frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 23 May 2025
DOI 10.3389/fmars.2025.1603090

OPEN ACCESS

EDITED BY
Zhibin Yu,
Ocean University of China, China

REVIEWED BY
Chengbo Wang,
University of Science and Technology of China, China
Xuerong Cui,
China University of Petroleum (East China), China
Jianmin Yang,
Sun Yat-sen University, China

*CORRESPONDENCE
Yanping Zhou
ypzhou@qust.edu.cn

RECEIVED 31 March 2025
ACCEPTED 30 April 2025
PUBLISHED 23 May 2025

CITATION
Cheng W, Chen H, Jiang J, Li S, Wang J and Zhou Y (2025) Recognition and classification techniques of marine mammal calls based on LSTM and expanded causal convolution. *Front. Mar. Sci.* 12:1603090. doi: 10.3389/fmars.2025.1603090

Recognition and classification techniques of marine mammal calls based on LSTM and expanded causal convolution

Wanlu Cheng[1,2], Hao Chen[3], Jiaming Jiang[1,2], Shuang Li[1,2], Jingjing Wang[1,2] and Yanping Zhou[1*]

[1]School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, China, [2]Shandong Key Laboratory of Deep Sea Equipment Intelligent Networking, Qingdao, China, [3]School of Mechanical Engineering, Ilmenau University of Technology, Ilmenau, Germany

Marine mammal calls play a vital role in navigation, localization, and communication. Effectively classifying these calls is essential for ecological monitoring, species conservation, and military biomimetic applications. However, traditional machine learning methods struggle to capture complex acoustic patterns, while most existing deep learning approaches rely solely on frequency-domain features and require large datasets, which limits their performance on small-scale marine mammal datasets. To address these challenges, we propose a hybrid architecture combining a time-attention Long Short-Term Memory (LSTM) network and a multi-scale dilated causal convolutional network. The model comprises three modules: (1) a frequency-domain feature extraction module employing dilated causal convolutions at multiple scales to capture multi-resolution spectral information from Mel spectrograms; (2) a time-domain feature extraction module that inputs Mel-frequency cepstral coefficients (MFCCs) into an LSTM enhanced with a time-attention mechanism to highlight key temporal features; and (3) a classification module leveraging transfer learning, where a pre-trained neural network is fine-tuned on real marine mammal call data to improve performance. Extensive experiments were conducted on vocalizations from four marine mammal species. Our proposed method outperformed existing baseline models across four evaluation metrics: accuracy, precision, recall, and F1 score, with improvements of 3%, 7%, 2%, and 4%, respectively. The results confirm the effectiveness of combining frequency- and time-domain features along with attention mechanisms and transfer learning. This hybrid approach enhances the accuracy and robustness of marine mammal call classification, especially under limited data conditions.

KEYWORDS

marine mammals, marine mammal call recognition and classification, transfer learning, LSTM, expansive causal convolutional networks

1 Introduction

The calls of marine mammals are crucial for communication, localization, and navigation within the marine ecosystem. These sounds not only contain rich information and features but also reflect their behavioral patterns, population distribution, and surrounding ecological conditions. By recognizing, monitoring, and analyzing marine mammal calls, we can gain a deeper understanding of their lifestyles, population numbers, and whether conservation measures are necessary Duan et al. (2022). Therefore, researching and developing techniques for the recognition and classification of marine mammal calls is of significant importance for protecting the marine environment, preserving marine biodiversity, and advancing the construction of maritime power.

The ability to accurately identify and classify these calls can also help us understand the impacts of environmental changes, including noise pollution and climate change, on marine life. This is especially important as some species, such as whales and dolphins, rely heavily on sound for navigation and communication in the ocean's vast and often murky waters. Furthermore, understanding marine mammal calls provides invaluable data for monitoring and enforcing conservation strategies, ensuring the protection of endangered species, and maintaining the overall balance of the marine ecosystem. Given the importance of marine mammals to the ecological health of the oceans, the research and development of call recognition and classification technologies are not only pivotal for the protection of these animals but also for promoting biodiversity and strengthening maritime power. Advancements in these technologies have significant implications for the conservation of marine environments, the sustainable use of ocean resources, and the preservation of marine biodiversity for future generations.

Research on marine mammal call recognition and classification primarily involves two approaches. The first approach involves manually extracting features such as MFCC Vimal et al. (2021) and spectral centroid Zhang (2021), followed by manual classification. For example, Nanaware et al. Nanaware et al. (2014) used the ISHMAEL algorithm and PAMGUARD algorithm for passive acoustic detection and manual classification of calls from six species of marine mammals. Clemins et al. (2006)proposed the Greenwood function cepstral coefficient (GFCC) and generalized perceptual linear prediction (GPLP) models for extracting features to classify animal calls across various species. While these manual classification methods perform well when dealing with a limited number of marine mammal calls, they struggle when the calls of different species have similar spectrograms. Moreover, manual extraction methods fail to capture the inherent features of the audio itself. The second approach extracts features from marine mammal calls and uses traditional machine-learning models for classification, which includes extracting the MFCC Liu et al. (2024a), Mel spectrogram Tang et al. (2023) and other audio data features. Ibrahim et al. (2016) proposed a method for extracting features of the upward calls of North Atlantic right whales using low-frequency cepstral coefficients and discrete wavelet transforms (DWTs), with Support Vector Machines (SVM) used as the final classifier. Zhong and Cai (2019) introduced a method that combines MFCC, linear frequency cepstral coefficients (LFCC), and time-domain features, using SVM for classification and recognition. Furthermore, the complexity and diversity of marine mammal calls pose significant challenges for traditional machine learning-based methods in recognition and classification tasks. First, these methods require manual feature selection, which demands a deep understanding of the domain from researchers. Second, due to the varied patterns and fluctuations in marine mammal calls, traditional machine-learning approaches often have limited generalization capabilities. These methods struggle to capture patterns in data that are highly variable or previously unseen. Additionally, when dealing with large-scale datasets, traditional machine learning methods tend to require long training times.

Recently, deep learning methods have achieved remarkable results in marine mammal call recognition and classification tasks. Compared to traditional machine learning approaches, deep learning methods do not require manual feature labeling. Second, deep learning methods are able to more effectively extract features from data samples, leading to improved classification performance. Therefore, deep learning methods outperform traditional machine learning approaches for classification tasks involving the complex and diverse calls of marine mammals. Huang et al. (2022) extracted MFCC from real marine mammal vocal signals and transformed them into heat maps for classification. They used a combination of convolutional neural networks (CNNs) and MFCC images to classify marine mammal calls. Li et al. (2022) used marine mammal calls as input to a neural network model and employed a Convolutional Neural Network Gated Recurrent Unit(CNN-GRU) structure to extract acoustic features and perform classification. These methods generally rely on a single classification approach or a single feature extraction technique for marine mammal call recognition and classification, which often fails to fully capture the complexity of the audio signals. This can result in information loss and, consequently, hinder the overall classification performance. Furthermore, training classifiers with CNNs typically demands a large amount of labeled data, which is often challenging to obtain in the field of marine mammal call recognition.

To address the aforementioned challenges, this paper proposes a marine mammal calls recognition and classification method based on time-attention LSTM and a multi-scale dilated causal convolutional network. This approach aims to resolve issues such as single-feature extraction, information loss, and insufficient data samples in existing algorithms for marine mammal calls recognition and classification. By combining time-attention LSTM with multi-scale dilated causal convolutional networks, the method enhances the accuracy and robustness of marine mammal calls recognition and classification. The main contributions of this paper are as follows: 1) Integration of temporal and frequency domain feature extraction. We propose a method that integrates temporal and frequency domain feature extraction modules to capture audio signal information from multiple perspectives. This improves the model's robustness in complex environments and enhances its

ability to capture diverse features. 2) Optimized feature extraction networks. The method employs different networks for different feature extraction modules, improving the overall feature extraction capability and final classification accuracy. The frequency domain feature extraction module uses a multi-scale dilated causal convolutional network to extract frequency-domain features, while the temporal feature extraction module uses an LSTM network combined with a time-attention mechanism to emphasize critical temporal features. 3) Addressing data scarcity through transfer learning. The proposed method addresses the issue of limited training samples for marine mammal calls data. A classifier was pre-trained using the AudioSet dataset, and transfer learning was applied to adapt the pre-trained model to the classification of marine mammal calls, achieving the final classification results.

To evaluate our model, we collect call data from four marine mammal species, killer whale, sperm whale, pilot whale, and bottlenose dolphin, to form a dataset with sample sizes of 2,960, 2,007, 2,441, and 2,101, respectively. Extensive experiments were conducted on a self-labeled dataset, and the results show that the model achieves accuracy, precision, recall, and F1 scores all exceeding 95%. Compared to five other baseline models, our approach achieved the best detection performance across all four metrics.

2 Related work

Mammal call recognition and classification technology is a branch of audio pattern recognition widely applied in bioacoustic research and ecological conservation. With the advancement of audio pattern recognition and machine learning technologies, researchers have employed various methods to enhance the accuracy and efficiency of call recognition and classification. These approaches range from manual feature extraction to traditional machine learning and further to modern deep learning methods, covering a wide range of techniques and application scenarios. This has brought significant progress and innovation to the field of mammal call recognition and classification.

In the early manual feature extraction methods, researchers analyzed audio signals using techniques such as spectral analysis, time-domain features, and spectrogram correlation analysis. For example, through spectral analysis, researchers could transform audio signals into spectrograms to examine the energy distribution of different frequency components and extract features related to mammal calls. However, these methods often require substantial manual intervention and specialized knowledge. Moreover, manual feature extraction was inefficient and challenging to apply to large-scale data or process acoustic signals in complex environments. With the development of machine learning technologies, traditional methods like SVM were introduced into call recognition and classification. These methods significantly improved processing efficiency and accuracy by automating feature selection and constructing classification models. For instance, Roch et al. (2008) proposed a method to determine whether clicks were produced by beaked whales, short-finned pilot whales, or dolphins. This method used the Teager Energy Operator to locate individual clicks, then applied cepstral analysis to construct feature vectors for these clicks, which were classified using Gaussian Mixture Models (GMMs) and SVMs. Ali K. Ibrahim et al. Ibrahim et al. (2016) utilized MFCC and DWTs to extract features of the North Atlantic right whale's upcalls, achieving a classification accuracy of 92.27% using SVM and K-Nearest Neighbor (KNN) algorithms. Zhong Mingtuo et al. Zhong and Cai (2019) proposed a method combining MFCC, LFCC, and time-domain features for marine mammal call recognition. After feature extraction and fusion using these three methods, SVM was employed for classification, resulting in a 5.5% improvement in accuracy over traditional methods. Zhao et al. (2023) integrated four machine learning models and input MFCC features into the combined model. This approach improved accuracy by 3.03% compared to individual models. The integrated model leveraged the strengths of multiple individual models to enhance classification accuracy and robustness, making it more capable of handling various noise types and changing environmental conditions. Similar works include Henaff et al. (2011); Zubair et al. (2013); Esfahanian et al. (2017).While these traditional machine learning-based methods for marine mammal call classification have achieved some success, they also have limitations. First, traditional machine learning models often struggle with nonlinear and high-dimensional data, leading to overfitting or underfitting. Second, they are inefficient when handling large-scale datasets and fail to fully leverage extensive data for training. Third, these models typically require manual parameter tuning, with significant experiments and expertise needed to identify optimal feature combinations and model parameters.

