

BIODCASE 2026 - TASK 5 - CROSS-DOMAIN MOSQUITO SPECIES CLASSIFICATION

Technical Report

Oliver Gstöttenbauer

Andrii Ushakov

Artem Shestakov

Johannes Kepler University Linz Johannes Kepler University Linz Johannes Kepler University Linz
 Linz, Austria Linz, Austria Linz, Austria

ABSTRACT

We describe our system for the BioDCASE 2026 Task 5 Cross-Domain Mosquito Species Classification challenge [1]. The main challenge of this task is domain shift: the test recordings come from devices and conditions that the model never saw during training. We train the probe with losses that pull the different domains together, so it relies less on the recording channel. We then combine several device-invariant features with an agreement gate: it changes the main prediction only when independent features agree and are confident. Our system reaches an unseen-domain balanced accuracy of 0.362, against the baseline’s 0.175. Architecture: frozen Perch 2.0 embedding $\rightarrow$ a residual-MLP probe trained with MMD and CORAL domain alignment $\rightarrow$ an agreement gate that overrides the probe only where three independent device-invariant features agree.

1. INTRODUCTION

Idea is to remove the device effect in two ways. First, we use features that keep the wingbeat but drop the channel. Second, we train the probe so that the domains look alike. We then trust these device-invariant features only when they agree with each other.

2. SYSTEM

2.1. Overview

The system has four parts that share the same shape: a frozen feature, a small probe, and an average over random seeds. Three of the four features are device-invariant, which means they are built to ignore the recording channel. The fourth feature, Perch, is a strong general model and gives the default answer. An agreement gate then decides when to follow the device-invariant features instead of the default. We never fine-tune the large models; we only train the small probes.

2.2. Representations

Perch 2.0 [2] is a large model trained on many animal sounds. Its 1536-dimensional embedding is the strongest single feature we have, so it gives the base prediction. Harmonic feature (102-d) is computed by hand from the clip. It keeps the pitch and the harmonic comb of the wingbeat, but it drops the slow shape of the spectrum, which mostly carries the device color. So it is device-invariant by design.

Bird-MAE [3] is a second large model, trained on bird sound with masked autoencoding. Its 1024-dimensional embedding gives a view that is independent from Perch.

Background-whitened spectrum (257-d) removes the device color directly. For each clip we look at the quiet frames, where only the background is present, and measure the device color there. We then subtract this color from the loud frames, where the mosquito is present. What is left is the mosquito spectrum without the channel.

2.3. Probe and training losses

Each probe is a small residual MLP. It normalizes the input, maps it to 256 units, passes it through two residual blocks, and ends in a species head and a small domain head. We train it with the objective

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + 0.5 \mathcal{L}_{\text{genus}} + 0.5 \mathcal{L}_{\text{SupCon}} + \mathcal{L}_{\text{CORAL}} + 0.25 \mathcal{L}_{\text{MMD}} + 0.01 \mathcal{L}_{\text{dom}}. \quad (1)$$

Here $\mathcal{L}_{\text{CE}}$ is the species cross-entropy and $\mathcal{L}_{\text{genus}}$ also asks for the right genus. The supervised contrastive loss $\mathcal{L}_{\text{SupCon}}$ [4] pulls same-species clips together. The two alignment losses do the key work: CORAL [5] matches the covariance of the embeddings across domains, and MMD [6] matches their distributions with a Gaussian kernel. Together they make the domains look alike, so the probe cannot use the channel as a shortcut. This recipe follows the Domain-Robust Bioacoustic Learning idea [7]. The MMD weight 0.25 was the main knob: too little leaves the shortcut, too much hurts the species signal.

2.4. Agreement gate

Perch gives the default prediction. The gate changes it only when all three device-invariant features (harmonic, Bird-MAE, background-whitened) predict the same species, and all three are confident, with a minimum softmax above 0.3. In that case the new prediction is the mean of the three. The gate helps in two ways. It recovers species that are easy in sound but shifted by the device, such as *Aedes aegypti*. And the third feature acts as a veto: it stops the gate from firing on hard species, where two features alone might agree by chance and lower the score.

