CROSS-DOMAIN MOSQUITO SPECIES CLASSIFICATION WITH INPUT-LEVEL MIXUP

Technical Report

Ewen Le Callet

Polytech Nantes
Nantes, France
ewen.le-callet@polytech-nantes.fr

ABSTRACT

This report describes our submission to the BioDCASE 2026 Task 5 challenge on cross-domain mosquito species classification (9 species, 5 domains). We build on the MTRCNN baseline and systematically evaluate eight domain generalization techniques across six experimental phases. Our key finding: input-level cross-domain Mixup with $\alpha = 0.2$ is the single most effective approach, improving $\text{BA}_{\text{unseen}}$ by 69% over baseline (0.156 $\rightarrow$ 0.263). Domain-adversarial training, manifold Mixup, focal loss, SWA, BYOL pretraining, and larger architectures all degraded unseen-domain performance. The final submission uses a single MTRCNN model (221K parameters) trained with SpecAugment and cross-domain Mixup.

Index Terms— Mosquito species classification, domain generalization, Mixup, bioacoustics, cross-domain

1. INTRODUCTION

The BioDCASE 2026 Task 5 challenge addresses mosquito species classification from flight sound recordings across different acoustic domains. The core difficulty is domain generalization: models must classify 9 mosquito species from audio recorded in unseen conditions (microphones, environments, protocols). The development dataset [1] contains 271,380 two-second 8 kHz clips across 5 domains with severe imbalance: D5 dominates (265,652 clips, 97.9%), while D4 has only 200 clips (0.07%). The official split designates D1 and D5 as seen domains (training and validation data available), and D2, D3, D4 as unseen domains (only in the test set). Evaluation uses species balanced accuracy on unseen domains ($\text{BA}_{\text{unseen}}$) as the primary metric, with Domain Shift Gap ($\text{DSG} = |\text{BA}_{\text{unseen}} - \text{BA}_{\text{seen}}|$) as secondary.

Our approach is deliberately simple: we systematically tested popular domain generalization methods and found that most hurt performance. The final system uses a single lightweight model with one augmentation technique—cross-domain Mixup with $\alpha = 0.2$.

2. METHOD

2.1. Model Architecture

We use the MTRCNN classifier [1] (221,267 parameters): input BatchNorm over 64 mel bins; three parallel CNN branches with kernel sizes 3, 5, and 7 in the time dimension, each with 3 convolutional stages (Conv $\rightarrow$ BatchNorm $\rightarrow$ ReLU $\rightarrow$ AvgPool) with increasing dilation; masked mean+max temporal pooling with frequency projection; and concatenated branch features (192-dim) projected through Linear(32) to species (9 classes) and domain (5 classes) classifier heads. Inputs are 64-bin log-mel spectrograms (FFT size 1024, hop 160 samples = 20 ms at 8 kHz). Training uses random 2-second crops; evaluation uses full sequences with padding masks.

2.2. Cross-Domain Mixup

Standard Mixup [2] interpolates pairs of samples (x_i, y_i) and (x_j, y_j) with a mixing coefficient λ drawn from a $\text{Beta}(\alpha, \alpha)$ distribution: $\tilde{x} = \lambda x_i + (1 - \lambda)x_j$, and similarly for labels $\tilde{y} = \lambda y_i + (1 - \lambda)y_j$. We add a cross-domain constraint: each sample is mixed only with a partner from a different recording domain. Per-sample λ_i values (rather than one per batch) increase mixing diversity. Domain labels are also mixed to provide soft targets. We found $\alpha = 0.2$ optimal: lower values produce gentler interpolations that better preserve individual sample characteristics while still providing domain regularization. Cross-domain Mixup forces the model to infer species identity from features that survive domain mixing, naturally promoting domain invariance at the input level—before domain-specific acoustic signatures are extracted by the CNN branches.

2.3. SpecAugment and Training Details

SpecAugment [3] applies 2 time masks (max 30 frames) and 2 frequency masks (max 15 mel bins) to each spectrogram. Training uses AdamW (lr=1e-3, weight decay=1e-4), batch size 64, for up to 100 epochs with early stopping (min 10 epochs, patience 5). The checkpoint with best validation species balanced accuracy is selected. Training takes approximately 30 minutes on an NVIDIA A800 80GB GPU. No external data or pretrained models were used.

