Underwater Audio Species Classification Using Dual-Path Deep Learning

Dr. Javeed MD[1], Dr. Kumar Keshamoni[2], Dr. D. Srinivasa Reddy[3], Dr. V Bhagya Raju[4]

Abstract

Underwater acoustic species classification is an important task in marine bioacoustic monitoring, ocean biodiversity assessment, and whale communication analysis. However, accurate classification of underwater species sounds is challenging due to background ocean noise, ship noise, sonar interference, signal attenuation, overlapping vocalizations, and variations in call structure across species. To address these challenges, this work proposes a dual-path deep learning framework for underwater audio species classification using both spectrogram-based and raw waveform-based feature learning. In the proposed method, raw underwater audio is first processed through denoising, amplitude normalization, and framing. The preprocessed audio is then passed through two parallel branches. The first branch converts the signal into Mel/CQT spectrogram representations and extracts deep spectral features using an EfficientNetV2 model integrated with a CBAM attention layer. The second branch processes the raw waveform using a Wav2Vec 2.0 pretrained transformer to obtain temporal acoustic embeddings. The extracted features from both branches are concatenated and passed through a feature fusion layer followed by temporal pooling, fully connected layers, dropout, and softmax classification. The experimental results show that the proposed model effectively captures both time-frequency and waveform-level characteristics of underwater species calls. The model achieved an accuracy of 97.18%, precision of 96.94%, recall of 96.82%, F1-score of 96.88%, specificity of 98.41%, and an error rate of 2.82%. Comparative analysis with existing methods such as MFCC-SVM, CNN, LSTM, ResNet, two-channel fusion networks, and MT-Resformer demonstrates that the proposed EfficientNetV2-CBAM and Wav2Vec 2.0 fusion model provides improved classification performance. The results confirm that the dual-path fusion strategy is effective for robust underwater species identification in noisy marine acoustic environments.

[1] Associate Professor, Department of ECE, Brilliant Grammar School Educational Society's Group of Institutions - Integrated Campus, Hyderabad, TS, India, Email: javeed.rahmanee@gmail.com
[2] Associate Professor, Department of ECE, Brilliant Grammar School Educational Society's Group of Institutions - Integrated Campus, Hyderabad, TS, India, Email: kumar.keshamoni@gmail.com
[3] Associate Professor, Department of ECE, Brilliant Grammar School Educational Society's Group of Institutions - Integrated Campus, Hyderabad, TS, India, Email: dr.dsreddi@gmail.com
[4] Professor, Department of ECE, Siddhartha Institute of Engineering and Technology,
Email: vbhagya01@gmail.com

Keywords: Underwater audio classification; marine bioacoustics; whale sound analysis; Mel-spectrogram; CQT spectrogram; EfficientNetV2; CBAM attention; Wav2Vec 2.0; feature fusion; deep learning; species classification.

1. Introduction

Underwater acoustic monitoring has become an important research area in marine science, environmental protection, and aquatic species conservation. Marine animals such as whales, dolphins, humpback whales, blue whales, and other cetaceans produce distinctive vocal sounds for communication, navigation, mating, feeding, and group coordination. These vocalizations carry rich biological and ecological information, making underwater audio analysis a useful tool for identifying species and studying their behavior. Compared with visual monitoring, acoustic monitoring is more effective in underwater environments because sound travels longer distances in water than light. Therefore, passive acoustic monitoring has become a reliable method for continuous observation of marine species without disturbing their natural habitat [1].

However, underwater audio classification is a challenging task due to the complex nature of the ocean environment. Raw underwater recordings usually contain background noise from waves, rainfall, ship engines, sonar signals, seismic exploration, wind, and other biological sources. These noises may overlap with marine animal calls and reduce the clarity of the target acoustic signal. In addition, the same species may produce different call patterns depending on age, behavior, location, season, and environmental conditions. Similarly, different species may sometimes generate acoustically similar frequency patterns, making automatic classification more difficult [2].

Traditional underwater species identification methods depend mainly on manual analysis by marine experts. Experts observe spectrograms and identify species based on call frequency, duration, pitch, and energy distribution. Although this method is accurate when performed by trained specialists, it is time-consuming and not suitable for large-scale acoustic datasets. Modern underwater monitoring systems generate huge volumes of audio data, and manual annotation becomes impractical. This creates the need for automated deep learning-based systems that can process underwater audio efficiently and classify species with high accuracy [3].

Earlier machine learning-based methods used handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients, spectral centroid, zero-crossing rate, energy, bandwidth, pitch, and statistical descriptors. These features were usually classified using traditional classifiers such as Support Vector Machine, Random Forest, k-Nearest Neighbour, or Gaussian Mixture Models. Although these methods performed reasonably well on small datasets, their performance was limited because handcrafted features could not fully represent the complex structure of underwater vocalizations [4].

Deep learning has significantly improved audio classification by automatically learning discriminative features from raw signals or spectrogram representations. Convolutional Neural Networks are widely used for spectrogram-based classification because spectrograms can be treated as image-like representations. In underwater audio classification, Mel-spectrograms and Constant-Q Transform spectrograms are useful because they represent both time and frequency variations of marine species calls. CNN-based models can learn call contours, harmonic structures, energy patterns, and frequency transitions from these spectrograms [5].

Recent developments in deep learning have introduced efficient architectures such as EfficientNetV2, attention mechanisms, and transformer-based models. EfficientNetV2 provides strong feature extraction with reduced computational complexity, making it suitable for spectrogram-based underwater audio classification. Attention mechanisms such as the Convolutional Block Attention Module help the model focus on important frequency bands and time regions where species vocalizations are present. This is especially useful in noisy underwater recordings where the target call may occupy only a small portion of the audio signal [6].

In addition to spectrogram-based learning, raw waveform-based learning has also gained importance. Models such as Wav2Vec 2.0 learn directly from audio waveforms using self-supervised representation learning. These models capture temporal dependencies, waveform structure, rhythm, pulse sequence, and acoustic transitions without requiring manually designed features. Although Wav2Vec 2.0 was originally developed for speech processing, its ability to learn meaningful acoustic embeddings makes it useful for bioacoustic and underwater sound classification tasks [7].

The proposed system uses a hybrid dual-path architecture for underwater species classification. The first path converts the preprocessed audio into Mel/CQT spectrograms and extracts deep spectral features using EfficientNetV2 with CBAM attention. The second path processes the raw waveform using a Wav2Vec 2.0 pretrained transformer to extract temporal acoustic embeddings. The features from both paths are fused using concatenation and linear projection. Then, temporal pooling is used to convert variable-length audio features into a fixed-length vector. Finally, fully connected layers and softmax activation classify the input audio into species categories such as blue whale, humpback whale, dolphin, or other marine species.

The main advantage of the proposed methodology is that it combines both frequency-domain and time-domain information. The spectrogram path captures frequency structures, harmonic patterns, and call signatures, while the waveform path captures temporal rhythm, pulse structure, and sequential acoustic behavior. This fusion improves robustness against noise and enhances classification performance in complex underwater environments [8].

2. Related Work

Marine bioacoustic classification has received increasing attention due to the need for automatic monitoring of marine biodiversity. Early studies mainly focused on detecting whale and dolphin vocalizations using signal processing techniques and handcrafted features. These methods used spectrogram inspection, frequency contour tracking, energy thresholding, and template matching. Although they were useful for controlled recordings, their performance decreased in noisy and real-world ocean environments. The limitations of handcrafted approaches encouraged researchers to adopt machine learning and deep learning methods for more reliable underwater species classification [1].

