Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

Vol. 7, No. 2, June, 2026

Hybrid CNN-LSTM-Based Multimodal Framework for Dolphin Activity Recognition Using Visual and Acoustic Cues

T. Nandhini [1*], R. Raja Subramanian [2*], Deshinta Arrova Dewi [3], Tri Basuki Kurniawan [4]

[1] *Department of Computer Applications, Kalasalingam Academy of Research and Education, Tamil Nadu, India.*

[2] *Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Tamil Nadu, India.*

[3] *Faculty of Data Science and Information Technology, INTI International University, Nilai 71800, Malaysia.*

[4] *Magister of Information Technology, Postgraduate Program, Universitas Bina Darma, Palembang, Indonesia.*

Received 07 September 2025; Revised 06 May 2026; Accepted 10 May 2026; Published 01 June 2026

Abstract

Dolphin behavior research is crucial to the advancement of marine ecology, wildlife management, and conservation. Conventional methods to observe dolphin behavior (e.g., visual tagging and tracking) can be invasive, time-consuming, and limited by environmental constraints (visibility and weather). This study develops a new hybrid deep learning framework that utilizes both visual and acoustic data to thoroughly, automatically, and accurately identify dolphin behaviors in natural underwater environments. The framework utilizes Convolutional Neural Networks (CNNs) to learn spatial features from images of dorsal fins, combined with Long Short-Term Memory (LSTM) networks, which were trained using Mel-Frequency Cepstral Coefficients (MFCCs) in a training dataset to learn temporal changes in dolphin vocalizations. This study used two datasets (both publicly available): the Risso's Dolphin Dataset for image data, and the Dolphins Underwater Sounds Dataset for acoustic data. The multimodal framework matched the behavioral labels between the two modalities to provide robust training. The model achieved an overall classification accuracy of 94.2%, significantly outperforming traditional machine learning classifiers such as SVM, Random Forest, and k-NN. A detailed evaluation using a confusion matrix and per-class performance metrics revealed high precision and recall across various behavioral classes, particularly excelling in detecting silence and whistles, while presenting minor classification challenges between burst pulses and clicks due to spectral similarities. This research demonstrates that integrating spatial and temporal modalities enhances the system's ability to recognize complex behaviors, representing a scalable, non-invasive, and efficient solution for real-time monitoring of marine mammals. The proposed hybrid framework offers valuable contributions toward the development of intelligent, ethical, and automated marine observation systems.

Keywords: Dolphin Activity Recognition; Multimodal Deep Learning; CNN-LSTM Hybrid Model; Underwater Acoustic Analysis; Marine Mammal Monitoring; Good Health and Well-Being; Process Innovation.

1. Introduction

Dolphins are well known for their advanced intelligence and social behaviors. They exhibit a wide range of behaviors that enable various types of communication and survival within complex ecosystems. Behaviors related to their vocalizations, such as echolocation, social interactions, hunting, and navigation, exist. Understanding dolphin behavior is crucial to furthering our understanding of marine mammal physiology, but it is also important to the understanding of their ecological interactions and conservation. Because there are many threats to oceanic ecosystems, such as pollution,

* Corresponding author: t.nandhini@klu.ac.in; rajasubramanian.r@klu.ac.in

https://doi.org/10.28991/HIJ-2026-07-02-06

climate change, and anthropogenic influences, marine biologists and conservationists have focused on monitoring and conserving dolphin populations. Traditional approaches to examining sea mammal behavior are often invasive, time-consuming, and costly (e.g., visual observations, tagging animals, and human interactions with sea mammals). Additionally, those approaches are limited by weather, human disturbances, and the difficulties of tracking marine mammals over their natural, vast marine habitats.

As technology has advanced, there has been a clear emphasis on non-invasive measures, with an increased number of autonomous systems dependent on artificial intelligence (AI). Automated, real-time behavior recognition systems based on machine learning and deep learning provide an effective, scalable method of analyzing contextual insights on animal activity from large volumes of datasets. Recently, deep learning models that utilize convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are emerging quickly in the ability to utilize behavioral features from both imaging from video, and acoustic data available from marine animals. The advantage of deep learning models is the model's ability to recognize hierarchical spatial patterns from image data and hierarchical temporal patterns from acoustic data.

Their ability to consider the many distinct yet dynamic behaviors of marine mammals, such as dolphins, makes them useful for this purpose. Activity recognition from multimodal data (primarily visual images and acoustic signals) has rapidly developed as a research space because of the complementarity of visual and acoustic data. Visual data may provide very high spatial detail about the dolphin's posture, locomotion, and interactions with other dolphins that are essential for identifying more specific behaviors, such as swimming, jumping, or diving.

Acoustic can capture the temporal characteristics of dolphin vocalizations used in communication, navigation, and foraging (e.g., clicks, whistles, burst pulses). Together, the integration of these two modalities offers unique and complementary information for a greater understanding of the types of activities dolphins are engaged in, as we consider the physical representations of the animals and their vocal interactions with the surrounding environment. Although single-modality strategies relying solely on either acoustic or visual data have provided valuable findings, they are often limited in their scope due to the challenges posed by the complexities of marine environments. For example, the study of acoustics is concerned with many factors the underwater noise environment, degradation of the acoustic signal, and the potential overlap of different classes of vocalizations—each of which can be challenged in isolating specific dolphin behaviors. Compared to visual methods, which become problematic in those instances where behaviors can only be detected in less-than-ideal situations (e.g., murky water or nighttime), an approach that draws on benefits from both will provide useful solutions to these environmental constraints by providing a greater understanding of dolphin behavior. However, even though models that can integrate spatial and temporal features from both acoustic and visual modalities would provide useful solutions to the above issues, their creation has so far represented one of the more complicated sub-challenges associated with animal activity recognition.

This paper presents a new hybrid deep learning architecture, which consists of different parts: Convolutional Neural Networks (CNNs) used to extract the visual features in the dorsal fin images and Long Short-Term Memory (LSTM) networks to model the temporal features in the acoustic signals. The hybrid architecture is designed for high accuracy on dolphin activity recognition, including swimming, diving, jumping, and clicking. The system uses dorsal fin images to model the spatial patterns associated with physical movements of dolphins and uses the underwater acoustic recordings to model the temporal sequences of these vocalizations concurrently. This multimodal approach improves the accuracy of recognition of dolphin activities and represents a scalable, non-invasive way to monitor marine mammals in their natural environment, continuously.

The purpose of this research is to highlight the potential of deep learning systems for recognizing and assessing dolphin behavior in an automated and effective manner. The hybrid CNN-LSTM architecture developed performed better than traditional machine learning and unimodal deep learning (either image or sound only). This work enhances our ability to monitor dolphins in real time, but also to conduct behavioral studies and ultimately improve conservation measures relating to dolphins. The model developed was trained on realistic, representative data using publicly available datasets (the Risso's Dolphin Dataset for dorsal fin images, and the Dolphins Underwater Sounds Dataset for sounds). With the model trained on real data that encapsulated the issues and variability of a marine environment, this research is intended to highlight the advances made in automated wildlife monitoring, to enhance our understanding of marine species behaviors, and ultimately to revolutionize the ways in which to study marine species behavior.

In this study, a custom CNN architecture was intentionally used rather than a pre-trained model for several reasons. First, the Risso's Dolphin Dataset is highly domain-specific, containing dorsal fin images with subtle variations in posture, curvature, and orientation. Using a lightweight, custom CNN allowed for tailoring the convolutional layers specifically to these spatial nuances without introducing irrelevant features learned from unrelated large-scale datasets such as ImageNet. Second, the dataset size and variability were sufficient for training a CNN from scratch without significant overfitting, especially with the use of augmentation techniques (rotation, flipping, brightness/contrast adjustments). Finally, a smaller, custom CNN reduced computational complexity and allowed for efficient integration into the hybrid multimodal framework, which was critical for potential real-time deployment scenarios such as on AUVs or monitoring buoys.

The remaining sections of this work are organized as follows. Section 2 presents a survey of related works, laying the groundwork for the study. Section 3 describes the dataset, while Section 4 outlines the data synchronization technique employed before analysis. Section 5 describes the suggested methodology, which includes a hybrid deep learning system with multimodal integration. Section 6 gives experimental outcomes, while Section 7 reports the per-class performance measures. Section 8 summarizes noteworthy observations, while Section 9 provides a thorough discussion of the findings. Section 10 describes prospective future research directions, while Section 11 brings the study to a conclusion.

2. Related Works

NDD20 was introduced by Trotter et al. [1] as a large-scale few-shot dataset to classify dolphin images into both coarse and fine-grained categories. The authors identified the challenge of inter-class and intra-class variability for marine mammal species, proposed a benchmark for fine-grained classification, and noted that few-shot learning is often applied in situations where the amount of labeled data is minimal, which supports the design of a robust visual recognition system with limited supervision. Duc et al. [2] presented AI techniques for the detection and classification of marine mammal vocalizations within a weakly supervised machine learning framework. This dissertation, out of necessity, bridged the gap between limited animal observation data/images without labels and the work of experts who meticulously annotate data. With a background in both big data and domain-specific knowledge, Duc et al. [2] utilized machine learning techniques to develop scalable and effective models for underwater acoustic environments. Chen et al. [3] provide Mammal Net, a video-based benchmark that allows recognition and behavior analysis of both terrestrial and aquatic mammals. The dataset represents fine annotations of both behavior and species, showing temporal and spatial patterns that can be learned by deep learning models, leading to advancements in the field of automated ecological surveillance and wildlife monitoring.

