

ADDRESSING CLASS IMBALANCE IN CROSS-DOMAIN MOSQUITO BIOACOUSTIC SPECIES CLASSIFICATION WITH BALANCED MULTI-TASK LEARNING

Technical Report

Meiting Lyu¹, Ruicheng Liu¹, Xu Shen², Zongye Chen¹, Jiatong Wu¹
Ziyu Wang¹, Wei Liu¹, Gongping Huang¹

¹ Wuhan University, Electronic Information School, Wuhan, China

² Waseda University, Japan

Email: lmt041109@gmail.com (M. Lyu)

ABSTRACT

The BioDCASE 2026 Cross-Domain Mosquito Species Classification (CD-MSC) challenge presents two difficulties: domain shift between recording environments, and severe class imbalance caused by unequal species abundance in field collections. To this end, we propose a multi-task system that combines balanced data augmentation with domain-robust representation learning. We mitigate species imbalance via class-balanced sampling and batch augmentation before training. Within the network, MixStyle modules generate cross-domain features after pooling and transition layers to alleviate domain shift. A Squeeze-and-Excitation block following the DenseNet backbone further refines channel feature responses. The shared embedding feeds three learning targets and a domain-adversarial head trained with gradient reversal. At inference time, predictions from complementary model variants are combined through a confidence-gated cascade fusion strategy. Through these combined techniques, our system achieves a substantial improvement in classification accuracy on unseen recording domains compared to the official baseline. Our system is submitted to the BioDCASE 2026 Task 5 challenge.

Index Terms— mosquito species classification, class imbalance, domain-adversarial training, multi-task learning, bioacoustics

1. INTRODUCTION

Bioacoustic mosquito species classification (MSC) recognizes mosquito species using wingbeat audio, delivering a scalable, non-invasive approach for vector monitoring and disease prevention [1] [2] [3]. The BioDCASE 2026 Cross-Domain Mosquito Species Classification (CD-MSC) task targets this practical scenario, which demands strong cross-domain generalization across diverse field recording domains. Variations introduced by distinct field locations, recording hardware and ambient noise create domain shifts that obscure mosquito bioacoustic features, causing vanilla classification models to overfit domain-specific artifacts rather than discriminative wingbeat features [4].

Domain adversarial training (DAT) [5] is widely adopted to learn domain-invariant representations. In practice, mosquito field collections are markedly imbalanced: common species such as *Aedes albopictus* contribute orders of magnitude more clips than rare *Anopheles* species that are of high epidemiological interest. Training without correcting for this skew produces classifiers that

almost exclusively predict frequent classes, and applying gradient reversal indiscriminately to rare-species samples can erode their species-discriminative features.

The contribution of this paper is a system that jointly addresses domain shift and class imbalance through four complementary components:

1. *Imbalance-aware data preparation.* Class-balanced batch sampling, short-clip concatenation, and targeted spectral augmentation reshape the input distribution and concentrate regularisation on under-represented species.
2. *A domain-generalisable backbone.* DenseNet121 backbone is extended with MixStyle-based feature statistic mixing and Squeeze-and-Excitation recalibration, which suppress domain-specific cues while preserving species-discriminative spectral patterns.
3. *A multi-task learning objective.* We adopt a multi-task learning objective that jointly optimizes class-balanced focal loss, CosFace metric learning, a genus-level auxiliary task, domain-adversarial training with gradient reversal, and supervised contrastive learning on a shared embedding.
4. *Complementary model fusion.* At inference, a confidence-gated cascade combines model variants that differ in architecture, training seed, and spectral resolution, further improving generalisation to unseen domains.

2. SYSTEM DESCRIPTION

The system processes each audio clip through the following stages in order: feature extraction, augmentation, backbone encoding, multi-head classification, and cascade ensemble fusion. Figure 1 illustrates the proposed framework.

2.1. Feature Extraction

Each recording is resampled to 8 kHz mono, peak-normalised, and converted to a log-mel spectrogram (512-sample Hann window, 80-sample hop, 128 mel bands covering 200–4000 Hz). The 200 Hz cutoff discards the low-frequency hum that sits below the wingbeat range without losing any of the harmonics that carry species information. The resulting spectrogram $\mathbf{X} \in \mathbb{R}^{T \times 128}$ is then treated as a single-channel image and passed through three simple clean-up steps.