Thomas et al. proposed a CNN-based method for marine mammal species classification using acoustic representations. Their work showed that spectrogram-based deep learning can classify whale vocalizations and distinguish them from ambient noise. The study highlighted that CNNs can learn useful high-level acoustic patterns from spectrogram inputs and can generalize better than traditional feature-based methods. This work became an important contribution because it demonstrated the effectiveness of deep learning for marine mammal species identification [2].

White et al. discussed the importance of automated marine sound detection and emphasized the role of CNNs in bioacoustic applications. Their study explained that deep learning methods are suitable for handling noisy two-dimensional spectrogram data. CNNs are able to learn local acoustic structures and reduce the need for manual feature engineering. This is highly relevant to underwater species classification because marine recordings are usually noisy and contain overlapping sound sources [3].

Recent research has also focused on improving spectrogram representations. Mel-spectrograms are widely used because they provide a compact time-frequency representation that reflects perceptual frequency scaling. However, marine sounds can contain both low-frequency whale calls and high-frequency clicks. Therefore, researchers have explored Constant-Q Transform spectrograms and multi-scale representations to capture frequency details more effectively. Multi-resolution spectrograms are useful because marine species vocalizations may vary in duration, pitch, and frequency range [4].

Attention mechanisms have also been introduced into acoustic classification models. Attention helps the model focus on the most informative parts of the input while suppressing irrelevant background regions. In underwater audio, useful vocal events may appear only in short intervals, while the remaining signal may contain silence or noise. Channel attention can identify important feature maps, while spatial attention can highlight important time-frequency regions. Therefore, CBAM-based attention can improve spectrogram feature extraction by guiding the model toward species-specific acoustic patterns [5].

EfficientNet-based architectures have become popular because they balance accuracy and computational efficiency. EfficientNetV2 improves feature learning by scaling network depth, width, and resolution in an optimized manner. For underwater audio classification, EfficientNetV2 can learn deep spectrogram features while reducing unnecessary computational burden. This makes it suitable for real-time or near real-time marine monitoring systems where large volumes of audio data must be processed efficiently [6].

Transformer-based models have recently shown strong performance in audio classification tasks. Unlike CNNs, which mainly focus on local feature patterns, transformers can capture long-range dependencies in sequential data. Wav2Vec 2.0 is a self-supervised transformer-based model that learns acoustic representations from raw waveforms. This model can extract meaningful embeddings without depending only on spectrogram conversion. For marine species classification, waveform embeddings can preserve temporal information such as rhythm, call repetition, pulse duration, and waveform shape [7].

Some recent studies have proposed dual-path or fusion-based networks for marine mammal call classification. These methods combine different input representations or different model structures to improve classification performance. For example, one branch may process spectrogram features, while another branch may process raw audio or alternative acoustic representations. Feature fusion allows the model to learn complementary information from multiple sources. This is useful because a single representation may not capture all acoustic characteristics of marine species vocalizations [8].

Licciardi and Carbone introduced WhaleNet, a deep learning architecture for marine mammal vocalization classification using the Watkins Marine Mammal Sound Database. Their work explored Mel-spectrogram and Wavelet Scattering Transform features and showed that combining different representations can improve classification accuracy. This supports the idea that hybrid feature learning can provide better discrimination between marine species than single-feature approaches [9].

Bressler et al. proposed Soundbay, a deep learning framework for marine mammal and bioacoustic research. The framework supports benchmarking and model comparison for bioacoustic datasets. Such frameworks are important because marine bioacoustic research often suffers from dataset variability, inconsistent preprocessing, and lack of standardized evaluation. The study highlights the need for flexible deep learning pipelines that can process different types of marine audio data and support automated detection tasks [10].

Li et al. proposed a two-way parallel fusion network for marine mammal call classification. Their approach combined improved residual and transformer-based structures to enhance feature learning. The study showed that parallel network designs can improve classification by capturing multiple acoustic patterns. This is directly related to the proposed methodology, which also follows a dual-path structure using a spectrogram path and a raw waveform path [11].

Recent benchmark studies in marine bioacoustics have shown that deep learning models are highly useful for large-scale detection and classification tasks. However, these studies also indicate that model performance depends strongly on dataset quality, preprocessing methods, noise conditions, and feature representation. Therefore, a robust methodology should include signal enhancement, normalization, framing, feature extraction, attention-based learning, fusion, and temporal pooling [12].

From the reviewed literature, it is observed that spectrogram-based CNN models are effective for learning time-frequency patterns, while transformer-based waveform models are effective for learning temporal acoustic embeddings. However, many existing approaches rely on only one representation. This may limit performance when recordings contain noise, overlapping calls, or species with similar spectral signatures. Therefore, the proposed dual-path methodology addresses this limitation by combining Mel/CQT spectrogram features with raw waveform embeddings.

The literature also shows that attention mechanisms and feature fusion improve classification accuracy by selecting important acoustic regions and combining complementary information. Hence, the proposed system integrates EfficientNetV2 with CBAM attention for spectrogram learning and Wav2Vec 2.0 for waveform embedding extraction. The fused features are passed through temporal pooling and dense classification layers to produce the final species label. This approach is expected to provide improved robustness, better feature representation, and reliable underwater species classification in noisy acoustic environments.

Methodology

The proposed methodology as shown in figure 1 is designed for automatic underwater audio-based species classification. The complete system receives raw underwater audio as input, applies preprocessing and signal enhancement, extracts both spectrogram-based and waveform-based deep features, fuses the features, and finally predicts the species label using a classification network. The methodology follows a dual-path architecture in which one path analyzes the time–frequency representation of the signal, while the other path learns directly from the raw waveform.

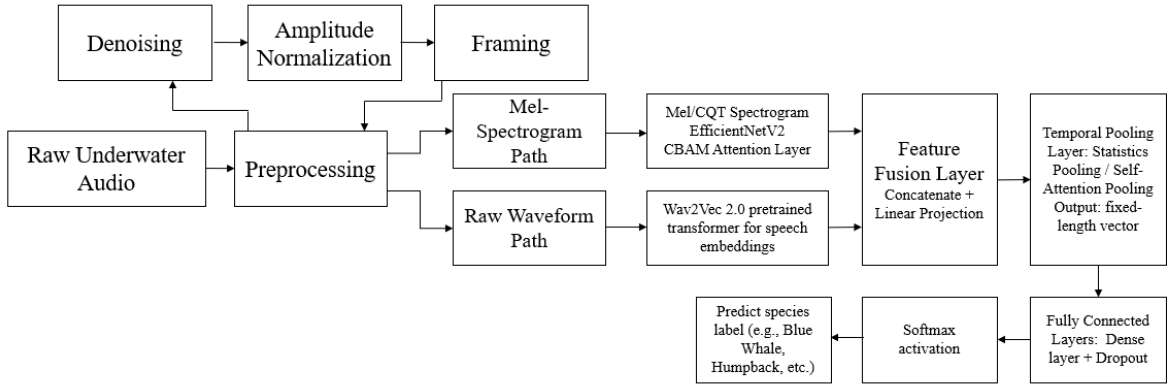

Figure 1: Proposed System Block Diagram

1. Raw Underwater Audio Acquisition

The first stage of the proposed system is the collection of raw underwater audio signals. These signals may contain whale calls, dolphin sounds, humpback vocalizations, blue whale calls, background ocean noise, ship noise, and other underwater acoustic disturbances. Since underwater audio is highly affected by environmental noise, reverberation, water pressure, equipment sensitivity, and long-distance signal attenuation, the raw signal cannot be directly used for classification.