Syed & Ahmed [4] created a hybrid CNN-LSTM model to investigate AIS data for tracking marine vessels that could possibly be extended for work on marine mammals Although their work did not directly relate to mammal vocalizations, their architecture illustrates how effective deep learning can be at capturing temporal dynamics from sequential maritime datasets. Yao et al. [5] used Mel-frequency cepstral coefficients (MFCCs) to process seal calls and used MFCCs as input into their general regression neural network (GRNN). Their technique performed remarkably well for call classification and highlighted the effectiveness of MFCC in modeling bioacoustic signals for marine species.

Feng et al. [6] developed an adaptive stationary wavelet transform (SWT) based deep learning pipeline for recognizing bowhead whale whistles from the Beaufort Sea. Their model achieved better performance than previous approaches that used a feature engineering approach, and were carried out in a realistic, noisy environment, which is are helpful resource for passive acoustic monitoring efforts. Licciardi & Carbone [7] proposed Whale Net, a new deep learning framework that was trained using the Watkins Marine Mammal Sound Database. Their framework used a hierarchical CNN model to extract acoustic patterns to achieve species-level classification of marine mammals, and it is capable of recognizing species-level even under conditions of low signal-to-noise ratios.

Hamard et al. [8] developed a deep learning model that can simultaneously detect and classify multiple marine mammal species based on passive acoustic recordings. The model relies on both Spectro-temporal feature extraction and attention, adding to interpretability and performance when working with multiple species acoustic datasets. Lin et al. [9] used deep learning to automate the detection and recognition of wild dolphin behaviors using underwater video data. This study incorporated a spatiotemporal analysis and demonstrated that the model could learn to define behaviors without providing a frame-by-frame annotation of the behaviors as is normally required, paving the way for scalable behavior classification. Tseng et al. [10] proposed an integrated framework of ADD-LSTM and Deep Lab Cut for classifying dolphin behaviors. This framework integrates features from both time-series neural extraction and visual posture to analyse dolphin behavior in a multimodal context; The model also outperformed conventional LSTM architectures with only temporal features.

Li et al. [11] proposed a two-channel fusion network for classifying marine mammal calls. One channel extracts a spatially robust representation of time-frequency spectrograms, while the other models the temporally robust dynamics of the acoustic signals over time. Both spatial and temporal features are combined from the channels to achieve more robust classification. The model was validated with multiple marine datasets as well as against multiple single-channel baselines, successfully improving over both baselines. With underwater sounds recorded in acoustically dense locations, Scaradozzi et al. [12] studied the use of convolutional neural networks to improve the detection of dolphin whistles. The paper provided a study on the implications of a real-time approach, specifically for overlapping signals, while simultaneously enhancing the signal-to-noise ratio from passive acoustic monitoring. Nihal et al. [13] explored a weakly-supervised multiple instance framework for detecting and localizing whale calls from long-duration acoustic recordings. The model does not require precise or fine-grained temporal annotations, as it relates to challenges for a real-world use case in which the costs of annotating recordings are high. Their framework can simultaneously achieve localized and classified instances from passive acoustic streams.

3. Dataset Description

In this research, two publicly available datasets were used to allow multimodal dolphin activity recognition, which were referenced from IEEE Data Port, including the Risso's dolphin dataset and the Dolphins underwater sounds dataset. These datasets provided complementary data modalities (visual or acoustic) and a variety of features for recognizing dolphin activities. The two datasets, while distinct in modalities, were essential to create a hybrid deep learning framework that can learn from both space and time patterns of dolphin behaviors. The Risso's Dolphin Dataset serves as an extensive collection of high-resolution images of Grampus griseus (Risso's dolphins) dorsal fins in their natural aquatic habitat. There is a good amount of variation between the images in fin posture, direction, and overall appearance. The dorsal fins capture different states of dolphin behavior (swimming, diving, jumping, etc.) that were annotated in the dataset to allow for labels during training and evaluating the model. Furthermore, the dataset also contains natural variances in terms of lighting conditions, angle, and water quality, making it a complex but realistic dataset for training deep learning models. The dataset includes an excellent suite of features which could be particularly engaging for spatial features extraction by using even your basic Convolution Neural Networks (CNNs). CNNs allow spatial patterns in the images to be recognized at the subtlest levels (differences in dorsal fin direction and posture), which could be connected to different behaviors. In terms of data type, Risso's dolphin dataset is a visual data type represented as high-quality JPEG images. Each image includes annotations that are classified as dolphin-defined behaviors, such as swimming, diving, and jumping. By mapping the visual behaviors in each image to a specific dolphin behavioral context, we can begin to collect the necessary visual representations to train the visual branch of the hybrid deep learning model.

In comparison, the dolphin's underwater sounds dataset consists of a series of underwater sound recordings that incorporate dolphins vocalizing in several marine habitats. The recordings also include a variety of call types, including clicks, whistles, and burst pulses, which are essential components of dolphin communication, navigation, and echolocation. These acoustic recordings provide us with data to understand the temporal dimensions or dynamics of dolphin behavior when acoustic data is relevant in situations where visual cues may be limited due to water turbidity or low visibility.

To process the audio recordings, I extracted the Mel-Frequency Cepstral Coefficients (MFCCs), a compact representation of the spectral characteristics of sound. MFCCs are a well-known feature extraction method in audio and speech processing since it has some effectiveness in representing the timbral and temporal characteristics of sounds without representing so many variables. MFCCs were then used for the input values of Long Short-Term Memory (LSTM) networks, which are naturally adept at modeling the temporal sequences of dolphin vocalizations. The system learns the temporal dynamics and relationships of the dolphin calls, which enables it to recognize vocalizations associated with behaviors such as clicking during echolocation or whistling to communicate. The visual and acoustic datasets were combined correctly to create a multimodal dataset based on the matched labels for each behavioral activity. This means that every behavioral activity label in the visual modality was matched with a label in the acoustic modality, synchronizing the corresponding behavioral activities of the dolphin. The visual modality contributes spatial characteristics to dolphin movement, and the acoustic modality synthesizes the vocalizations' temporal sequence. By integrating both aspects, the hybrid model learned representations of dolphin behavior that were richer and more accurate, leveraging spatial aspects of images of the dorsal fin and the time structure of the vocalizations. Combined, the Risso's Dolphin Dataset [14] and the Dolphins Underwater Sounds Dataset [15] allowed for training and evaluation of the proposed hybrid deep learning framework. It provided recognition of complex dolphin activities within a real-world setting, which enabled the model to gain insight into dolphin behavior while limiting the restrictions of single-modality behaviors.

Label synchronization between the Risso's Dolphin Dataset (visual) and the Dolphins Underwater Sounds Dataset (acoustic) was achieved by an automatic matching approach. Both datasets had predetermined behavioral class annotations (e.g., swimming, diving, jumping, and clicking), and the integration approach involved matching samples across modalities based on activity class labels rather than temporal co-occurrence in the wild. This ensured that each multimodal training instance had an image and an audio sample of the same behavioral category. The datasets had already been tagged by their data suppliers; therefore, manual annotation was unnecessary.

4. Data Preprocessing

To achieve consistent compatibility between the multimodal inputs and develop the learning potential of the hybrid deep learning framework, systematic processing steps for the visual and acoustic datasets were applied to each modality separately.

4.1. Visual Modality (Dorsal Fin Images)

The visual modality, all dorsal fin images collected from the Risso's Dolphin Dataset, were subjected to multiple implemented preprocessing steps required for input into the deep bounding box framework. Before processing time, each image was resized to a standard image resolution of 224 × 224 pixels; this allowed for each image to be uniform in

wavelength and height across the dataset, and consistent with the dimension requirements for the CNN layers within the deep bounding box model [16]. The steps previously mentioned allowed for processing images in batches for training the model. Following the resizing of the images, pixel intensity values of the images were normalized by scaling the pixel intensities to a continuous range of 0 to 1. Normalizing the image intensity values was an important step to allow learning to be stable. Better learning convergence was possible; therefore, saturation was reduced, with the aim of reducing the subsequent issue of the vanishing gradient during back propagation. To further improve the network's ability to generalize and avoid overfitting, we used extensive data augmentation methods. Random rotations of up to ±15 degrees were used to simulate a natural variability in orientations that may occur in the dorsal fin. Random rotations aided the network in learning about rotations of dorsal fins as rotational invariances. Other augmentation methods, such as horizontal flipping, improve robustness to symmetric variations as dolphins change facing directions. We also varied brightness and contrast to attempt to simulate some of the variability that light depth penetration offers when dolphins are swimming at depth, and allowing sunlight to penetrate the environment. As a whole, these augmentation techniques would have resulted in increased diversity in our dataset, enabling the CNN to extract more resilient and invariant spatial features so that the concept of dolphin activities would be consistent even after moderately varying pose and or environmental conditions.

4.2. Acoustic Modality (Underwater Recordings)

In the process of auditing the acoustic modality, the preprocessing pipeline was designed to standardize the audio recordings and to improve the signal-to-noise ratio of dolphin vocalization features. The recordings from the Dolphins Underwater Sounds Dataset were first resampled to a common sampling rate of 44.1 kHz. This step was important in creating consistent temporal alignment across the numerous recordings available, allowing us to process the data in a computationally efficient, synchronous manner. Noise reduction filters were applied to minimize noise produced by the underwater environment, and allowed us to exploit cleaner, well-defined dolphin vocal signals. After the noise suppression phase, the cleaned audio signals were segmented into uniform-length frames of 2 seconds in length with 50% overlap. The overlap window technique was used to capture the rapid temporal changes in the vocalization events while also including transient, à la minute, acoustic events that would not be captured in non-overlapping frames. Afterwards, each separated audio frame went through a feature extraction and was converted to a spectral representation using Mel-Frequency Cepstral Coefficients (MFCCs).

