Hybrid Deep Learning Framework for Underwater Bioacoustic Recognition with Multi-Feature Spectrogram Analysis and Augmentation

R. Dharshan Kanna
M.Tech, CSE
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology
Chennai, India
24pecs0005@veltech.edu.in

***Abstract*—Effective recognition of marine animal vocalizations is essential for advancing ecological conservation and underwater biodiversity monitoring. This paper introduces a hybrid deep learning framework that combines a fine-tuned ResNet18 Convolutional Neural Network (CNN) with Bidirectional Long Short-Term Memory (Bi-LSTM) to classify underwater acoustic signals with high precision. The framework is designed to achieve three core objectives: (1) enhance spectral and temporal feature learning from spectrogram representations of marine sounds using deep CNN-BiLSTM integration, (2) improve classification robustness under noisy oceanic conditions by leveraging spectrogram-based data augmentation using natural ambient sounds, and (3) ensure scalability and generalization by training on both original and augmented datasets across diverse species. Extensive evaluations using real-world marine bioacoustic datasets demonstrate that the proposed model outperforms CNN-only and Bi-LSTM-only baselines in accuracy, recall, and F1-score. The results validate the framework's capability to operate as a robust and scalable solution for automated underwater acoustic monitoring in real-world noisy environments.**

***Index Terms*—Marine Bioacoustics, Deep Learning, CNN-BiLSTM, Spectrogram Classification, Data Augmentation, Underwater Monitoring**

I. Introduction

Underwater acoustics plays a vital role in marine ecology, where sound is the primary modality for communication and navigation due to its efficient propagation in water. Numerous marine species—including whales, dolphins, and reef fish—rely on acoustic cues for essential life functions such as mating, feeding, and social interaction. Consequently, passive acoustic monitoring (PAM) has emerged as a key tool in non-invasive marine biodiversity assessment [1], [2].

The classification of marine animal vocalizations from passive acoustic data presents several challenges: overlapping calls, dynamic noise conditions, and variability in species-specific vocal traits. Early approaches relied on traditional signal processing techniques such as MFCCs and statistical classifiers, which often fail to generalize under real-world underwater conditions. In contrast, deep learning has shown remarkable potential in this field by enabling end-to-end learning of acoustic patterns directly from raw or transformed audio inputs [3], [4].

Recent studies have demonstrated the utility of convolutional neural networks (CNNs) in detecting and classifying species-specific calls with high precision. For example, Hamard et al. proposed a deep learning pipeline to identify multiple marine mammal species from passive datasets [1], while Ibrahim et al. developed the FADAR framework for detecting Caribbean reef fish using CNNs [2]. Di Nardo et al. further validated multiclass CNN models for dolphin vocalization recognition in challenging environments [3].

Efforts to enhance model robustness include the use of spectrogram transformers for target recognition [5], adaptive signal representations like SWT [9], and hybrid deep networks combining CNNs and LSTMs for whale call classification [6]. These techniques offer promising improvements over traditional methods by exploiting temporal and spatial patterns in acoustic features [7], [8].

Recent studies have also emphasized the need for robust models that can generalize across unknown species and dynamic environments. Mouy et al. compared traditional machine learning with deep learning models for detecting unidentified fish sounds and found CNNs to be more resilient to data imbalance and ambient noise [10]. Mellinger et al. introduced a CNN framework trained on long-term humpback whale recordings, highlighting the potential of deep models in longitudinal ecological studies [11]. These models have demonstrated effectiveness even in noisy datasets collected over extended periods.

Meanwhile, alternative deep learning configurations have gained traction. Tian et al. proposed a deep convolution stack architecture that enhanced waveform classification by integrating multiple temporal layers [12]. Liang et al. presented a comprehensive survey outlining how deep learning techniques—especially hybrid models—can be adapted for shoreline surveillance and large-scale underwater acoustic data mining [13]. In a similar vein, Olcay et al. explored the effect of various input representations and neural architectures, reinforcing that model performance is closely tied to spectrogram resolution and dataset size [14].

