DOMAIN-AWARE AUDIO FUSION FOR BIODCASE 2026 CROSS-DOMAIN MOSQUITO SPECIES CLASSIFICATION

Technical Report

Haoyang Zhou

University of Bremen
Computer Science undergraduate program
Bremen, Germany
zhou1@uni-bremen.de; backup: karlzhou001@163.com

ABSTRACT

This technical report describes four submitted systems for the BioDCASE 2026 Task 5 Cross-Domain Mosquito Species Classification challenge. The systems address closed-set mosquito species recognition under acquisition-domain shift. Starting from the released CD-MSC MTRCNN baseline, we use domain-aware model selection, long-recording sliding-window inference, FreqMix/DIR-like robust training, and probability fusion with public audio-pretrained representations. The best submitted development-test system fuses a domain-bucket MTRCNN ensemble with PANNs CNN14 and AST AudioSet representations, reaching 0.6982 balanced accuracy and 0.5636 unseen-domain balanced accuracy on the released development test split. No external labelled mosquito data, model checkpoints, or evaluation-set labels are used.

Index Terms— BioDCASE 2026, mosquito species classification, domain generalisation, audio classification, model fusion

1. INTRODUCTION

The BioDCASE 2026 Cross-Domain Mosquito Species Classification task evaluates mosquito species recognition under domain shift. The official ranking is primarily based on balanced accuracy on unseen-domain clips, while the domain shift gap is used as a secondary metric. This makes the task different from ordinary in-domain audio classification: a system that recognises frequent training-domain clips can still perform poorly if it fails on species-domain combinations not observed during training.

The submitted systems are built on top of the released CD-MSC baseline repository and development dataset. All systems predict one of the nine closed-set species for each evaluation clip. The evaluation filenames are treated as sample identifiers.

2. DATA AND DEVELOPMENT PROTOCOL

The development dataset contains five acquisition domains and nine mosquito species with strong class and domain imbalance. The official development test split contains both seen-domain and unseen-domain samples for species-specific domain generalisation analysis. We report the following development-test metrics for every submitted system: overall balanced accuracy (BA), seen-domain balanced accuracy (BA_{seen}), unseen-domain balanced accuracy (BA_{unseen}), and the domain shift gap

$$DSG = |BA_{\mathrm{unseen}} - BA_{\mathrm{seen}}|. \quad (1)$$

The system selection criteria were based on the released development split and additional domain-aware local analysis. Since the final official ranking emphasises unseen-domain performance, one submitted system is selected for maximum development BA_{unseen}, while another is selected for best overall development BA.

3. SYSTEMS

3.1. Base acoustic model

The neural acoustic backbone follows the released MTRCNN baseline: 8 kHz waveform input, 64-bin log-mel spectrogram features, and a lightweight multitemporal convolutional architecture with a species head and an auxiliary domain head. We trained and evaluated variants using domain-aware sampling and robust augmentation. The strongest MTRCNN variant used FreqMix-style spectral mixing and DIR-like domain robustness during training.

3.2. Domain-bucket inference

The first submitted system, `Zhou_UB_task5_1`, uses a domain-bucket ensemble. Predictions from the FreqMix/DIR-like MTRCNN model are routed through domain-conditioned development-selected weights. This system is conservative and keeps the prediction distribution close to the MTRCNN family.

3.3. Sliding-window aggregation

The second submitted system, `Zhou_UB_task5_2`, applies long-recording sliding-window inference. Per-window species probabilities are averaged to produce clip-level predictions. A lightweight tabular domain classifier trained on log-mel summary statistics is then used for domain-conditional probability adjustment. This system mainly targets the mismatch between short training clips and longer or noisier evaluation recordings.

3.4. Audio-pretrained fusion

The third and fourth submitted systems fuse the domain-bucket MTRCNN output with public audio-pretrained representations. We used PANNs CNN14 and AST AudioSet-based models as complementary acoustic descriptors. The fusion is a weighted probability average, where weights are selected on the development test split:

Table 1: Development-test metrics for the submitted systems.

System	BA	BA_{seen}	BA_{unseen}	DSG
Zhou_UB_task5_1	0.6548	0.8553	0.5205	0.3348
Zhou_UB_task5_2	0.6179	0.8151	0.5107	0.3045
Zhou_UB_task5_3	0.6982	0.8546	0.5636	0.2910
Zhou_UB_task5_4	0.7001	0.8536	0.5565	0.2971

- `Zhou_UB_task5_3`: base = 0.38, PANNs = 0.62, AST = 0.00, selected for best development BA_{unseen}.
- `Zhou_UB_task5_4`: base = 0.32, PANNs = 0.56, AST = 0.12, selected for best development BA.

The motivation is that public audio-pretrained models provide broader acoustic invariances than the small mosquito-specific training set alone, while the domain-bucket MTRCNN retains task-specific species structure.

4. RESULTS

Table 1 reports development-test performance for the four submitted systems. The PANNs/AST fusion systems provide the strongest unseen-domain scores among the packaged submissions.

The results show a clear gap between the original MTRCNN-style domain-bucket system and the audio-pretrained fusion systems. The best unseen-domain system improves development BA_{unseen} from 0.5205 to 0.5636. The best overall system achieves 0.7001 BA while retaining competitive unseen-domain performance.

5. SUBMISSION DETAILS

The final submission package contains four TXT prediction files, one for each system, plus corresponding YAML meta-information files. Every TXT file follows the Task 5 format `file_id,predicted_species_id` and contains one prediction for each released evaluation clip. Public pretrained audio representations are used only as generic acoustic feature sources. No external labelled mosquito recordings are used.

6. CONCLUSION

The submitted systems focus on domain generalisation rather than in-domain accuracy alone. The strongest approach combines task-specific domain-aware MTRCNN predictions with audio-pretrained models through development-selected probability fusion. This simple fusion gives the best local unseen-domain performance while remaining lightweight and reproducible.

7. REFERENCES

[1] BioDCASE 2026 Task 5: Cross-Domain Mosquito Species Classification. [Online]. Available: https://biodcase.github.io/challenge2026/task5

[2] Y. Hou et al., "BioDCASE 2026 CD-MSC Baseline," 2026. [Online]. Available: https://github.com/Yuanbo2020/CD-MSC

[3] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[4] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech*, 2021.