In recent years, deep learning has garnered widespread attention in the field of marine mammal call recognition and classification. Researchers have begun applying CNN Alzubaidi et al. (2021), Recurrent Neural Networks (RNN) Sherstinsky (2020), and their variants, such as LSTM Duan (2022), to audio pattern recognition tasks, significantly improving the performance of call recognition and classification. Deep learning-based models can automatically extract high-level features from raw audio data and capture more complex spatiotemporal dependencies. For instance, Caleb Buchanan et al. Buchanan et al. (2021) designed a method to automatically detect bottlenose dolphin clicks. This method extracted, fused, and classified features from grayscale images, binary pixel values, and raw images of the audio signals. It successfully automated the recognition of bottlenose dolphin clicks, reducing the workload of manual feature extraction while improving processing efficiency and reliability. Ali K. Ibrahim et al. Ibrahim et al. (2021) proposed a classifier based on a Multi-Model Deep Learning (MMDL) algorithm to detect the upcalls of whales. This algorithm integrated CNN and Stacked Autoencoders (SAE) and demonstrated superior performance compared to traditional machine learning algorithms. Cai Wenyu et al. Cai et al. (2022) introduced a Multi-Channel Parallel (MDF-PNet) model comprising four branches: Mel spectrogram, MFCC, LFCC, and mean MFCC branches. A fully connected layer was used to fuse the results from the different branches. This approach employed neural

networks based on transfer learning in each branch to accelerate convergence and combine complementary features from four different perspectives. Feng et al. (2023) introduced adaptive wavelet transforms to extract features of bowhead whale whistles and used a CNN-LSTM model for recognition. This method leveraged the strengths of adaptive wavelet transforms and the CNN-LSTM model to more effectively extract and classify features. Murphy et al. (2022) proposed a method using residual learning networks for classifying marine mammal calls based on acoustic data from the William A. Watkins Marine Mammal Sound Database. Tabak et al. (2022) introduced a CNN model based on the ResNet18 architecture for classifying bat calls, achieving an accuracy of 92%. Liang et al. (2025) developed a method combining automatic detection and unsupervised clustering to extract acoustic features from PAM data and effectively remove noise, enabling whale calls recognition without manual annotation, with an average accuracy of 84.83%. White et al. (2022) employed transfer learning and deep convolutional neural networks to detect multiple marine sound sources, achieving high-accuracy identification of various sound sources such as odontocete whistles and ship noise, demonstrating the potential for ecological information extraction from large-scale PAM data to support marine mammal conservation. Schneider et al. (2024) combined convolutional neural networks with clustering methods to automatically detect and classify manatee calls from long-term recordings, further enabling individual identification and population size estimation, highlighting the feasibility of species monitoring through acoustics in visually limited environments. Liu et al. (2024b) proposed the XCFSMN framework based on knowledge distillation for efficient marine mammal sound source detection and classification, significantly improving model inference efficiency while maintaining accuracy. Liang et al. (2024) introduced an automatic detection method based on weighted spectral entropy, combining adaptive filtering, time-frequency transformations, and likelihood ratio detection to effectively enhance the detection performance of marine mammal tonal calls under low signal-to-noise ratio conditions. Di Nardo et al. (2025) proposed a CNN-based passive acoustic classification method that integrates spectral edge filtering to improve noise robustness, achieving high-precision identification of four types of bottlenose dolphin calls, providing an effective tool for dolphin behavior research and conservation.Similar works include Lü et al. (2024); Yin et al. (2025); Best (2022); Yang et al. (2023); Li et al. (2024).

From the analysis of the aforementioned studies, it can be concluded that current methods primarily rely on either a single classification approach or a single feature extraction method for marine mammal call recognition and classification. Moreover, most of these methods focus solely on feature extraction in the frequency domain, potentially failing to fully capture the complexity of audio signals that contain significant information in both the time and frequency domains. This may result in the loss of important information and negatively impact the final classification performance, resulting in lower recognition accuracy. Additionally, training classifiers using conventional convolutional neural networks requires a large amount of labeled data. However, obtaining large-scale labeled call datasets in the field of marine mammal call recognition is often extremely challenging.

3 System model

Feature extraction is the core step in the model's processing of audio data. This paper adopts a marine mammal call recognition and classification method based on feature fusion. The overall approach is illustrated in Figure 1, which adopts a dual-branch parallel structure. The two branches are independent during the feature extraction process, and they extract frequency-domain and time-domain features based on their respective advantages. The features extracted by both branches are concatenated and fused to form a complete feature representation, which is then input into a classifier to obtain the prediction result. Specifically, the frequency-domain branch uses a multi-scale dilated causal convolution network, a structure capable of capturing spectral features of audio signals at multiple scales, effectively enhancing the model's ability to model long-range dependencies in the frequency spectrum. The time-domain branch introduces an LSTM network with a time attention mechanism, which focuses on key information segments in the time series, improving the model's ability to perceive important temporal features.

The flowchart of the marine mammal calls recognition and classification method based on feature fusion is shown in Figure 2. Part a is the data processing section, which preprocesses the raw marine mammal calls data through operations such as clipping and denoising to obtain the corresponding Mel spectrograms. Part b is the feature extraction section, mainly consisting of frequency-domain feature extraction, time-domain feature extraction, and feature fusion. Part c is the classifier initialization section, where a convolutional neural network model with optimal classification performance is trained on the AudioSet dataset. The parameters of the encoder CNN convolutional neural network are initialized using a pre-trained network from the AudioSet audio tagging task, followed by fine-tuning. Part d is the transfer learning section, which builds upon part c by applying transfer learning to fine-tune the pre-trained neural network using real marine mammal calls data, replacing the original convolutional neural network for classification. The entire model is implemented through three main modules: first, the frequency-domain feature extraction module, which obtains spectral feature information of the signal; then, the time-domain feature extraction module, which extracts temporal information of the audio signal; and finally, the classification output module, which uses transfer learning to replace the original network with the pre-trained neural network and perform classification of marine mammal calls to meet specific task requirements.

3.1 Based frequency domain feature extraction module based on multi-scale dilated causal convolution

We preprocessed the raw audio signals of marine mammal calls by clipping, denoising and windowing, a 25 ms Hamming window

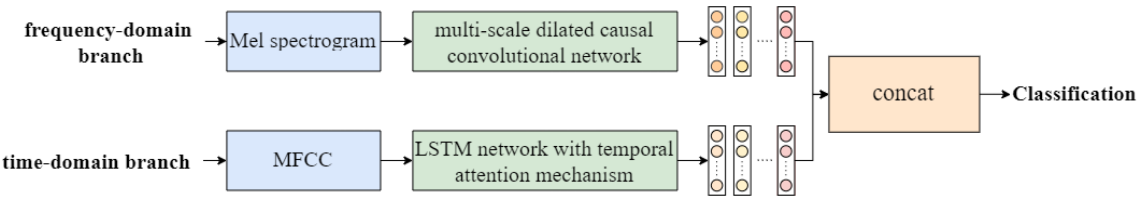

FIGURE 1
Overall framework diagram.

was applied to the audio signal for windowing. Each frame of audio signals was then subjected to a Fast Fourier Transform (FFT) to convert the time-domain signals into the frequency domain, obtaining the frequency-domain information for each frame. Additionally, we applied a Mel filter bank to the spectrogram, merging the energy in different frequency ranges with weighting to obtain the Mel spectrograms of killer whales, sperm whales, pilot whales, and bottlenose dolphins. Figures 3a-d show the spectrograms and Mel spectrograms of killer whales, sperm whales, pilot whales, and bottlenose dolphins.

The calls of marine mammals exhibit various forms, frequencies, durations, and rhythms, reflecting their different needs in communication, foraging, and navigation. Since Mel spectrograms can effectively reduce the dimensionality of spectral data, extract key information from audio data, and thus reduce data redundancy, we use Mel spectrograms to observe the frequency distribution of marine mammal calls, as well as the energy distribution of different frequency components. These spectrograms present complex spectral shapes to showcase the spectral characteristics of marine mammal calls at different time periods.

