MOSQUITO SPECIES CLASSIFICATION VIA DOMAIN-ADVERSARIAL FINE-TUNING OF THE AUDIO SPECTROGRAM TRANSFORMER

Technical Report

Eva Boneš and Matija Marolt

University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, 1000 Ljubljana, Slovenia
{eva.bones},{matija.marolt}@fri.uni-lj.si

ABSTRACT

We describe our submissions to the BioDCASE 2026 Cross-Domain Mosquito Species Classification task (Task 5) [1]. We explore two distinct paradigms to address severe domain shift across five recording setups under a Leave-One-Domain-Out protocol. Our first system is a lightweight baseline that uses probabilistic YIN (pYIN) fundamental frequency (f_0) summary statistics paired with a Random Forest classifier. Our second system fine-tunes a pre-trained Audio Spectrogram Transformer (AST) regularized via a Domain-Adversarial Neural Network (DANN) framework to enforce domain-invariant representations. On the official challenge split, the pitch baseline yields a narrower Domain Shift Gap (DSG $= 0.296$, $BA_{\text{unseen}} = 0.241$), whereas the high-capacity DANN-AST model yields a higher cross-domain accuracy but a wider gap ($BA_{\text{unseen}} = 0.314$, DSG $= 0.460$).

***Index Terms*—** Mosquito classification, BioDCASE, Domain Adaptation, Audio Spectrogram Transformer, Probabilistic YIN

1. INTRODUCTION

BioDCASE 2026 Task 5 requires classifying mosquito species from wingbeat recordings under a Leave-One-Domain-Out (LODO) evaluation protocol: for each species, one recording domain is held out entirely from training. The primary metric is balanced accuracy on those unseen domains (BA_{unseen}). The challenge thus explicitly tests cross-domain generalisation, not merely standard classification accuracy.

To address this challenge [1], we developed two separate systems representing different trade-offs between representation capacity and domain robustness. System 1 relies on tightly constrained feature engineering via robust pitch tracking and high-confidence statistical filtering. System 2 leverages a massive pre-trained Audio Spectrogram Transformer (AST) [2] backbone with a domain adversary connected through a Gradient Reversal Layer (GRL) [3] to explicitly suppress domain-specific deep features.

2. SYSTEM 1: PITCH-STATISTICS BASELINE

Our first system is a deliberately minimal baseline designed to establish the classification performance attainable from fundamental frequency (f_0) tracking alone.

Given an audio recording, the signal is resampled to 44.1 kHz and passed through a 4th-order Butterworth band-pass filter (300–3800 Hz) to isolate the mosquito wingbeat band. We then apply Harmonic–Percussive Source Separation (HPSS), retaining exclusively the harmonic component to suppress broadband environmental noise. We utilize the probabilistic YIN (pYIN) algorithm with a frame length of 1024 and a hop length of 256 ($\approx 5.8\,ms$) to extract a per-frame f_0 track within the bounds of $[300, 1100]$ Hz. The pYIN hyper-parameters were tuned on a 90-file annotated subset setting $\beta = (1, 4)$, boltzmann $= 4$, switch_prob $= 0.1$, and no_trough_prob $= 0.05$.

To isolate clean signals, we retain voiced, in-band frames (voicing probability > 0.15, 300–1100 Hz) and apply a confidence filter keeping only the top 20% frames by voicing probability. This removes low-confidence frames where domain noise and octave errors tend to concentrate. From these filtered frames, we compute six summary statistics: median, mean, standard deviation, minimum, maximum, and interquartile range. Temporal attributes such as signal duration or count of voiced regions are intentionally excluded, as they correlate strongly with specific recording setups rather than species biology and would degrade domain transfer.

A Random Forest classifier consisting of 500 trees maps the six summary features to a single species label per file. The model was trained on the combined training and validation data. To counter the severe class imbalance present in the dataset (where domains are heavily dominated by three species, and three minority species represent $< 1\%$ of samples), we utilized balanced class weighting. Files yielding zero voiced frames default automatically to the majority class prediction.

3. SYSTEM 2: DOMAIN-ADVERSARIAL AST

Our second submission consists of an end-to-end deep learning framework utilizing the `MIT/ast-finetuned-audioset-10-10-0.4593` backbone from HuggingFace, which is an AST pre-trained on AudioSet [4] with a hidden size of 768.

3.1. Architecture and Classification Heads

All audio is converted to 16 kHz mono, log-mel spectrograms (128 mel bins, 10.24 s clips, matching the AST's expected input format), and normalised with the dataset statistics used during AST pre-training. Rather than using `pooler_output`, we mean-pool *all* token outputs of the AST's last hidden state (patch tokens + Classification Token), which consistently provided richer representations in early tests.