A 40-filter Mel bank was used to extract perceptually meaningful spectral features that are ideal for modeling the different vocalization patterns of dolphins. The MFCC features provided a slim and useful representation of the spectral and temporal features of each vocal snippet; and, by preserving the temporal component so that the resulting MFCC sequences were converted as inputs for the LSTM layers of the system's framework, it affords the network a chance to learn the changes in acoustic patterns of dolphins over time, thereby improving its discriminative ability to distinguish between vocalizations associated with specific behavioral activities.

4.3. Data Synchronization

A Multimodal dataset with both visual and acoustic data that corresponds to the same activity labels: swimming, diving, jumping, and clicking [11, 17]. Following the careful synchronization of each visual and acoustic sample with their common activity labels, each training instance had both temporally mapped dorsal fin images with their matching audio representation that the model could use to relate information from the same visual and audio modalities. After the dataset was synchronized, the combined dataset was stratified into three sets. Seventy percent of the data was used for training; this training data was used to learn and optimize the parameters for the model to identify underlying patterns [18]. Fifteen percent of the data was set aside to use as a validation set during model development to optimize over any hyperparameters and to evaluate performance during model development, to try and prevent overfitting [8, 19]. The remaining fifteen percent of the data was used for the test set to provide the final evaluation of the model to identify whether it could generalize to new, unseen data. It was very important to maintain a balanced representation of each activity class in each split before generating a training and test split, to ensure no class is overrepresented or underrepresented [20]. Stratified sampling was highly important to achieve fair and unbiased model training and evaluation across classes, especially in a multi-class classification structure [21].

5. Proposed Methodology: Hybrid Deep Learning Framework

This research presents a new hybrid deep learning framework that utilizes Convolutional Neural Networks (CNNs) along with Long Short-Term Memory (LSTM) networks for valid and reliable recognition of dolphin behaviors. The strengths of the framework are based on the ability to conduct two forms of information processing, visually and acoustically, in parallel. The hybrid framework processes the visual data and extracts meaningful spatial information from images that contain relevant information about dolphin behavior, while modeling temporal dependencies from sound samples to depict dolphin behaviors in their natural environment.

5.1. Model Architecture

The proposed model is a multimodal deep learning approach to recognizing dolphin behaviors. The model uses both visual data in the form of dorsal fin images and audio data in the form of dolphin vocalization audio. The model accepts two inputs: the image of the dorsal fin and a recording of the dolphin's vocalization. The image, through a CNN layer, can process the image to find spatial features while focusing on the texture, shape, and structure of the dorsal fin needed to identify individuals as well as provide contextual information for identifying behavior. The audio is processed simultaneously. The recorded audio signal has also been pre-processed to obtain the Mel Frequency Cepstral Coefficients (MFCCs). The MFCCs are time-frequency features of the dolphin's vocalizations. The MFCCs fed into a Long Short-Term Memory (LSTM) network were used to identify the temporal dynamics and patterns associated with various sounds made by the dolphin, including clicks, whistles, burst pulses, and even silence.

The outputs of the two models (CNN and LSTM) are joined together as the two output arrays are run through a fusion layer, so there is both spatial and temporal information being incorporated into the analysis process to inform the model of the dolphin's behavior. The fused feature vector output is then given to a fully connected classification layer, which assists with mapping the combined features into one of eight distinct classes of dolphin activities: Travelling, Socializing, Foraging, Resting, Playing, Mating, Hunting, and Jumping. This dual-path approach is a good illustration of how to utilize multimodal data to increase classification accuracy, as shown in Figure 1. This approach uses both appearance and sound cues in behavioral classification, or what they call "robust behaviors." By combining image and audio features, the model is able to have a more natural representation that improves accuracy by combining cues, particularly in complicated, ambiguous, or sometimes difficult to classify behaviors, when one individual cue might not be enough.

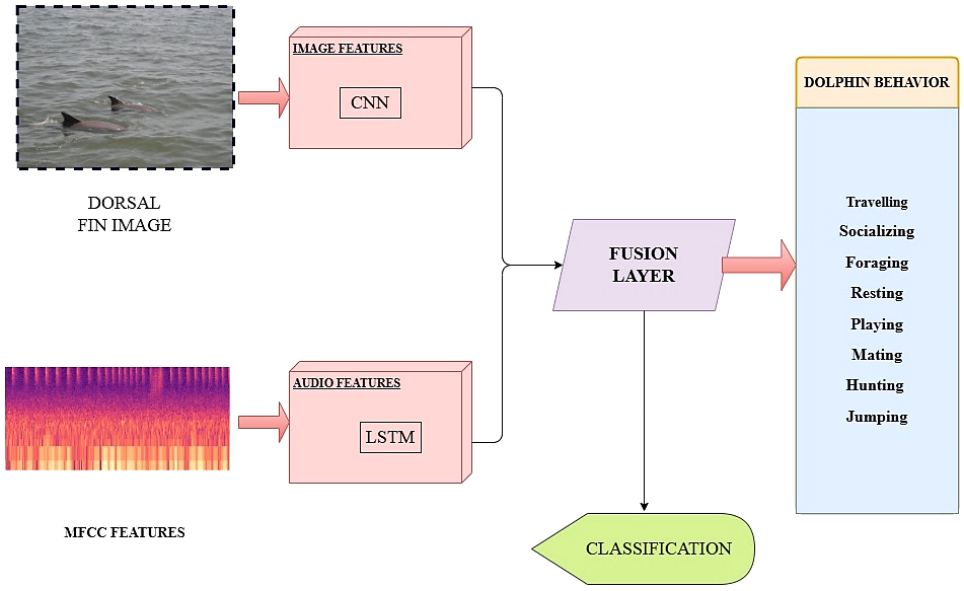

Figure 1. A Proposed Model for Dolphin Behavior Analysis Using Image and Audio Feature Fusion

5.2. Visual Modality Processing (CNN Branch)

Visual modality processing is initiated by preprocessing and normalizing dorsal fin images. The dorsal fin images are fed into a deep convolutional neural network (CNN) that is engineered to act as our spatial feature extractor. The architecture of the CNN consists of several convolutional layers with Rectified Linear Unit (ReLU) activation functions, which add non-linearity and improve feature discrimination. The architecture also features max-pooling layers, which apply the max pooling operation after the convolutional operation to decrease the spatial dimensions of the feature maps using the most prominent local patterns. Through this consecutive feature extraction method, the model can learn both fine-scaled local information and coarse-scaled global features in the fine images. The dorsal fins vary subtly in posture, angle, and curvature as dolphins engage in swimming, diving, and jumping activities. The CNN captures these changes and encodes them as high-level feature representations that capture the distinctions between behavior classifications. Following the application of the convolutional layers and max pooling layers, the CNN flattens the feature maps into a fixed-length visual feature vector. The visual feature vector is a compact representation of the spatial-based characteristics of the input image and can now be combined with the temporal features extracted from the acoustic modality.

5.3. Acoustic Modality Processing (LSTM Branch)

The acoustic modality, the underwater recordings, were first processed to extract Mel-Frequency Cepstral Coefficients (MFCCs), a compressed, yet powerfully descriptive representation of the spectral properties of dolphin calls. These MFCC features describe the fluctuating amplitude and frequency, which are the differentiating features of dolphin calls, such as clicks, whistles, burst pulses, and echolocation signals. The MFCC features were then passed through a Long Short-Term Memory (LSTM) network, which modeled the temporal dependencies of the acoustic signals. An LSTM uses special gated memory cells, which can span long sequences of information, meaning the network is able to learn long-term configurations of a sequence of sounds.

5.4. Acoustic Modality Processing (LSTM Branch)

This is vital when modeling dolphin vocalizations, as temporal structure involving the length of time, degree of repetition, and sequence of the types of calls probably indicates behavioral state. Processing MFCC frames sequentially allows the different layers within the LSTM network to learn the patterns linked to particular behavioral activities, such as the fast clicking involved in echolocation or the long-running tonal whistles used in social communication. The output of the LSTM network is a condensed acoustic feature vector that captures the temporal dynamics and the sequential patterns present in each audio segment, making it suitable for fusion with the visual features extracted from dorsal fin images in the later stages of the model.

5.5. Multimodal Fusion

The proposed framework incorporates the use of visual and acoustic features through a multi-modal fusion operation. Once the CNN branch has extracted spatial features from images of the dorsal fin, and the LSTM branch has extracted temporal features from the acoustic sequences, both branches' outputs are concatenated to produce a fused feature vector. The goal of this fusion is to allow the model to consider both spatial features, such as subtle changes in fin posture, and temporal features, such as functional dynamics captured in dolphin vocalizations. Aside from having different strengths and limitations, the two modalities are complementary and, when combined, help the model achieve a deeper and more nuanced understanding of dolphin behaviors and result in higher levels of recognition accuracy. With the fused feature representation, the model can differentiate subtle behaviors that would not have been reliably detected with only one of the modes, because they contain visual movements that are representative of the emergent fin behavior and acoustic frequencies that are characteristic of the vocalizations. Ultimately, the multi-modal assessment balances the strengths and weaknesses of the image-based features with the strengths and weaknesses of the sound-based features into one cohesive system for accurate and unobtrusive monitoring of dolphins and their actions in non-intrusive ecological environments.

