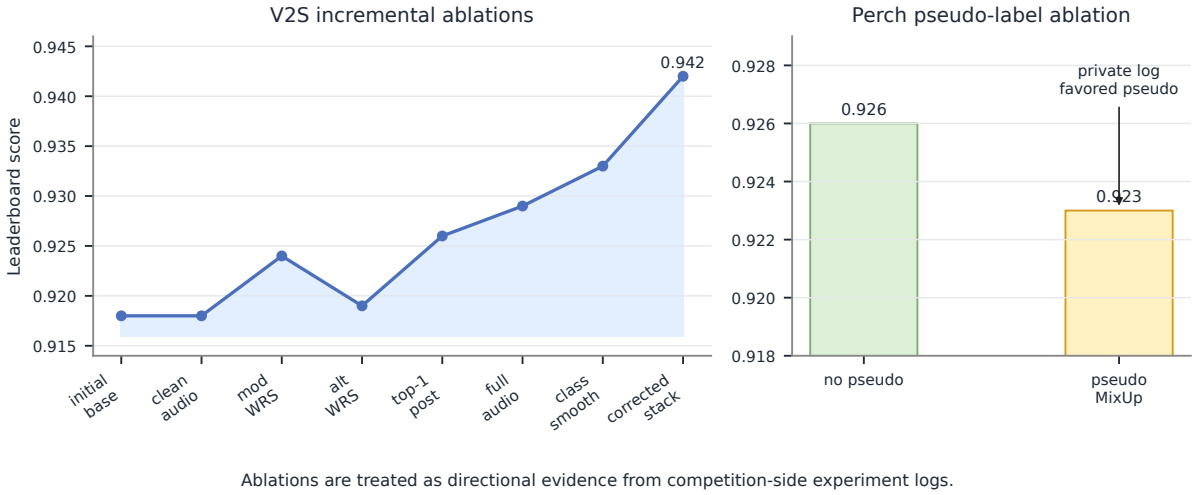

Figure 5: Ablation summary from competition-side experiment logs.

V2S stack after that bug was corrected.

The row-order alignment row reports two leaderboard values from a sensitivity check. The slash denotes a paired comparison, not an improvement direction: changing the alignment/order handling moved LB from 0.917 to 0.914, showing that row-level ordering was a fragile failure mode.

Figure 5 visualizes the same trend. The V2S results improved through a sequence of small changes, especially broader audio usage, class-type smoothing, and a late baseline-path correction. The smoothing and pseudo-label experiments were both split-sensitive, and the final corrected V2S stack row included a bug fix rather than an isolated new trick. We therefore treat these rows as engineering evidence rather than strict causal measurements.

6. Discussion and Limitations

The final solution was not driven by a single novel module. Its strength came from combining several imperfect but complementary views of the audio. The SED branches were local and spectrogram-based, while the Perch/ProtoSSM branch was embedding-based, metadata-aware, and sequence-aware. Rank fusion was the practical glue between these branches because it reduced the impact of probability calibration differences.

The ablation logs also show why we treated pseudo labeling cautiously. Although pseudo-labeled MixUp did not improve public LB for the Perch single model, it looked better on private in our records. We therefore kept it as a restricted branch-specific procedure instead of using pseudo labels as a general second-round training signal.

We applied the same caution to temporal smoothing. The final SED post-processing used different kernels for event-like and texture-like classes, but a later check showed opposite movement on public and private scores. This reinforced our view that such constants should be treated as split-sensitive calibration knobs rather than universally transferable hyperparameters.

The large final V2S jump in the ablation table also required caution. We later traced part of the earlier baseline path to a bug. We do not claim that any single new weighting trick caused the gain; it should be read as the cumulative effect of the preceding V2S improvements after the baseline was corrected.

Finally, the private leaderboard was favorable to our three-branch submission. The 0.957 private score should be interpreted with this shake in mind. Our post-hoc interpretation is that the cleaner three-branch ensemble generalized slightly better than the larger four-branch ensemble on the hidden split.