The processed Mel spectrograms are input into a multi-scale dilated causal convolutional neural network for frequency-domain feature extraction. The frequency-domain feature extraction module based on multiscale dilated causal convolution is shown in Figure 4. This module consists of the input section, the multi-scale dilated causal convolution feature extraction section, and the feature vector fusion section. The input section provides the Mel spectrogram obtained from the input audio signal. The multi-scale dilated causal convolution feature extraction section combines the advantages of dilated convolution and causal convolution. This section adopts a multi-scale approach, performing convolution operations at different scales to capture various levels of features from the data. Each scale consists of three convolution layers, with each convolution layer including one-dimensional dilated causal convolution (1D Conv), an activation function (ReLU), and a

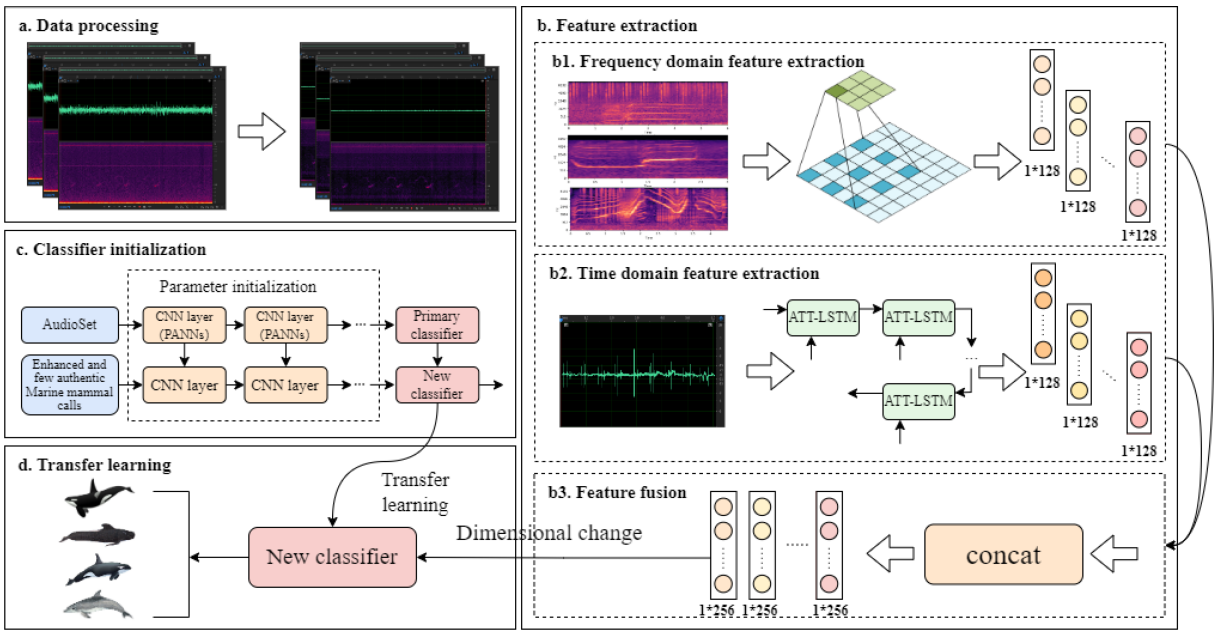

FIGURE 2
Flowchart of the marine mammal call recognition and classification method based on feature fusion, where **(a)** represents data preprocessing, **(b)** (including b1, b2, b3) represents the feature extraction process, **(c)** represents the pre-training process of the classifier, and **(d)** represents the transfer learning process.

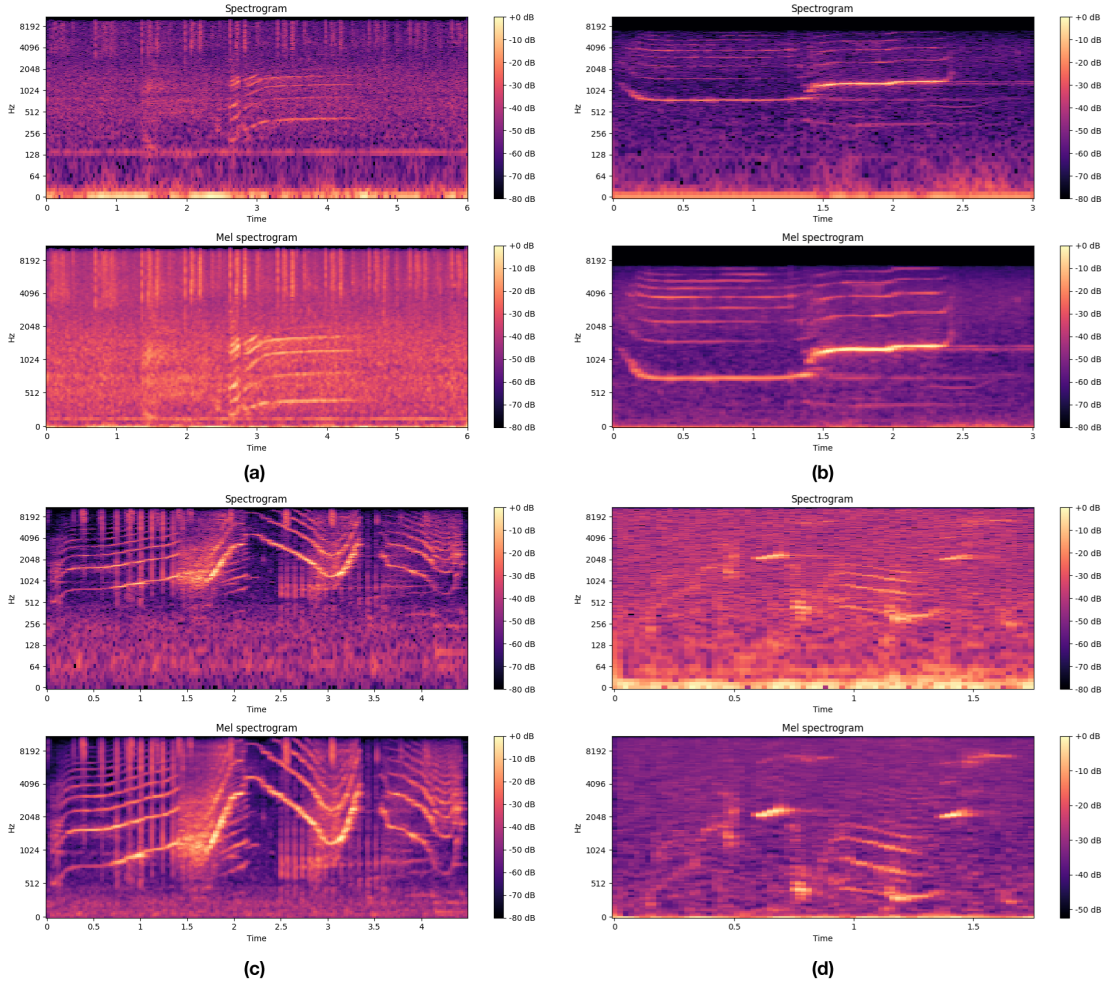

FIGURE 3
Spectrograms and Mel spectrograms of four marine mammals. **(a)** Killer Whale. **(b)** Sperm Whale. **(c)** Pilot Whale. **(d)** Bottlenose dolphin.

Dropout layer to prevent overfitting. The dilation factors for the three scales of dilated causal convolution are 1, 2, and 3, respectively. The calculation is as follows:

Assume there are M different dilation factors $d_1,\ldots,d_M$, the outputs for each scale are $y^{(1)},\ldots,y^{(M)}$, the final output vector u is calculated by Equation 1:

$$u = \sum_{m=1}^{M} \alpha_m \cdot y^{(m)} \tag{1}$$

Where M is the number of different dilation factors, and in this chapter, $M = 3$, $\alpha_m$is the weight parameter for the $m - th$ scale, and $y^{(m)}$ is the output of the $m - th$ scale's dilated causal convolution. By using multiple different dilation factors, this module is able to perform multi-scale feature extraction, which helps capture both short-term and long-term features, enhancing the ability to recognize complex time series patterns.

The model utilizes multiple different dilation factors, allowing the module to extract features at multiple scales, capturing features over different time ranges. This helps capture patterns and information across various time scales and enables the simultaneous capture of short-term and long-term features, thereby improving the recognition of complex time series patterns.

3.2 Time-domain feature extraction module based on ATT-LSTM

We employ a multi-scale dilated causal convolutional neural network module to extract the frequencydomain features of the audio data. However, this approach may overlook certain time-domain features. Furthermore, LSTM, which excel in processing time-series data, offer advantages such as capturing long-term dependencies, mitigating gradient vanishing, and handling sequences of variable lengths. By incorporating a time attention mechanism into the LSTM, we combine the sequential modeling capability of LSTM with the dynamic weighting property of time attention, which can significantly improve the identification and processing of important time-domain features. Therefore, this paper proposes using LSTM with a time attention mechanism for time-domain feature extraction of audio signals to complement the

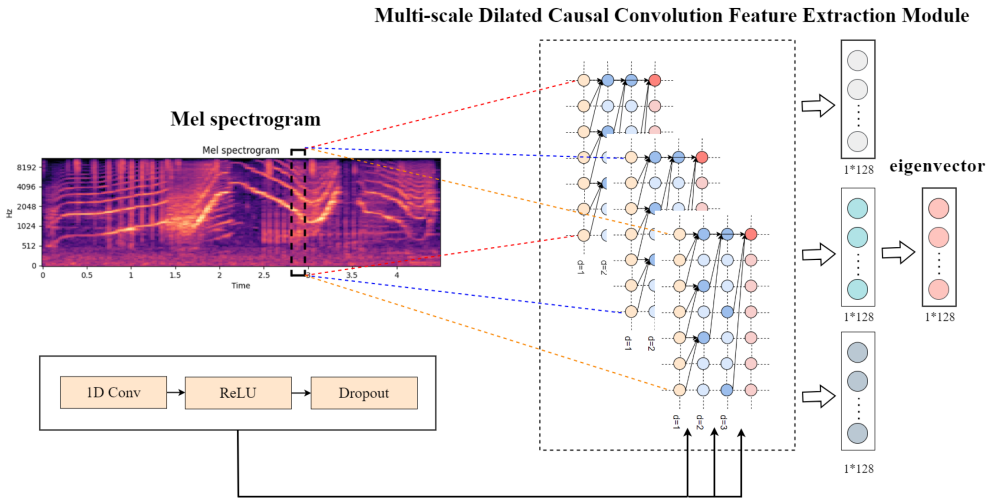

FIGURE 4
Structure diagram of the frequency-domain feature extraction module based on multi-scale dilated causal convolution.

frequency-domain features. First, we preprocess the raw marine mammal call data through clipping, denoising, and MFCC processing to generate a series of MFCC feature vectors, which represent the frequency spectrum characteristics of the audio signal at different time steps. These MFCC feature sequences are then input into an LSTM for time-domain feature extraction. The LSTM addresses long-term dependency issues by introducing a mechanism known as "gates," allowing it to better capture long-term dependencies in sequence data. After that, a time attention mechanism is introduced at the LSTM's output layer, which calculates the importance weight of each time step feature and applies a weighted process to the LSTM output sequence features. The structure of the time-domain feature extraction module using ATT-LSTM is shown in Figure 5. It consists of a forget gate, an input gate, an output gate, a cell state for controlling updates to the unit state, and the time attention mechanism. These gates control the flow of information within the LSTM unit, allowing it to selectively forget, add, or output information when processing sequence data. In the LSTM output layer, the time attention mechanism dynamically adjusts the feature weights, allowing the model to automatically identify and prioritize the most critical time points in the audio signal, reducing the impact of irrelevant information on model performance and thus improving classification accuracy and robustness.