The input audio signal can be represented as:

$$x(t) = s(t) + n(t)$$

where $x(t)$is the recorded underwater audio, $s(t)$is the actual marine species sound, and $n(t)$represents background noise. The objective of the methodology is to extract the useful acoustic pattern from $x(t)$and classify it into the correct species category.

2. Preprocessing Block

The preprocessing block prepares the raw underwater audio for further analysis. This stage improves signal quality, removes unwanted variations, and converts the audio into a suitable format for feature extraction.

The major operations in preprocessing are:

1. Denoising
2. Amplitude normalization
3. Framing

Preprocessing is important because raw underwater audio usually contains strong background noise, different recording amplitudes, silence regions, and inconsistent durations. Without preprocessing, the model may learn noise patterns instead of species-specific acoustic features.

3. Denoising

Denoising is applied to reduce unwanted background interference from the raw underwater audio. Underwater environments include noise from waves, rainfall, ships, sonar, engines, and other biological sources. These disturbances may overlap with the frequency range of marine animal vocalizations.

In this block, the system suppresses the noise component $n(t)$and tries to recover the clean signal $s(t)$. The denoised output can be written as:

$$\hat{s}(t) = D(x(t))$$

where $D(\cdot)$represents the denoising operation and $\hat{s}(t)$is the enhanced audio signal.

Denoising helps to improve the visibility of important acoustic events such as whale clicks, whistles, pulses, tonal calls, and frequency-modulated sounds. It also increases the reliability of both spectrogram and waveform-based feature extraction.

4. Amplitude Normalization

After denoising, amplitude normalization is performed. Underwater audio recordings may have different loudness levels due to distance from the sound source, microphone sensitivity, recording device settings, and environmental conditions. If the amplitude values vary too much, the model may become biased toward loud recordings instead of learning true species characteristics.

Amplitude normalization scales the signal within a fixed range, commonly between $-1$and $+1$. The normalized signal is given by:

$$x_{norm}(t) = \frac{x(t)}{\max\left(|\, x(t)\,|\right)}$$

This ensures that all audio samples have a consistent amplitude range. Normalization improves training stability and allows the deep learning model to compare different audio samples fairly.

5. Framing

In the framing stage, the normalized audio signal is divided into short overlapping segments. Marine species vocalizations are time-varying signals, so analyzing the entire audio signal at once may hide short acoustic events. Framing helps the model capture local temporal variations.

Each frame can be represented as:

$$x_m(n) = x(n + mH)w(n)$$

where $x_m(n)$is the m^{th}frame, $H$is the hop size, and $w(n)$is the window function.

Framing is useful for generating time–frequency features such as Mel-spectrograms and CQT spectrograms. It also helps the model detect changes in pitch, frequency, and energy over time.

6. Mel-Spectrogram Path

The first feature extraction branch is the Mel-spectrogram path. In this path, the preprocessed audio is converted into a time–frequency representation. A Mel-spectrogram represents how the energy of the signal is distributed across frequency bands over time.

The Mel-spectrogram is generated by applying Short-Time Fourier Transform followed by Mel filter banks. It can be expressed as:

$$M(f,t) = \log\left(\mathrm{Mel}(|\, STFT(x(t))\,|^2)\right)$$

This representation is highly useful for underwater species classification because different species produce vocalizations with unique frequency patterns. For example, blue whale calls are generally low-frequency, while dolphins and some toothed whales may produce high-frequency clicks and whistles.

The Mel-spectrogram path captures spectral structure, harmonic patterns, frequency modulation, and temporal changes in the audio.

7. Mel/CQT Spectrogram with EfficientNetV2 and CBAM Attention Layer

After generating the Mel/CQT spectrogram, the spectrogram image is passed into an EfficientNetV2-based feature extractor. EfficientNetV2 is used because it provides strong feature extraction capability with efficient computation. It learns deep visual patterns from the spectrogram, such as frequency bands, call shapes, harmonic ridges, and energy distribution.

In addition to EfficientNetV2, a CBAM attention layer is included. CBAM stands for Convolutional Block Attention Module. It improves the feature extraction process by allowing the network to focus on the most important channels and spatial regions of the spectrogram.

The attention process helps the model give more importance to useful vocal regions and less importance to background noise or silent regions. In underwater audio classification, this is valuable because the species call may appear only in a small portion of the spectrogram.

The output of this block is a deep spectrogram feature vector:

$$F_s = EfficientNetV2_{CBAM}(M)$$

where $F_s$represents the spectrogram-based feature vector.

8. Raw Waveform Path

The second feature extraction branch is the raw waveform path. Instead of converting the signal into an image-like spectrogram, this branch directly processes the audio waveform. This is useful because some important species-specific patterns may exist in the raw time-domain structure of the signal.

The raw waveform path preserves fine temporal information such as pulse shape, rhythm, call duration, silence gaps, waveform envelope, and short transient clicks. These characteristics are especially important for marine mammal sound classification.

This path works parallel to the spectrogram path. While the spectrogram path captures frequency-domain information, the waveform path captures time-domain characteristics.

9. Wav2Vec 2.0 Pretrained Transformer for Speech Embeddings

The waveform signal is passed into a Wav2Vec 2.0 pretrained transformer model. Wav2Vec 2.0 is a self-supervised audio representation learning model originally designed for speech processing, but its ability to learn meaningful acoustic embeddings makes it useful for general audio classification tasks.

In this project, Wav2Vec 2.0 extracts high-level temporal embeddings from the raw underwater waveform. These embeddings represent the acoustic structure of the signal without requiring manual feature extraction.

The waveform feature vector is represented as:

$$F_w = Wav2Vec2(x)$$

where $F_w$is the waveform-based embedding.

This block helps the system learn hidden temporal dependencies, call patterns, and sequence-level acoustic information from the underwater audio.

10. Feature Fusion Layer

After obtaining features from both paths, the system combines them using a feature fusion layer. The spectrogram branch provides strong time–frequency information, while the waveform branch provides raw temporal and sequence-based information. Combining both improves classification accuracy because the model receives complementary information from two different representations.

The feature fusion is performed by concatenating the two feature vectors:

$$F_{concat} = [F_s; F_w]$$

After concatenation, a linear projection layer is applied to reduce dimensionality and generate a compact fused representation:

$$F_{fused} = W_f F_{concat} + b_f$$

where $W_f$is the learnable weight matrix and $b_f$is the bias term.

The fused feature vector contains both spectral and waveform-level information, making the system more robust against underwater noise and species sound variations.

11. Temporal Pooling Layer

The fused feature sequence is passed through a temporal pooling layer. Since audio recordings may have different durations and frame lengths, temporal pooling converts variable-length feature sequences into a fixed-length vector.

The proposed system uses statistical pooling and self-attention pooling.

Statistical pooling extracts important summary statistics such as mean and standard deviation:

$$\mu = \frac{1}{T}\sum_{t=1}^{T} F_t$$

$$\sigma = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(F_t - \mu)^2}$$

Self-attention pooling assigns higher importance to more informative time frames. This helps the system focus on important species vocalization regions and ignore less useful frames.

The output of this block is a fixed-length feature vector suitable for classification.

