Engineering Applications of Artificial Intelligence 127 (2024) 107320

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A ship-radiated noise classification method based on domain knowledge embedding and attention mechanism

Lu Chen, Xinwei Luo *, Hanlu Zhou

Key Laboratory of Underwater Acoustic Signal Processing (Southeast University), Ministry of Education, Nanjing, 210096, China

ARTICLE INFO

Keywords:
Ship-radiated noise classification
Cyclostationary analysis
Fusion features
Hierarchical underwater acoustic transformer
Attention mechanism

ABSTRACT

Ship classification based on machine learning (ML) has proven to be a significant underwater acoustic research direction. One of the critical challenges rests with how to embed domain signal knowledge into ML models to obtain suitable features that highly correlate with the classification and create better predictors. In this paper, a novel ML-based ship classification model, Hierarchical Underwater Acoustic Transformer (HUAT), is proposed to improve the classification performance. Firstly, the Detection of Envelope Modulation on Noise (DEMON) spectra of ship-radiated noise signals are estimated by cyclostationary analysis. The motivation for using a DEMON-based preprocessing scheme is that valuable propeller information can be revealed by exploiting the second-order cyclostationarity of ship-radiated noise signals. Secondly, the useful features of DEMON spectra are enhanced using a multi-head self-attention module, and the potential features of the Mel spectrograms are extracted employing a Convolutional Neural Network (CNN) module. The two kinds of features are fused to provide ship classification patterns. The challenge of feature learning in the deep classification model is reduced by leveraging domain-related classification knowledge. Finally, the Swin Transformer, based on shifted window self-attention mechanism, is used to learn high-level feature representations and conduct ship classification. Experimental results show that the HUAT model achieves excellent classification performance on ship-radiated noise datasets, ShipsEar and DeepShip. And its classification efficiency is better than the model based on traditional Transformer architecture. In addition, the proposed method provides technical support for the underwater intelligent system capable of automatically sensing sailing vessels and recognizing vessel types.

1. Introduction

Sound waves are the only physical medium that can travel long distances in the ocean (Li, 2012). They are considered the best information carriers for sensing and recognizing underwater targets. In the area of underwater acoustic signal processing, the classification of ship-radiated noise is an essential study direction with significant economic and military implications. Ship-radiated noise classification models have the potential to be utilized in the automatic detection of maritime traffic, contributing to the assurance of maritime traffic safety (Wenz, 1972). Furthermore, these models can find application in ship detection and warning, thereby enhancing support and security for an increasingly diverse range of marine activities (Robert J. Urick, 1983). With the rapid growth of machine learning (ML) technologies, the research trend in ship classification has gradually changed from approaches based on manual features to ML methods. Ship classification based on ML includes data preprocessing, feature extraction, and classification processes.

Ship-radiated noise received by passive sonar carries the characteristic information about ships, which includes mechanical noise, propeller noise, and hydrodynamic noise (Li, 2012; Robert J.Urick, 1983). The cavitation noise of the propeller contains some physical parameters of vessels, such as the number of propeller blades and the spindle speed of a ship target, which are critical for ship classification (Hanson et al., 2008). DEMON analysis is a commonly used technique for obtaining the envelope spectrum of propeller cavitation noise at low frequencies by demodulating the broadband signal. The classical strategy of DEMON analysis is the multirate subband demodulation method (Clark et al., 2010). Correct subband and weighting factors are essential for achieving DEMON spectra with a high signal-to-noise ratio, which requires defining bespoke parameters for each signal. Besides that, the other widely used features for ship signal representation are the Low Frequency Analysis and Recording (LOFAR) spectrum, the Mel Frequency Cepstral Coefficient (MFCC), the Gammatone Frequency Cepstral

* Corresponding author.
E-mail address: luoxinwei@seu.edu.cn (X. Luo).

https://doi.org/10.1016/j.engappai.2023.107320
Received 2 August 2023; Received in revised form 27 September 2023; Accepted 14 October 2023
Available online 21 October 2023

Coefficient (GFCC), the Mel frequency spectrogram, etc. In recent years, these features have usually been combined with an ML model for ship classification (de Moura and de Seixas, 2015; Wang et al., 2019). However, these studies usually use traditional signal processing methods for improving signal features, which are then directly fed into the classifier (Kamal et al., 2021; F. Liu et al., 2021). Most of them do not perform further extraction of potential features based on the peculiarity of the signal features. Which may cause the classification ability of ML models cannot be effectively utilized.