The calculation process of each part inside the ATT-LSTM module is as follows: The forget gate of the original LSTM decides which information should be discarded at the current time step. Its output ranges from 0 to 1, where 0 indicates completely forgetting and 1 indicates completely remembering. It depends on the current input and the previous hidden state. The calculation Equation 2 is :

$$f_a = \sigma(W_f * [h_{t-1}, x_t] + b_f) \tag{2}$$

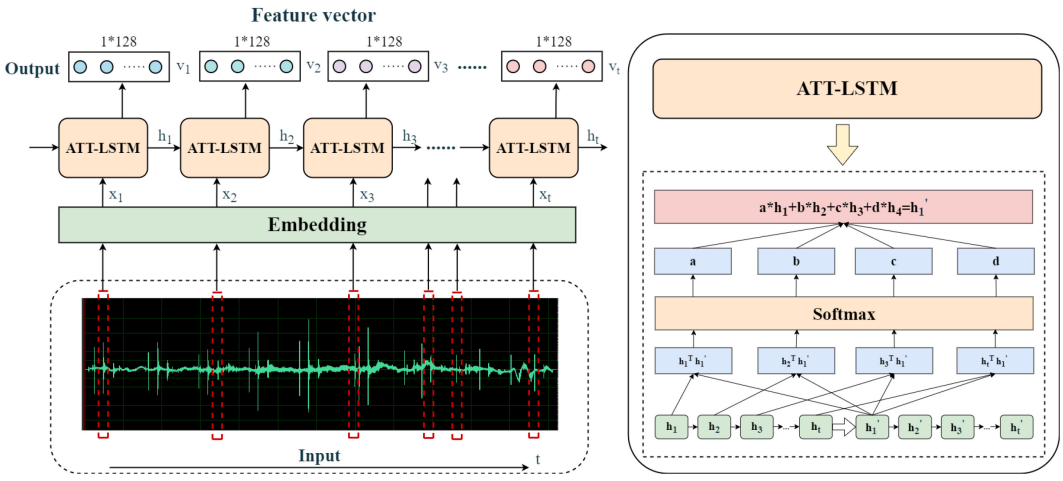

FIGURE 5
Structure diagram of the time-domain feature extraction module based on ATT-LSTM.

Where W_f is the weight matrix of the forget gate, $[h_{t-1}, x_t]$ is the hidden state from the previous time step, h_{t-1} is the vector formed by concatenating the current input $x_t$and the previous hidden state, b_f is the bias term, and σ is the Sigmoid function.

The calculation of the input gate involves the current input and the previous hidden state. It determines what information should be added to the cell state at the current time step. Its output range is between 0 and 1, where 0 means completely ignored and 1 means fully retained. The calculation Equation 3 is :

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \tag{3}$$

Where W_i is the weight matrix of the input gate, $[h_{t-1}, x_t]$ is the hidden state from the previous time step, and the vector formed by concatenating the current input x_t and the previous hidden state h_{t-1}, while b_i is the bias term.

The cell state is responsible for passing and storing information between different time steps. It is jointly regulated by the forget gate and the input gate, allowing the model to maintain appropriate memory when processing long sequences. The calculation Equation 4 is :

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{4}$$

Where f_t is the output of the forget gate, C_{t-1} is the cell state from the previous time step, and $\widetilde{C}_t$ is the new candidate cell state. Its calculation formula is $\widetilde{C}_t = \tanh(W_c * [h_{t-1}, x_t] + b_c)$, tanh is the hyperbolic tangent function, W_c is the weight matrix used to calculate the new candidate cell state, $[h_{t-1}, x_t]$ is the vector formed by concatenating the previous hidden state h_{t-1} and the current input x_t, and b_c is the bias term.

The output gate, which is the output of the LSTM, determines what information should be output at the current time step based on the current input and the previous hidden state. Its output will be passed to the hidden state of the next time step. The calculation Equation 5 is as follows :

$$v_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \tag{5}$$

Where W_o is the weight matrix of the output gate, $[h_{t-1}, x_t]$ is the vector formed by connecting the previous hidden state h_{t-1} and the current input x_t, and b_o is the bias term. The hidden state at the current time step can be represented by Equation 6:

$$h_t = v_t * \tanh(C_t) \tag{6}$$

In addition, we introduced a time attention mechanism to adjust the impact weight of all hidden layers H at the input time step on the hidden state $h_t^{'}$ at the output time step. This weight is calculated using cosine similarity by Equation 7:

$$e_t = \frac{h_t' \cdot H}{\|h_t'\| \cdot \|H\|} \tag{7}$$

Then, the Softmax function is used to calculate the contribution of each hidden layer to the output hidden layer based on the similarity values, as shown in Equation 8:

$$S_t = \frac{e_t}{\sum_{j=1}^{t} e_j} \tag{8}$$

Finally, we calculate the weighted sum of the contribution of each hidden layer to obtain the final output value of the hidden layer, as shown in Equation 9:

$$v_i^{'} = \sum_{j=0}^{i} S_j * h_j \tag{9}$$

After the feature extraction of the audio signal by the ATT-LSTM module, the output vector $v_1^{'}, v_2^{'} \ldots, v_t^{'}$ is concatenated to obtain the ATT-LSTM time-domain feature extraction vector. Finally, the Concat function is introduced to fuse the frequencydomain feature vector and the time-domain feature vector, as shown in Equation 10:

$$h = Concat\ (u, v) \tag{10}$$

3.3 Audio classification output module

The audio classification output module adopts the core idea of PANNs (Prototypical Audio Neural Networks), a neural network model used for audio classification and related tasks, which is pre-trained using the AudioSet dataset Kong et al. (2020). PANNs transfer pre-trained models from the computer vision field to the audio domain to address audio classification tasks. The training process mainly consists of two steps: First, pretraining on a large-scale audio dataset, aiming to learn general feature representations from audio data; Second, fine-tuning, where the pretrained model is applied to a specific audio classification task and its parameters are adjusted to adapt to the task-specific dataset and requirements.

The audio classification output module we constructed is shown in Figure 6. Since marine mammal call datasets are usually scarce, this part adopts transfer learning to address the issue of insufficient samples. First, a convolutional neural network model with optimal classification performance is trained using the AudioSet dataset. Then, the network trained on the AudioSet dataset for audio labeling tasks is used to initialize the parameters of the encoder CNN, followed by fine-tuning. The model is retrained using augmented data and a small amount of real data to replace the AudioSet dataset. This pre-trained model has already learned generalized audio features from a large amount of audio data, so it can be fine-tuned with a small amount of marine mammal call data to adapt to the specific task requirements. Finally, transfer learning is applied to mitigate the problem of insufficient samples by using the trained neural network to train on real marine mammal call data, replacing the original classification module. The final classification is obtained using the Softmax function to determine the category of the sound data, as shown in Equation 11:

$$p_i = \frac{\exp\ (z_i)}{\sum_{j=1}^{n} \exp\ (z_j)} \tag{11}$$

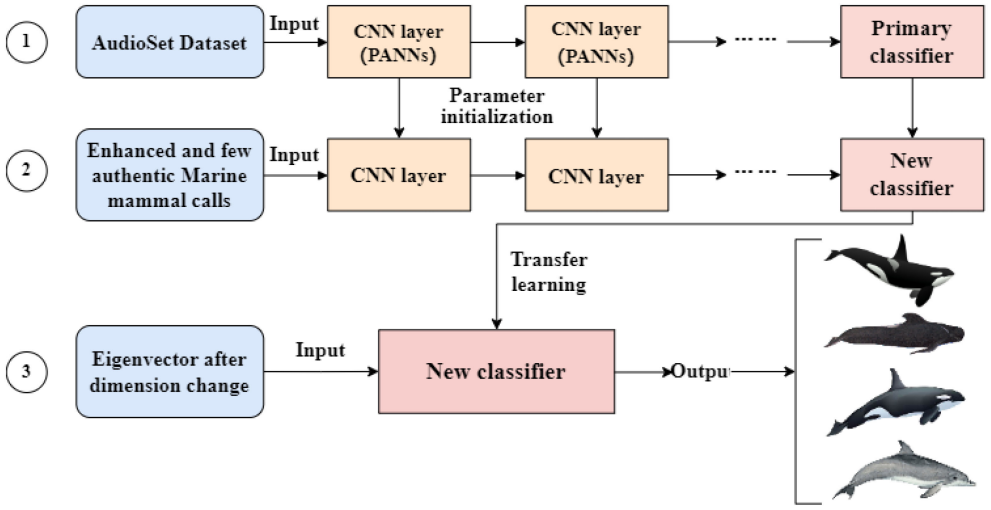

FIGURE 6
Construction diagram of marine mammal call classifier.

Where z_i represents the feature embedding for the $i-th$ class. Finally, we minimize the following cross-entropy loss function to train our model, as shown in Equation 12:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} log(p_{ij}) \tag{12}$$

Where N is the number of samples, C is the number of classes, y_{ij} is the true label indicating whether sample i belongs to class j, and p_{ij} is the probability predicted by the model that sample i belongs to class j.

4 Experimental analysis

4.1 Experimental data

The dataset used in this study was collected from the Watkins Marine Mammal Sound Database Sayigh et al. (2016) and the Whale FM website. For the experiments, we selected audio recordings of four marine mammal species that are frequently active in the waters surrounding China. These species include killer whales, sperm whales, pilot whales, and bottlenose dolphins. Due to the imbalance in the number of audio samples for each species and the varying quality of the recordings, manual inspection, segmentation, and denoising were required. Ultimately, we obtained the following numbers of calls samples: 1,500 samples of killer whales, 512 samples of sperm whales, 1,123 samples of pilot whales, and 966 samples of dolphins. The duration of each audio sample ranges from 1 second to 1 minute. The data sources for each category are listed in Table 1.