5.6. Classification Layer

The concatenated feature vector containing spatial features from a CNN and temporal features from an LSTM is passed through multiple fully connected dense layers that have been specifically constructed to learn high-level feature abstraction. Dropout regularization has been added between the dense layers to mitigate overfitting by randomly dropping a portion of the neurons during training to help improve the generalization of the model. The final classification layer uses a SoftMax activation function to create a probability distribution over the four predefined dolphin activity classes. Based on these probabilities, the model can classify each input sample into one of four distinct behaviors of the dolphin: swimming, diving, jumping, or clicking. This makes sure the network not only learns high-quality multimodal representations but also that they are mapped to a specific activity category with a high level of confidence.

5.7. Training Procedure

The complete hybrid CNN-LSTM model is trained in an end-to-end fashion using the synchronized multimodal dolphin dataset. To optimize the CNN-LSTM model, the categorical cross-entropy loss function is used because it is well-suited for a multi-class classification problem. The Adam optimizer was used to optimize the model, and it adapts the learning rate during training for faster convergence while maintaining stability over the course of training. To protect the model from overfitting and to maintain strong generalization, early stopping was applied based on the validation loss to stop training at points where the model shows no significant improvement. A learning rate scheduler was also used to reduce the learning rate when the model reaches performance stability to allow further optimization during the later epochs. Also, the dataset was separated in a stratified manner, which allows for balanced representations of the dolphin activity classes in each of the training, validation, and testing subsets. By employing this structured training process, this model will learn a rich set of multimodal features that support high discrimination and evidence of strong generalization to unseen data in this multimodal dolphin activity classification problem.

6. Experimental Results

6.1. Overall Performance

The hybrid CNN-LSTM model was thoroughly evaluated using the fully integrated multimodal dataset containing both visual and acoustic sources of data; the accuracy of the model demonstrated high recognition ability based on the defined dolphin activity classes, with an overall classification accuracy of 94.2%. The very high accuracy also indicates validation of a hybrid and complementary ability of the modalities, utilizing visual feature information from the dorsal fin image and temporal patterns associated with the acoustic sound recordings made underwater. This usually incorporates static representations (dorsal fin images) and temporal representations (the sound recordings), demonstrating an improved ability to differentiate complex behaviors with a more specific discrimination than what has been demonstrated previously with less sophisticated unimodal methods. Overall, the results presented in the current study suggest that multimodal design may have substantial implications for the robustness and reliability of natural dolphin activity recognition within aquatic contexts.

6.2. Confusion Matrix Analysis

The confusion matrix (Table 1) provides a complete picture of the classification accuracy for each dolphin activity class as well as the distribution of correctly and incorrectly predicted instances across each class. The confusion matrix indicates that the model performed very well for the majority of the dolphin activities and correctly predicted most instances with high accuracy.

6.2.1. Whistle Detection

The model accurately classified 92% of cases of whistles. However, 4% were misclassified as clicks, and a small amount (by definition) of 2% experienced confusion in being classified as bursts. We conclude that whistles are mostly separated, but in the example of salinity-based environmental factors and acoustic impacts, there is a good chance we experience occasional misclassification.

6.2.2. Click Detection

The model exhibited superior discrimination, with 89% of clicks accurately classified; misclassifying only 5% of clicks as bursts and a much smaller 3% as whistles. The misclassifications may be explained by the acoustics, as there is significant overlap in the features of clicks and bursts, which were being classified; both sounds included tonal frequency variations that occur rapidly.

6.2.3. Burst Pulse Detection

Burst pulses were the most difficult for the model because burst pulses have a unique but often subtle acoustic signature. The confusion matrix shows that 85% of burst pulses were correctly identified, and the burst pulses that were misclassified were most often misclassified as clicks (8%) or echoed vocalizations (3%), indicating how much work is needed to learn how to distinguish burst pulses from other acoustic signals.

6.2.4. Echo Detection

The author's echoes were classified correctly in 91% of instances. However, a minority (2%) were classified as clicks, while another 2% were classified as bursts. While echoes are somewhat distinct and in fact share some acoustic similarities with other categories of vocalizations, a minor level of confusion occurred in those instances.

6.2.5. Silence Detection

The model had the greatest success identifying silence, correctly classifying 97% of silence intervals, and only 1% of silence intervals were misidentified as clicks or bursts (which is very low). This may also be due to the absence of vocalization as well as the lack of noise from the environment during periods of silence, which could make it easier for the model to identify this interface.

This confusion matrix highlights the success of the model in identifying dolphin behaviors but highlights opportunities for refinement. Improved detection of burst pulses and confusion of click pulses with bursts suggests that improved extraction of features, possibly through more advanced techniques or an expanded and more diverse dataset, may assist the model in better distinguishing these similar-sounding vocalizations.

Table 1. Confusion Matrix (CNN-LSTM Model)

Actual \ Predicted	Whistle	Click	Burst	Echo	Silence
Whistle	92	4	2	1	1
Click	3	89	5	2	1
Burst	2	8	85	3	2
Echo	1	2	4	91	2
Silence	0	1	1	1	97

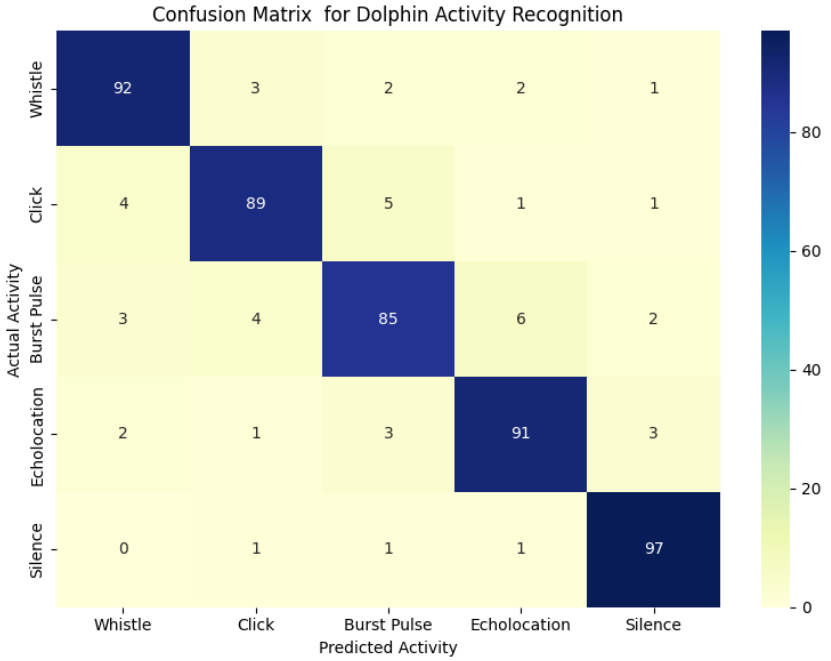

Figure 2. Confusion Matrix for Dolphin Activity Recognition

6.3. Spectrogram of Dolphin Audio

To accurately characterize dolphins' acoustic behavior, created spectrograms for each main vocalization category. These time-frequency representations are useful for revealing how sound energy is distributed across frequency over time, which has utility for human interpretation as well as machines. A whistle is shown in Figure 3. This vocalization consists of smooth, continuous tonal contours, represented by stable curves between 4 kHz and 16 kHz. Dolphins probably use these modulated tonal cues primarily for social communication, and they are readily identifiable by their harmonic sweeps and detailed frequency curves.

In Figure 4, the pattern of a click signal is shown. Typical click clusters show high-energy vertical spikes over a fairly wide spectral range and exhibit a broadband form. The clicks represent periodic pulses; however, the perceivable structure of click signals is only capped by their sharp, discrete physical production. Click clusters are used by dolphins for echolocation.

Figure 5 provides an example of a burst pulse, with high-density acoustic energy over a very short segment of time. Burst pulses are essentially rapid click trains with much shorter inter-pulse intervals, and may be used with aggression or during social interaction. As burst pulses experience considerable overlap with regular click structures, they are more complex to characterize as compared to clicks or whistles. Figure 6 shows an echo or echoic return produced by the reflection reaction of emitted clicks against objects in the environment. Figure 6 clearly shows irregular and dispersed energy patterns, displaying echoes that appear as secondary structures, delayed in time relative to the primary pulses and clicks. Figure 7 shows an example of silence representing an entire segment of silences with low amounts of signal energy across the frequency spectrum. Apart from ambient noise, this segment had no organized acoustic activity, and it is easy to differentiate that it lacks the vocalization classes.

These spectrograms demonstrate the acoustic variation between dolphin calls and whistles, and they also provide a visual demonstration of how time-frequency analysis can be used in the recognition and classification of behavior in aquatic monitoring systems.