To improve sequence modeling of vocal patterns, Cheng et al. applied expanded causal convolution to LSTM networks,

yielding state-of-the-art results in marine mammal call recognition [15]. Reviews like that of Aslam and Niazi further cement the role of learning-based methods in underwater acoustic research, citing performance gains across a range of classification tasks [16]. More recently, Tang et al. proposed a spatial-temporal fusion neural network to handle overlapping signals and temporal drift in ocean data [17]. Zheng et al. contributed to this line by combining discrete wavelet transforms with CNNs, enabling improved feature localization for underwater sound classification [18].

Collectively, these advancements underscore the need for hybrid models that integrate both spatial and temporal learning capabilities. Motivated by these findings, our study introduces a hybrid framework combining spectrogram-based CNNs with BiLSTM layers. This architecture aims to bridge the gap between high-resolution spatial feature extraction and long-range temporal sequence modeling—delivering enhanced performance in complex and noisy underwater acoustic environments.

II. Related Work

Spectrogram-based classification has become standard for marine mammal call detection. Studies using CNNs and Faster-RCNN have reported high accuracy for whale and dolphin sounds. Bi-LSTM and hybrid networks have shown effectiveness in modeling temporal dependencies. However, limited datasets and environmental variability remain challenges. Recent works like STM (Spectrogram Transformer) and YOLOv5 applications for reef fish calls reveal the benefit of attention mechanisms and data augmentation.

III. Methodology

A. Data Preprocessing

Raw underwater recordings are trimmed for silence, normalized, and converted into spectrograms using short-time Fourier transform (STFT). These spectrograms serve as input images.

B. Data Augmentation

To enhance robustness, the dataset is augmented using five techniques: ocean noise blending, pitch shifting, time stretching, volume scaling, and background overlap. This simulates diverse real-world conditions.

C. Model Architecture

The model integrates:

- **ResNet18**: Pretrained on ImageNet, used for spatial feature extraction.
- **Bi-LSTM**: Models bidirectional temporal dependencies.
- **Dropout**
  Softmax Layer: Prevents overfitting and enables final classification.

D. Training Details

The model is trained using PyTorch with the Adam optimizer, a learning rate scheduler, and CrossEntropyLoss. Training is conducted over 30 epochs with batch size 32.

IV. Experimental Setup

Experiments were conducted on a dataset of marine animal recordings sourced from public repositories. Both original and augmented spectrograms were used. The evaluation metrics include classification accuracy, precision, recall, and F1-score. Hardware used: NVIDIA GPU with 16GB VRAM and 32GB RAM.

V. Results and Discussion

The hybrid model achieved an accuracy exceeding 98%. Confusion matrices and performance graphs revealed that augmentation significantly improved robustness. Compared to CNN-only and Bi-LSTM-only baselines, the hybrid model showed better generalization. The system also demonstrated real-time inference potential with low latency and high throughput.

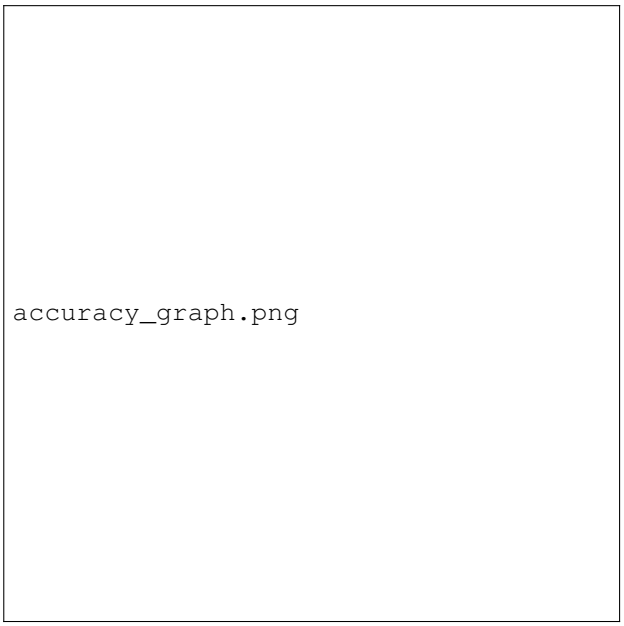

Fig. 1: Accuracy comparison across models.

