DOMAIN-BALANCED ADVERSARIAL TRAINING FOR CROSS-DOMAIN MOSQUITO SPECIES CLASSIFICATION

Technical Report

Sulagna Saha

Mila – Quebec AI Institute
Montréal, Canada
saha23s@mtholyoke.edu

Aaron Elcheson

Mila / unaffiliated
aaron.elk5@gmail.com

ABSTRACT

We describe our submission to the BioDCASE 2026 Cross-Domain Mosquito Species Classification (CD-MSC) task. The dominant difficulty in this dataset is not the classifier but an extreme *training-domain* imbalance: one recording condition (D5) accounts for 99.4% of the training clips, so standard domain-generalisation (DG) methods never encounter the minority conditions they must generalise to. We show that *domain-balanced resampling* is the prerequisite that unlocks every subsequent DG method, and build a system around a multi-task MTRCNN backbone trained with balanced sampling, conditional domain-adversarial training (cDANN), and a domain-invariant contrastive loss (DiCL), with an optional frequency-band-selective augmentation (FBS-Mix). Because the unseen-domain metric is high-variance, our submitted system is a single *combined ensemble* that integrates all of our development work—four DG recipes over three seeds and four leave-one-domain-out (LODO) folds, a full-data model, and two conditional-DANN models—by equal-weight softmax averaging, yielding one robust, reproducible classifier. It attains species $\mathrm{BA}_{\mathrm{unseen}} \approx 0.25$ on the official per-species held-out partition. We also report a methodological caution: the unseen-domain metric is high-variance, so a single best-on-validation model overfits it; ensembling is the variance-robust choice.

***Index Terms*—** domain generalisation, bioacoustics, mosquito, class imbalance, ensembling

1. INTRODUCTION

Passive acoustic monitoring of mosquito wingbeats enables scalable species surveillance, but classifiers degrade sharply under *domain shift*: a model trained on one microphone/environment transfers poorly to another. The BioDCASE 2026 CD-MSC task [1] formalises this with nine species recorded across five opaque conditions (D1–D5). The official ranking is determined primarily by species balanced accuracy on unseen domains ($\mathrm{BA}_{\mathrm{unseen}}$), with the domain shift gap $\mathrm{DSG}=|\mathrm{BA}_{\mathrm{seen}}-\mathrm{BA}_{\mathrm{unseen}}|$ as a secondary metric (smaller preferred) and $\mathrm{BA}_{\mathrm{seen}}$ reported for reference.

The defining property of the development data is an extreme training-domain imbalance: D5 holds 212,339 of 213,647 training clips (99.4%); D1–D4 together provide only 1,308 clips (Table 1; Fig. 1, left). This single fact governs the task: any DG method that assumes balanced domain coverage is starved of the very signal it needs. We therefore make balanced resampling the foundation of the system and evaluate all method choices under a leave-one-domain-out (LODO) protocol over D1–D4 (D5 is never held out: doing so would leave 1,308 training clips, a data-starvation rather than a domain-shift regime).

Table 1: Training-domain distribution. D5 is 99.4% of training data.

	D1	D2	D3	D4
Train clips	634	230	364	80
% of train	0.30	0.11	0.17	0.04

D5: 212,339 train clips (99.38%). Total train: 213,647.

2. SYSTEM DESCRIPTION

Audio is resampled to 8 kHz and converted to 64-bin log-mel spectrograms (512-sample Hann window, 80-sample hop, 0–4 kHz), standardised per mel bin using training-split statistics.

Backbone. We use the multi-task MTRCNN [2] (Fig. 2): three parallel convolutional branches (temporal kernels 3/5/7), each three Conv–BN–ReLU–Pool stages, masked mean+max pooling over time, and a shared embedding feeding two linear heads (species, domain).

Domain-balanced sampling. Each training clip is sampled with weight $1/\mathrm{count}(d_i)$ over its domain d_i, equalising domain frequency per epoch. This is the single most important component: it upsamples D1–D4 by 200–2,600$\times$ so that DG objectives receive minority-domain signal. Without it, DANN reverses only a "D5-vs-rest" gradient and barely moves $\mathrm{BA}_{\mathrm{unseen}}$; with it, the discriminator must separate all five conditions and the reversed gradient removes all condition-specific directions.

Domain-adversarial training (DANN). A gradient-reversal layer [3] between the embedding and the domain head pushes the embedding to be domain-uninformative, with the standard $\lambda(p)$ ramp.