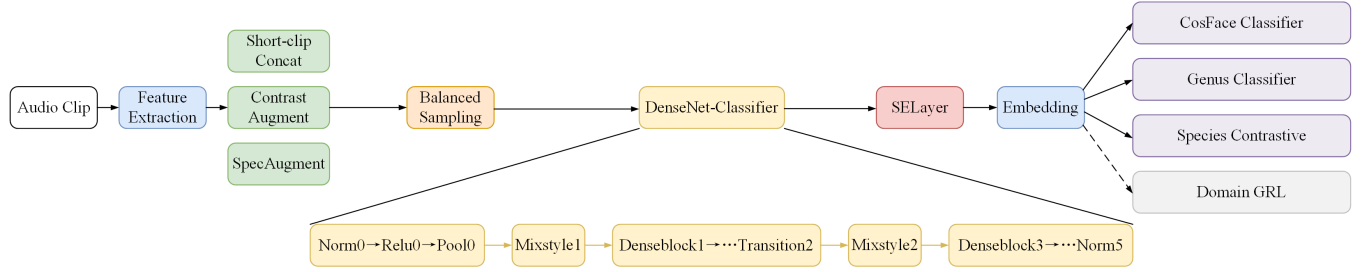


Figure 1: Overview of the proposed framework. Data flows from feature extraction through augmentation and balanced sampling into the DenseNet121 encoder, which is followed by SE recalibration and four task heads.

First, we clip all amplitudes above the 99.5th percentile, eliminating the bias caused by sporadic loud impulsive noises against the statistical distribution of quiet audio segments. Second, with a 0.5 probability during training, we estimate and subtract the noise floor of each mel band. Each recording site features distinct constant background hum; eliminating this site-specific hum prevents the model from relying on background noise as a shortcut to infer the recording domain instead of identifying the mosquito species responsible for the audio clip. Third, clips longer than 200 frames (2 s) are cropped to a fixed 200-frame window. We sample the window at random offsets in training for lightweight temporal augmentation. Validation and test clips only undergo peak clipping with no extra processing. This setup preserves original recording conditions for credible evaluation.

2.2. Imbalance and Domain-Aware Augmentation

Beyond the generic normalisation above, we apply three *training-only* augmentations, each purposefully designed to counter a distinct symptom of the class-imbalance and domain-shift problems introduced in Section 1:

Short-clip concatenation Rare species are disproportionately represented by very short recordings, which both waste padding capacity and starve the encoder of temporal context. Clips shorter than 65 frames are concatenated along the time axis with a randomly selected clip of the *same* species (the original clip is kept unchanged if no candidate exists). This doubles the effective duration of short rare-species clips, reduces zero-padding in mini-batches, and increases intra-class variability without introducing cross-species label noise.

Contrast augmentation The whole spectrogram is scaled by a factor drawn uniformly from $[0.5, 1.5]$, emulating the amplitude variability introduced by different microphones, gain settings, and source distances across recording sites. By desensitising the encoder to overall loudness, this augmentation acts at the input level in concert with MixStyle and domain-adversarial training (Section 2.4), which pursue the same domain-invariance goal inside the network.

SpecAugment One time mask (up to 10 frames) and one frequency mask (up to 16 bins) are applied at randomly chosen positions. For the three most under-represented *Anopheles* species (*An. dirus*, *An. minimus*, *An. stephensi*; species indices 5, 7, 8), both mask counts are doubled, concentrating the strongest regularisation exactly where the model is most prone to memorising spurious cues from the smallest training pools.

The three augmentations are applied sequentially, in the order listed above, immediately after feature extraction and before mini-

batches are assembled; validation and test data bypass this stage entirely.

2.3. Class-Balanced Sampling

The training set remains heavily imbalanced across the nine mosquito species even after the augmentations above. As the final step of data preparation, every training sample is assigned a sampling weight using the effective-number formulation [6]:

$$w_c = \frac{1 - \beta}{1 - \beta^{n_c}}, \quad \beta = 0.999, \quad (1)$$

where n_c is the number of training samples for species c .