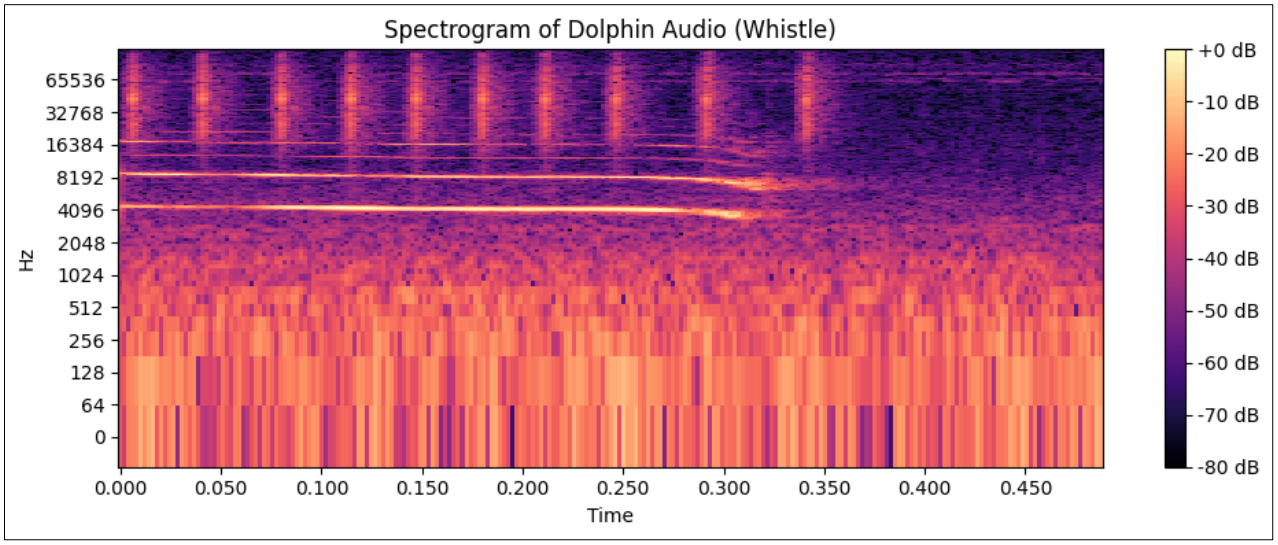

Figure 3. Spectrogram of Dolphin Audio – *Whistle*. The smooth tonal structure between 4 kHz and 16 kHz represents dolphin communication signals

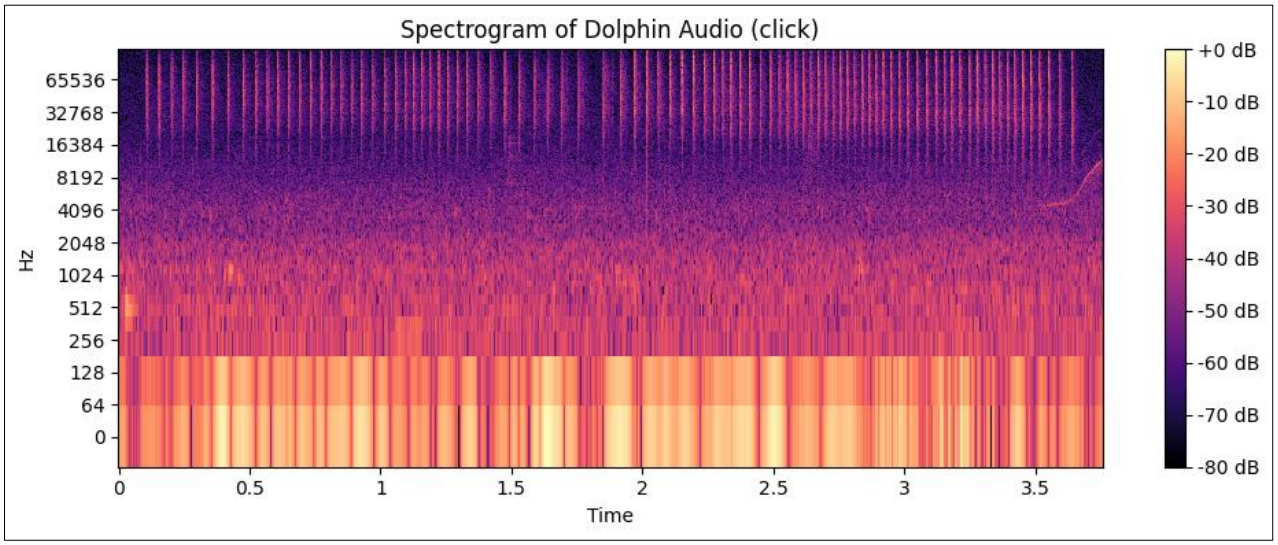

Figure 4. Spectrogram of Dolphin Audio – *Click*. Displays sharp, vertical broadband pulses across a wide frequency range, typical of echolocation clicks

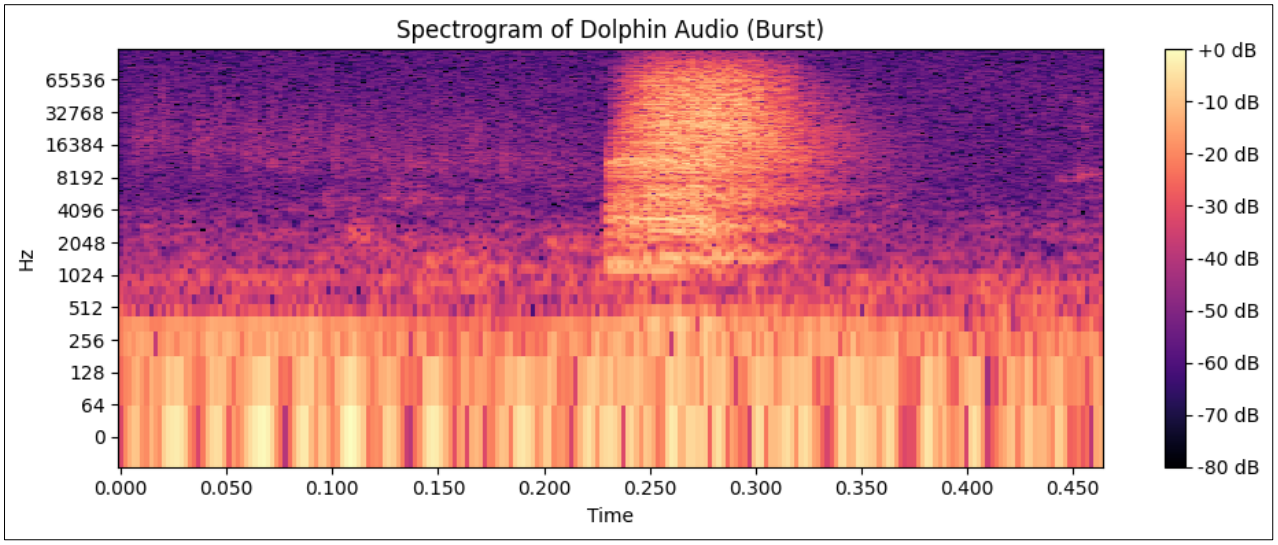

Figure 5. Spectrogram of Dolphin Audio – *Burst Pulse*. Shows densely packed acoustic pulses over a short duration, often overlapping with click patterns

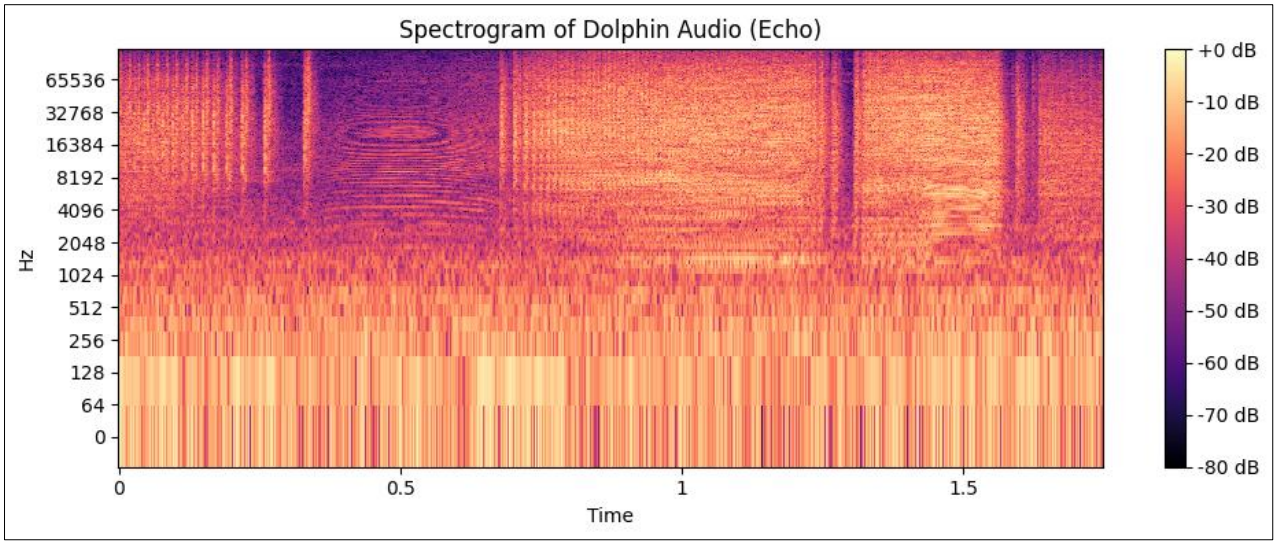

Figure 6. Spectrogram of Dolphin Audio – *Echo*. Illustrates delayed reflections of echolocation signals, appearing as irregular patterns following initial clicks