Domain-invariant contrastive loss (DiCL). Following the mosquito-specific formulation in [4], same-species clips from *different* conditions form positive pairs, pulled together in a two-layer projection head ($\rightarrow 128$) with temperature $\tau=0.2$ (raised from 0.07 because minority batches yield few positive pairs). The total loss is $\mathcal{L}_{\mathrm{sp}} - \lambda(p)\,\mathcal{L}_{\mathrm{dom}} + 0.1\,\mathcal{L}_{\mathrm{DiCL}}$.

FBS-Mix (optional). A per-frequency variance analysis (Fig. 1, right) shows mel bins 0–8 ($\sim$36–325 Hz) are condition-dominated while bins 9–35 ($\sim$361–1,360 Hz) carry the device-

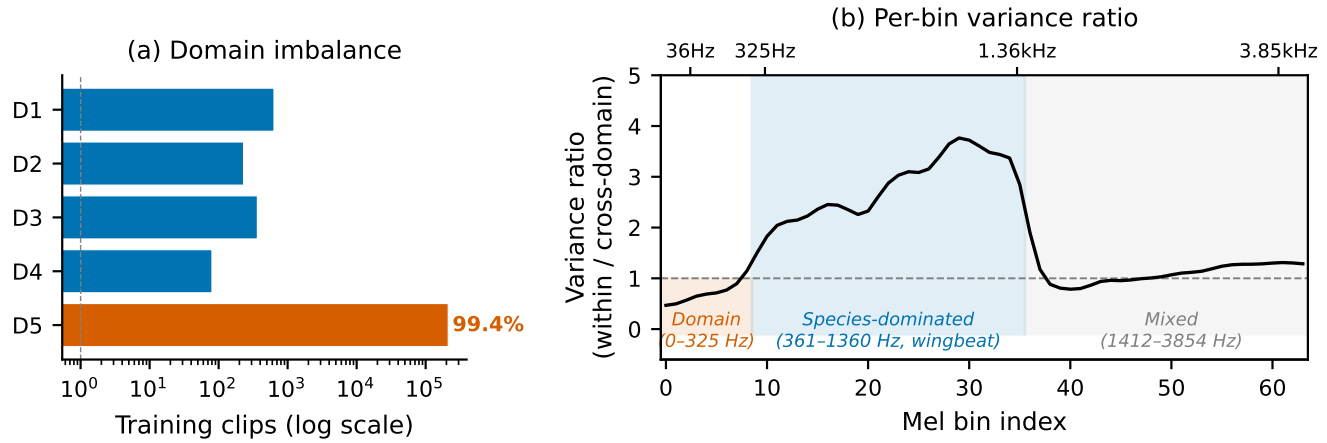

Figure 1: **Left:** training-domain distribution—D5 dominates (99.4%), starving DG methods of minority-condition signal. **Right:** per-mel-bin ratio of within-condition to cross-condition variance. Bins 9–35 (∼361–1,360 Hz, ratio>1) carry the device-invariant wingbeat; bins 0–8 are condition-dominated. This motivates FBS-Mix, which mixes only the condition band.

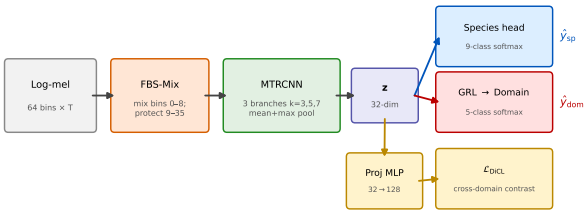

Figure 2: System overview: log-mel front-end → three-branch MTRCNN (temporal kernels 3/5/7) → masked mean+max pooling → shared embedding feeding species and domain heads, trained with domain-balanced sampling, cDANN, and DiCL.

invariant wingbeat. The Frequency-Band-Selective Mix augmentation blends instance statistics of the condition band only (AdaIN with $\lambda \sim \mathrm{Beta}(0.1, 0.1)$, statistics over valid frames), leaving the wingbeat core untouched. Unlike MixStyle—which corrupts the wingbeat and yields no gain—FBS-Mix reduces DSG at no inference cost.

3. SUBMITTED SYSTEM AND STRATEGY

Metric definition matters. The eval set is scored on the *official per-species* partition: a clip is "unseen" when its domain is that species' held-out domain (still present in training via other species). This differs from a whole-domain LODO measure; we report all numbers on the official partition to match the challenge.