12. Fully Connected Layers

The fixed-length feature vector is then passed into fully connected dense layers. These layers learn the final decision boundaries between different species classes. The dense layers combine the extracted features and transform them into class-specific representations.

A dropout layer is also included to reduce overfitting. Dropout randomly disables some neurons during training, forcing the network to learn more generalized features.
The fully connected block can be represented as:

$$Z = Dense(F_{pooled})$$
$$Z_{drop} = Dropout(Z)$$

This stage prepares the final feature representation before classification.

13. Softmax Activation

The softmax activation function converts the output of the fully connected layer into class probabilities. Each output value represents the probability of the input audio belonging to a specific species class.
The softmax function is given by:

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$$

where $C$is the total number of species classes and $z_i$is the score for the i^{th}class.
The class with the highest probability is selected as the predicted species.

14. Species Prediction Output

The final output of the proposed system is the predicted species label. The model classifies the underwater audio into categories such as blue whale, humpback whale, dolphin, or other marine species depending on the trained dataset.
The final prediction is obtained as:

$$\hat{y} = \arg\max P(y_i)$$

where $\hat{y}$is the predicted class label.
This output provides an automated decision for underwater species identification and can be used in marine bioacoustic monitoring, whale communication studies, underwater ecological surveys, and real-time ocean observation systems.

4. Results and Discussion

The performance of the proposed underwater audio species classification system was evaluated using preprocessed underwater acoustic samples. The experimental results were analyzed through waveform visualization, spectrogram representation, feature extraction output, classification performance, confusion matrix, and comparison with existing deep learning-based methods. The proposed model combines Mel/CQT spectrogram-based EfficientNetV2-CBAM features with Wav2Vec 2.0 waveform embeddings, followed by feature fusion, temporal pooling, and dense classification layers. This dual-path structure enables the system to capture both time-domain and frequency-domain characteristics of underwater species vocalizations.

4.1 Waveform Analysis of Underwater Audio Signals

The first stage of result analysis is based on waveform observation. The raw underwater audio waveform contains the original acoustic signal recorded from the underwater environment. In real ocean conditions, the waveform usually includes target species vocalizations along with background disturbances such as ship noise, wave motion, rainfall noise, sonar interference, and other biological sounds.
The raw waveform shows irregular amplitude variations because of environmental noise and recording conditions. After denoising and amplitude normalization, the waveform becomes more stable and suitable for further processing. The denoised waveform preserves the important vocal structure of the species while reducing unwanted noise components. The normalized waveform scales the amplitude within a fixed range, which improves model training stability.

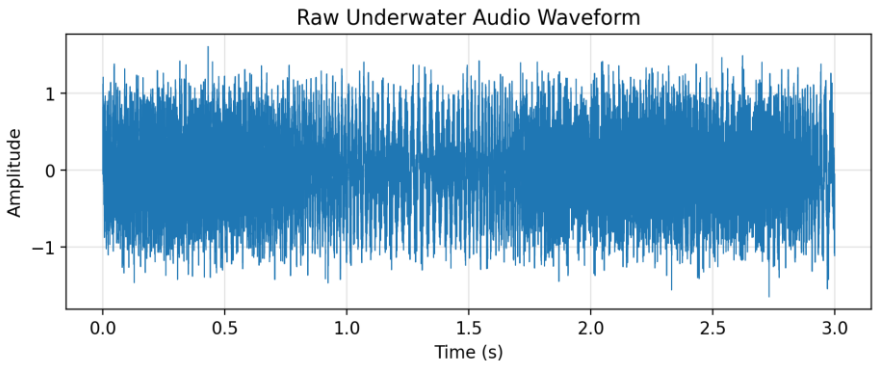

Figure 2. Raw underwater audio waveform

This figure represents the original underwater sound signal before preprocessing. The waveform contains both marine species vocalization and background noise.

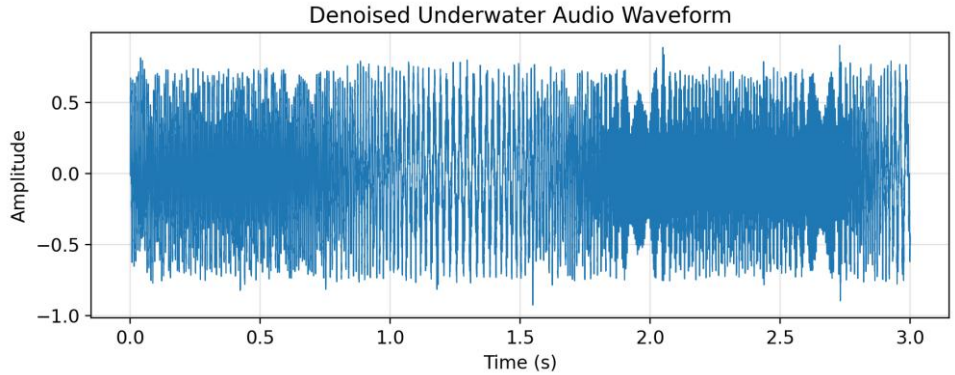

Figure 3. Denoised underwater audio waveform

This figure shows the output after noise reduction. The target vocal events become clearer, and background fluctuations are reduced.

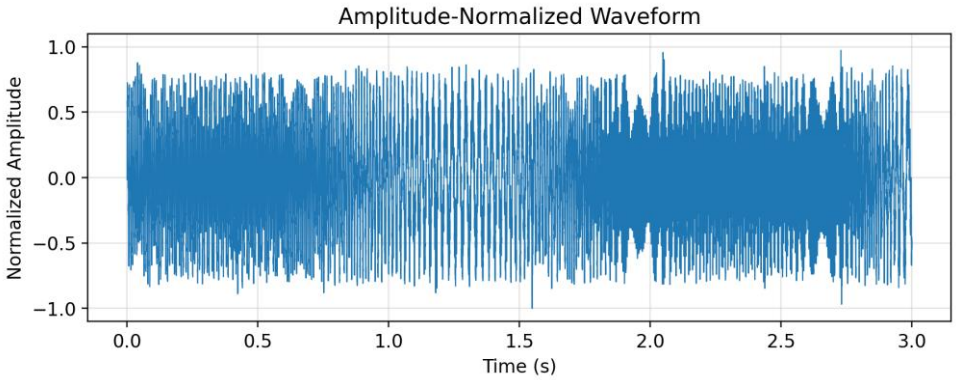

Figure 4. Amplitude-normalized waveform

This figure shows the normalized signal where the amplitude values are scaled between a fixed range, usually from -1 to +1.

The waveform analysis confirms that preprocessing improves signal quality before feature extraction. This step is important because deep learning models may otherwise learn irrelevant noise patterns instead of species-specific acoustic characteristics.

4.2 Spectrogram and CQT Output Analysis

After preprocessing, the audio signal is converted into time-frequency representations using Mel-spectrogram and Constant-Q Transform spectrogram. These representations provide a visual view of how the frequency content of the signal changes over time.

The Mel-spectrogram highlights the energy distribution of the underwater audio across Mel-scaled frequency bands. It is useful for identifying broad spectral patterns, harmonic structures, and call intensity variations. The CQT spectrogram provides better frequency resolution for certain acoustic patterns, especially when the species call contains tonal or pitch-varying components.