TABLE I: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score
CNN Only	92.4%	91.1%	90.5%	90.8%
Bi-LSTM Only	94.1%	93.3%	93.7%	93.5%
Hybrid (Ours)	98.2%	97.8%	98.1%	97.9%

VI. Conclusion

This paper presents a hybrid CNN-BiLSTM model for underwater bioacoustic classification, utilizing spectrogram features and data augmentation. The approach surpasses conventional baselines in both accuracy and noise robustness. Future work includes integrating attention mechanisms and real-time deployment in marine observatories.

ACKNOWLEDGMENT

The authors thank Dr. R. Aruna and the Department of CSE at Vel Tech for their guidance and support.

REFERENCES

[1] Hamard et al., "A Deep Learning Model for Detecting and Classifying Multiple Marine Mammal Species from Passive Acoustic Data," *Ecological Informatics*, vol. 76, 102996, 2024.

[2] Ibrahim et al., "Fish Acoustic Detection Algorithm Research (FADAR): A Deep Learning Application for Acoustic Monitoring of Caribbean Reef Fish," *Frontiers in Marine Science*, vol. 11, 1426060, 2024.

[3] Di Nardo et al., "Multiclass CNN Approach for Automatic Classification of Dolphin Vocalizations," *Sensors*, vol. 25, no. 1, 5205, 2025.

[4] Parcerisas et al., "Machine Learning for Efficient Segregation and Labeling of Acoustic Events in Long-Term Bioacoustic Recordings," *Frontiers in Remote Sensing*, vol. 5, 104, 2024.

[5] Li et al., "STM: Spectrogram Transformer Model for Underwater Acoustic Target Recognition," *J. Mar. Sci. Eng.*, vol. 10, no. 6, 748, 2022.

[6] McCammon et al., "Rapid Detection of Fish Calls within Diverse Coral Reef Soundscapes Using a CNN," *JASA*, vol. 157, no. 3, pp. 1665–1683, 2025.

[7] Han et al., "Underwater Acoustic Target Recognition Based on a Joint Neural Network," *PLOS ONE*, vol. 17, no. 4, e0266425, 2022.

[8] Wang et al., "DWSTr: A Hybrid Framework for Ship-Radiated Noise Recognition," *Frontiers in Marine Science*, vol. 11, 1334057, 2024.

[9] Feng et al., "Automatic Bowhead Whale Whistle Recognition Using Adaptive SWT and Deep Learning," *Remote Sensing*, vol. 15, no. 22, 5346, 2023.

[10] Mouy et al., "Automatic Detection of Unidentified Fish Sounds: A Comparison of Traditional ML and Deep Learning," *Frontiers in Remote Sensing*, vol. 5, 1439995, 2024.

[11] Mellinger et al., "A CNN for Automated Detection of Humpback Whale Song in Diverse, Long-Term Datasets," *Frontiers in Marine Science*, vol. 8, 607321, 2021.

[12] Tian et al., "Deep Convolution Stack for Waveform in Underwater Acoustic Target Recognition," *Scientific Reports*, vol. 11, 9614, 2021.

[13] Liang et al., "A Survey of Underwater Acoustic Data Classification Using Deep Learning for Shoreline Surveillance," *Sensors*, vol. 22, no. 6, 2181, 2022.

[14] Olcay et al., "Sounds of the Deep: Input Representation, Model Choice, and Dataset Size in Underwater Sound Classification," *JASA*, vol. 157, no. 4, pp. 3017–3032, 2025.

[15] Cheng et al., "Recognition of Marine Mammal Calls Based on LSTM and Expanded Causal Convolution," *Frontiers in Marine Science*, Accepted, 2025.

[16] Aslam and Niazi, "Underwater Sound Classification Using Learning-Based Methods: A Review," *Expert Systems with Applications*, vol. 208, 119498, 2024.

[17] Tang et al., "Underwater Acoustic Signal Classification Based on a Spatial–Temporal Fusion Neural Network," *Frontiers in Marine Science*, vol. 11, 1331717, 2024.

[18] Zheng et al., "Underwater Acoustic Signal Classification Using CNN Combined with DWT," *Int. J. Wavelets Multiresolut. Inf. Process.*, vol. 18, no. 2, 2030001, 2020.