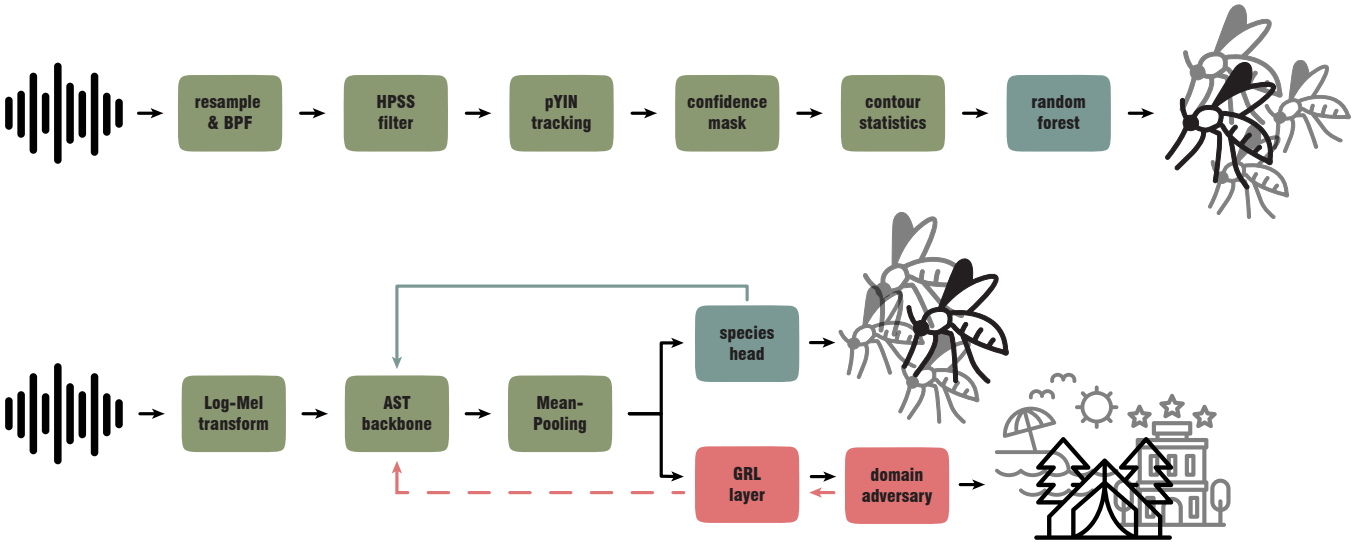

Figure 1: Architectural comparison of the two proposed systems. System 1 (top) utilizes a deterministic feature-engineered pipeline based on fundamental frequency (f_0) summary contours to achieve domain robustness. System 2 (bottom) employs an end-to-end ASTs backbone with parallel task heads. Solid arrows denote the forward propagation of features. Colored return loops denote backpropagation paths: the solid teal line represents the target species gradient, whereas the dashed coral line explicitly denotes the domain-adversarial gradient inversion imposed by the GRLs.

Two individual heads branch off from this pooled representation:

- **Species head**: a shallow MLP $768 \rightarrow 256$ (GELU, Dropout 0.3) $\rightarrow 9$ classes.
- **Domain adversary**: a deeper MLP $768 \rightarrow 512 \rightarrow 256 \rightarrow 5$ domains (GELU, Dropout 0.6 per layer), preceded by a GRL. The adversary is made deeper to ensure it can competitively penalise the backbone for leaking domain traits.

3.2. DANN Objective and Optimization

The combined loss for System 2 is formulated as:

$$\mathcal{L} = w_s \cdot \mathcal{L}_{\text{species}} + w_d \cdot \mathcal{L}_{\text{domain}}$$

where $\mathcal{L}_{\text{species}}$ is cross-entropy on the 9 species (label smoothing $\epsilon = 0.1$) and $\mathcal{L}_{\text{domain}}$ is cross-entropy on the 5 domains. The GRL negates the domain gradient before it reaches the backbone, punishing the network for maintaining domain distinctness.

The GRL strength α follows a sigmoid ramp [3]:

$$\alpha(p) = \frac{2}{1 + e^{-5\,p}} - 1, \quad p = \frac{t - t_{\text{warmup}}}{T - t_{\text{warmup}}}$$

where t is the current epoch, $T = 30$ total epochs, and $t_{\text{warmup}} = 3$. We set loss weights $w_s = 3.0$ and $w_d = 1.0$, with a batch size of 16. Optimization used AdamW (weight decay $\lambda = 10^{-2}$) with differential learning rates (backbone LR $= 2.5 \times 10^{-5}$; classification heads LR $= 5 \times 10^{-5}$) and a per-group linear warm-up followed by cosine decay. Max gradient clipping norm was set to 1.0.

