Q-MoFusion: A Quantum Classifier for Mosquito Species Classification (Student Abstract)

Vishesh Kumar, Ahana Chanda, Poulomi Bhattacharya, Akshay Agarwal

Trustworthy BiometraVision Lab, IISER Bhopal, India
{vishesh22,poulomi24,akagarwal}@iiserb.ac.in, ahanachanda21@gmail.com

Abstract

Automated mosquito species identification is critical for combating vector-borne diseases. We introduce Q-MoFusion, a novel hybrid quantum-classical framework that fuses deep features from pre-trained Audio Spectrogram Transformer (AST) and Whisper models using a Variational Quantum Circuit (VQC). Our approach significantly outperforms individual backbones and prior state-of-the-art benchmarks, demonstrating superior accuracy and robustness, particularly on imbalanced classes. Q-MoFusion demonstrates the potential of hybrid quantum computing to enhance bioacoustic surveillance for addressing critical public health challenges.

Introduction

Mosquito-borne diseases cause over 700,000 deaths annually (Organization 2024), a global health crisis amplified by climate change. Effective public health responses hinge on rapid species identification; however, traditional manual surveillance is slow, costly, and requires specialized expertise, hindering timely interventions (Semwal et al. 2022). Automated approaches have sought to overcome these limitations. While early acoustic methods analyzed wingbeat frequencies (Kahn, Celestin, and Offenhauser 1945), they were hampered by poor audio quality (Chen et al. 2014). This led to optical and image-based CNN solutions (Goodwin et al. 2021), but these often require capturing the insect, making them impractical for large-scale, non-invasive surveillance.

Acoustic monitoring remains a promising avenue for real-time deployment. Leveraging recent advances in audio foundation models, we introduce Q-MoFusion, a novel hybrid quantum-classical framework. Q-MoFusion uniquely fuses complementary feature representations from the AST (Gong, Chung, and Glass 2021) and OpenAI's Whisper (Radford et al. 2023) within a Variational Quantum Circuit (VQC), harnessing quantum machine learning to capture complex data correlations. Our contributions include this novel hybrid architecture, a robust feature extraction and regularization strategy, and its validation on a geographically diverse dataset, demonstrating its potential for global public health applications.

Proposed Methodology: Q-MoFusion

Our proposed Q-MoFusion framework is a hybrid quantum-classical model that consists of three main stages: (1) a classical feature extraction stage using frozen foundation models, (2) a hybrid quantum-classical fusion head for classification, and (3) a robust regularization to prevent overfitting.

Overall Architecture: The model processes a raw audio waveform through two parallel, frozen backbones, AST and Whisper, to extract and aggregate fixed-size feature vectors. These classical vectors are projected, concatenated, and then fused within a Variational Quantum Circuit (VQC), which serves as the core mechanism for fusion. A final feed-forward network then classifies the VQC's output to produce the final prediction.

Backbone Feature Extraction: To generate fixed-size feature representations, each audio sample is first padded or truncated to 30 seconds. This clip is then passed through two frozen, pre-trained backbones to extract complementary embeddings. From the AST, we take the final hidden state embedding corresponding to the [CLS] token, which serves as a global summary of the input spectrogram. Concurrently, from the Whisper encoder, we apply mean-pooling across the entire final hidden state sequence to create an aggregate representation. This dual-extraction process yields two distinct feature vectors, $v_{AST} \in \mathbb{R}^{D_{AST}}$ and $v_{WHP} \in \mathbb{R}^{D_{WHP}}$, for each audio sample.

Hybrid Quantum-Classical Fusion Head: The trainable fusion head, designed to integrate the classical feature vectors v_{AST} and v_{WHP} using a VQC.

Classical Pre-processing: The backbone feature vectors are first projected into a shared latent dimension d using two separate linear layers. The resulting vectors are concatenated to form a unified classical representation, $z_{classic} \in \mathbb{R}^{2d}$. This vector is then passed through another linear layer to project it down to a dimension of N_q, the number of qubits in our quantum circuit. This final classical vector, $\phi \in \mathbb{R}^{N_q}$, serves as the input parameter for the VQC.