According to prior studies (Barros et al., 2022; F. Liu et al., 2021; Luo et al., 2021), acquiring signal features that contain critical information is essential for classification performance and is more likely to benefit from the ML model. How to obtain useful signal features and effectively embed them into the ML model to achieve improved classification performance is still an issue that remains worth thinking about. This paper analyzes the cyclostationary behavior of ship-radiated noise and produces high-quality DEMON spectra. They are fused with the Mel spectrograms into fusion features as the input of our classifier. Before feature fusion, the DEMON spectra and Mel spectrograms underwent additional processing utilizing two different automatic feature extractors. A multi-head self-attention module is utilized to enhance useful features at specific cycle frequency of DEMON spectra, and a CNN model is employed to extract the potential features of the Mel spectrogram. The combination of the DEMON spectrum, which represents envelope information of the ship's propeller noise, and the Mel spectrogram, which reflects the time-frequency properties of the signal, can accurately depict the potential features of the ship-radiated noise.

The Transformer model (Vaswani et al., 2017), based on attention mechanism, is one of the mainstream frameworks in recent years. It has been widely used in the fields of natural language processing (NLP) (Raffel et al., 2020), computer vision (CV) (Dosovitskiy et al., 2020), and audio classification (Noumida and Rajan, 2022) and shown excellent performance. However, the application of Transformer architecture is still uncommon in the UATR domain. This paper uses the Swin Transformer based on shifted window self-attention as the classifier. It is a lightweight Transformer architecture, which saves a large amount of training time and resources. Additionally, the attention-based Transformer is better at capturing the local and global information of the input data when compared to CNNs and RNNs. It acquires the potential features of inputs more effectively and has achieved exceptional performance in various tasks. The main contributions of this paper are summarized as follows:

- This paper analyzes the cyclostationarity of ship-radiated noise signals and acquires high-quality DEMON spectra.
- This paper proposes a novel feature fusion method to embed domain signal knowledge into the classifier. A multi-head self-attention module is used to enhance useful features of DEMON spectra, and a CNN model is employed to extract the potential features of the Mel spectrogram before feature fusion.
- This paper proposes a Hierarchical Underwater Acoustic Transformer (HUAT) model for ship classification that uses the Swin Transformer, which is good at extracting global and local information from inputs, as the classifier.

The rest of this paper is organized as follows: Section 2 analyzes the related work. Section 3 describes the cyclostationarity of ship-radiation noise signals and cyclostationary analysis methods. Section 4 presents the details of our ship classification method. Section 5 gives the dataset processing method, model parameters, and analysis of experimental results. Section 6 extracts the conclusion. Section 7 discusses the limitations of our work and future work.

2. Related work

2.1. Applications of cyclostationary analysis

Cyclostationary analysis has numerous applications in fault diagnosis, condition monitoring, spectrum sensing, and biomedical fields (Napolitano, 2016). Chen (2020) analyzes the cyclostationarity of vibration signals and employs a Convolutional Neural Network (CNN) model to improve the classification performance of rolling element bearing faults. Zhu et al. (2005) research the one-to-third-order cyclostationarity of gearbox vibration signals, which can be effectively used to detect the wearing state of gearboxes. Kazemi et al. (2014) demonstrate that vital sign detection using cyclostationary analysis is insensitive to signal-to-noise levels and can automatically mask significant amounts of noisy information. They process Doppler radar signals modulated by physiological motion using the cyclostationary method to capture the hidden periodicity in heartbeat and respiration rates accurately. Additionally, the ship-radiated noise exhibits stationary second-order cyclostationary behavior (Fırat and Akgül, 2018). With the aid of spectral correlation, a powerful technique for assessing cyclostationarity, Barros et al. (2022) acquire the cyclic spectra of ship-radiation signals. Then, cyclic spectra are used as input to a CNN model for classifying ships. These methods usually input the cyclic spectra to ML models directly for downstream tasks, which cannot fully explore the potential information contained in the ship-radiated noise that consists of mechanical noise, propeller noise, and hydrodynamic noise (Li, 2012; Robert J.Urick, 1983). Therefore, this paper proposed a feature fusion strategy for effectively extracting ship-radiated noise signal features.

2.2. Machine learning-based ship-radiated noise classification methods

Most of the early ML-based ship classification studies utilized statistical learning techniques based on Support Vector machines (SVM), such as Single-class SVM (de Moura and de Seixas, 2015) and BAT + SVM (Sherin B. M. and Supriya M. H., 2015). Moura et al. (2015) use LOFAR spectra of ship-radiated noise as the input of a single-class SVM to recognize different ship types. Sherin and Supriya (2015) utilize an SVM classifier for four types of ship noise signal classification. They use a new algorithm to optimize parameters and achieve higher classification accuracy than the SVM classifier based on other kernel optimization methods. However, these methods have difficulty accurately extracting the potential features of signals. And SVM classifiers are usually unsatisfactory in solving multi-classification problems.