3. EXPERIMENTS

We conducted 17 experiments across six phases, systematically evaluating techniques including domain-adversarial training (DANN) [4], focal loss, manifold Mixup, SWA (Stochastic Weight Averaging), BYOL self-supervised pretraining [5], larger architectures (MTRCNNLarge 1.72M,

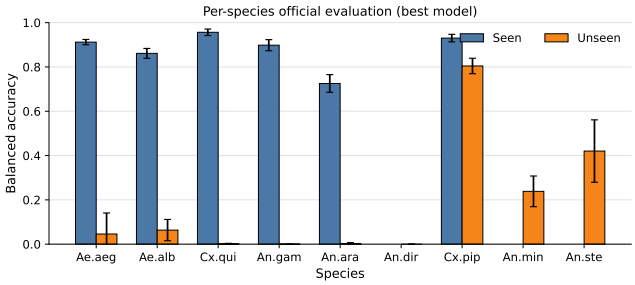

Figure 1: Per-species balanced accuracy on seen (D1/D5) vs. unseen (D2/D3/D4) domains for the best model (Mixup α=0.2, seed 42). Species absent from seen domains (An. dirus, An. minimus, An. stephensi) show zero seen-domain accuracy but substantial unseen-domain performance, demonstrating cross-domain generalization.

ResNet18 11.2M), and multi-seed ensembles. Table 1 presents the complete comparison. * marks techniques that degraded generalization relative to the Mixup-only configuration.

3.1. Key Observations

D3 improvement. The most dramatic improvement is in D3 (Columbia forest): 0.265 → 0.603 (+128%). Cross-domain Mixup appears particularly effective when domain shift involves different environmental acoustics rather than just microphone differences.

D4 remains hard. D4 (200 training clips) shows negligible improvement (0.118 → 0.122). With only ∼80 D4 samples in the training set, even cross-domain Mixup cannot compensate for extreme data scarcity.

Alpha sensitivity. BA_{unseen} follows a clear trend: α=0.2 (0.263) > α=0.3 (0.248) > α=0.25 (0.176) > α=0.5 (0.132). Low alpha creates gentler interpolations that better preserve individual sample identity.

Val Domain BA as diagnostic. Validation domain balanced accuracy inversely correlates with BA_{unseen}. Our best model (BA_{unseen}=0.263) achieved Val Domain BA ≈ 0.55. Manifold Mixup (BA_{unseen}=0.153) had Val Domain BA ≈ 0.72—the high domain discriminability confirms that domain information persists in the features.

4. DISCUSSION

4.1. Why Input-Level Mixup Works When Other Methods Fail

Manifold Mixup is a trap for DG. Manifold Mixup achieved the highest validation BA (0.775) but the worst BA_{unseen} (0.153). Mixing in feature space occurs after the CNN branches extract domain-specific features. Domain information is already entangled with content; mixing post-hoc does not remove domain signatures.

DANN conflicts with learning. Gradient reversal degraded BA_{unseen} (0.248 → 0.169). The adversarial objective likely conflicts with species classification when domain and class are correlated (different species distributions per domain).

BYOL amplifies domain bias. Self-supervised pretraining on 271K clips (97.9% D5) learns D5-specific acoustic features as part of the representation. Val Domain BA increases from 0.55 to 0.73, confirming more domain information, not less. Fine-tuning cannot easily undo this bias.

Larger architectures overfit. MTRCNNLarge and ResNet18 improve BA_{seen} but degrade BA_{unseen} (0.248 → 0.197 and 0.214). More parameters memorize domain-specific patterns rather than learning domain-invariant features.

SWA is mistimed. Best checkpoints (epochs 13–14) occur before SWA averaging starts (epoch 15). SWA averages post-convergence weights, missing the optimal region.

Ensembles dilute the best model. The single seed-42 model (BA_{unseen}=0.263) outperforms a 3-seed ensemble (0.222). Weaker seeds reduce the average, and domain generalization is seed-sensitive due to data loading order determining which cross-domain pairs are formed.