The specific steps of data preprocessing are as follows: First, manual denoising was performed. To reduce the interference of environmental noise and other non-target factors on classification performance, we used Adobe Audition to manually denoise the raw audio data. During this process, we combined spectrogram analysis with manual listening to locate and remove segments containing background noise, non-marine mammal calls, and other irrelevant signals. For the retained valid audio segments, we standardized the sampling rate to 16 kHz to ensure consistency and quality of the input for subsequent modeling. Second, audio segmentation was conducted. Since the durations of the original audio clips varied greatly, we further used Adobe Audition to uniformly segment the denoised audio data. The audio clips were trimmed into segments ranging from 1 to 5 seconds in length, and those shorter than 1 second were discarded to ensure that each sample contained sufficient information. Finally, considering the issue of class imbalance in the dataset, we adjusted the clipping ratio of each category to maintain a relatively balanced number of samples per class as much as possible during segmentation. (In future studies, we plan to introduce data augmentation techniques such as generative adversarial networks to further expand the number of samples in underrepresented classes and systematically improve model performance under imbalanced data conditions.) After these data processing steps, the number of audio samples for Killer Whales, sperm whales, pilot whales, and bottlenose dolphins was expanded to 2960, 2007, 2441, and 2101, respectively, to achieve better marine mammal sound detection performance.

4.2 Evaluation metrics

In this paper, we use four evaluation metrics to assess the performance of the model: Accuracy (A), Precision (P), Recall (R), and F1 Score (F1), as shown in Equations 13-16: Accuracy (A) refers to the ratio of the number of correctly classified samples to the total number of samples.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

TABLE 1 Description of the datasets used.

Dataset Name	Watkins Marine Mammal Sound Database	Whale FM Website
Killer Whale		✓
Sperm Whale	✓	
Pilot Whale		✓
Bottlenose Dolphin	✓	

where, TP stands for True Positive (the number of samples correctly classified as positive), TN stands for True Negative (the number of samples correctly classified as negative), FP stands for False Positive (the number of samples incorrectly classified as positive), and FN stands for False Negative (the number of samples incorrectly classified as negative). Precision (P) refers to the proportion of samples predicted as positive that are actually positive. It reflects the accuracy of the classifier in predicting positive classes.

$$P = \frac{TP}{TP + FP} \tag{14}$$

Recall (R) refers to the proportion of actual positive samples that are correctly predicted as positive. It reflects the classifier's ability to cover positive samples.

$$R = \frac{TP}{TP + FN} \tag{15}$$

The F1 score is the harmonic mean of precision and recall.

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{16}$$

4.3 Experimental details

In the experiment, this paper implements a marine mammal call recognition and classification system based on Pytorch. To train our model, we randomly split the dataset into training and testing sets in an 80:20 ratio for each class. In the data preprocessing stage, we use Mel spectrograms to extract audio features for the frequency domain feature extraction stage, where the audio sample rate is set to 16,000 Hz, and the number of Mel filters is set to 64 to provide finer frequency resolution. In the model training stage, we set the dilation factors of the dilated causal convolution at three scales to be 1, 2, and 3, with three layers at each scale, and the number of layers of the LSTM network is set to 2. We implement the alignment of the output features from the dilated causal convolution layers and LSTM network layers to achieve feature fusion. Furthermore, the training objective is to minimize the distance between the predicted and actual samples. The experiment trains the audio samples for 30 batches using the Adam algorithm with a learning rate of 0.001, and a warm-up cosine scheduler is employed. The experiment is conducted on a Linux machine with an Intel(R) Core(TM) i9-10920X CPU, 24 GB of RAM, and an NVIDIA RTX 3090 GPU.

4.4 Experimental results and analysis

To evaluate the effectiveness of the marine mammal call recognition and classification method based on ATT-LSTM and multi-scale dilated causal convolution, we investigate the following three questions:

- RQ1: Can our method achieve better results than other methods in the marine mammal sound detection and classification task?
- RQ2: What is the contribution of the time-domain feature extraction module and frequency-domain feature extraction module used in our method to the model?
- RQ3: How do hyperparameter settings affect the performance of our method?
- RQ4: How does our method perform compared to the baseline method under different noise conditions?

RQ1: Can our method achieve better results than other methods in the marine mammal sound recognition and classification task?

To validate the effectiveness of our method in the marine mammal sound detection and classification task, we selected five high-performance baseline models for comparison experiments and conducted classification studies on the calls of four marine mammals: the bottlenose dolphin, Killer Whale, sperm whale, and pilot whale. Model 1 uses an improved residual network-based model, Res2Net Gao et al. (2019), This improves the network's feature extraction performance by representing features at multiple scales. Model 2 adopts TDNN (Time Delay Neural Network) Martinez et al. (2022), which captures temporal dependencies in input sequences by introducing time delays. Model 3 employs Ecapa-TDNN (Emphasized Channel Attention, Propagation, and Aggregation Time Delay Neural Network) Desplanques et al. (2020), an improved version of TDNN, which incorporates a channel attention mechanism and a feature aggregation strategy. Model 4 utilizes an Attention-DenseNet-based model Xie et al. (2023), combines the DenseNet121 network architecture with a self-attention module and a center loss function to enhance sound recognition. Model 5 adopts the DRCNN (Deep Residual Convolutional Neural Network) model Feng and Cheng (2023), an acoustic model built on deep residual convolutional neural networks. We applied these baseline models for marine mammal sound recognition and classification.

We conducted experiments on the proposed method and five baseline models using four marine mammal sound datasets we constructed and compared the classification performance of different methods using four performance metrics. Each data point represents the average value of the four detection metrics, and the results are shown in Table 2.

As shown in Table 2, our method achieved the highest scores across all evaluation metrics. Specifically, it attained an accuracy of 97%, which is 3% higher than the best-performing baseline model, AttentionDenseNet, indicating a stronger overall capability in correctly classifying sample categories. The precision reached 99%, significantly outperforming other models-7% higher than

TDNN and 9% higher than DRCNN—demonstrating a very low false positive rate when identifying samples as a specific call type. The recall was 96%, also surpassing all comparison models by at least 2% compared to the highest recall among them (Attention-DenseNet), showing superior ability in correctly retrieving target samples. The F1 score was 0.97, reflecting a well-balanced and stable performance by considering both precision and recall. This score is much higher than traditional models such as Res2Net (83%) and ECAPA-TDNN (89%), and also outperforms the more complex Attention-DenseNet (93%), highlighting our method's notable advantages in classification stability and consistency. We attribute this to the fact that our model can integrate features of marine mammal sounds from both the time domain and frequency domain perspectives. In contrast to methods that only extract frequency domain features of audio or use a single method to extract multiple features, our method provides a more comprehensive capture of audio features.

Overall, the proposed method based on ATT-LSTM and multi-scale dilated causal convolution demonstrates significant advantages over other baseline methods and achieves the best detection performance across all four evaluation metrics.

In addition, our proposed method for marine mammal sound recognition and classification based on ATT-LSTM and multi-scale dilated causal convolution outperforms the five baseline methods in all four metrics. Compared to the best-performing baseline model, Attention-DenseNet, our method achieves absolute improvements of 3% in accuracy and 4% in F1 score. Compared to the Ecapa-TDNN model, our method achieves improvements of 5% and 8% in the two metrics. We attribute this to the fact that our model can integrate features of marine mammal sounds from both the time domain and frequency domain perspectives. In contrast to methods that only extract frequency domain features of audio or use a single method to extract multiple features, our method provides a more comprehensive capture of audio features.

Overall, the proposed method based on ATT-LSTM and multi-scale dilated causal convolution demonstrates significant advantages over other baseline methods and achieves the best detection performance across all four evaluation metrics.

To validate the effectiveness of our proposed method in marine mammal sound classification, we compared the classification performance of the model for different marine mammal sounds using the four evaluation metrics, as shown in Table 3. Specifically,

TABLE 2 Comparison of marine mammal call recognition performance.

Model Name	Accuracy	Precision	Recall	F1 Score
Res2Net	0.87	0.84	0.83	0.83
TDNN	0.89	0.92	0.89	0.90
Ecapa-TDNN	0.92	0.88	0.91	0.89
DRCNN	0.93	0.90	0.92	0.91
Attention-DenseNet	0.94	0.92	0.94	0.93
Our method	**0.97**	**0.99**	**0.96**	**0.97**

Bold values indicate the experimental results of the proposed method in this paper.

TABLE 3 Classification performance of the model on marine mammal calls.

Marine mammal species	Accuracy	Precision	Recall	F1 Score
Bottlenose Dolphin	0.96	0.97	0.98	0.97
Killer Whale	0.98	1.0	0.97	0.98
Sperm Whale	0.98	0.98	0.92	0.95
Pilot Whale	0.95	1.0	0.98	0.99

our method achieved more than 95% in both accuracy and F1 score for the detection of each marine mammal sound, indicating that our model can efficiently recognize and classify the sounds of the four marine mammals with robustness. Furthermore, our method achieved a precision of 1 for two types of marine mammal sounds, which means the model correctly identified all positive class results. For each marine mammal, our method performed better than 97% in all four metrics for detecting the sounds of the bottlenose dolphin, Killer Whale, and Pilot Whale. However, the detection performance for sperm whale sounds was not as good (with recall only exceeding 92%). We analysed the reason and found that during the data clipping process, certain sperm whale audio recordings contained noticeable environmental noise, which somewhat affected the model's classification performance for sperm whale sounds. Therefore, there is still room for improvement in both the model's recognition and classification of sperm whale sounds and the overall model performance.

In addition, we have plotted the confusion matrix for the recognition accuracy of the four marine mammal sounds using our method to visually observe the model's performance in classifying the sounds of the four marine mammals, as shown in Figure 7 The rows of the matrix represent the predicted categories of the marine mammals, while the columns represent the true categories. The black boxes on the diagonal represent the

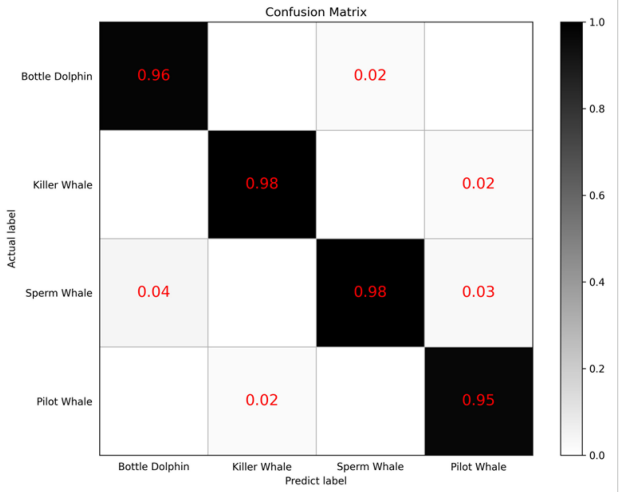

FIGURE 7
Confusion matrix of the feature fusion model.

proportion of correctly predicted categories, and the color depth indicates the degree of classification success, with darker colors indicating better classification performance. We found that the model achieved an average classification accuracy of approximately 97% for the four marine mammal sounds, with classification accuracy for each sound data being above 95%. Additionally, for the misclassified results, the confusion matrix shows a relatively uniform distribution. This further indicates that our model achieves a good fit for all four marine mammal sounds.