Fully connected neural networks, one of the basic ML architectures, typically combine with signal feature extraction methods for automatic classification (Khishe and Mohammadi, 2019; Qiao et al., 2021; Wang et al., 2019). The passive sonar target recognition method proposed by Khishe and Mohammadi (2019) first uses the Mel frequency cepstral coefficient (MFCC) to extract signal features. Then it uses the Salp Swarm Algorithm (SSA) to train a fully connected neural network for target recognition to avoid the network falling into local optimization and improve the model convergence speed. Wang et al. (2019) use GFCC and MEMD to extract the multi-dimensional features of signals and propose an improved deep neural network (DNN) to recognize ships, marine mammals, and background noise. However, the single structural form of fully connected neural networks restricts the ability to extract potential features from samples.

At present, many studies use deep learning models like CNNs and Recurrent Neural Networks (RNNs) to construct end-to-end classification models (Kamal et al., 2021; F. Liu et al., 2021), which directly map waveforms or spectrograms to their labels (Hu et al., 2021; Luo et al., 2021; Zhou and Yang, 2020). The CNNs are typically used in conjunction with manual feature extraction methods. Luo et al. (2021) combine the ResNet model and multiple STFTs to recognize the type of ship. Hu et al. (2021) use depthwise separable convolution and dilated

convolution to extract the features of one-dimensional time domain signals for passive underwater acoustic target recognition. Zhou and Yang (2020) use the CNN network for denoising and representation of ship signals, and the random forest is the classifier for target recognition of denoised data. Experimental results show that the CNN-based method can learn a better denoising representation of the underwater acoustic data than the traditional methods. CNNs are good at capturing local information but usually ignore global features. Setting a larger convolution kernel or stacking more layers to obtain global information will substantially increase the model's complexity.

RNNs are skilled in capturing abstract feature information from sequence data, which has attracted the attention of researchers for ship classification (Kamal et al., 2021; F. Liu et al., 2021). The training of RNNs has proved problematic because the backpropagated gradients will grow or shrink at each time stamp, so over many time stamps, they typically explode or vanish (Bynagari, 2020). Most studies use the variant architectures of RNNs to build target recognition networks, such as Long Short-Term Memory (LSTM) (Gers et al., 2000), Bi-directional Long Short-Term Memory (Bi-LSTM), etc. F. Liu et al. (2021) use the constructed 3-D features and the convolutional recurrent neural network for ship target recognition. Kamal et al. (2021) propose a model that combines CNN with LSTM for target recognition in shallow sea acoustic signals and uses one layer of selective attention mechanism to select useful features. The proposed method is one of the few methods using the attention mechanism in the ship classification domain. Since the input of the current moment of RNNs needs to depend on the previous output, this dependency structure is unfriendly to massively parallel computing.

The classical attention-based Transformer model (Vaswani et al., 2017) has shown promising results in the areas of NLP (Liu et al., 2022; Raffel et al., 2020), CV (Dosovitskiy et al., 2020; Z. Liu et al., 2021), and audio classification (Gong et al., 2021; Noumida and Rajan, 2022). The Transformer architecture breaks through the parallel computing bottleneck of RNNs. And the global information extraction ability of the attention-based Transformer architecture is much better when compared with CNNs. However, the applications of Transformer architecture in ship classification filed is still rare. The UATR-transformer (Feng and Zhu, 2022) is a convolution-free architecture for underwater acoustic target recognition. The proposed method establishes a mapping relationship between Mel spectrogram and ship types and achieves high classification accuracy. The method does not take the unique ship-radiated noise characteristic into account. Therefore, the final classification results are easily disturbed by the quality of the spectrogram.

3. DEMON spectrum analysis based on cyclostationarity

Cyclostationarity is a set of processes that exhibit periodicity in their statistics (Antoni et al., 2017). The periodic rotation of the propeller gives the cavitation noise cyclostationarity. The modulation process of ship-radiated noise $x(t)$ can be expressed as:

$$x(t) = (1 + m(t))s(t) + n(t) = (1 + Kcos(2\pi f_0 t))s(t) + n(t) \tag{1}$$

where $x(t) = (s_{t_1}, s_{t_2}, \ldots, s_{t_n})$ is the time-domain representation of the ship-radiated noise signal segment, s_{t_i} is the amplitude of signal $x(t)$ in time t_i, n is the signal length, a 5 s segment waveform of $x(t)$ is shown in Fig. 1 (a). $m(t) = Kcos(2\pi f_0 t)$ is the propeller modulation function, f_0 is the fundamental frequency of propeller, K is the modulation depth, $s(t)$ is the noise generated by the rupture of the air bubble caused by the propeller rotation, and $n(t)$ is background noise. Both $s(t)$ and $n(t)$ are assumed to be stable random noise.