2.4. Model Architecture

DenseNet121 backbone. The feature tensor is treated as a single-channel image $(B, 1, T, 128)$ and passed to a DenseNet121 [7] pretrained on ImageNet. The first convolutional layer is adapted from three-channel RGB to single-channel by summing the pretrained weights along the channel dimension, preserving the magnitude of the initial feature response. A batch normalisation layer operating on the frequency axis is prepended to standardise the input distribution before the backbone.

MixStyle domain augmentation. To encourage domain-invariant feature learning, MixStyle modules [8] are inserted at two locations inside the DenseNet: after the first pooling stage (`pool0`) and after the second transition block (`transition2`). During training, each module randomly re-styles the instance-level mean and variance of the feature map by mixing them with those of a randomly permuted sample in the batch:

$$\hat{\mathbf{x}} = \sigma_{\text{mix}} \cdot \frac{\mathbf{x} - \mu(\mathbf{x})}{\sigma(\mathbf{x})} + \mu_{\text{mix}}, \quad (2)$$

where μ_{mix} and σ_{mix} are convex combinations (with $\lambda \sim \text{Beta}(1, 1)$) of the statistics of the original and the permuted sample.

Squeeze-and-Excitation layer. After the DenseNet dense block, a Squeeze-and-Excitation module [9] with channel reduction ratio 16 re-weights each output channel by a scalar gate learned from global average pooling.

Embedding projection. The pooled feature is compressed to a 128-dimensional embedding through a two-layer MLP: Linear $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ Dropout(0.5) $\rightarrow$ Linear $\rightarrow$ BN, followed by L2 normalisation. This shared embedding is consumed by three independent heads.

Species head (CosFace). During training, the cosine similarity between the L2-normalised embedding and the L2-normalised class weight of the ground-truth species is reduced by the margin m before scaling [10]. Let θ_c denote the angle between the L2-normalised embedding and the L2-normalised weight of class c , and let y be the ground-truth class index. The logit for class c is then computed as:

$$\text{logit}_c = \begin{cases} s(\cos \theta_c - m) & c = y \\ s \cos \theta_c & c \neq y. \end{cases} \quad (3)$$

where s is the scale factor (set to 30) and m is the margin (set to 0.2). This formulation enlarges the angular decision boundary between adjacent species. At inference time, the margin is not applied.

Genus head. A linear layer maps the embedding to three genus logits (*Aedes*, *Culex*, *Anopheles*). The genus head acts as a coarse-grained auxiliary task, sharing informative phylogenetic structure across species within the same genus and providing a taxonomically consistent training signal.

Domain head with gradient reversal. A two-layer MLP classifies the recording domain (D1–D5). A gradient reversal layer (GRL) [5] with coefficient $\alpha=1.0$ is placed between the embedding and the domain MLP. α controls the strength of the gradient reversal (positive values fully reverse the domain gradient). During back-propagation, the domain gradient is negated before reaching the shared backbone, incentivising the encoder to discard domain-identifying cues. To protect rare species whose species-discriminative features have not yet been adequately learned, a *selective mask* zeroes the domain loss contribution of samples from *Anopheles dirus* (the least-represented species, index 5):

$$\mathcal{L}_{\text{domain}} = \frac{\sum_i m_i \ell_i^d}{\sum_i m_i}, \quad m_i = \mathbf{1}_{\{\text{species}_i \neq 5\}}. \quad (4)$$

2.5. Multi-Task Training Objective

The total loss is a weighted sum of four terms:

$$\mathcal{L} = \alpha_s \mathcal{L}_{\text{species}} + w_g \mathcal{L}_{\text{genus}} + \lambda_d \mathcal{L}_{\text{domain}} + \alpha_c \mathcal{L}_{\text{contrast}}, \quad (5)$$

with fixed weights $\alpha_s=1.0$, $\lambda_d=0.5$, $\alpha_c=0.5$.

Species loss. Focal loss [11] ($\gamma=2.0$, label smoothing $\epsilon=0.1$) is applied to the CosFace logits. Each sample’s contribution is additionally scaled by the same effective-number class weight w_c defined in (1).