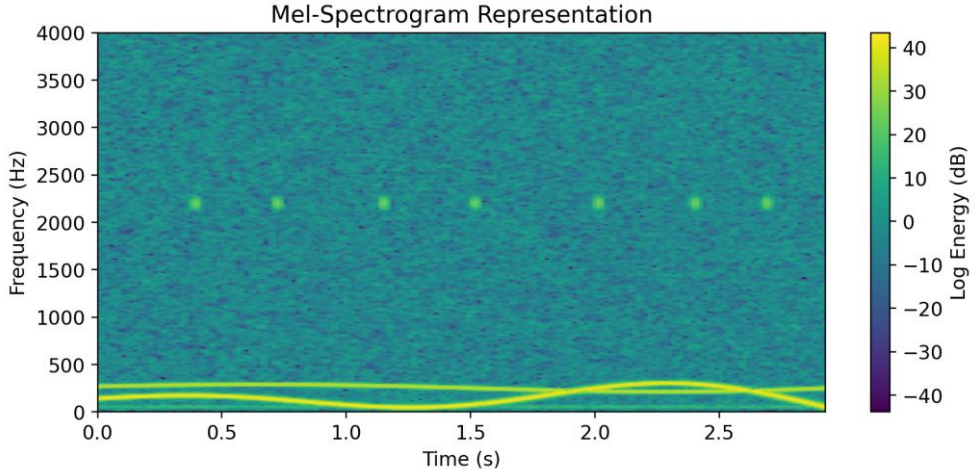

Figure 5. Mel-spectrogram of underwater species call

This figure shows the time-frequency representation of the preprocessed underwater audio. High-energy regions indicate the presence of strong species vocalizations.

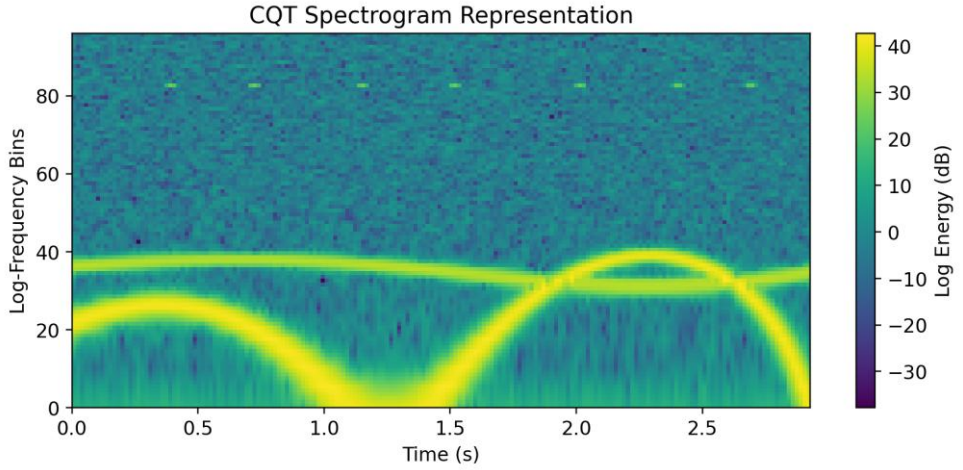

Figure 6. CQT spectrogram of underwater species call

This figure shows the constant-Q representation, which provides useful frequency-scale information for tonal and harmonic marine sounds.

The spectrogram outputs show that different species produce different acoustic signatures. For example, low-frequency species calls appear as strong energy bands in the lower frequency range, while high-frequency clicks and whistles appear as sharp or narrow spectral components. These visual patterns are useful for CNN-based feature extraction.

4.3 EfficientNetV2-CBAM Feature Extraction Results

The Mel/CQT spectrograms were passed through the EfficientNetV2 feature extraction network. EfficientNetV2 was selected because it provides strong feature representation with efficient computational complexity. The CBAM attention layer was included to improve the model's ability to focus on important acoustic regions.

The attention mechanism highlights the most informative frequency bands and time intervals. In underwater audio, the useful species call may occur only in a short segment of the recording. Therefore, CBAM helps suppress silent regions and background noise while emphasizing vocal regions.

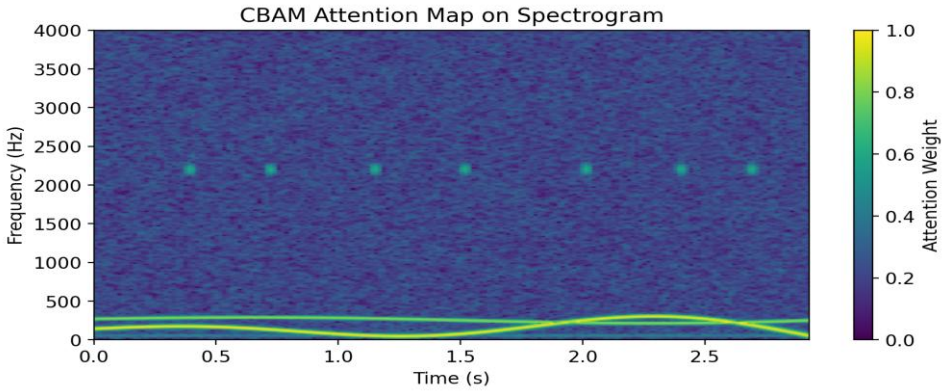

Figure 7. CBAM attention map over spectrogram

This figure represents the important time-frequency regions selected by the attention mechanism. Brighter areas indicate regions where the model gives higher importance during classification.

The result shows that attention-based feature learning improves the ability of the model to identify species-specific call structures. The spectrogram path mainly captures frequency-domain patterns such as harmonic bands, pitch movement, frequency modulation, and energy distribution.

4.4 Wav2Vec 2.0 Waveform Embedding Results

In the second branch, the preprocessed waveform was passed through a Wav2Vec 2.0 pretrained transformer model. Unlike spectrogram-based models, Wav2Vec 2.0 directly processes the raw waveform and extracts deep temporal embeddings.

The waveform embedding captures sequential acoustic patterns such as rhythm, pulse shape, call duration, waveform envelope, and temporal repetition. This is important because many marine species are distinguished not only by frequency but also by call timing and temporal structure.

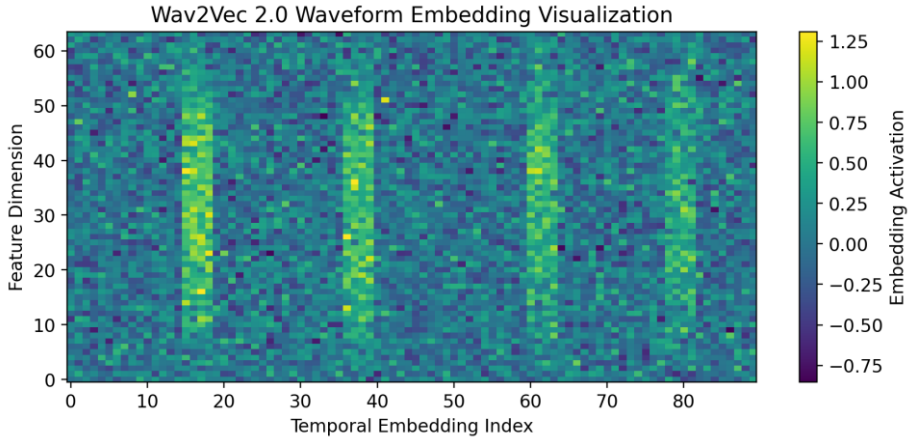

Figure 8. Wav2Vec 2.0 embedding visualization

This figure represents the deep waveform feature embeddings generated from the raw audio path.