The first-order statistic of the signal $x(t)$, with a mean of:

$$m_x(t) = E[x(t)] = E[(1 + Kcos(2\pi f_0 t))s(t) + n(t)] = m_s(1 + Kcos(2\pi f_0 t)) + m_n \tag{2}$$

where m_s and m_n are constants, assume the period is T_0, then $m_x(t) = m_x(t + T_0)$, indicating that the mean value of the ship-radiated noise is a periodic function and that the ship-radiated noise conforms to first-order

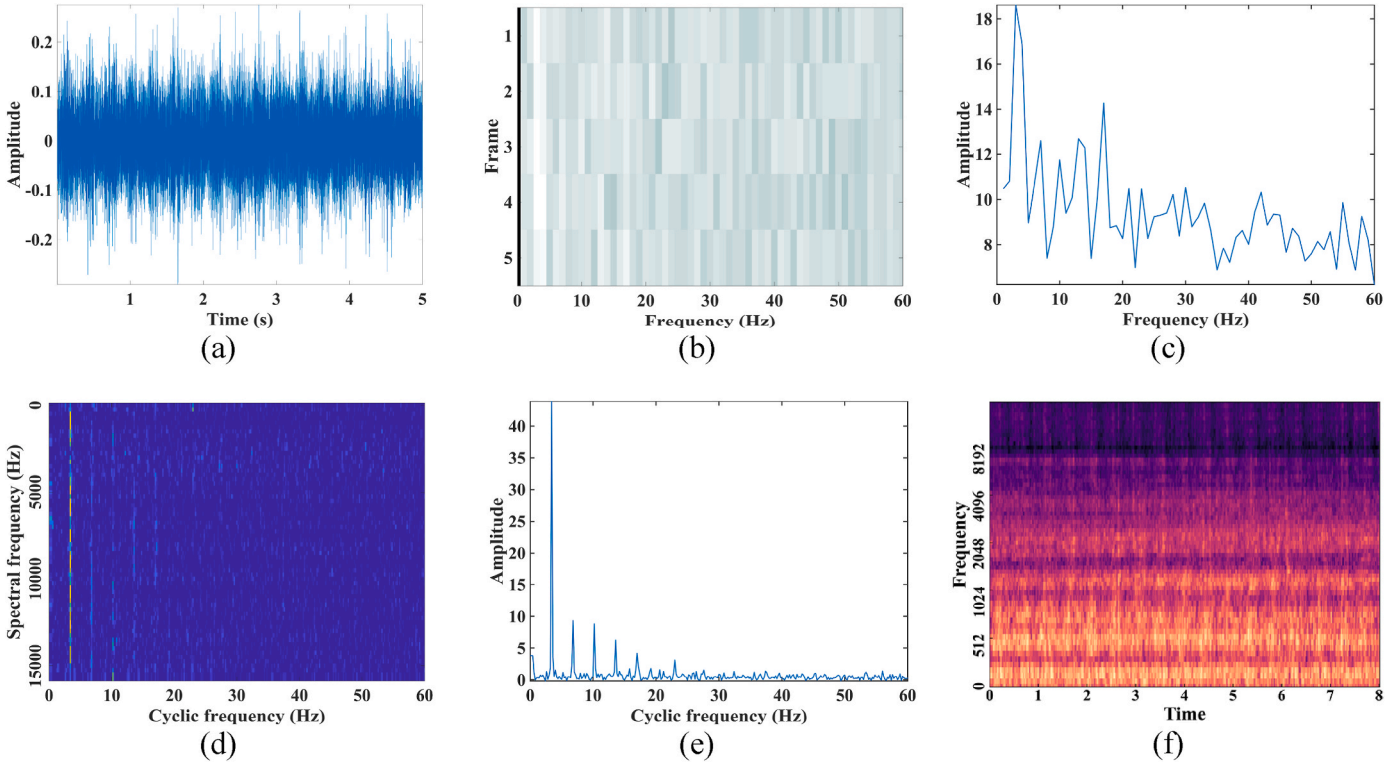


Fig. 1. Examples of signal DEMON spectrum and Mel spectrogram: (a) original 5s segment, (b) DEMON spectrum obtained using multirate subband demodulation method, (c) envelope spectrum of (b), (d) DEMON spectrum acquired using FAM algorithm, (e) enhanced envelope spectrum of (d), and (f) Mel spectrogram.

cyclostationarity.

The second-order statistic of the signal $x(t)$, with an autocorrelation function of:

$$\begin{aligned} R_x(t,\tau) &= E[x(t)x(t+\tau)] \\ &= E[[(1+m(t))s(t)+n(t)][(1+m(t+\tau))s(t+\tau)+n(t+\tau)]] \\ &= E[m(t)m(t+\tau)]R_s(\tau)+E[m(t)](R_s(\tau)+m_s m_n) \\ &+E[m(t+\tau)](R_s(\tau)+m_s m_n)+2m_s m_n+R_s(\tau)+R_n(\tau) \end{aligned} \tag{3}$$

where $R_s(\tau)$ and $R_n(\tau)$ are the autocorrelation functions of $s(t)$ and $n(t)$, respectively, and m_s and m_n are the mean values of $s(t)$ and $n(t)$, respectively. As shown in Eq. 3, $R_x(t,\tau)$ is a periodic function, demonstrating that the ship-radiation noise conforms to second-order cyclostationarity. Therefore, cyclostationary signal processing tools can be used for feature extraction.