RQ2: What is the contribution of the time-domain feature extraction module and frequency-domain feature extraction module used in our method to the model?

As shown in Figures 8–11, we compare the four quantitative metrics under three conditions: using both feature extraction modules, using only the time-domain feature extraction module, and using only the frequency-domain feature extraction module. The results show that our method, which integrates both time-domain and frequency-domain features of marine mammal calls, achieves the best recognition and classification performance. Compared to using only the time-domain or frequency-domain feature extraction branches, there is a significant performance improvement. Specifically, compared to using only the time-domain feature extraction method, our method improves by 5%-10% across all four metrics for the four audio datasets. Compared to using only the frequency-domain feature extraction method, our method improves by 3%-8%, achieving optimal detection performance. This demonstrates that integrating two types of feature extraction modules complements the single features extracted, achieving the "1 + 1>2" effect.

Furthermore, we compared the contributions of the time-domain and frequency-domain modules to the model's performance. We found that the model is more significantly influenced by the frequency-domain feature extraction branch than by the time-domain feature extraction branch, which is closely related to the characteristics of audio data. The frequency-domain information of marine mammal call data records the distribution of the audio signal in terms of frequency, which is crucial for recognizing audio data. On the other hand, the time-domain information records rapid changes and transient features in the audio data, representing more detailed information. Therefore, integrating both time-domain and frequency-domain features is beneficial for improving the model's ability to recognize and classify marine mammal calls.

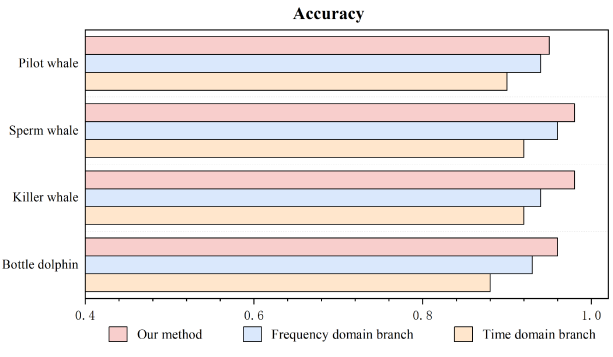

FIGURE 8
Accuracy of different feature extraction models.

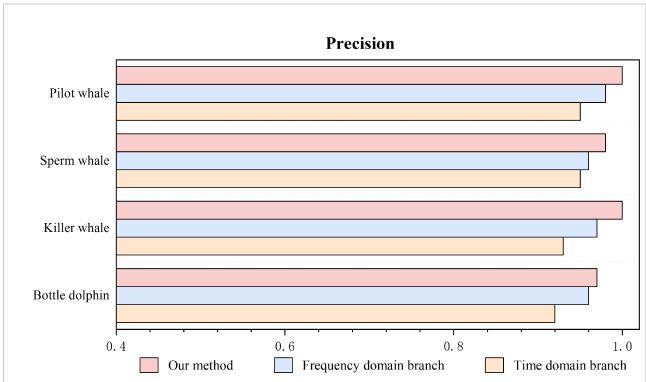

FIGURE 9
Precision of different feature extraction models.

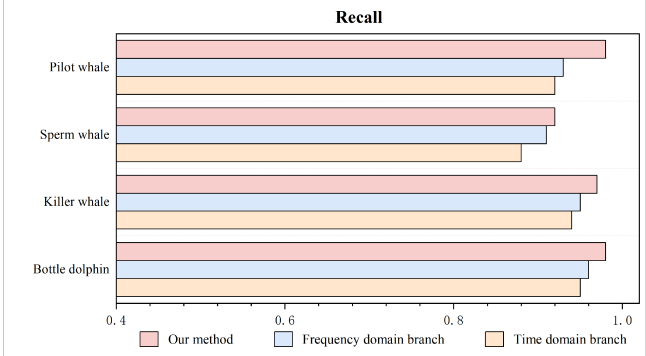

FIGURE 10
Recall of different feature extraction models.

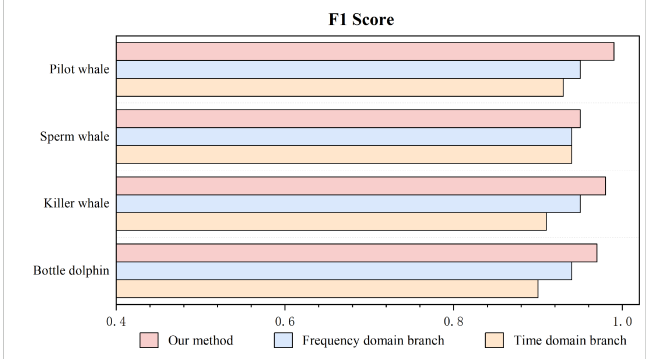

FIGURE 11
F1 Score of different feature extraction models.

RQ3: How do hyperparameter settings affect the performance of our method?

Tuning and hyperparameters play a key role in the feature extraction performance of our method. We conducted experiments to investigate the impact of hyperparameter settings on our method's performance, including combinations of different

dilation factors, the number of attention heads, and the choice of pre-trained models. To optimize these hyperparameters, we used Neural Network Intelligence (NNI) for fine-tuning.

(1) Different dilation factor combinations and number of attention heads.

To explore the effect of dilation factor settings on frequency-domain feature extraction, we designed sensitivity comparison experiments to assess the model's sensitivity to different receptive fields. We analyzed whether the current dilation factor combination of [1, 2, 3] is optimal. We compared it with four other dilation configurations, keeping the other network parameters unchanged and training and testing on the same dataset. Table 4 shows the impact of different dilation factor combinations on our model's performance. Adjusting the dilation factors from small to large first expanded the receptive field, allowing the network to aggregate more frequency-domain context, thus improving accuracy. When the dilation factor combination of [1, 2, 3] was used, the model achieved the best performance with an accuracy of 97%. However, as the dilation factor was further increased, the classification accuracy dropped because excessively large dilation factors resulted in an overly large receptive field, causing information sparsity and loss of details, which hindered feature interaction and led to a decline in accuracy.

Table 4 also demonstrates the performance of our method with different numbers of attention heads. We designed comparison experiments to analyze whether the current number of attention heads is optimal. We compared it with four other configurations of different head numbers while keeping other network parameters unchanged. From Table 4 and Figure 12, we can see that increasing the number of attention heads improves our method's accuracy, as more attention heads capture more feature information and enhance the expression ability of features. When the number of attention heads reached 8, the model achieved the best accuracy of 97%. However, when the number of attention heads exceeded 8, the model's performance no longer improved because the additional attention heads overlapped with existing information, leading to diminishing returns.

(2) Pre-trained model selection.

To verify the impact of the pre-trained model on the performance of this task, we designed a comparative experiment in which different pre-trained models were trained using the same pre-trained dataset, and then compared in terms of classification accuracy and F1-score after transfer. This further validates the adaptability and effectiveness of the selected pre-trained model. In the experiment, we selected five pretrained models: ResNet34 Cheng and Yu (2024), VGG16 Singh et al. (2023), DenseNet121 Tan et al. (2023), EfficientNet-B0 Smelyakov et al. (2022), and PANNs-CNN14, and uniformly pre-trained them using the AudioSet dataset to eliminate the impact of training data differences on the results. Then, using the same process, we trained and evaluated these models on the target task and subsequently fine-tuned them on our marine mammal calls dataset, assessing their classification accuracy and F1-score on the test set. The experimental results are shown in Table 5. The results indicate that PANNs-CNN14 achieved the best classification performance in this task, with an accuracy of 97.0% and an F1-score of 97%, demonstrating its advantages and adaptability in audio feature modeling. Therefore, we chose PANNs-CNN14 as the pre-trained model.

RQ4: How does our method perform compared to baseline methods under different noise conditions?

Audio data is often affected by various types of noise, leading to a decline in model classification performance. To further evaluate the robustness of the model in practical applications, we designed noise contrast experiments to verify the model's stability and anti-interference ability under noisy conditions. Before that, we collected some marine environmental noise data and extracted noise (including seawater noise and submarine noise) from the marine mammal call audio data to enhance the noisy data. We used the five high-performing baseline models from the baseline experiment for comparison. To simulate real-world environments, we added noise with different intensities to the original audio data, where the noise intensity was measured by the Signal-to-Noise Ratio (SNR). We conducted experiments with noise intensities of 30 dB, 20 dB, 10 dB, 0 dB, and -10 dB. The accuracy comparison of different models under the influence of marine noise is shown in Figure 13. Compared to the best-performing Attention-DenseNet model, our method achieves an average accuracy that is about 5.6% higher. This indicates that, under the same signal-to-noise ratios (SNRs), our model outperforms other models. From the slope in the

TABLE 4 Hyperparameter comparison.

Dilation Factor Combination	Accuracy	Number of Heads	Accuracy
[1,1,1]	0.92	2	0.90
[1,2,2]	0.95	4	0.95
[1,2,3]	0.97	6	0.94
[2,3,4]	0.94	8	0.97
[3,4,5]	0.91	10	0.95

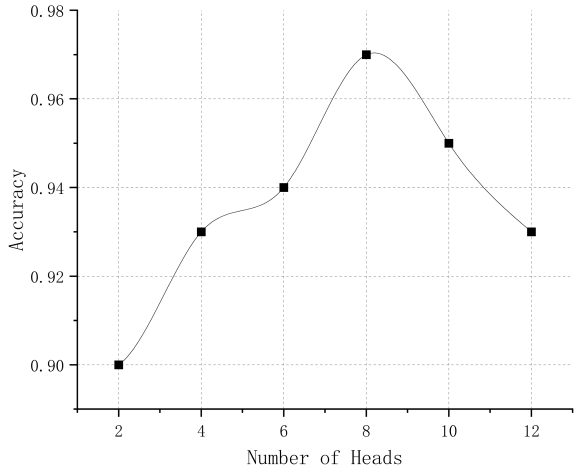

FIGURE 12
Curve of the effect of the number of attention heads on temporal domain features.