Genus loss. Cross-entropy with label smoothing ($\epsilon=0.1$) on the genus head. Its weight w_g decays linearly from 1.0 to 0.5 over the first 50 epochs and is clamped at 0.5 thereafter:

$$w_g = \max\left(0.5, 1.0 - \frac{e - 1}{0.5 T}\right), \quad (6)$$

where e is the current epoch and T the total number of epochs.

Supervised contrastive loss. Following [12], a supervised contrastive loss with temperature $\tau=0.05$ is applied on the L2-normalised embeddings, pulling same-species embeddings together and pushing different-species embeddings apart in hyperspherical space.

2.6. Inference

Confidence-Gated Cascade Fusion. At inference time we combine three complementary models through a three-round cascade fusion strategy [13]. The three models differ in architecture, training seed, and Mel filter-bank resolution: finer resolution (128–256 bins) and coarser resolution (64 bins).

Round 1 (high precision, conservative). The designated model’s prediction is accepted only if its softmax confidence exceeds a species-specific threshold $\tau_s \in [0.5, 0.6]$, retaining only trustworthy predictions.

Round 2 (balanced). Unresolved samples proceed to a second designated model, whose prediction is accepted unconditionally, striking a trade-off between precision and recall.

Round 3 (high recall, fallback). Remaining samples are resolved by a third designated model without any confidence constraint, ensuring full coverage. For species on which all models perform poorly (e.g., *Anopheles dirus*), the first two rounds are skipped entirely.

When multiple candidates within the same round disagree, ties are broken in favour of the model with the highest per-species precision on the held-out validation data.

3. EXPERIMENTAL SETUP

Dataset. We use the BioDCASE 2026 Task 5 development data comprising nine species across five recording domains (D1–D5): *Aedes aegypti*, *Ae. albopictus*, *Culex quinquefasciatus*, *Anopheles gambiae*, *An. arabiensis*, *An. dirus*, *Cx. pipiens*, *An. minimus*, and *An. stephensi*. We follow the official train/validation split.

Training. AdamW optimizer [14] (learning rate 10^{-3} , weight decay 10^{-4}), batch size 64, up to 100 epochs. ReduceLROnPlateau reduces the learning rate by factor 0.5 after 6 consecutive epochs without improvement in validation species balanced accuracy (BA), with a minimum lr of 10^{-6} . Early stopping is applied with a minimum of 15 epochs and patience 10. Training uses a single NVIDIA RTX 3090 GPU.

Metrics. We report species-level balanced accuracy (macro-average recall) overall and per domain, as well as BA on seen versus unseen domains and the Domain-Species Gap (DSG) metric defined by the challenge.

4. RESULTS

4.1. Backbone Architecture Comparison

We compared seven CNN backbones under baseline training (cross-entropy only, no class-balance, no domain adaptation). Table 1 shows DenseNet121 achieves the highest seen-domain BA and best unseen-domain BA, justifying its selection.

4.2. Final System Performance

Table 2 reports the performance of our single model (model_final.pth, seed 2048). The single model generalizes well to seen domains (67.61% BA) but degrades on unseen domain D4 (11.49% BA).

By combining three complementary models through our confidence-gated cascade fusion (Section 2.6), we achieve substantial gains across all metrics, as shown in Table 3. The ensemble improves test-set overall balanced accuracy from 51.59% to 56.02%

Table 1: Cross-domain performance of different CNN backbones (baseline setup). Values are balanced accuracy (BA) and Domain-Species Gap (DSG) on test split, mean $\pm$ std over ten seeds.