6.1. Limitations

This work has several limitations. First, the ablation logs were collected during competition development and were not always run under a single fixed offline protocol. They are useful for engineering interpretation, but they should not be read as statistically rigorous causal estimates. Second, local validation remained imperfect under domain shift, as illustrated by the class-type smoothing check where removing the step improved private-side score while slightly hurting public LB. A stronger study would retrain the same branches under several independent fold definitions and compare them on held-out soundscape regions. Third, the pseudo-labeling procedure was intentionally narrow. It helped in our private-side record, but it did not provide a clean public-LB gain, so broader semi-supervised training would need more careful confidence filtering and leakage control. Finally, the final 0.957 private score should be interpreted with leaderboard shake in mind. The hidden split was favorable to our three-branch blend, and we do not claim that the exact weights and post-processing constants are universally optimal.

7. Supplementary Material

A supplementary Kaggle notebook accompanies this working note draft and produces the CSV summaries and figures needed for an expanded data-analysis section, including class-imbalance plots, soundscape label-density plots, train-vs-soundscape class-frequency scatter plots, and spectrogram examples.

8. Acknowledgments

We thank Kaggle, the BirdCLEF+ 2026 organizers, and the LifeCLEF community for hosting the competition and releasing a dataset that supports practical bioacoustic research. We also thank the many data contributors whose recordings made this task possible.

9. AI-assisted Workflow Statement

AI-assisted tools, including ChatGPT/Codex, were used during the development of this working note to assist with drafting, editing, code/document organization, and the preparation of script-based figures. All figures were generated from author-written scripts or competition/data outputs rather than generative image models. All experiments were designed and configured by the authors, all results were manually verified, and all scientific interpretations and conclusions reflect the authors' own judgment. The authors take full responsibility for the accuracy, integrity, and originality of the work presented in this paper.

10. Conclusion

We presented an 18th place BirdCLEF+ 2026 solution based on a three-branch rank ensemble. The solution combined two SED branches with different mel views and a Perch/ProtoSSM branch with metadata priors and temporal correction. The final rank fusion preserved per-class ordering while

reducing calibration mismatch, and file-level top-window post-processing improved consistency across 60 second soundscapes. Competition-side ablations suggest that several small engineering choices, rather than a single dominant trick, were responsible for the final performance.

References

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

[2] S. Kahl, T. Denton, L. Sugai, L. Piatti, W. H. Arruda Moreno, M. M. Barbosa, M. Eibl, C. Z. Fieker, C. M. Garcia, D. L. Hokama Sousa, J. E. d. A. Júnior, R. C. Kridler, M. Lasseck, A. V. d. Melo, H. Müller, M. d. O. Neves, M. G. d. Reis, L. K. Sampaio Teles, K.-L. Schuchmann, J. L. M. M. 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, in: Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, 2026.

[3] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: Proceedings of the 36th International Conference on Machine Learning, 2019.

[4] 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, 2021.

[5] S. Adavanne, T. Virtanen, Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network, arXiv preprint arXiv:1710.02998 (2017).

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

[7] B. Ghani, T. Denton, S. Kahl, H. Klinck, Global birdsong embeddings enable superior transfer learning for bioacoustic classification, Scientific Reports 13 (2023) 22876.

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

[9] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, Specaugment: A simple data augmentation method for automatic speech recognition, arXiv preprint arXiv:1904.08779 (2019).

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017.

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

[12] Google Research, Perch: Pretrained bioacoustic embeddings, https://github.com/google-research/perch, 2024. Accessed during BirdCLEF+ 2026 system development.

A. Data Diagnostics from the Executed Kaggle Notebook

Table 5 summarizes the main values extracted from the executed notebook. The figures in this appendix are the paper-ready outputs exported from the same notebook.

Table 5
Dataset diagnostics from the executed Kaggle notebook.

Quantity	Value
Target classes	234
Focal training recordings	35,549
Labeled soundscape windows	1,478
Median focal recordings per class	125
Classes with fewer than 10 focal recordings	25
Primary-label imbalance ratio	499.0
Mean positive labels per labeled soundscape window	4.22
Train-vs-soundscape frequency Spearman r	-0.427