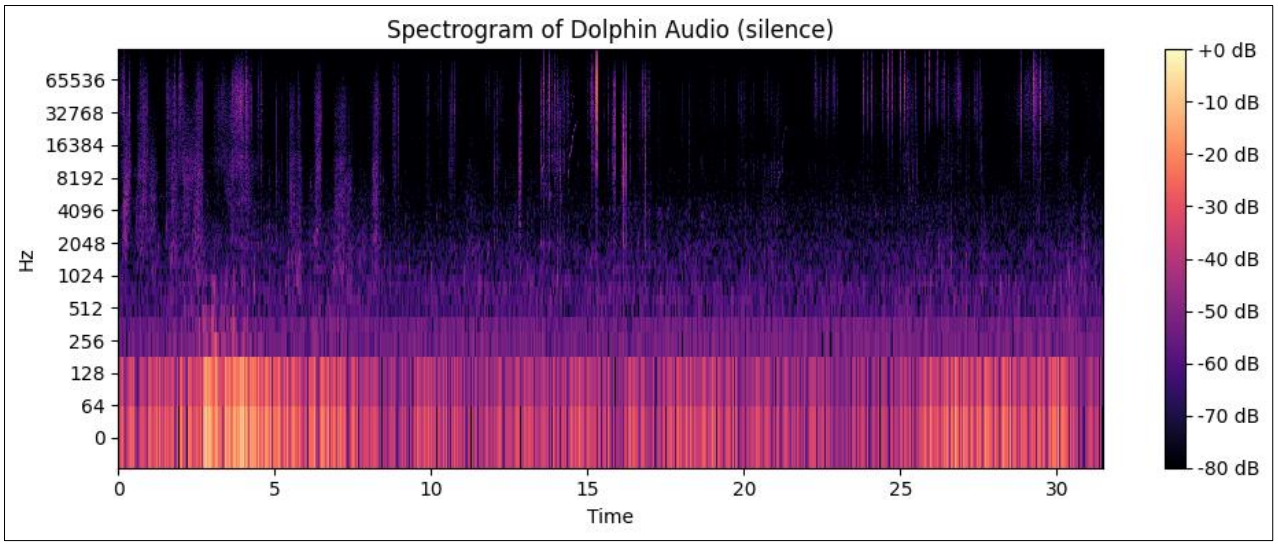

Figure 7. Spectrogram of Dolphin Audio – *Silence*. Represents the absence of structured vocalizations, showing minimal energy across the spectrum

7. Per-Class Performance Metrics

Figure 8 represents the per-class performance metrics (precision, recall, and F1-score) for each dolphin activity presented in Table 2. The results showed that the model did an effective job of differentiating between dolphin activities, especially silence. For silence, the model achieved a precision of 98.0%, a recall of 97.0%, and an F1-score of 97.5%, indicating the model accurately classified periods of no vocalizations. There was likely a clear distinction in acoustic signals where silence was easier to classify than the more complex vocalizations. In comparison, for burst pulses, the model performed a little lower and had a precision of 87.6%, a recall of 85.0%, and an F1-score of 86.3%. The less successful classification was likely due to spectrum overlap with other dolphin vocalizations and the model being sensitive in regards to clicks. Since burst pulses and clicks have acoustic similarities, there are occasions when the model confuses burst pulses with clicks during classification.

The model showed good performance for the other activity classes (whistle, click, and echolocation). For whistle detection, the precision was 93.8%, recall 92.0%, and F1-score 92.9%, indicating that the model was able to identify these vocalizations correctly 93.8% of the time. Click had a precision of 91.7%, recall 89.0%, and F1-score 90.3%, indicating good generalization, as there were also issues with detecting clicks together with other acoustic events. Echolocation activity was classified well as well, with precision 92.7%, recall 91.0%, and F1-score 91.8%. Again, this indicates that the model was able to accurately classify echolocation signals compared to other dolphin calls, likely due to the unique temporal and spectral patterns of echolocation. Overall, the model performed well for all activity classes; in particular, the detection of silence was performed very well, and the model performed reasonably well when detecting the challenging task of burst pulses. These results indicate that the value of combining visual and acoustic data was successful overall and will lead to robust dolphin behavior recognition.

Table 2. Variation of Performance by Dolphin Activity

Dolphin Activity	Precision (%)	Recall (%)	F1-Score (%)
Whistle	93.8	92	92.9
Click	91.7	89	90.3
Burst Pulse	87.6	85	86.3
Echolocation	92.7	91	91.8
Silence	98	97	97.5

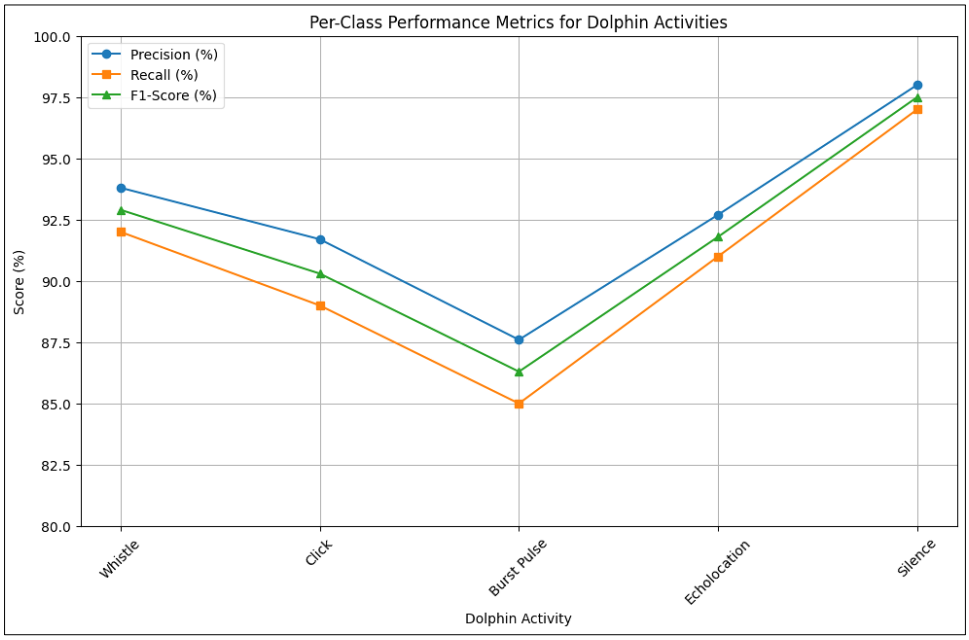

Figure 8. Per–Class Performance Metrics for Dolphin Activities

7.1. Linking Dolphin Acoustic Signals to Daily Behavioral Patterns

Dolphins are intelligent marine mammals that use a variety of acoustic signals, including clicks, whistles, burst sounds, and silence, to communicate and carry out daily activities. Each of these signals is paired to some extent with behavioral actions, and understanding those pairings helps the researcher interpret dolphin behaviors based on acoustic data. Dolphins produce clicks when they are traveling. The echolocation sounds they produce help navigate through the ocean and find out whether other dolphins are nearby. After dolphins produce a rhythmic clicking in order to create a map of their immediate surroundings, they can navigate better and avoid collisions when traveling. While socializing, dolphins produce signature whistles to communicate their identity to each other. Signature whistles function as names in dolphin pods; each dolphin produces a distinct signature whistle to denote it within a group. When dolphins are playing or competing against each other, they will also employ burst-pulsed sounds during social interactions to display their recognition of each other's presence and express strong feelings (i.e., excitement, antagonism, mild aggression, etc.). While foraging, dolphins will produce clicks to search for prey; they depend on the echoes of those clicks to find the location, size, and movement of the prey. This echolocation is an essential skill to help dolphins hunt prey in low-visibility waters since prey may often be camouflaged or hidden from view. Dolphins typically do not make vocalizations when resting. This time of acoustic silence is used to save energy and avoid detection by predators. Dolphins can only rest one hemisphere of their brain at a time, and when they do this, they remain semi-alert and can respond to danger if required. Playing is a highly social and cognitive behavior seen in dolphins. Playfulness usually involves burst sounds and whistles, and dolphins may perform physical acts, including rotating sideways through the water, playing at mock fighting, and playing with objects.

The sounds are useful for group bonding and learning. Dolphins produce whistles and burst sounds during mating. These sounds are often part of the courtship behavior, and they are usually in association with what can be described as synchronized swimming, body terrain, or light touches that facilitate reproduction. Hunting behaviors utilize a high rate of bursts and fast clicking sounds that may utilize the high speed of sound to form a precise echolocation of the prey to track or catch during a chase. This is a more active and concentrated state of foraging. Jumping behavior is a highly energetic display of bursts of pulsating sounds. Jumping behavior can also be produced with another dolphin in what's called a mount or leap. As with swimming through the air, this behavior may also indicate excitement or be a non-verbal form of communication amongst dolphins. Jumping behavior has both a social and functional role; the physical action may also be used to dislodge parasites or indicate a signal to the pod. Table 3 provides a systematic framework for how

dolphins relate different types of sounds to behaviors, allowing for classification of behaviors based on acoustics. This is particularly useful for datasets with sound recordings in which there may be behavioral monitoring or real-time activity recognition, which is important for marine research.

Table 3. Dolphin Activity-Sound-Behavior Mapping

S. No.	Activity		Sound Type	Behavior
1		TRAVELLING	CLICKS	Echolocation for navigation and detecting obstacles and other pod members during movement.
2		SOCIALIZING	SIGNATURE WHISTLE	Identification and social communication within the group.
			BURST SOUNDS	Expressing excitement, playfulness, or mild aggression in interactions.
3		FORAGING	CLICKS	Detecting and tracking prey using echolocation.
			ECHO	Interpreting reflected sound to locate prey (size, shape, distance).
4		RESTING	SILENCE	Reduced vocal activity; conserving energy and staying alert to surroundings.
5		PLAYING	BURST SOUNDS, WHISTLES	Engaging in fun or learning behaviors, often with physical contact or vocal mimicry.
6		MATING	WHISTLES, BURST SOUNDS	Mating communication or courtship communication, usually coupled with gentle movements or body gestures
7		HUNTING	CLICKS	Tracking prey with precise echolocation clicks during chase
8		JUMPING	BURST SOUNDS	Energetic leaps often linked to excitement or communication

7.2. Comparison with Traditional Machine Learning Models

The hybrid CNN-LSTM model was compared to traditional machine learning models using MFCC features only for the acoustic modality. As shown in Table 4, the proposed deep learning approach outperformed traditional models such as Support Vector Machine (SVM), Random Forest (RF), and k-Nearest Neighbors (k-NN) by a considerable margin in accuracy and average F1-score. All produced lower performance with every metric compared to the deep-learning-based method. The hybrid CNN-LSTM model produced an overall classification accuracy of 94.2%, an average F1-score of 91.7% and was effective at learning underlying complex patterns from spatial and temporal data streams. The SVM model had an accuracy of 85.6% and an average F1-score of 82.3%, while the Random Forest had an accuracy of 83.4% and an average F1-score of 80.1%. The k-NN classifier performed the poorest with an accuracy of 80.2% and an average F1-score of 77.5%.