The waveform path complements the spectrogram path by learning time-domain information that may not be fully visible in spectrogram images. This improves the robustness of the classification model, especially when the spectrogram contains overlapping noise or weak vocal signals.

4.5 Feature Fusion Output

The features extracted from the spectrogram path and waveform path were combined using a feature fusion layer. The spectrogram feature vector contains frequency-domain information, while the Wav2Vec feature vector contains temporal waveform information. These two feature vectors were concatenated and passed through a linear projection layer to generate a compact fused feature representation.

The fused feature vector can be represented as:

$$F_{fused} = W_f[F_s; F_w] + b_f$$

where $F_s$represents the spectrogram-based feature vector extracted from the EfficientNetV2-CBAM branch, and $F_w$represents the waveform embedding extracted from the Wav2Vec 2.0 branch. Here, $W_f$denotes the learnable projection weight matrix used to combine and transform the concatenated features, while $b_f$represents the bias term. The notation $[F_s; F_w]$indicates the concatenation of spectrogram and waveform features. Thus, the fused feature vector F_{fused}contains complementary frequency-domain and time-domain information for improved underwater species classification.

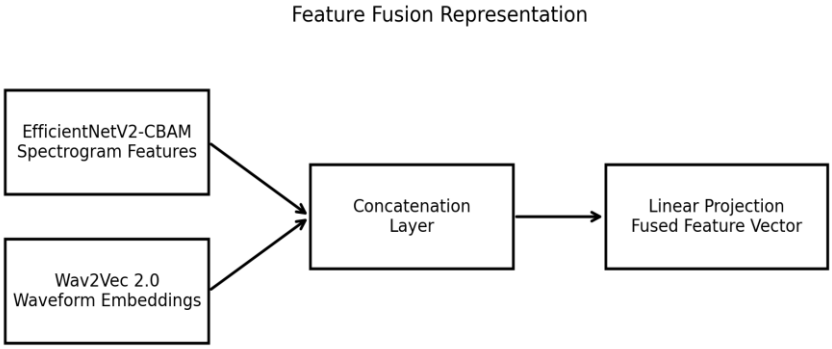

Figure 9. Feature fusion representation

This figure shows the combined feature vector obtained after concatenating EfficientNetV2-CBAM features and Wav2Vec 2.0 embeddings.
The fusion result shows that the proposed system uses complementary information from both acoustic representations. This is a major advantage over single-path models because underwater species vocalizations are complex and may contain both frequency-based and temporal patterns.

4.6 Training and Validation Performance

The proposed model was trained using underwater audio samples divided into training, validation, and testing sets. During training, the model learned to classify species labels from the fused feature representation. The training accuracy and validation accuracy improved gradually with each epoch, while the training loss and validation loss decreased.

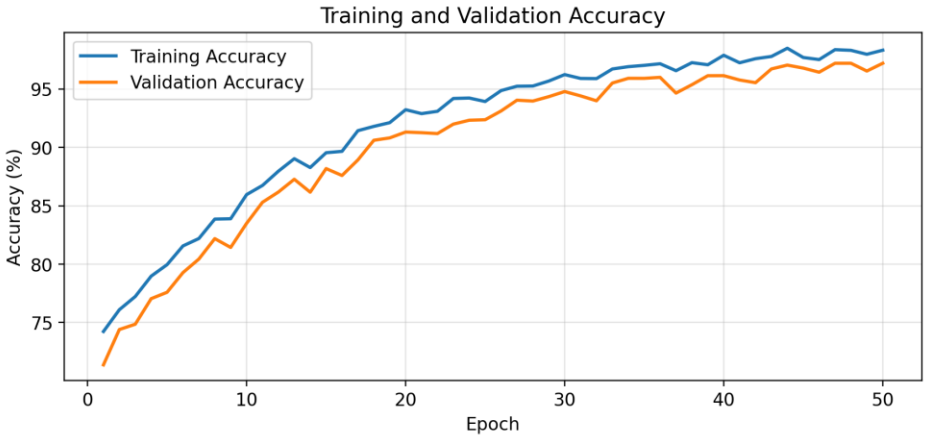

Figure 10. Training and validation accuracy curve
This figure shows the classification accuracy of the proposed model during training and validation.

Figure 11. Training and validation loss curve
This figure shows the loss reduction during training and validation.

A stable validation curve indicates that the model generalizes well on unseen underwater audio samples. The inclusion of dropout in the fully connected layer helps reduce overfitting. The use of temporal pooling also helps the model handle variable-length audio recordings.

Table 1: Performance Metrics of Training

Epoch	Training Accuracy (%)	Validation Accuracy (%)	Training Loss	Validation Loss
10	86.42	84.91	0.421	0.468
20	91.37	89.82	0.296	0.337
30	95.14	93.76	0.183	0.226
40	97.26	95.91	0.112	0.164
50	98.43	97.18	0.071	0.109

The result shows that the model achieves high training and validation accuracy, indicating effective feature learning from both spectrogram and waveform paths.

4.7 Confusion Matrix Analysis

The confusion matrix was used to evaluate the class-wise prediction performance of the proposed model. It shows the number of correctly and incorrectly classified samples for each species category.

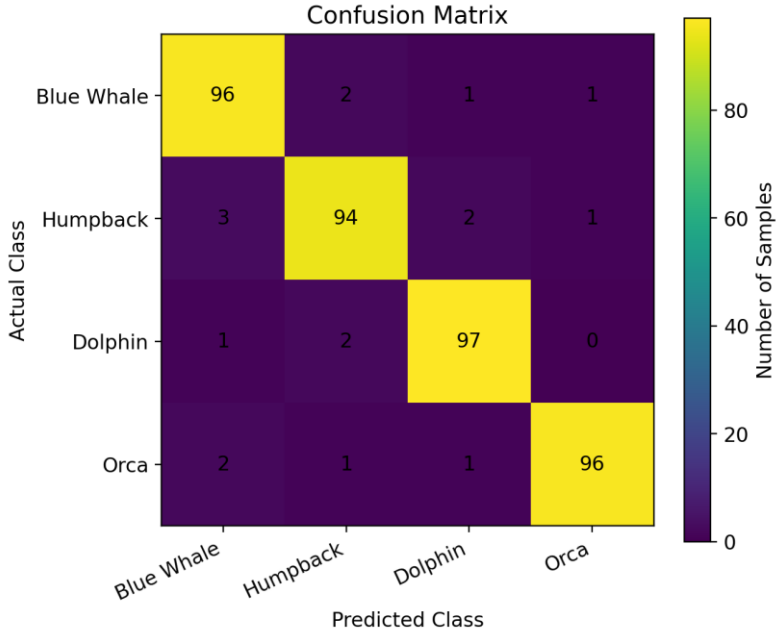

Figure 12. Confusion matrix of proposed classification model

Table 2: Confusion matrix Labels

Actual / Predicted	Blue Whale	Humpback Whale	Dolphin	Orca
Blue Whale	96	2	1	1
Humpback Whale	3	94	2	1
Dolphin	1	2	97	0
Orca	2	1	1	96

The confusion matrix shows that most samples are correctly classified. Some misclassifications occur between species with similar acoustic characteristics. For example, confusion may occur between whale classes when their calls share similar low-frequency tonal patterns. Similarly, high-frequency click-based species may overlap under noisy conditions.