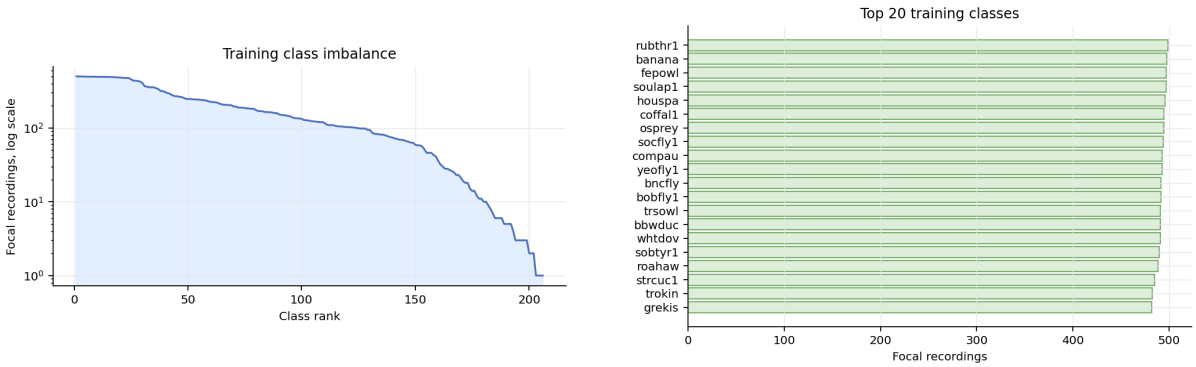

Figure 6: Focal-training class imbalance and the most frequent classes.

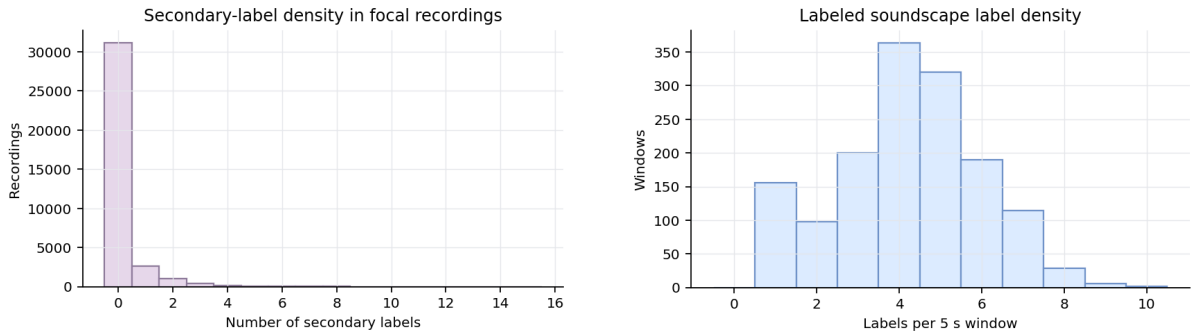

Figure 7: Label-density contrast between focal recordings and labeled soundscape windows.

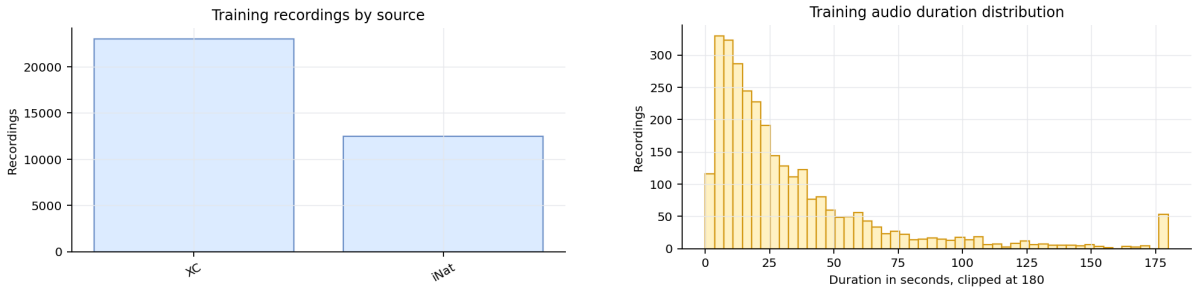

Figure 8: Training source distribution and sampled duration distribution.

