WingBeats and Snapshots: Fusing Sound and Vision for Mosquito Monitoring (Student Abstract)

Ahana Chanda, Akshay Agarwal

Trustworthy BiometraVision Lab, IISER Bhopal, India
ahanachanda21@gmail.com, akagarwal@iiserb.ac.in

Abstract

Accurate identification of mosquito species is crucial for controlling vector-borne diseases, yet visual or acoustic methods alone are often insufficient. We propose a multimodal deep-learning framework that combines high-resolution images with wingbeat audio using a SwinV2 vision transformer and an Audio Spectrogram Transformer, thereby capturing complementary cues. On a six-species dataset, it achieves 97% accuracy, comparable to the best single-modality baseline, and is designed to improve robustness under noise or environmental variation, demonstrating the value of integrating multiple data sources for reliable mosquito surveillance.

Introduction

Mosquito surveillance is crucial for controlling vector-borne diseases, but species identification is still unnecessarily difficult (Al Maruf et al. 2025; Hagiwara et al. 2022). Vision-based ID is slow and error-prone, while wingbeat acoustics are fast but fragile in the presence of noise and overlapping frequencies. Single-modality deep-learning models excel on clean data but fail when faced with poor lighting, damaged samples, or noisy audio.

Multimodal approaches exist, but few genuinely fuse both image and acoustic signals into a single model. We present a unified framework using a vision transformer for images and AST for audio, learning shared representations that single-modality models overlook. This enhances robustness, particularly in the presence of noise, making it suitable for large-scale mosquito monitoring in real-world settings.

By combining complementary cues instead of treating modalities as rivals, our model achieves stronger generalisation and reliability for practical surveillance.

Proposed Algorithm

The proposed mosquito species identification model utilises a SwinV2 vision transformer for images, an Audio Spectrogram Transformer (AST) for wingbeat audio, and a fusion module that integrates logits from both modalities for final predictions.

Image and Audio Encoders: The SwinV2-base transformer (Liu et al. 2021), pretrained on ImageNet-22k, captures visual patterns distinguishing morphologically similar species. The AST model (Gong, Chung, and Glass 2021) converts five-second audio clips into Mel spectrograms, mapping frequencies to human hearing for clearer species-specific cues. Both encoders output class logits for fusion.

Multimodal Fusion: The fusion model combines image and audio logits through learnable weighting parameters instead of fixed concatenation. A weighted sum of the two logits is passed through a fully connected layer, enabling the model to adaptively depend on the more reliable modality for each sample. Both backbones are fine-tuned jointly, so fusion influences feature learning rather than acting as a post-processing step.

Training and Implementation: Each sample includes a 192×192 normalized image and a five-second audio clip resampled to 16 kHz, padded or trimmed as needed. The model is trained end-to-end using the Adam optimizer with a learning rate of 1×10^{-4}, and fixed seeds are used for reproducibility. Performance is evaluated using metrics such as accuracy, precision, recall, and the F1 score.

Experimental Results and Analysis

Dataset: The image dataset comprises 3000 high-resolution samples across six species (400 training, 100 testing per species) (Karim, Mahmud, and Khan 2024), resized to 192×192 pixels and normalized. The audio dataset contains 486 wingbeat recordings of the same species (Kiskin et al. 2021), split 4:1 for training/testing, and standardized to five-second segments via trimming or zero-padding, with a resampling rate of 16 kHz. Despite its smaller size, it captures species-specific frequency-temporal patterns, though real-world audio may vary more.

Results and Analysis: Table 1 presents the performance of image, audio models on individual modality data and fused modality data using the proposed fusion model. The image-based models (ViT and SwinV2) and the audio-based models (Wav2Vec2 and AST) are first evaluated independently. It is observed that on the image modality, SwinV2 perform significantly better than the traditional ViT model. Similarly, on the audio modality, the AST model yields approximately perfect detection accuracy, which is 9.5% higher than that of the Wav2Vec2 algorithm.