Input-level Mixup destroys domain info early. Cross-domain Mixup operates on raw spectrograms—before any domain-specific feature extraction. By blending D5 and D1–D4 spectrograms, the model cannot rely on domain-specific acoustic signatures. The low Val Domain BA (0.55, close to the 0.5 floor for 5 balanced classes) confirms that features are substantially domain-invariant.

4.2. Limitations

D4 performance (0.122) remains near chance despite Mixup, indicating that extreme data scarcity requires approaches beyond augmentation (e.g., prototypical networks, few-shot learning). Training is unstable: identical configurations with seed 42 produced BA_{unseen} values of 0.263 vs. 0.172 in different runs, likely due to data loading order affecting cross-domain pair formation probabilities. The seen/unseen accuracy trade-off (BA_{seen} drops from 0.881 to 0.766) may be undesirable in deployments where seen-domain performance matters.

4.3. Future Directions

MixStyle [6] (mixing feature statistics across samples after BatchNorm) is a promising next step: it separates domain style from content at the feature level without the information entanglement problem of manifold Mixup. Physically-meaningful audio augmentations (random EQ, bandpass filtering, colored noise, room impulse responses) could better simulate real acoustic domain shifts compared to SpecAugment. Domain-balanced sampling would force equal representation of D1–D5, preventing gradient domination by D5.

5. CONCLUSION

We presented a simple, effective approach for cross-domain mosquito species classification: a lightweight MTRCNN model trained with cross-domain Mixup (α = 0.2) and SpecAugment. Through systematic experimentation (17 configurations, 6 phases), we showed that this combination achieves BA_{unseen} = 0.263 (+69% over baseline), substantially outperforming more complex alternatives. Our key

Table 1: Development test set results. BA_{unseen} on D2/D3/D4, BA_{seen} on D1/D5. * = degraded generalization vs. Mixup-only.

Phase	Model	Val BA	BA_{seen}	BA_{unseen}	DSG	D2	D3	D4
1	Baseline (MTRCNN)	0.852	0.881	0.156	0.726	0.155	0.265	0.118
2	Cross-domain Mixup α=0.2	0.761	0.766	0.263	0.503	0.132	0.603	0.122
2	Cross-domain Mixup α=0.3	0.743	0.768	0.248	0.519	0.324	0.604	0.071
3	+ DANN *	0.740	0.764	0.169	0.595	0.325	0.275	0.125
3	+ DANN + FocalLoss *	0.744	0.774	0.191	0.582	0.342	0.279	0.159
4	MTRCNNLarge (1.72M) *	0.781	0.824	0.197	0.627	0.401	0.266	0.149
4	ResNet18 (11.2M) *	0.726	0.859	0.214	0.644	0.285	0.251	0.037
5	Manifold Mixup α=0.3 *	0.775	0.773	0.153	0.620	0.370	0.613	0.189
5	Cross-domain+class Mixup	0.758	0.761	0.217	0.544	0.364	0.408	0.118
5	Ensemble (3 seeds) *	—	0.786	0.222	0.564	0.054	0.392	0.127
5	Mixup α=0.2 + SWA *	0.759	0.750	0.172	0.578	0.392	0.279	0.091
6	BYOL pretrain + Mixup *	0.744	0.759	0.162	0.597	0.446	0.260	0.044
6	Retrain on train+val *	—	0.760	0.142	0.618	0.270	0.291	0.064

finding is that input-level Mixup is the single most effective domain generalization technique for this task, while manifold Mixup, DANN, focal loss, SWA, BYOL pretraining, and larger architectures all degrade performance. The final submission uses no ensembles, external data, or pretrained models.

6. REFERENCES

[1] Y. Hou et al., "BioDCASE 2026 challenge baseline for cross-domain mosquito species classification," arXiv:2603.20118, 2026.

[2] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in Proc. ICLR, 2018.

[3] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in Proc. Interspeech, 2019.

[4] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," JMLR, vol. 17, no. 1, 2016.

[5] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," in Proc. NeurIPS, 2020.

[6] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with MixStyle," in Proc. ICLR, 2021.