Quantum Processing (VQC): The quantum circuit, implemented in Pennylane, operates on N_q qubits and performs three key steps:

1. **Data Encoding:** The classical vector ϕ is encoded into the quantum state of the qubits using an 'AngleEmbedding' layer, $U_{embed}(\phi)$. This operation rotates each qubit

Model	Accuracy	M-F1	W-F1	M-ROC	W-ROC	M-PR	W-PR
AST	70.41	45.48	66.84	95.95	96.58	54.37	79.82
Whisper	80.63	52.12	79.71	**96.96**	**98.18**	56.78	85.51
Ours	**84.29**	**58.88**	**83.61**	96.15	97.88	**61.42**	**87.36**

Table 1: Comparison of proposed Q-MoFusion with the benchmark approaches. The best score is **bold-faced**. W- and M- mean weighted and macro values, respectively.

on the Bloch sphere based on the corresponding feature value in ϕ, preparing the initial quantum state $|\psi_{in}\rangle = U_{embed}(\phi)|0\rangle^{\otimes N_q}$.

2. **Variational Layers:** A series of learnable quantum gates, parameterized by weights θ, are applied. We use 'BasicEntanglerLayers', denoted $U_{var}(\theta)$, which consist of single-qubit rotation gates and CNOT gates to create entanglement between qubits. This parameterized circuit allows the model to learn complex transformations on the encoded data: $|\psi_{out}\rangle = U_{var}(\theta)|\psi_{in}\rangle$.
3. **Measurement:** Classical information is extracted by measuring the final quantum state. We compute the expectation value of the Pauli-Z operator (σ_z) for each of the N_q qubits. This yields a classical output vector $q_{out} \in \mathbb{R}^{N_q}$, where each element represents the measurement result $\langle\psi_{out}|\sigma_z^i|\psi_{out}\rangle$ for the i-th qubit.

Final Classification: The vector q_{out} extracted from the VQC is passed through a final classical classification head, consisting of layer normalization, a GELU activation, Dropout, and a linear layer that outputs logits for the N_{cls} mosquito classes.

Implementation Details

The model is implemented using PyTorch for the classical components and PennyLane for the quantum circuit simulation. The pre-trained backbone models were sourced from the Hugging Face Transformers library.

Training Hyperparameters: The fusion head was trained for a maximum of 50 epochs with an early stopping patience of 7 epochs on the validation accuracy. We used a batch size of 64. The model parameters were optimized using the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 0.01.

Architectural Details: The classical projection dimension d was set to 256. For the quantum circuit, we used $N_{\text{qubits}} = 16$ qubits and the ansatz consisted of $L = 8$ entangling layers. The dropout rate in the final classifier was set to 0.2, and the Mixup hyperparameter α was set to 0.2. The VQC was simulated using PennyLane's high-performance 'default.qubit' simulator. All experiments were conducted on a single NVIDIA A100 GPU.

Result and Analysis

Dataset: We use the HumBugDB dataset (Kiskin et al. 2021), following the benchmark protocol from BEANS (Hagiwara et al. 2023). We selected all species with at least 100 samples and grouped the remainder into an "OTHERS" category. This resulted in 14 classes, including a "non-mosquito" category. The dataset of $9,049$ audio samples was randomly split into training (60%), validation (20%), and testing (20%) sets with stratification.

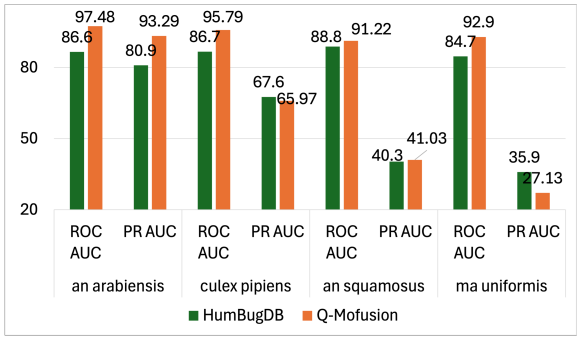

Figure 1: Comparative performance analysis of our proposed Q-MoFusion model against the state-of-the-art HumBugDB baseline on a per-class basis.