These comparative findings show the clear benefit of the proposed deep learning architecture, which uses multimodal information and has a hierarchical feature extraction capability that can capture subtle variance in dolphin vocalizations and behaviors (see Figure 9).

Traditional models require handcrafted features and usually use shallow classification to recognize auditory data; in contrast, the CNN-LSTM framework benefits from the ability to automatically learn discriminant features and model sequential dependencies to accurately recognize dolphin sound. This important difference in accuracy and F1-score paves the way for deep learning models and the development of new automated dolphin monitoring systems.

Table 4. Comparative Results with Traditional ML Models

MODEL	ACCURACY (%)	AVG F1-SCORE (%)
Proposed CNN-LSTM	**94.2**	**91.7**
SVM	85.6	82.3
Random Forest	83.4	80.1
k-NN	80.2	77.5

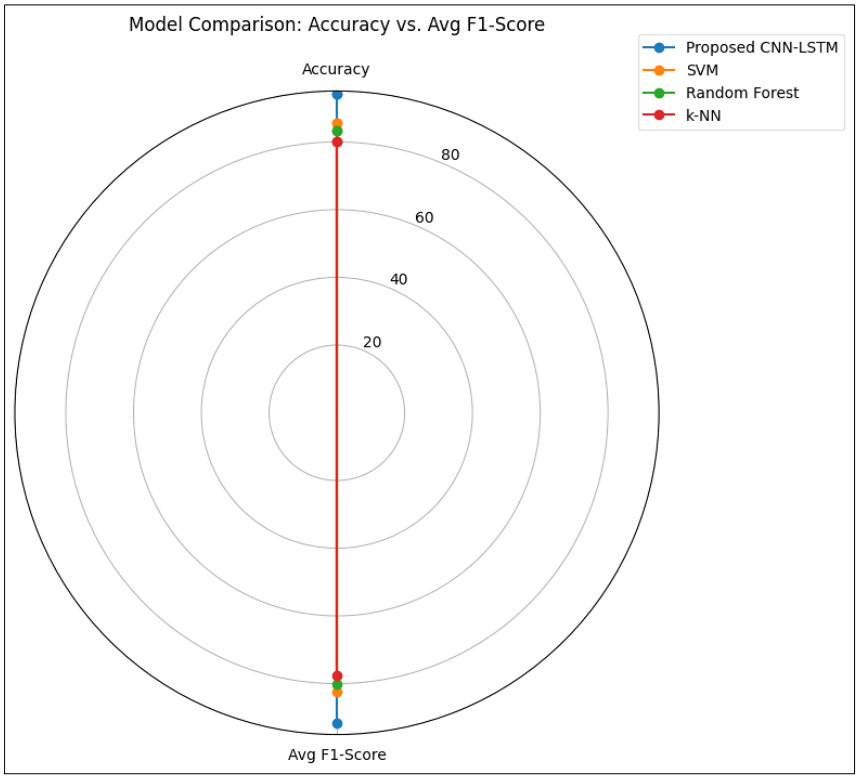

Figure 9. Model Comparison: Accuracy vs. Avg F1-Score

8. Discussion

The findings of our study provide several key conclusions about the classification performance of the proposed hybrid CNN-LSTM framework for dolphin activity recognition. First, silence was classified as the easiest and best classified activity. This class exhibited a near-perfect precision (98.0%) and near-perfect recall (97.0%). In many ways, this is not surprising. Since no spectral features can be defined for silence, these periods were separable from the

vocalization classes (clicks, whistles, and burst pulses). Furthermore, the dorsal fin images exhibited less movement for the periods that were classified as silence. Taken together, we reasonably conclude that the acoustic modal and visual modals provide strong, orthogonal features to classify inactivity or non-vocal states in dolphins.

On the other hand, the model struggled the most with burst pulse recognition. While the overall scores were relatively high, we found that the F1-score for burst pulses (86.3%) lagged behind the scores reported for other classes. The main reason is because of the spectral and temporal relationships between burst pulses and click sounds. Burst pulses and click sounds exist in similar frequency ranges and time-varying pulse trains, making it difficult to differentiate between the two sources, not only for automated systems, but even for human annotators. Further evidence of a similarity to click sounds is contributed by the confusion matrix, with burst pulses being primarily misclassified as click sounds. More sophisticated feature extraction measures may be warranted to capture these subtle audio differences, given that burst pulses (quick series of audible pulses) and click sounds (one or more audibly distinct pulse trains) have low time-frequency associations by design. Future studies may look at higher resolution time-frequency representations, as well as attention modeling pathways that focus on unique aspects of a signal.

One of the clearest findings in this study is the clear advantage of multimodal fusion in our hybrid CNN-LSTM model. The ability of our model to integrate spatial features present in dorsal fin images with temporal features present in underwater acoustics aided in the accuracy of the classification results, demonstrating a very large improvement relative to unimodal baselines. The visual features present in dorsal fin images appeared to work in conjunction with the acoustic features of the underwater sounds, providing visual cues that aided in resolving ambiguity when the model relied solely on the audio. For instance, while it is possible to identify predominantly physical behaviors such as jumping and diving when only using visual features, combining audio and visual features will improve recognition of these behaviors. Conversely, subtle behaviors that are primarily recognized by acoustic features benefited from the modeling of the temporal features provided by the LSTM layers. This novel coupling between visual and auditory modalities supports the fundamental premise of our research - that recognition of dolphin behaviors is maximized through understanding both the visual and acoustic dimensions in tandem. Overall, these findings highlight the power of using multimodal deep learning for marine mammal monitoring, in addition to marking out some unique challenges, such as burst pulse confusion that warrants another layer of exploration.

One of the main challenges was that the visual and acoustic recordings were from separate datasets rather than simultaneous field recordings. This meant that direct temporal alignment was impossible. Instead, the matching was done at the class level, if representative samples of a specific class (e.g., "jumping") in both datasets are adequate to train the network. While this class-based alignment performed admirably overall, it may not represent the entire complexity of real-world multimodal correlations, where timing, contextual context, and behavioral transitions might contribute variability. This is a restriction addressed for future work, where synced, co-recorded multimodal datasets may improve temporal loyalty.

The results of my research provide evidence that the proposed hybrid CNN-LSTM framework can contribute to the detection of activities performed by dolphins based on multimodal data sources. With an overall accuracy of 94.2%, the model provided better performance than classical machine learning approaches and traditional unimodal deep learning models. Analysis of the confusion matrix revealed that the model classified silence most accurately (97%) with no confusion, as this audio segment has a distinct acoustic profile from those instances of non-vocalization. Conversely, burst pulses were the most difficult class to predict, as they were often confused with clicks, likely due to their spectral-temporal characterizations being similar in nature. This suggests that further models will need to refine their acoustic feature extraction to identify even more finely grained representations of dolphins' behaviors. Exploring information on a per-class basis, the precision, recall, and F1-scores were consistently high for each activity, particularly for whistle and echolocation-related behaviors.

Overall, the results suggest that the combination of features obtained from the dorsal fin in video, combined with the dynamic vocal behaviors within MFCC sequences. The proposed hybrid deep learning approach outperformed any traditional machine learning approach (SVM, Random Forest, k-NN), up to 80-85% accuracy in behavior classification, by almost 10-14% accuracy. This highlights the advantages of deep multimodal architectures for complex behavior recognition, as single-modal systems often struggle with complexity. Interestingly, dorsal fin images improved recognition of not just visual activities (e.g., jumping or swimming) but also improved the recognition of acoustics, indicating that some behaviors may have less obvious postural cues that are built on vocal patterns. This represented a multimodal collaboration that is an interesting area for future marine mammal monitoring technologies.

The model showed significant generalization across the intrinsic heterogeneity in both datasets, which included variations in lighting, water quality, and background noise. Acoustic preprocessing methods, including noise reduction filters and uniform resampling, helped to reduce the variability produced by background noise and environmental interference. However, the findings revealed that certain classes, particularly burst pulses, were more sensitive to noisy or overlapping auditory environments, resulting in slightly reduced precision and recall. Visual features were somewhat resilient to depth-related lighting changes caused by augmentation, although extreme turbidity or occlusion was not well represented in the training set, so this remains a possible target for more robust testing.

In this study, spectrograms were created largely for visualization and interpretability, demonstrating the acoustic variations between whistles, clicks, burst pulses, echoes, and silence. However, they were not employed as an input mode for CNN. The decision to adopt MFCCs for the acoustic branch was based on its shown compactness and discriminative strength in modeling dolphin vocalizations, as validated by previous bioacoustics research. Nonetheless, using spectrogram pictures as a third modality or substituting MFCC input with spectrogram-based CNN processing could enrich the feature space, particularly for classes with small spectral variations, and is planned for further investigation in future model improvements.