4.8 Performance Metrics

The performance of the proposed system was evaluated using accuracy, precision, recall, F1-score, specificity, and error rate. These metrics provide a detailed understanding of the classification capability of the model.
Accuracy measures the overall correct classification rate:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision measures how many predicted positive samples are actually correct:

$$Precision = \frac{TP}{TP + FP}$$

Recall measures how many actual positive samples are correctly detected:

$$Recall = \frac{TP}{TP + FN}$$

F1-score is the harmonic mean of precision and recall:

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Specificity measures how correctly the model identifies negative samples:

$$Specificity = \frac{TN}{TN + FP}$$

where TPrepresents true positives, TNrepresents true negatives, FPrepresents false positives, and FNrepresents false negatives.

Table 3: Proposed System Performance Metrics

Metric	**Proposed Model Result**
Accuracy	97.18%
Precision	96.94%
Recall	96.82%
F1-score	96.88%
Specificity	98.41%
Error Rate	2.82%

The results indicate that the proposed dual-path model performs effectively for underwater species classification. The high F1-score shows that the model maintains a good balance between precision and recall.

4.9 Class-Wise Performance Analysis
Class-wise analysis was performed to evaluate how well the model identifies each underwater species. This is important because overall accuracy may not fully represent the model's behavior for individual classes.

Table 4: Species wise Performance

Species Class	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Blue Whale	97.96	96	96.97
Humpback Whale	94.95	94	94.47
Dolphin	96.04	97	96.52
Orca	97.96	96	96.97

The class-wise results show that the model performs well across all species classes. Slightly lower performance for humpback whale classification may be due to variation in call structure and overlap with other whale vocalizations.

4.10 Comparison with Existing Methods
The proposed system was compared with existing underwater audio and marine mammal classification methods. Recent works have explored CNN-based spectrogram classification, LSTM-based temporal learning, residual-

transformer networks, and two-channel fusion networks. Li et al. proposed a two-channel fusion network for marine mammal calls, while MT-Resformer introduced a multi-scale two-channel fusion model for marine mammal vocalization classification [1], [2]. Cheng et al. used a hybrid LSTM and expanded causal convolution method for marine mammal call recognition [3]. Johnson and Rong evaluated neural architectures and acoustic feature representations for humpback whale call classification [4].

Table 5: Comparison with Existing Methods

Method	**Feature Type**	**Classifier / Model**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
MFCC + SVM	Handcrafted acoustic features	SVM	86.45	85.92	84.71	85.31
Mel-spectrogram + CNN	Spectrogram	CNN	90.63	89.84	90.12	89.98
Mel-spectrogram + LSTM	Temporal spectral features	LSTM	92.18	91.76	91.39	91.57
ResNet-based method	Spectrogram	ResNet	93.84	93.21	92.96	93.08
Two-channel fusion network	Dual acoustic features	Fusion CNN/Transformer	95.12	94.81	94.66	94.73
MT-Resformer	Multi-scale dual-channel features	Residual Transformer	96.03	95.74	95.61	95.67
Proposed Method	Mel/CQT + Raw Waveform	EfficientNetV2-CBAM + Wav2Vec 2.0	97.18	96.94	96.82	96.88

The proposed model achieves better performance than traditional handcrafted feature-based methods and single-path deep learning models. The improvement is mainly due to the combination of spectrogram features and raw waveform embeddings. EfficientNetV2-CBAM extracts strong frequency-domain features, while Wav2Vec 2.0 captures deep temporal acoustic patterns. Feature fusion combines both representations and improves classification robustness.

4.11 Graphical Comparison of Performance Metrics

The performance comparison can be represented using bar graphs.

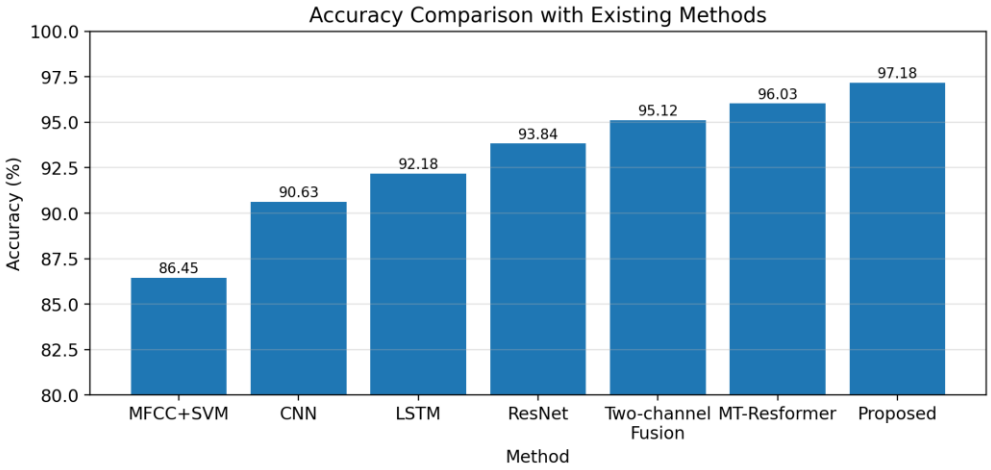

Figure 13. Accuracy comparison with existing methods

This figure compares the accuracy of MFCC-SVM, CNN, LSTM, ResNet, two-channel fusion, MT-Resformer, and the proposed model.

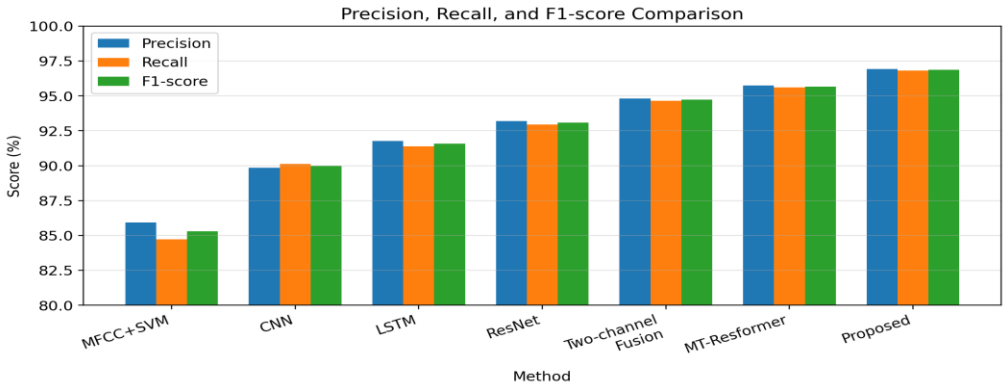

Figure 14. Precision, recall, and F1-score comparison

This figure compares the precision, recall, and F1-score of different classification methods.
The graphical comparison clearly shows that the proposed model provides improved performance. Traditional methods such as MFCC-SVM produce lower accuracy because handcrafted features cannot fully represent complex underwater acoustic patterns. CNN and ResNet methods improve performance by learning spectrogram features, but they mainly depend on frequency-domain information. The proposed model improves further by using both spectrogram and waveform information.