A powerful tool for extracting hidden cyclostationary information from signals is Spectral Correlation (SC) (Antoni et al., 2017), which is calculated as follows:

$$SC_x(\alpha,f) = \lim_{N\to\infty}\sum_{n=-N}^{N}\sum_{m=-\infty}^{\infty} R_x(t_n,\tau_m)e^{-j2\pi t_n\alpha}e^{-j2\pi\tau_m f} \tag{4}$$

where α is the cyclic frequency (modulation frequency), f is the spectral frequency (carrier frequency), $t_n = \frac{n}{F_s}$, $\tau_m = \frac{m}{F_s}$, F_s is the sampling frequency.

Cyclic spectral estimation methods include the Cyclic Periodogram Method (Boustany and Antoni, 2005), the FFT Accumulation Method (FAM) (Roberts et al., 1991), the Strip Spectral Correlation Algorithm (SSCA) (Simic and Simic, 1999), and the Fast Spectral Correlation Algorithm (Antoni et al., 2017), etc. These algorithms have greatly facilitated the application of cyclostationarity in various areas of signal processing.

This paper uses the FAM algorithm to estimate the Cyclic spectra of the ship-radiated noise and compares them with the DEMON spectra obtained using the multirate subband demodulation method (Clark et al., 2010). For a ship-radiated noise signal of length 5s, the DEMON spectra obtained using the two methods are shown in Fig. 1.

Where Fig. 1 (e) is the Enhanced Envelope Spectrum (EES) (Antoni et al., 2017), which is calculated as follows:

$$SC_{o,x} = \frac{SC_x(\alpha,f)}{\sqrt{SC_x(0,f).SC_x(0,f-\alpha)}} \tag{5}$$

$EES(\alpha) = \int_{f_1}^{f_2} |SC_{o,x}|df$ (6) where $SC_{o,x}$ is the normalized form of $SC_x(\alpha,f)$, the spectrum shown in (d) is the spectrum after normalization, and $[f_1,f_2]$ is the given frequency band range. $SC_{o,x}$ is part of the inputs of the proposed HUAT model.

As can be seen from Fig. 1, the DEMON spectrum obtained by the multirate subband demodulation method is highly disturbed by noise. Additionally, the frequency domain resolution is low due to the short length of the original signals, which makes it challenging to extract valuable information from the ship-radiated noise signals. It is worth noticing that the FAM algorithm based on spectral correlation can suppress noise interference to a large extent and extract the fundamental frequency and harmonics of propeller noise from original signals. These are important ship-radiated noise signal domain knowledge for improving ship classification performance.

4. Methodology

Fig. 2 shows the flow chart of the proposed HUAT model. Firstly, DEMON spectra and Mel-frequency spectrograms of ship-radiated noise signal segments $[x_1(t),x_2(t),\ldots,x_N(t)]$ are calculated, where N is the number of samples. They are the inputs of the HUAT model. The DEMON spectrum is calculated using the FAM algorithm mentioned in Section 3. And Section 4.1 gives the calculation process of Mel-frequency spectrograms. Then multi-head self-attention and a two-layer convolution with a small kernel are used to extract valuable features from DEMON spectra and Mel spectrograms. They are concatenated together as the input feature of our classifier and finally get predicted labels for signals. The classifier of the HUAT model is the Swin Transformer (Z. Liu et al., 2021). And the output of our model is the ship type label corresponding to the ship-radiated noise signal. Tables 1 and 2 in Section 5.1 present the label of different signals.

4.1. Mel spectrogram and feature fusion method

Another feature involved in our feature fusion approach is the Mel spectrogram. The ship-radiated signals $[x_1(t),x_2(t),\ldots,x_N(t)]$ are processed with a 1024-frame Hanning window and a 320-frame shift, then computed in the Fast Fourier Transform (FFT) (Cooley and Tukey, 1965) of each segment and filtered by a 64-dimension log Mel filter bank. The $T\times F$-dimension Mel spectrograms are part of the input to the classifier. Fig. 1 (f) shows the Mel spectrogram, X_{mel}, of the original signal $x(t)$ in Fig. 1 (a).

As shown in Fig. 3, two automatic feature extractors are used to extract potential features from the DEMON spectrum and the Mel spectrogram before feature fusion. The DEMON spectrum contains propeller shaft frequency, blade frequency, and noise information

Table 1
Class description of ShipsEar dataset.