Analysis: We evaluated our proposed Q-MoFusion model against two strong individual backbone models: the AST (Gong, Chung, and Glass 2021) and Whisper (Radford et al. 2023). The comprehensive results are summarized in Table 1.

Our proposed Q-MoFusion model significantly outperforms its backbones, achieving the highest accuracy of 84.29%, Macro F1-score 58.88%, and Precision-Recall (PR) AUC 61.42%. The superior Macro F1 score confirms its enhanced performance on minority classes in our imbalanced dataset. While the Whisper baseline shows excellent class separation with a higher ROC AUC, our model's superior PR AUC indicates a more practical and effective balance between precision and recall. These results validate that our quantum-classical fusion strategy creates a more robust classifier for this task.

Furthermore, a per-species comparison against the HumBugDB benchmark shown in Figure 1 confirms our model's advancement on challenging minority classes. Q-MoFusion demonstrates superior discriminative power by achieving significantly higher ROC AUC scores across all listed species (e.g., improving Anopheles arabiensis from 86.6 to 97.48). This class-specific performance uplift directly contributes to our model's superiority.

Apart from that, to validate the effectiveness of our Q-MoFusion strategy, we evaluated a late fusion strategy using $\alpha \cdot AST\_logits + (1 - \alpha) \cdot Whisper\_logits$ with $\alpha = 0.5$. This late fusion yields 79.51% accuracy, which is 4.78% lower than the Q-MoFusion.

Conclusion

We introduced Q-MoFusion, a novel hybrid model that successfully fuses features from classical AST and Whisper backbones within a quantum circuit. Our model validates the significant potential of hybrid quantum-classical systems for solving complex, real-world classification tasks in public health. Future work will explore advanced quantum circuit architectures and deployment on real quantum hardware.

Acknowledgements

The authors acknowledge partial support for V. Kumar through the Visvesvaraya PhD Fellowship, Government of India.

References

Chen, Y.; Why, A.; Batista, G.; Mafra-Neto, A.; and Keogh, E. 2014. Flying insect classification with inexpensive sensors. *Journal of insect behavior*, 27: 657–677.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Goodwin, A.; Padmanabhan, S.; Hira, S.; Glancey, M.; Slinowsky, M.; Immidisetti, R.; Scavo, L.; Brey, J.; Sai Sudhakar, B. M. M.; Ford, T.; et al. 2021. Mosquito species identification using convolutional neural networks with a multitiered ensemble model for novel species detection. *Scientific reports*, 11(1): 13656.

Hagiwara, M.; Hoffman, B.; Liu, J.-Y.; Cusimano, M.; Effenberger, F.; and Zacarian, K. 2023. BEANS: The Benchmark of Animal Sounds. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5.

Kahn, M. C.; Celestin, W.; and Offenhauser, W. 1945. Recording of Sounds Produced by Certain Disease-Carrying Mosquitoes. *Science*, 101(2622): 335–336.

Kiskin, I.; Sinka, M.; Cobb, A. D.; Rafique, W.; Wang, L.; Zilli, D.; Gutteridge, B.; Dam, R.; Marinos, T.; Li, Y.; et al. 2021. HumBugDB: A Large-scale Acoustic Mosquito Dataset. In *Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Organization, W. H. 2024. Vector-borne Diseases. https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases. Accessed: September 26, 2024.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*, 28492–28518. PMLR.

Semwal, A.; Melvin, L. M. J.; Mohan, R. E.; Ramalingam, B.; and Pathmakumar, T. 2022. AI-enabled mosquito surveillance and population mapping using Dragonfly robot. *Sensors*, 22(13): 4921.