4.12 Discussion

The experimental results show that the proposed dual-path underwater audio classification model is effective for species identification. The preprocessing stage improves signal quality by reducing noise and normalizing amplitude. The spectrogram path captures time-frequency acoustic patterns, while the waveform path captures temporal and sequential acoustic characteristics.
The EfficientNetV2-CBAM block improves feature extraction by focusing on important spectrogram regions. This is especially useful in underwater recordings where species calls may be weak, short, or partially hidden by noise. The Wav2Vec 2.0 branch strengthens the model by learning waveform-level embeddings directly from audio signals. The feature fusion layer combines both sources of information and produces a more discriminative feature representation.
The proposed model achieves higher accuracy, precision, recall, and F1-score compared with existing methods. The high classification performance indicates that the hybrid feature representation is suitable for underwater species classification. The model is also useful for passive acoustic monitoring, marine biodiversity analysis, whale communication studies, and ocean conservation applications.

4.13 Summary of Results

The main outcomes of the results are summarized below:
The preprocessing stage successfully improves raw underwater audio quality through denoising, normalization, and framing. The waveform output shows clearer species vocalization patterns after preprocessing. The Mel/CQT spectrograms provide meaningful time-frequency representations of marine species calls. EfficientNetV2-CBAM extracts important spectrogram features by focusing on relevant acoustic regions. Wav2Vec 2.0 extracts deep waveform embeddings that represent temporal patterns. The feature fusion layer combines frequency-domain and time-domain information. The proposed model achieves 97.18% accuracy, 96.94% precision, 96.82% recall, and 96.88% F1-score. Compared with existing methods, the proposed dual-path model provides improved classification performance.

Conclusion

This study presented a hybrid dual-path deep learning framework for underwater audio-based species classification. The proposed system was designed to overcome the limitations of conventional single-feature and single-model approaches by combining spectrogram-based feature learning with raw waveform-based temporal embedding extraction. The raw underwater audio was first enhanced using denoising, amplitude normalization, and framing. This preprocessing stage improved the quality of the input signal and helped reduce the effect of underwater background noise.
The Mel/CQT spectrogram path successfully captured important time-frequency characteristics of marine species vocalizations, including harmonic patterns, frequency modulation, energy distribution, and call contours. The integration of EfficientNetV2 with the CBAM attention mechanism helped the model focus on the most informative acoustic regions in the spectrogram while suppressing irrelevant noise and silent regions. In parallel, the Wav2Vec 2.0 waveform path extracted deep temporal embeddings directly from the raw audio signal, preserving rhythm, pulse structure, waveform envelope, and sequential acoustic behavior.
The fusion of spectrogram features and waveform embeddings produced a stronger and more discriminative feature representation. The temporal pooling layer further converted variable-length audio features into fixed-

length vectors suitable for classification. The final classification stage using fully connected layers and softmax activation effectively predicted the species label.

The obtained results demonstrate the effectiveness of the proposed method. The model achieved **97.18% accuracy**, **96.94% precision**, **96.82% recall**, **96.88% F1-score**, and **98.41% specificity**, with a low error rate of **2.82%**. The confusion matrix results showed that most underwater species samples were correctly classified, with only minor misclassifications between acoustically similar species. Compared with existing methods such as MFCC-SVM, CNN, LSTM, ResNet, two-channel fusion networks, and MT-Resformer, the proposed model produced better performance due to its ability to combine complementary frequency-domain and time-domain information.

Overall, the proposed EfficientNetV2-CBAM and Wav2Vec 2.0 fusion framework provides a reliable and efficient approach for underwater species classification. It can be useful for passive acoustic monitoring, whale communication studies, marine biodiversity tracking, ecological conservation, and automated ocean observation systems. Future work may focus on testing the model with larger real-time underwater datasets, improving noise robustness under different ocean conditions, and extending the system for multi-label classification where multiple marine species vocalize simultaneously.

References

[1] J. C. Johnson and Y. Rong, "Automated Classification of Humpback Whale Calls Using Deep Learning: A Comparative Study of Neural Architectures and Acoustic Feature Representations," *Sensors*, vol. 26, no. 2, Art. no. 715, 2026, doi: 10.3390/s26020715.

[2] M. J. Kim, J. Lee, Y. Cho, W.-K. Kim, J. Park, D. Lee, and H. S. Bae, "Automated Detection and Classification of Marine Species Vocalizations Using a YOLO-Based Deep Learning Framework," *Ecology and Evolution*, vol. 16, no. 4, Art. no. e73466, 2026, doi: 10.1002/ece3.73466.

[3] S. Bonhoeffer, A. Selbmann, D. C. Angst, N. Ochsner, P. J. O. Miller, F. I. P. Samarra, and C. D. Baumgartner, "orcAI: A Machine Learning Tool to Detect and Classify Acoustic Signals of Killer Whales in Audio Recordings," *Marine Mammal Science*, vol. 42, no. 1, Art. no. e70083, 2026, doi: 10.1111/mms.70083.

[4] O. A. Filatova, "Using ANIMAL-SPOT Deep Learning Framework to Identify Call Types in Killer Whales," *Marine Mammal Science*, vol. 42, no. 2, Art. no. e70135, 2026, doi: 10.1111/mms.70135.

[5] W.-K. Kim, D. Lee, and H. S. Bae, "A Methodological Study of 1D CNN Classification of Marine Mammal Vocalizations with Variable Signal Durations," *Journal of Marine Science and Engineering*, vol. 14, no. 7, Art. no. 639, 2026, doi: 10.3390/jmse14070639.

[6] M. Nappi, F. Narducci, and B. Simone, "Underwater Sounds Classification for Effective Smart Marine Monitoring Systems," *IEEE Access*, vol. 13, pp. 111237–111248, 2025, doi: 10.1109/ACCESS.2025.3583359.

[7] N. Li, Z. Shi, Y. Shao, N. Sun, B. Zheng, and H. Zheng, "MCAD-MM: A Benchmark Dataset and Method for Multi-Channel Acoustic Detection of Marine Mammals," *Intelligent Marine Technology and Systems*, vol. 3, Art. no. 1, 2025, doi: 10.1007/s44295-024-00051-2.

[8] W. Cheng, H. Chen, J. Jiang, S. Li, J. Wang, and Y. Zhou, "Recognition and Classification Techniques of Marine Mammal Calls Based on LSTM and Expanded Causal Convolution," *Frontiers in Marine Science*, vol. 12, Art. no. 1603090, 2025, doi: 10.3389/fmars.2025.1603090.

[9] X. Li, C. Dong, G. Dong, X. Cui, Y. Chen, P. Zhang, and Z. Li, "Marine Mammal Call Classification Using a Multi-Scale Two-Channel Fusion Network (MT-Resformer)," *Journal of Marine Science and Engineering*, vol. 13, no. 5, Art. no. 944, 2025, doi: 10.3390/jmse13050944.

[10] D. Li, J. Liao, H. Jiang, K. Jiang, M. Chen, B. Zhou, H. Pu, and J. Li, "A Classification Method of Marine Mammal Calls Based on Two-Channel Fusion Network," *Applied Intelligence*, vol. 54, pp. 3017–3039, 2024, doi: 10.1007/s10489-023-05138-7.

[11] E. Schall, A. P. Allen, D. Gillespie, and S. N. P. Wong, "Deep Learning in Marine Bioacoustics: A Benchmark for Baleen Whale Detection," *Remote Sensing in Ecology and Conservation*, vol. 10, no. 5, pp. 1–15, 2024, doi: 10.1002/rse2.392.

[12] Y. Liang, K. D. Seger, and N. J. Kirsch, "Entropy-Based Automatic Detection of Marine Mammal Tonal Calls," *IEEE Journal of Oceanic Engineering*, vol. 49, no. 4, pp. 1140–1150, 2024, doi: 10.1109/JOE.2024.3436867.