Label	Vessel type	Dataset size
Class A	fishing boats, trawlers, mussel boats, tugboats and dredgers	341
Class B	motorboats, pilot boats and sailboats	301
Class C	passenger ferries	801
Class D	ocean liners and ro-ro vessels	494
Class E	background noise recordings	200

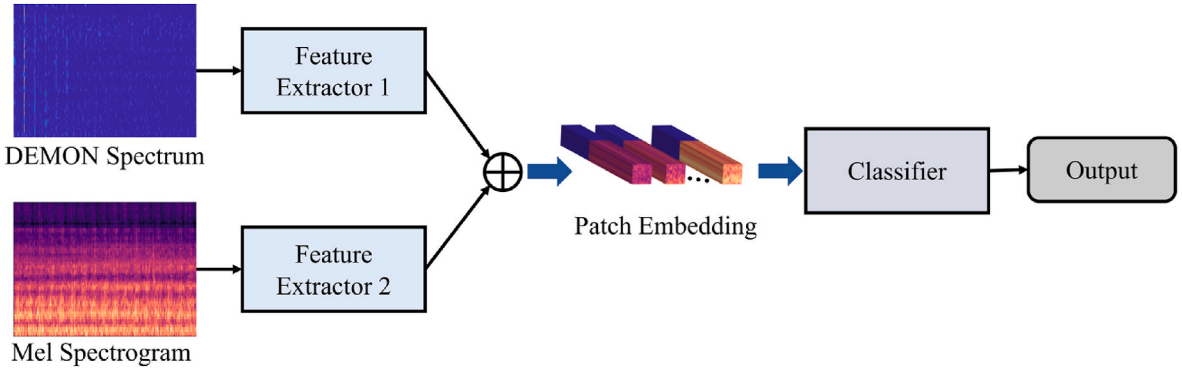


Fig. 2. The flow chart of HUAT model.

Table 2
Class description of DeepShip dataset: the Data size before "/" is the total data segments, and 2000 of them are used for model training and testing.

Label	Dataset size	Label	Dataset size
Cargo	6491/2000	Passenger Ship	8716/2000
Tug	7377/2000	Tanker	6883/2000

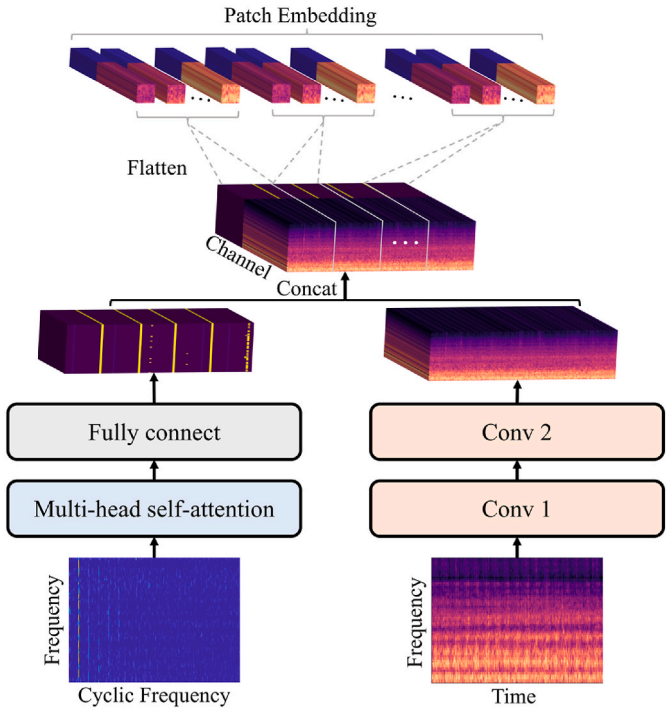


Fig. 3. Flow chart of automatic feature extraction and fusion method.

(Hanson et al., 2008). Therefore, a multi-head self-attention with eight heads is applied to allow the model to notice the modulation information better. It is helpful to obtain the useful features of the DEMON spectrum. Then a fully connected layer is used to adjust the feature channels. The calculation process is shown as follows:

$$Att_i = softmax\left(\frac{W_i^q X_{dem} * W_i^k X_{dem}}{\sqrt{W_i^v X_{dem}}}\right) \tag{7}$$

$$Att_{output} = concat(Att_1, Att_2, \ldots, Att_h) \tag{8}$$

$Feat_{dem} = FC(Att_{output})$ (9) where $X_{dem} = SC_{o,x}$, W_i^q, W_i^k, and W_i^v are the learnable parameters, $h = 8$ is the head number, $softmax()$, $concat()$, and $FC()$ are activation function, concatenate operation and fully connected layer, respectively.

The subgraph (b) of Fig. 4 is the multi-head self-attention feature $Feat_{dem}$ of the DEMON spectrum shown in (a). The feature diagram indicates that the features of some cycle frequencies are enhanced. As shown in the overlapped chart Fig. 4 (c) of (a) and (b), the enhanced part exactly corresponds to the frequencies with the most concentrated spectral energy of DEMON, while other noises are weakened.