TABLE 5 Comparison of different pre-trained models.

Pre-trained Model	Accuracy	F1 Score
VGG16	0.90	0.89
ResNet34	0.91	0.90
DenseNet121	0.91	0.90
EfficientNet-B0	0.92	0.91
PANNs-CNN14	0.97	0.97

graph, we can see that as the SNR decreases, the accuracy of the Res2Net, TDNN, and DRCNN models drops at a similar rate, with Res2Net consistently having lower accuracy than the other models. The Ecapa-TDNN model shows a slower decline in accuracy when the SNR is greater than 10dB, but its accuracy decreases more rapidly when the SNR drops below 10 dB. Additionally, compared to the five baseline models, our model has a slower accuracy decline and higher accuracy across all SNR levels. Therefore, it can be concluded that, as the SNR decreases, our model is more stable and less affected by noise interference, demonstrating better robustness. When the SNR is below 0 dB, the accuracy of all methods falls below 50%, which we attribute to the noise covering part of the original call audio, causing a sharp decline in performance for all models. Overall, under the same noise conditions, our proposed model performs better and is less affected by noise, indicating its superior stability.

Cross-dataset validation: To ensure that the model is trained and tested on multiple different data partitions, thereby making the evaluation results more stable and comprehensive while reducing randomness, we introduced K-fold cross-validation. The dataset was divided into k parts, with one part used as test data and the remaining k-1 parts as training data. We adopted 5-fold cross-validation to evaluate the performance of five baseline models and our proposed model on the marine mammal call recognition task. This is equivalent to performing cross-dataset validation among five "different but related" sub-datasets, allowing us to test the model's stability under changing data distributions. The accuracy, standard deviation, and other metrics of each model are shown in Table 6.

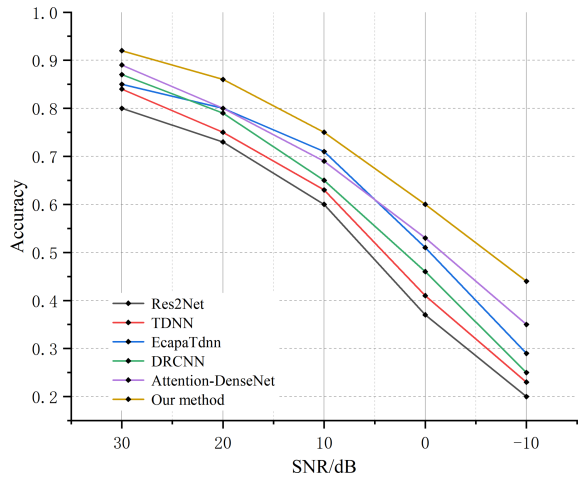

FIGURE 13
Comparison of accuracy of different models under the influence of marine noise.

The experimental results demonstrate that our proposed method exhibits significant advantages in both classification performance and model stability. In the 5-fold cross-validation, our method achieved an average accuracy of 0.97, with all folds maintaining an accuracy above 0.96. This performance is significantly better than that of the other compared models. Specifically, compared to the second-best model, Attention-DenseNet, our method improved classification accuracy by 3.2%; when compared to the baseline model Res2Net, the improvement reached 11.5%. This performance advantage is primarily attributed to our proposed feature extraction approach, which integrates multi-scale dilated causal frequencydomain features with adaptive attention-based LSTM time-domain features, significantly enhancing the model's adaptability to complex marine acoustic environments.

In terms of model stability, our method also performed exceptionally well, with a standard deviation of only 0.005, which is substantially lower than that of other models. While Attention-DenseNet and DRCNN also exhibited relatively good stability, their fluctuation ranges were still larger than that of our method. Further analysis revealed that the traditional TDNN model, lacking an effective feature selection mechanism, was more sensitive to changes in data distribution. These experimental results fully validate that our proposed method not only achieves superior classification accuracy but also offers excellent robustness.

Computational Complexity Analysis: To further evaluate the practical efficiency of each model during the training phase, this paper compares the training time of the proposed method with five baseline models under the same experimental environment. First, we unified the training environment, parameter settings, and training dataset for all models. In the experiment, 32 audio samples were trained per batch, with 100 epochs, using the Adam optimizer, a learning rate of 0.001, and a warm-up cosine scheduler for the process. The experiment was conducted on a Linux machine with an Intel(R) Core(TM) i9-10920X CPU, 24 GB RAM, and an NVIDIA RTX 3090 GPU. During the experiment, the total training time, from the beginning of training to final convergence, was recorded.

Figure 14 shows the training overhead of different classification models and their achieved accuracy. The maximum-minimum boundary bars represent the highest and lowest performance values across multiple tests, while the height of the bars displays the average performance of the tests. We conducted the experiment on the collected marine mammal calls dataset.

As shown in the figure, it can be seen that the proposed method, which integrates a dual-branch structure and transfer learning

TABLE 6 Accuracy, average, and standard deviation of each model under 5-fold cross-validation.

Model Name	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average	Standard Deviation
Res2Net	0.84	0.88	0.87	0.86	0.90	0.87	0.022
TDNN	0.88	0.90	0.89	0.89	0.89	0.89	0.012
Ecapa-TDNN	0.91	0.93	0.92	0.92	0.94	0.92	0.010
DRCNN	0.92	0.94	0.93	0.92	0.94	0.93	0.008
Attention-DenseNet	0.93	0.94	0.94	0.93	0.96	0.94	0.007
DRCNN	0.96	0.97	0.97	0.97	0.98	0.97	0.005

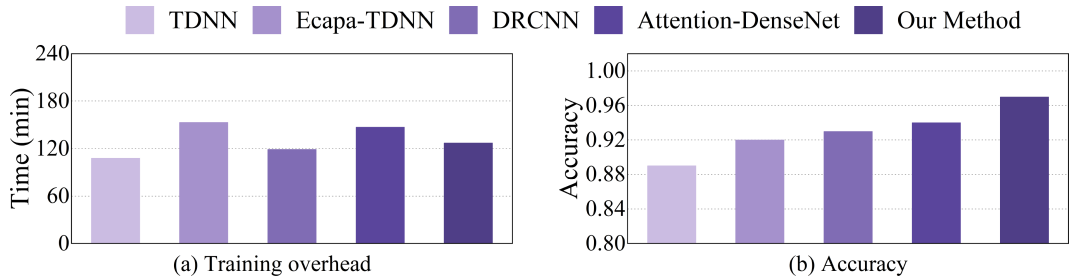

FIGURE 14
Training overhead and accuracy achieved by different classification models, where **(a)** represents the time complexity of the five models, and **(b)** represents the comparison of the model precision rates.

strategy, has slightly higher structural complexity compared to TDNN and DRCNN, but its accuracy is much higher than theirs. Moreover, its overall training time is shorter than that of ECAPA-TDNN and Attention-DenseNet, while its accuracy is higher, demonstrating a good balance of training efficiency and classification performance. This is due to the use of the transfer learning strategy in this paper, which addresses the issue of insufficient experimental data and significantly accelerates the model's convergence process, allowing it to achieve better performance within the same number of epochs. Therefore, the proposed method strikes a good balance between training time and performance.

5 Conclusion

This paper presents a marine mammal call recognition and classification method based on an ATT-LSTM and multi-scale dilated causal convolutional network. The method leverages the Mel spectrogram as input and applies multi-scale dilated causal convolutions to extract the frequency-domain features of the audio, substantially improving the network's ability to capture intricate frequency-domain patterns. Concurrently, the LSTM module delves into the temporal characteristics of the audio signals and introduces a timeattention mechanism to highlight crucial temporal features. This approach not only effectively complements the frequency-domain features but also significantly boosts the model's overall feature representation capability for audio data. Additionally, the method incorporates transfer learning to address the challenge of limited training samples for marine mammal calls, utilizing pre-trained models to enhance performance. Extensive experimental validation has demonstrated the feasibility and reliability of the proposed method. The results show that the proposed method achieved an accuracy of 97%, a precision of 99%, a recall of 96%, and an F1 score of 97%, representing performance improvements of 3%, 7%, 2%, and 4%, respectively, compared to the best baseline model. Even in conditions with varying SNR, the recognition and classification accuracy of this method exceeds that of other models, highlighting its robustness and resistance to noise interference. This method not only ensures the accuracy of neural network classifiers in sound recognition tasks but also achieves low-latency, high-accuracy recognition and classification of marine mammal call, making it highly suitable for practical applications in dynamic and noisy marine environments. In the future, we plan to introduce deep learning-based data augmentation techniques (such as generative adversarial networks) to further expand the number of samples in underrepresented classes, thereby systematically improving the model's performance under imbalanced data conditions.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

Ethics statement

The animal study was approved by The Committee of Science and Technology Ethics of Qingdao University of Science and Technology. The study was conducted in accordance with the local legislation and institutional requirements.

Author contributions

WC: Methodology, Software, Writing – original draft, Writing – review & editing. HC: Data curation, Writing – review & editing. JJ: Validation, Visualization, Writing – review & editing. SL: Investigation, Writing – review & editing. JW: Methodology, Supervision, Writing – review & editing. YZ: Funding acquisition, Methodology, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported in part by the National Natural Science Foundation of China under Grants 62171246, 62101298 and U24A20215.

Acknowledgments

The authors are grateful to the Watkins Marine Mammal Sound Database website and the Whale FM website for providing us with the audio of marine mammal calls needed for the experiment, and the support of Python.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021). Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *J. big Data* 8, 1–74. doi: 10.1186/s40537-021-00444-8

Best, P. (2022). *Automated detection and classification of cetacean acoustic signals. Ph.D. thesis* (Université de Toulon - Laboratoire LIS).