3. EXPERIMENTAL SETUP

Data and split. We use a per-species leave-one-domain-out split: for each species, one domain is held out and used only at test time. The validation set is almost all from one single domain.

Table 1: Unseen-domain balanced accuracy (BA_u) of the main configurations. Baseline from [1].

System	BA_u
Baseline (MTRCNN)	0.175
Perch base (8 seeds)	0.318
+ agreement gate (2 features)	0.341
+ agreement gate (3 features)	0.362
foreground-harmonic variant (in noise)	0.365

Training. We train every probe with AdamW and a cosine schedule, and we sample the domains evenly. The Perch base runs for 30 epochs, the other probes for 20. We train the Perch base with 8 seeds and each other probe with 3 to 5 seeds, then average their softmax outputs. We pick each checkpoint on the seen validation score; we never look at the test labels.

Evaluation. We report BA_u on the unseen test clips.

4. RESULTS

Table 1 shows the main numbers. The Perch base alone reaches 0.318. A gate that uses two features lifts it to 0.341. The full gate with three features reaches 0.362, which is our deployed system. A foreground-harmonic variant reaches 0.365, but this gain is within the seed noise (± 0.02 – 0.03).

We also look at the gate by species. The gain over the base comes mostly from *Aedes aegypti*, which rises from 0.40 to 0.81 because the three features agree on it. Easy species such as *Culex pipiens* (0.88) and *Anopheles minimus* (0.64) stay high. The hard *Anopheles* species stay low: *An. gambiae* (0.05) and *An. arabiensis* (0.19). These two sound almost the same, and their unseen clips have a strong shift, so no feature can separate them well. The third feature matters here too. With only two features, the gate over-fires and pulls *arabiensis* and *quinquefasciatus* down; the third feature vetoes those wrong agreements and keeps them at the base level. This is why three features (0.362) beat two (0.341).

5. CONCLUSION

Frozen embeddings, domain-alignment losses, and an agreement gate together about double the unseen balanced accuracy over the from-scratch baseline (0.362 versus 0.175). The gate is the main lever: it follows independent device-invariant features only when they agree, which recovers shifted species without hurting the rest. The remaining limits are structural, not about the features: we lack an unseen-domain validation signal, and some species overlap in sound. Both point to the same next step, a validation split that exposes the unseen regime.

6. REFERENCES

- [1] Y. Hou, V. Zdravkovic, M. Sinka, Y. Li, W. Wang, M. D. Plumbley, K. Willis, and S. Roberts, “Biodcase 2026 challenge baseline for cross-domain mosquito species classification,” 2026. [Online]. Available: <https://arxiv.org/abs/2603.20118>
- [2] B. van Merriënboer, V. Dumoulin, J. Hamer, L. Harrell, A. Burns, and T. Denton, “Perch 2.0: The bitter lesson for bioacoustics,” 2025. [Online]. Available: <https://arxiv.org/abs/2508.04665>
- [3] L. Rauch, I. Moummad, R. Heinrich, A. Joly, B. Sick, and C. Scholz, “Can masked autoencoders also listen to birds?” 2025. [Online]. Available: <https://arxiv.org/abs/2504.12880>
- [4] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, “Supervised contrastive learning,” 2021. [Online]. Available: <https://arxiv.org/abs/2004.11362>
- [5] B. Sun and K. Saenko, “Deep CORAL: Correlation alignment for deep domain adaptation,” in *Computer Vision – ECCV 2016 Workshops*. Springer, 2016, pp. 443–450.
- [6] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” *Journal of Machine Learning Research*, vol. 13, no. 25, pp. 723–773, 2012.
- [7] Y. Hou, Z. Liu, X. Shen, and S. Roberts, “Learning domain-robust bioacoustic representations for mosquito species classification with contrastive learning and distribution alignment,” 2025. [Online]. Available: <https://arxiv.org/abs/2510.00346>