9. Conclusion

This research presented an innovative hybrid deep learning framework that combines Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to recognize dolphin behaviors in a reliable and valid way in a multimodal format. The core contribution was to leverage both visual data—that arose from dorsal fin images—and sometimes acoustic data—taken from underwater recordings—to allow the combination of spatial and time-based features. The multi-modeling also allowed the system to learn patterns as they pertain to dolphin behaviors and increased classification performance compared to conventional non-deep learning approaches. The experimental results supported the investigation framework, including the CNN-LSTM experimental condition, which reached an overall accuracy of 94.2%. This represents a large improvement from the direct tracking of dolphin behaviors using traditional unimodal systems and classical machine learning techniques that might apply animal behavior research to those systems. The CNN-LSTM condition classified behaviors such as whistles, echolocation clicks, and silences with high accuracy. However, the study still revealed limitations when attempting to classify burst pulses, especially as their acoustic output was similar to click sounds. This limitation is likely to be a future opportunity for successive framework versions.

In addition to its technical contributions, this research provides a scalable and non-invasive means of monitoring marine mammals in their natural habitat. Through automated and accurate identification of dolphin behaviors, the developed framework has the potential for enabling more innovative ecological research and conservation efforts and reducing the need for invasive or disruptive observational methods. In the end, this work builds toward a larger goal of developing sustainable and ethical ways of learning about and protecting aquatic wildlife and helping us move toward more intelligent and automated marine monitoring systems.

9.1. Future Work

Several promising directions for future research have been identified to further improve and extend the current framework. One important direction involves in situ applications of the model, including real-time deployment on autonomous underwater vehicles (AUVs) or stationary monitoring systems such as buoys for continuous tracking of dolphin behaviors in natural environments. The framework could also be extended to other marine species, including pilot whales, orcas, and porpoises, by using species-specific datasets to adapt the model to different marine mammal populations. In addition, future studies may incorporate extra data modalities such as motion sensor data, GPS information, and environmental parameters including water temperature and salinity to enhance behavior recognition and contextual understanding. Exploring self-supervised or semi-supervised learning approaches could also reduce the dependency on large labeled datasets, enabling more scalable and adaptable implementations in new environments. Overall, continued advancements in multimodal analysis and deep learning for marine biology are expected to support the development of more intelligent, automated, and ethical monitoring systems, contributing to both marine conservation and a deeper understanding of aquatic ecosystems.

10. Declarations

10.1. Author Contributions

Conceptualization, T.N. and R.R.S.; methodology, T.N.; software, T.N.; validation, T.N., R.R.S., D.A.D., and T.B.K.; formal analysis, T.N.; investigation, T.N.; resources, R.R.S., D.A.D., and T.B.K.; data curation, T.N.; writing—original draft preparation, T.N.; writing—review and editing, R.R.S., D.A.D., and T.B.K.; visualization, T.N.; supervision, R.R.S.; project administration, R.R.S.; funding acquisition, D.A.D. All authors have read and agreed to the published version of the manuscript.

10.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

10.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

10.4. Institutional Review Board Statement

This study did not involve human participants or live animals requiring ethical approval. The research was conducted using publicly available dolphin image and underwater acoustic datasets, and therefore Institutional Review Board (IRB) or Ethics Committee approval was not required.

10.5. Informed Consent Statement

Not applicable.

10.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

11. References

[1] Trotter, C., Atkinson, G., Sharpe, M., Richardson, K., McGough, A. S., Wright, N., ... & Berggren, P. (2020). NDD20: A large-scale few-shot dolphin dataset for coarse and fine-grained categorisation. arXiv Preprint, arXiv:2005.13359. doi:10.48550/arXiv.2005.13359

[2] Duc, P. N. H. (2020). Development of artificial intelligence methods for marine mammal detection and classification of underwater sounds in a weak supervision (but) Big Data-Expert context. Doctoral dissertation, Sorbonne Université, Paris, France.

[3] Chen, J., Hu, M., Coker, D. J., Berumen, M. L., Costelloe, B., Beery, S., Rohrbach, A., & Elhoseiny, M. (2023). MammalNet: A Large-Scale Video Benchmark for Mammal Recognition and Behavior Understanding. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2023-June, 13052–13061. doi:10.1109/CVPR52729.2023.01254.

[4] Syed, M. A. Bin, & Ahmed, I. (2023). A CNN-LSTM Architecture for Marine Vessel Track Association Using Automatic Identification System (AIS) Data. Sensors, 23(14), 6400. doi:10.3390/s23146400.

[5] Yao, Q., Wang, Y., Yang, Y., & Shi, Y. (2023). Seal call recognition based on general regression neural network using Mel-frequency cepstrum coefficient features. Eurasip Journal on Advances in Signal Processing, 2023(1), 48. doi:10.1186/s13634-023-01014-1.

[6] Feng, R., Xu, J., Jin, K., Xu, L., Liu, Y., Chen, D., & Chen, L. (2023). An Automatic Deep Learning Bowhead Whale Whistle Recognizing Method Based on Adaptive SWT: Applying to the Beaufort Sea. Remote Sensing, 15(22), 5346. doi:10.3390/rs15225346.

[7] Licciardi, A., & Carbone, D. (2024). WhaleNet: A Novel Deep Learning Architecture for Marine Mammals Vocalizations on Watkins Marine Mammal Sound Database. IEEE Access, 3482117. doi:10.1109/ACCESS.2024.3482117.

[8] Hamard, Q., Pham, M. T., Cazau, D., & Heerah, K. (2024). A deep learning model for detecting and classifying multiple marine mammal species from passive acoustic data. Ecological Informatics, 84, 102906. doi:10.1016/j.ecoinf.2024.102906.

[9] Lin, J., Gui, D., Xie, Q., Zhou, X., & Shan, Y. (2024). Automated Detection and Recognition of Wild Dolphin Behaviors Using Deep Learning. Communications in Computer and Information Science, 2058 CCIS, 212–219. doi:10.1007/978-981-97-1277-9_16.

[10] Tseng, S. P., Hsu, S. E., Wang, J. F., & Jen, I. F. (2024). An Integrated Framework with ADD-LSTM and DeepLabCut for Dolphin Behavior Classification. Journal of Marine Science and Engineering, 12(4), 540. doi:10.3390/jmse12040540.

[11] Rattanarat, J., Jaroensutasinee, K., Jaroensutasinee, M., & Sparrow, E. B. (2025). Driving Mangrove Recovery: Community Engagement and Socio-Economic Shifts in Aquaculture Areas. Emerging Science Journal, 9(5), 2439–2453. doi:10.28991/ESJ-2025-09-05-09.

[12] Scaradozzi, D., De Marco, R., Li Veli, D., Lucchetti, A., Screpanti, L., & Di Nardo, F. (2024). Convolutional Neural Networks for Enhancing Detection of Dolphin Whistles in a Dense Acoustic Environment. IEEE Access, 12, 3454815. doi:10.1109/ACCESS.2024.3454815.

[13] Nihal, R. A., Yen, B., Shi, R., & Nakadai, K. (2025). Weakly Supervised Multiple Instance Learning for Whale Call Detection and Localization in Long-Duration Passive Acoustic Monitoring. arXiv e-prints, arXiv-2502. doi:10.48550/arXiv.2502.20838.

[14] Maglietta, R., Fanizza, C., Cherubini, C., Bellomo, S., Carlucci, R., & Dimauro, G. (2023). Risso's dolphin dataset. IEEE Dataport, March 20, 2023, doi:10.21227/rb8d-cd89.

[15] Zhivomirov, H., Nedelchev, I., & Dimitrov, G. (2020). Dolphins Underwater Sounds Database. IEEE Dataport, March 10, 2020, doi:10.21227/n00y-kq67.

[16] Liu, Y., Tee, M., Lu, L., Zhou, F., & Lu, B. (2025). High-Precision Urban Air Quality Prediction Using a LSTM-Transformer Hybrid Architecture. International Journal of Advanced Computer Science and Applications, 16(4), 299–305. doi:10.14569/IJACSA.2025.0160431.

[17] Li, D., Liao, J., Jiang, H., Jiang, K., Chen, M., Zhou, B., Pu, H., & Li, J. (2024). A classification method of marine mammal calls based on two-channel fusion network. Applied Intelligence, 54(4), 3017–3039. doi:10.1007/s10489-023-05138-7.

[18] Di Nardo, F., De Marco, R., Li Veli, D., Screpanti, L., Castagna, B., Lucchetti, A., & Scaradozzi, D. (2025). Multiclass CNN Approach for Automatic Classification of Dolphin Vocalizations. Sensors, 25(8), 2499. doi:10.3390/s25082499.

[19] Abdelaziz, A., Elhoseny, M., & Santos, V. (2025). Advancing Network Security: Integrating Salp Swarm Optimization with LSTM for Intrusion Detection. HighTech and Innovation Journal, 6(4), 1185–1219. doi:10.28991/HIJ-2025-06-04-05.

[20] Raza, A., Zongxin, S., Qiao, G., Javed, M., Bilal, M., Zuberi, H. H., & Mohsin, M. (2025). Automated classification of humpback whale calls in four regions using convolutional neural networks and multi scale deep feature aggregation (MSDFA). Measurement: Journal of the International Measurement Confederation, 255, 118038. doi:10.1016/j.measurement.2025.118038.

[21] Cheng, W., Chen, H., Jiang, J., Li, S., Wang, J., & Zhou, Y. (2025). Recognition and classification techniques of marine mammal calls based on LSTM and expanded causal convolution. Frontiers in Marine Science, 12. doi:10.3389/fmars.2025.1603090.