Buchanan, C., Bi, Y., Xue, B., Vennell, R., Childerhouse, S., Pine, M. K., et al. (2021). "Deep convolutional neural networks for detecting dolphin echolocation clicks," in *2021 36th International Conference on image and vision computing New Zealand (IVCNZ)* (IEEE), 1–6. doi: 10.1109/IVCNZ54163.2021.9653250

Cai, W., Zhu, J., Zhang, M., and Yang, Y. (2022). A parallel classification model for marine mammal sounds based on multi-dimensional feature extraction and data augmentation. *Sensors* 22, 7443. doi: 10.3390/s22197443

Cheng, Y., and Yu, W. (2024). "Research on resnet34 improved model," in *2024 9th international conference on intelligent informatics and biomedical sciences (ICIIBMS)*, vol. 9. (IEEE), 11–14. doi: 10.1109/ICIIBMS62405.2024.10792749

Clemins, P. J., Trawicki, M. B., Adi, K., Tao, J., and Johnson, M. T. (2006). "Generalized perceptual features for vocalization analysis across multiple species," in *2006 IEEE international conference on acoustics speech and signal processing proceedings*, vol. 1. (IEEE), I–I. doi: 10.1109/ICASSP.2006.1660005

Desplanques, B., Thienpondt, J., and Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*. 2020, 3830–3834. doi: 10.21437/Interspeech.2020-2650

Di Nardo, F., De Marco, R., Li Veli, D., Screpanti, L., Castagna, B., Lucchetti, A., et al. (2025). Multiclass cnn approach for automatic classification of dolphin vocalizations. *Sensors* 25, 2499. doi: 10.3390/s25082499

Duan, D. (2022). Detection method for echolocation clicks based on lstm networks. *Mobile Inf. Syst.* 2022, 4466037. doi: 10.1155/2022/4466037

Duan, D., Lü, L., Jiang, Y., Liu, Z., Yang, C., Guo, J., et al. (2022). Real-time identification of marine mammal calls based on convolutional neural networks. *Appl. Acoustics* 192, 108755. doi: 10.1016/j.apacoust.2022.108755

Esfahanian, M., Erdol, N., Gerstein, E., and Zhuang, H. (2017). Two-stage detection of north atlantic right whale upcalls using local binary patterns and machine learning algorithms. *Appl. Acoustics* 120, 158–166. doi: 10.1016/j.apacoust.2017.01.025

Feng, C., and Cheng, W. (2023). Speech recognition algorithm based on residual convolutional neural network. *Comput. digital Eng.* 51, 440–444. doi: 10.3969/j.issn.1000-386x.2020.11.044

Feng, R., Xu, J., Jin, K., Xu, L., Liu, Y., Chen, D., et al. (2023). An automatic deep learning bowhead whale whistle recognizing method based on adaptive swt: Applying to the beaufort sea. *Remote Sens.* 15 (22), 5346. doi: 10.3390/rs15225346

Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., and Torr, P. (2019). Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 652–662. doi: 10.1109/TPAMI.34

Henaff, M., Jarrett, K., Kavukcuoglu, K., and LeCun, Y. (2011). "Unsupervised learning of sparse features for scalable audio classification," in *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, vol. 11, 2011.

Huang, N. Y., Deyou, J., and Yang, C. (2022). "A study on the classification over vocalization of marine," in *2022 Western China Acoustic Academic Exchange Conference*. 2022.

Ibrahim, A. K., Zhuang, H., Cherubin, L. M., Erdol, N., O'Corry-Crowe, G., and Ali, A. M. (2021). A multimodel deep learning algorithm to detect north atlantic right whale up-calls. *J. Acoustical Soc. America* 150, 1264–1272. doi: 10.1121/10.0005898

Ibrahim, A. K., Zhuang, H., Erdol, N., and Ali, A. M. (2016). "A new approach for north atlantic right whale upcall detection," in *2016 international symposium on computer, consumer and control (IS3C)* (IEEE), 260–263. doi: 10.1109/IS3C.2016.76

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. (2020). Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 28, 2880–2894. doi: 10.48550/arXiv.1912.10211

Li, S., Liu, P., Yan, Q., Wang, K., Gan, W., and Wang, J. (2022). "A classification method for marine mammals based on acoustic features," in *2021–2022 Academic Conference of the Underwater Acoustics Branch of the Acoustical Society of China*.

Li, S., Yu, Z., Wang, P., Sun, G., and Wang, J. (2024). Blind source separation algorithm for noisy hydroacoustic signals based on decoupled convolutional neural networks. *Ocean Eng.* 308, 118188. doi: 10.1016/j.oceaneng.2024.118188

Liang, Y., Seger, K. D., and Kirsch, N. J. (2024). Entropy-based automatic detection of marine mammal tonal calls. *IEEE J. Oceanic Eng.* 49, 1140–1150. doi: 10.1109/JOE.2024.3436867

Liang, Y., Wang, Y., Chen, F., Yu, H., Ji, F., and Chen, Y. (2025). Automatic detection and unsupervised clustering-based classification of cetacean vocal signals. *Appl. Sci.* 15 (7), 3585. doi: 10.3390/app15073585

Liu, F., Li, G., and Yang, H. (2024a). Application of multi-algorithm mixed feature extraction model in underwater acoustic signal. *Ocean Eng.* 296, 116959. doi: 10.1016/j.oceaneng.2024.116959

Liu, X., Liu, X., Du, S., and Cheng, J. (2024b). "Hear you say you: An efficient framework for marine mammal sounds' classification," in *Proceedings of the AAAI conference on artificial intelligence* 38 (20), 22250–22257. doi: 10.1609/aaai.v38i20.30230

Lü, Z., Shi, Y., Lü, L., Han, D., Wang, Z., and Yu, F. (2024). Dual-feature fusion learning: An acoustic signal recognition method for marine mammals. *Remote Sens.* 16 (20), 3823. doi: 10.3390/rs16203823

Martinez, A. M. C., Spille, C., Roßbach, J., Kollmeier, B., and Meyer, B. T. (2022). Prediction of speech intelligibility with dnn-based performance measures. *Comput. Speech Lang.* 74, 101329. doi: 10.1016/j.csl.2021.101329

Murphy, D. T., Ioup, E., Hoque, M. T., and Abdelguerfi, M. (2022). Residual learning for marine mammal classification. *IEEE Access* 10, 118409–118418. doi: 10.1109/ACCESS.2022.3220735

Nanaware, S., Shastri, R., Joshi, Y., and Das, A. (2014). "Passive acoustic detection and classification of marine mammal vocalizations," in *2014 international conference on communication and signal processing* (IEEE), 493–497. doi: 10.1109/ICCSP.2014.6949891

Roch, M. A., Soldevilla, M. S., Hoenigman, R., Wiggins, S. M., and Hildebrand, J. A. (2008). Comparison of machine learning techniques for the classification of echolocation clicks from three species of odontocetes. *Can. Acoustics* 36, 41–47.

Sayigh, L., Daher, M. A., Allen, J., Gordon, H., Joyce, K., Stuhlmann, C., et al. (2016). "The watkins marine mammal sound database: an online, freely accessible resource," in *Proceedings of meetings on acoustics*, vol. 27. (AIP Publishing). doi: 10.1121/2.0000358

Schneider, S., Von Fersen, L., and Dierkes, P. W. (2024). Acoustic estimation of the manatee population and classification of call categories using artificial intelligence. *Front. Conserv. Sci.* 5, 1405243. doi: 10.3389/fcosc.2024.1405243

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena* 404, 132306. doi: 10.1016/j.physd.2019.132306

Singh, G., Guleria, K., and Sharma, S. (2023). "A transfer learning-based pre-trained vgg16 model for skin disease classification," in *2023 IEEE 3rd mysore sub section international conference (MysuruCon)* (IEEE), 1–6. doi: 10.1109/MysuruCon59703.2023.10396942

Smelyakov, K., Honchar, Y., Bohomolov, O., and Chupryna, A. (2022). Machine learning models efficiency analysis for image classification problem. In. *COLINS*. 2022, 942–959.

Tabak, M. A., Murray, K. L., Reed, A. M., Lombardi, J. A., and Bay, K. J. (2022). Automated classification of bat echolocation call recordings with artificial intelligence. *Ecol. Inf.* 68, 101526. doi: 10.1016/j.ecoinf.2021.101526

Tan, P. S., Lim, K. M., Tan, C. H., and Lee, C. P. (2023). Pre-trained densenet-121 with multilayer perceptron for acoustic event classification. *IAENG Int. J. Comput. Sci.* 50 (1), 07.

Tang, N., Zhou, F., Wang, Y., Zhang, H., Lyu, T., Wang, Z., et al. (2023). Differential treatment for time and frequency dimensions in mel-spectrograms: An efficient 3d spectrogram network for underwater acoustic target classification. *Ocean Eng.* 287, 115863. doi: 10.1016/j.oceaneng.2023.115863

Vimal, B., Surya, M., Darshan, M., Sridhar, V., Ashok, A., et al. (2021). "Mfcc based audio classification using machine learning," in *2021 12th international conference on computing communication and networking technologies (ICCCNT)* (IEEE), 1–4. doi: 10.1109/ICCCNT51525.2021.9579881

White, E. L., White, P. R., Bull, J. M., Risch, D., Beck, S., and Edwards, E. W. (2022). More than a whistle: Automated detection of marine sound sources with a convolutional neural network. *Front. Mar. Sci.* 9, 879145. doi: 10.3389/fmars.2022.879145

Xie, Z., Li, D., Sun, H., and Zhang, A. (2023). Deep learning techniques for bird chirp recognition task. *Biodiversity Sci.* 31, 22308. doi: 10.17520/biods.2022308

Yang, M., Shen, Z., Wang, Y., Chen, J., Han, W., and Yang, S. (2023). Remote anomaly detection for underwater gliders based on multi-feature fusion. *Ocean Eng.* 284, 115179. doi: 10.1016/j.oceaneng.2023.115179

Yin, J., Ding, J., Yang, Y., Yu, J., Ma, L., Xie, W., et al. (2025). Wave-induced motion prediction of a deepwater floating offshore wind turbine platform based on bi-lstm. *Ocean Eng.* 315, 119836. doi: 10.1016/j.oceaneng.2024.119836

Zhang, J. (2021). Music feature extraction and classification algorithm based on deep learning. *Sci. Programming* 2021, 1651560. doi: 10.1155/2021/1651560

Zhao, C. W., Biao, K., and Zhu, J. (2023). Acoustic identification of marine mammals based on stacking classification fusion. *J. Hangzhou Dianzi Univ.* 43, 7–13.

Zhong, M., and Cai, W. (2019). Marine mammal sound recognition based on feature fusion. *Electronic Sci. Tech.* 32, 32–37.

Zubair, S., Yan, F., and Wang, W. (2013). Dictionary learning based sparse coefficients for audio classification with max and average pooling. *Digital Signal Process.* 23, 960–997. doi: 10.1016/j.dsp.2013.01.004