Ensembling, done selectively. On the held-out test set we find two opposing effects. (i) Averaging in models with a strong seen-domain bias (frozen large-audio embeddings; a full-split model) raises overall and seen accuracy but *lowers* $\mathrm{BA}_{\mathrm{unseen}}$ and widens DSG, because softmax averaging reduces variance but not the shared seen-domain bias. (ii) Averaging *DG-only* configurations across seeds and folds is a variance-reduction win. A bootstrap over the 4,202 unseen test clips shows the unseen metric is high-variance (e.g. a single seed of our best config reads 0.290, 95% CI $[0.274, 0.304]$, but its three-seed average is 0.244): selecting the single best-on-validation model overfits the metric, while ensembling *DG-only* models reduces it.

Submitted system: combined ensemble. For a single, clean, reproducible submission we integrate all of our development work into one equal-weight softmax ensemble of 53 checkpoints: the four strongest DG recipes (DANN+DiCL at embedding/projection widths 32 and 128; DANN+DiCL+proj128; DANN+FBS-Mix) over three seeds and four LODO folds (48); the chosen recipe retrained on the *full* development data over three seeds (3); and the second author's two conditional-DANN (cDANN) models (2). Each member is normalised by its own training statistics. This deliberately folds in the full-data and cDANN models—which raise seen and overall accuracy at a small $\mathrm{BA}_{\mathrm{unseen}}$ cost—in exchange for a single robust deployable classifier rather than a noisy single-model point estimate. All component choices were fixed on development data; the evaluation set is unlabelled, so nothing is tuned to the leaderboard.

Inference. For each evaluation clip we average the species softmax of all members and take the arg-max as the predicted species.

Relation to whole-domain LODO. A common internal DG benchmark reports, for each held-out domain, the balanced accuracy of a *domain-specialist* (a model that never saw that domain), averaged over folds; our best recipe reaches $\approx$ 0.38–0.40 under that measure. The official metric is structurally harder and is the number we report: it scores a single deployable model on hidden-domain clips using the per-species held-out partition, so one cannot route each clip to its matching specialist, and the balanced mean includes sparse per-species combinations. The same models therefore read ≈ 0.25 officially, which is the value the organisers compute on the evaluation set.

4. RESULTS

Table 2 reports held-out test results on the official per-species partition. Balanced sampling is the dominant lever; the DG objec-

Table 2: Held-out development test, official per-species partition. $\mathrm{BA}_{\mathrm{unseen}}$ primary (higher better), DSG secondary (lower better). Top: components/ablations. Bottom: the submitted combined ensemble. †single seed—point estimate inflated by metric variance.

System	$\mathrm{BA}_{\mathrm{unseen}}$	$\mathrm{BA}_{\mathrm{seen}}$	DSG
Components / ablations			
MTRCNN + balanced	0.198	0.751	0.553
+ DANN	0.266	0.764	0.498
+ DANN + DiCL (e128)†	0.290	0.703	0.413
Full-split DG (3-seed)	0.250	0.768	0.517
Robust DG ensemble (48)	0.247	0.758	0.511
Frozen Perch embedding probe	0.214	0.804	0.591
cDANN	0.263	0.795	0.532
Submitted system			
Combined ensemble (53)	**0.244**	0.766	0.522

tives add cross-condition robustness; and the submitted ensemble gives the variance-robust operating point rather than the highest (but noisy) single-model point estimate.

5. CONCLUSION

The binding constraint is a 99.4% training-domain imbalance; domain-balanced resampling is the prerequisite that makes adversarial and contrastive DG objectives effective. Because the unseen-domain metric is high-variance, we submit a seed/fold ensemble of DG-only models rather than a single validation-best model, and exclude seen-biased components that trade $\mathrm{BA}_{\mathrm{unseen}}$ for overall accuracy.

6. ACKNOWLEDGMENTS

We thank Julien Boussard and Ricard Marxer for their advice on the process.

7. REFERENCES

[1] I. Kiskin *et al.*, "BioDCASE 2026 challenge baseline for cross-domain mosquito species classification," in *Proc. ICASSP*, 2026, arXiv:2603.20118.

[2] R. I. Fatan Kasyidi, "Multi-temporal-resolution CNN for mosquito species classification from wingbeat audio," in *Proc. ICASSP*, 2025.

[3] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[4] X. S. S. R. Yuanbo Hou, Zhaoyi Liu, "Learning Domain-Robust Bioacoustic Representations for Mosquito Species Classification with Contrastive Learning and Distribution Alignment," *arXiv:2510.00346*, 2025.