At the same time, a two-layer CNN with a small kernel size is used to extract the potential features of the Mel spectrogram. Inspired by the VGG network (Simonyan and Zisserman, 2015), several consecutive convolutions with small kernels are used instead of larger ones in the patch embedding stage. To ensure that the size of the input sequence remains an integer in the subsequent reshape process and to avoid information loss due to excessive clipping times, we set the two-layer CNN with a kernel size of 2×2 and a stride of 2. The calculation process is shown as Eq. (10):

$$Feat_{mel} = CNN_2(CNN_1(X_{mel})) \tag{10}$$

$$Feat_{fuse} = Feat_{dem} + Feat_{mel} \tag{11}$$

where $Feat_{mel}$ is the potential feature of Mel spectrogram, $Feat_{fuse}$ is the fusion feature of the DEMON spectrum and the Mel spectrogram.

Fig. 3 shows the process of feature fusion. The valuable information in the DEMON spectrum is primarily focused on the low-frequency band of the cyclic frequency. Therefore, we intercept the low-frequency part of the DEMON spectrum with length T and merge it with the Mel spectrum. The frequency dimensions of the DEMON spectrum and the Mel spectrum are identical. Consequently, the fused feature dimension is $T \times F \times C$, where $C = C1 + C2$ is the sum of the channel numbers of the DEMON and Mel features.

The difference between an audio spectrogram and an image is that the length and width dimensions of the spectrogram represent different information. And one dimension is usually much longer than another. We first split the longer dimension into F frame windows to better capture the potential information in a short frame. Then we calculate the average pooling of fusion features, $Feat_{fuse}$, and flatten them into $T/4 \times F/4$ sequences $Feat_{PaE}$ in the order of time windows, where a token in the output sequence is a patch, and the fusion feature sequences are called patch embeddings. They are the input sequences of the classifier.

4.2. Classifier architecture

The left part of Fig. 5 is an overview of the Swin Transformer (Z. Liu et al., 2021) architecture. The input is a set of fusion features from DEMON spectra and Mel spectrograms. To make the model capture the correlation between patches well, we add absolute position embedding Y_{PoE} to the patch embeddings like the BERT model (Devlin et al., 2019).

$Feat'_{PaE} = Feat_{PaE} + Y_{PoE}$ (12) where Y_{PoE} is trainable position embedding, $Feat'_{PaE}$ is the input of the Swin Transformer module.

The calculation process of the Swin Transformer (Z. Liu et al., 2021) block is shown in Eqs. (13) and (14). These processes will repeat n_1 to n_4 times in each module, respectively.

$$Y'_k = SWA(LN(Y_{k-1})) + Y_{k-1} \tag{13}$$

$$Y_k = MLP(LN(Y'_k)) + Y'_k \tag{14}$$

where Y'_k and Y_k represent the output of the shifted window multi-head self-attention (SWA) and multilayer perception (MLP) modules of block k, and $LN()$ is layer normalization.

The difference between the Swin Transformer (Z. Liu et al., 2021) and a traditional Transformer (Vaswani et al., 2017) is that the Swin Transformer block computes self-attention in each window. The right subgraph of Fig. 5 illustrates the shifted window approach of Swin Transformer. The window partitioning is shifted in layer $k-1$ and builds connections between non-overlapping windows in layer k. The shifted window approach moves the edge parts and joins them together, making the number of self-attention windows consistent in each layer. Most importantly, the SWA mechanism of the Swin Transformer reduces the computational parameters of multi-head attention. It only computes the attention matrix inside each $N \times N$ window, and the window size N can be customized. For an audio with tf patches, the computational complexity of SWA and global multi-head self-attention (GA) is shown in Eqs. (15) and (16):

$$\Omega(SWA) = tfH^2 + N^2H \tag{15}$$

$$\Omega(GA) = tfH^2 + (tf)^2H \tag{16}$$

where H is the hidden state dimension of patch embedding.