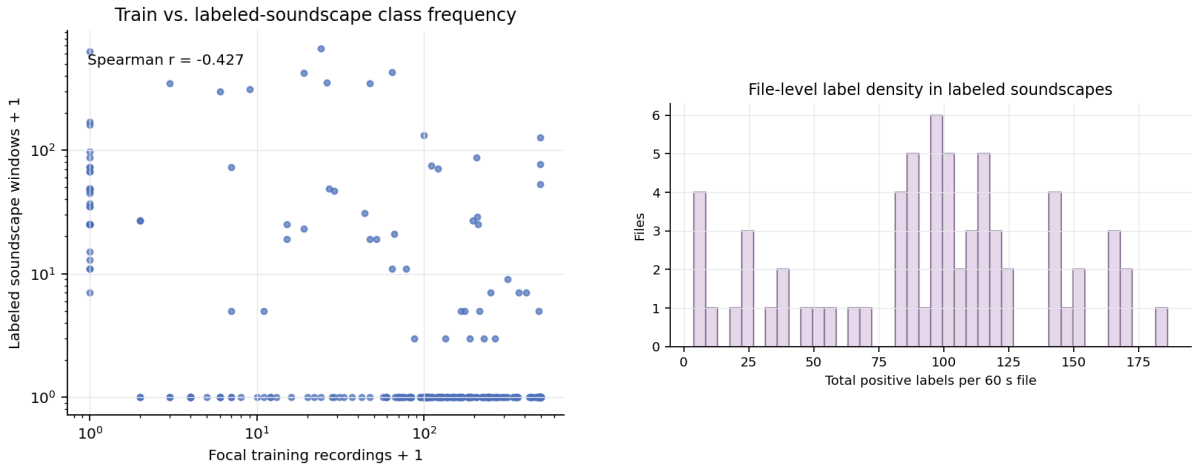

Figure 9: Class-frequency shift and file-level label density in labeled soundscapes.

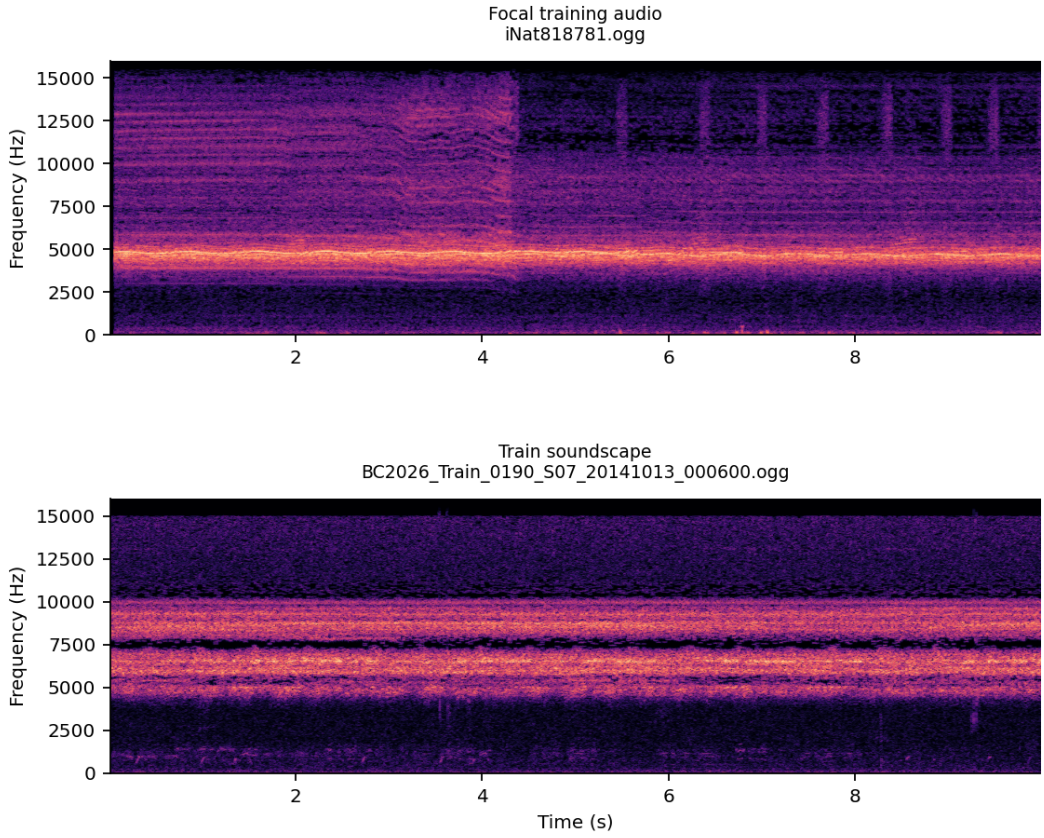

Figure 10: Representative spectrogram examples from one focal recording and one soundscape recording. These examples are used only as qualitative illustrations; the aggregate diagnostics in Table 5 and Figure 9 provide stronger evidence of dataset mismatch.