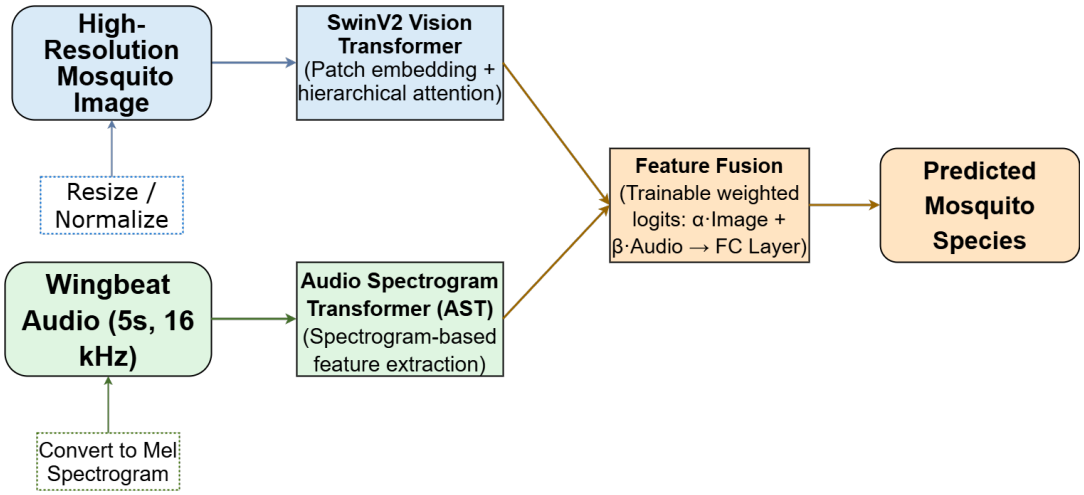

Figure 1: Proposed SwinV2–AST multimodal architecture for mosquito species identification.

Model	Accuracy	Precision	Recall	F1
ViT	97.0	0.97	0.96	0.97
SwinV2	99.2	0.99	0.99	0.99
Wav2Vec2	90.0	0.90	0.86	0.86
AST	99.5	0.99	0.99	0.99
Proposed Fusion	97.8	0.97	0.97	0.97

Table 1: Mosquito species detection performance of the individual and proposed fusion model under a noise-free setting.

Model	Accuracy	Precision	Recall	F1
SwinV2	72.00	0.79	0.72	0.69
AST	83.51	0.84	0.84	0.83
Fusion	**93.17**	**0.94**	**0.93**	**0.93**

Table 2: Robustness evaluation of best best-performing individual modality model and the proposed fusion model.

Inspired by the tremendous success of these two models, we proposed a multimodal architecture by fusing the highest-performing models, i.e., SwinV2 and AST, to achieve state-of-the-art performance. Surprisingly, the fusion does not show any advantage over individual models and even shows slightly lower performance than any individual model. The primary reason might be the contradictory nature of not only modalities but also models; however, both models independently yield close to perfect detection performance.

Robustness: The real world can be noisy, either in terms of the environment itself or the acquisition device induces noise in either or both modalities. Therefore, we assert that while the proposed fusion does not show any improvement in terms of detection capacity, it makes the system robust towards noise due to the large-scale pre-training of models. On the images, Gaussian noise with $\sigma = 0.15$ and salt-and-pepper noise at a 2% pixel corruption rate have been applied. Whereas, on audio, Gaussian noise is added to achieve an SNR of approximately 10 dB. These settings are used consistently for SwinV2, AST, and the proposed fusion model's robustness evaluation. The robustness performance of each model is reported in Table 2.

While each model, whether trained on an individual modality or combined, showed performance drops with noise, the drop with the fused model is significantly lower than that of the individual models. For example, the image model (Swinv2) shows a drop of 27.20% and the audio-only model (AST) yields a drop of 15.99%. However, their fusion only suffers a drop of 4.66%, reflecting that their fusion might not be better than individual models under clean testing, but is drastically more robust than them.

Explainability: It is important when model outputs inform public health decisions. We believe that the fusion model reduces the risk of being affected by noise in one modality by learning to balance visual and acoustic cues. Since both models are fine-tuned together, the fusion influences feature learning rather than acting as a simple post-processing step. The fusion model remained more stable under noise, suggesting that this weighting helps maintain consistent decisions when one input becomes unreliable.

Conclusion

In this research, image, audio, and multimodal models are benchmarked for the identification of mosquito species. It is observed that single-modality models performed well on clean data but are sensitive to noisy modalities. Meanwhile, the proposed fusion model stays reliable in both clean and noisy settings. With multimodal mosquito classification underexplored, these results are useful for real-world surveillance, where imperfect data is the norm. Future work should prioritise building a large-scale multimodal mosquito dataset to enable stronger architectures for practical monitoring.

References

Al Maruf, A.; Mahmudul Haque, M.; Ara Rumy, R.; Jahan Puspo, J.; and Aung, Z. 2025. TransembleNet: Enhancing vector mosquito

species classification through transfer learning-based ensemble model. *PLOS ONE*, 20(5): e0322171.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021. AST: Audio Spectrogram Transformer. In *Proceedings of Interspeech 2021*, 571–575.

Hagiwara, M.; et al. 2022. BEANS: The Benchmark of Animal Sounds. arXiv:2210.12300.

Karim, A.; Mahmud, M.; and Khan, R. 2024. Advanced vision transformers and open-set learning for robust mosquito classification: A novel approach to entomological studies. *PLoS Computational Biology*, 20(12): e1012654.

Kiskin, I.; Sinka, M.; Cobb, A. D.; Rafique, W.; Wang, L.; Zilli, D.; Gutteridge, B.; Dam, R.; Marinos, T.; Li, Y.; Msaky, D.; Kaindoa, E.; Killeen, G.; Herreros-Moya, E.; Willis, K.; and Roberts, S. J. 2021. HumBugDB: A Large-scale Acoustic Mosquito Dataset. NeurIPS 2021 Track on Datasets and Benchmarks. arXiv:2110.07607.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *IEEE/CVF ICCV*.