Data augmentation applied to System 2 training included SpecAugment [5] (two frequency masks, $F \sim \mathcal{U}[0, 16]$; two time masks, $T \sim \mathcal{U}[0, 32]$), log-mel gain jitter ($\Delta \sim \mathcal{U}[-0.3, 0.3]$), and circular random time shifts up to 12.5% of the clip duration.

4. RESULTS AND DISCUSSION

Both models were evaluated on the held-out test split (`Test_ids`). Table 1 summarizes the aggregate performance metrics and the resulting Domain Shift Gap (DSG), while Table 2 provides a detailed per-species breakdown for both submissions.

Table 1: Aggregate metrics on `Test_ids` for both proposed submissions.

Approach	BA_{seen}	BA_{unseen}	DSG
System 1: Pitch Baseline	0.538	0.241	**0.296**
System 2: Domain-Adversarial AST	0.774	**0.314**	0.460

Table 2: Per-species unseen-domain balanced accuracy and relative split proportions.

Species	Domain	Split %	System 1 (Pitch)	System 2 (DANN-AST)
Aedes aegypti	D3	5.0%	0.219	0.235
Aedes albopictus	D2	9.9%	0.200	0.731
Culex quinquefasciatus	D1	16.0%	0.326	0.020
Anopheles gambiae	D1	19.4%	0.214	0.211
Anopheles arabiensis	D1	43.2%	0.123	0.030
Anopheles dirus	D4	1.1%	0.050	0.000
Culex pipiens	D3	1.6%	0.706	1.000
Anopheles minimus	D2	2.4%	0.293	0.000
Anopheles stephensi	D4	1.4%	0.041	0.600
Mean		**100.0%**	**0.241**	**0.314**

System 1 isolates the discriminative capacity of fundamental frequency (f_0) tracking from broader acoustic textures. Feature ablation experiments showed that replacing the top-20% voicing con-

fidence filter with the full, unconstrained voiced track improved in-domain performance ($BA_{\text{seen}} = 0.583$) but degraded cross-domain transfer ($BA_{\text{unseen}} = 0.225$, DSG $= 0.358$). This performance drop confirms that low-confidence tracking regions are highly susceptible to encoding domain-specific noise and environmental artifacts. Restricting the feature space to high-confidence voiced windows effectively acts as an implicit domain regularizer. An error analysis indicates that the remaining cross-domain misclassifications in System 1 stem from biological overlaps in inter-species f_0 distributions rather than pitch estimation failures, as the underlying tracking achieved a detection rate exceeding 99%.

System 2 achieves a higher absolute cross-domain average ($BA_{\text{unseen}} = 0.314$) due to the strong representational capacity of the pre-trained attention layers, but it is substantially more prone to domain overfitting (DSG $= 0.460$).

The per-species breakdown highlights a clear trade-off between the two approaches. System 1 establishes a more robust performance floor under severe acoustic shifts. For instance, on domain D1, the pitch baseline achieves a recall of 0.326 for *Culex quinquefasciatus* and 0.123 for *Anopheles arabiensis*, outperforming System 2, which drops to 0.020 and 0.030, respectively. Explicitly tracking f_0 bypasses the background acoustic variations that degrade the deep network's learned features.

On the other hand, System 2 excels on rare classes when encountered in cleaner target domains, such as *Anopheles stephensi* in D4 ($BA_{\text{unseen}} = 0.600$ vs. 0.041) and *Aedes albopictus* in D2 ($BA_{\text{unseen}} = 0.731$ vs. 0.200). When background noise does not overlap with target signal characteristics, the deep model's broader spectro-temporal context scales much better than raw pitch summaries. However, regulating System 2 during training was critical; after reaching peak validation performance at epoch 7, the model began overfitting to the source domains as the backbone re-encoded domain-specific traits faster than the GRL adversary could suppress them.

5. SUBMISSIONS

We submit two distinct predictions to the official BioDCASE 2026 Cross-Domain Mosquito Species Classification challenge:

- **Submission 1**: Predictions derived from our pitch summary statistical model utilizing the Random Forest classifier detailed in Section 2.
- **Submission 2**: Predictions generated via our end-to-end Domain-Adversarial Audio Spectrogram Transformer architecture detailed in Section 3.

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] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proceedings of Interspeech*, 2021, pp. 571–575.

[3] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. s. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research (JMLR)*, vol. 17, no. 1, pp. 2096–2130, 2016.

[4] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[5] 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 *Proceedings of Interspeech*, 2019, pp. 2613–2617.