Backbone	BA _{seen}	BA _{unseen}	DSG
MTRCNN	0.879 $\pm$ 0.010	0.185 $\pm$ 0.014	0.694 $\pm$ 0.019
ResNet	0.903 $\pm$ 0.005	0.171 $\pm$ 0.035	0.732 $\pm$ 0.036
Inception V3	0.863 $\pm$ 0.009	0.187 $\pm$ 0.013	0.676 $\pm$ 0.015
MobileNet	0.826 $\pm$ 0.232	0.134 $\pm$ 0.030	0.693 $\pm$ 0.226
Xception	0.888 $\pm$ 0.015	0.149 $\pm$ 0.032	0.739 $\pm$ 0.033
EfficientNet	0.902 $\pm$ 0.008	0.113 $\pm$ 0.034	0.789 $\pm$ 0.029
DenseNet121	0.912 $\pm$ 0.008	0.194 $\pm$ 0.017	0.718 $\pm$ 0.019

Table 2: Single-model performance on validation and test splits.

Metric	Validation	Test
Species Accuracy	70.51 %	62.19 %
Species BA (overall)	74.95 %	51.59 %
BA – D1	85.12 %	33.37 %
BA – D2	100.0 %	73.90 %
BA – D3	77.78 %	60.74 %
BA – D4	50.00 %	11.49 %
BA – D5	67.40 %	67.62 %
BA – seen domains	—	67.61 %
BA – unseen domains	—	35.33 %
DSG	—	0.323

and unseen-domain BA from 35.33% to 40.16%, while reducing the Domain-Species Gap (DSG) from 0.323 to 0.299.

5. CONCLUSION

Our system significantly outperforms the official baseline on the BioDCASE 2026 cross-domain mosquito species classification task. The balanced multi-task learning framework effectively mitigates severe class imbalance without sacrificing domain invariance, achieving substantial gains on under-represented species.

6. REFERENCES

- [1] I. Kiskin, M. Sinka, A. D. Cobb, W. Rafique, L. Palmer-Young, R. Odonko, K.-L. Osei, M. wa Kyambadde, B. Torr, and S. J. Roberts, “Humbugdb: A large-scale acoustic mosquito dataset,” in *Proc. NeurIPS Workshop on Machine Learning for Public Health*, 2020.
- [2] L. Torres-Sorando and D. J. Rodriguez, “Models of spatio-temporal dynamics in malaria,” *Ecological modelling*, vol. 104, no. 2-3, pp. 231–240, 1997.
- [3] M. B. Meerwijk, “Phantom menace: dengue and yellow fever in asia,” *Bulletin of the History of Medicine*, vol. 94, no. 2, pp. 215–243, 2020.
- [4] Y. Hou, Z. Liu, X. Shen, and S. Roberts, “Learning domain-robust bioacoustic representations for mosquito species classification with contrastive learning and distribution alignment,” in *2026 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2026, pp. 15 207–15 211.
- [5] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-

Table 3: Ensemble system performance on validation and test splits.

Metric	Validation	Test
Species Accuracy	71.07 %	63.61 %
Species BA (overall)	74.57 %	56.02 %
BA – D1	78.87 %	42.21 %
BA – D2	72.16 %	71.35 %
BA – D3	73.02 %	58.86 %
BA – D4	50.00 %	11.82 %
BA – D5	69.90 %	70.03 %
BA – seen domains	—	70.03 %
BA – unseen domains	—	40.16 %
DSG	—	0.299

adversarial training of neural networks,” in *J. Mach. Learn. Res.*, vol. 17, 2016, pp. 1–35.

- [6] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, “Class-balanced loss based on effective number of samples,” in *Proc. IEEE/CVF CVPR*, 2019, pp. 9268–9277.
- [7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE/CVF CVPR*, 2017, pp. 4700–4708.
- [8] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, “Domain generalization with MixStyle,” in *Proc. ICLR*, 2021.
- [9] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE/CVF CVPR*, 2018, pp. 7132–7141.
- [10] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “CosFace: Large margin cosine loss for deep face recognition,” in *Proc. IEEE/CVF CVPR*, 2018, pp. 5265–5274.
- [11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proc. IEEE/CVF ICCV*, 2017, pp. 2980–2988.
- [12] P. Khosla, P. Tian, C. Wang, A. Sohl-Dickstein, D. Krishnan, A. Courville, and S. Bengio, “Supervised contrastive learning,” in *Proc. NeurIPS*, vol. 33, 2020, pp. 18 661–18 673.
- [13] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proc. IEEE/CVF CVPR*, 2001, pp. 511–518.
- [14] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. ICLR*, 2019.