The computational complexity of SWA is linear when N is fixed, but

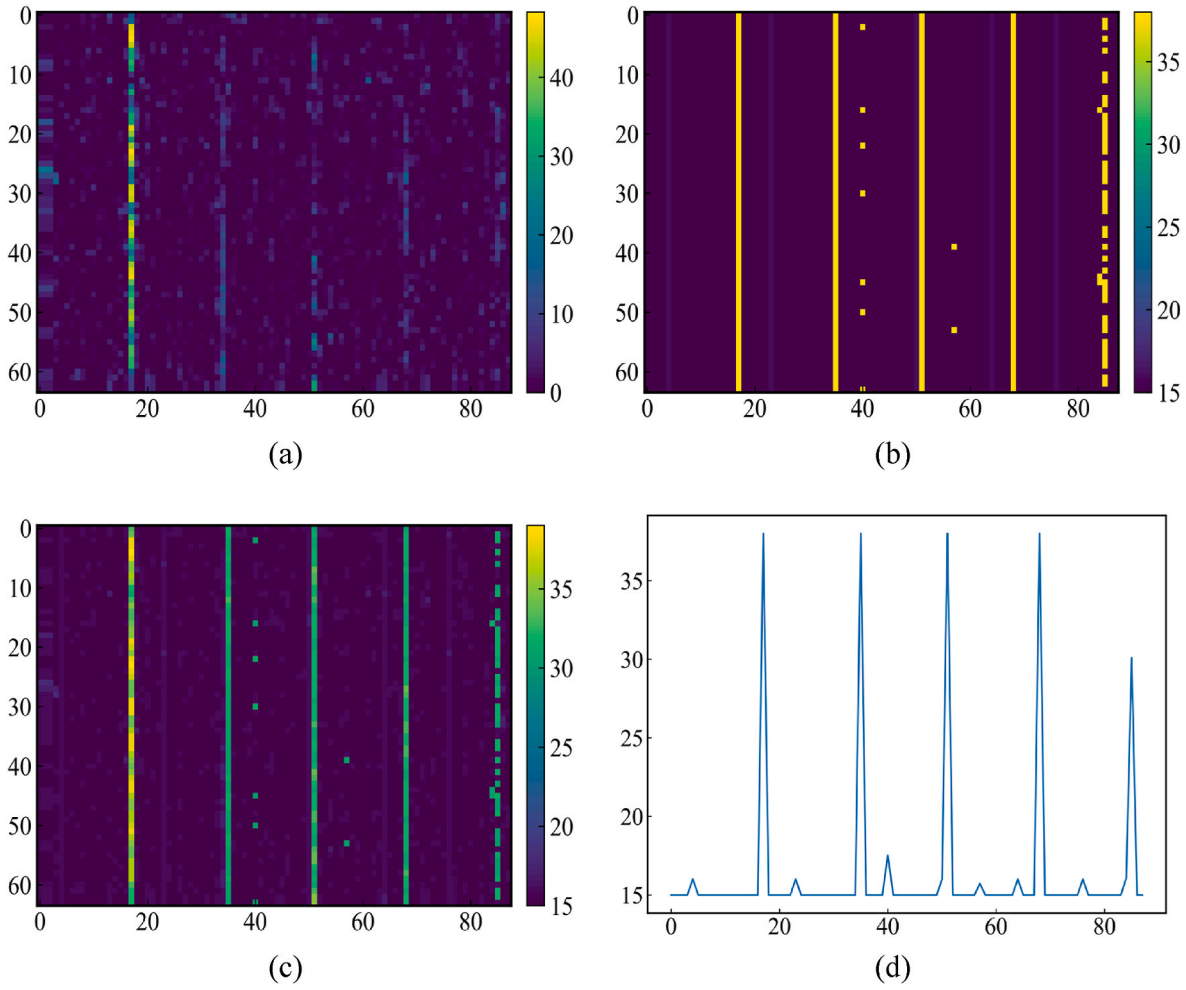


Fig. 4. The example of attention feature map visualization: (a) DEMON spectrum, (b) attention feature visualization, (c) overlapped image of (a) and (b), and (d) envelope spectrum of (b).

that of GA is square to the patch number. Except for that, we can customize the model depth and the number of repetitions of Swin Transformer blocks for each module. It is also possible to set a different number of heads for the multi-head attention mechanism of each module.

Patch merging after the Swin Transformer Block is a downsampling operation that merges adjacent 2×2 patches. After i times patch merging, it reduces the number of patches to $T/(4 \times 2i) \times F/\ (4 \times 2i)$ and increases the latent dimension to $8i \times F$. Then it reduces the latent dimension to $4i \times F$ by a fully connected layer. The Transformer module of the HUAT model includes three modules with Swin Transformer Block and Patch Merging and one module with Swin Transformer Block only, which are used to obtain the hierarchical representations of the inputs. The last two layers of the classifier are a normalization layer and a linear layer. They are used to map the dimension of the hidden layer to the number of categories for classification. Finally, use a softmax function to get the predicted labels.

5. Experimental results and analysis

In this section, we evaluate the performance of the HUAT model on the ShipsEar (Santos-Domínguez et al., 2016) and DeepShip (Irfan, 2021) datasets. We first introduce the data preprocessing and the model parameter details briefly. The classification performance of our model is then provided by analyzing and reporting a set of experimental results. We present the test results for each category of the datasets in Section 4.2.1. Additionally, the high-dimensional data visualization tool, t-SNE (Laurens van der Maaten, Geoffrey Hinton, 2008), is employed to project the high-dimensional features of the data into a two-dimensional space, allowing for visualization of the spatial distributions of the test set both before and after prediction by the HUAT model. These visualization results are presented in Fig. 7. To verify the influence of the fused features on the model's performance, we compare the classification results of the model using the DEMON spectrum, the Mel spectrogram, and the fusion features as input, respectively. Moreover, this paper compares the classification accuracy of the HUAT model with previous deep learning-based works, and the comparison results are shown in Section 4.2.4. Finally, Section 4.2.5 compares the performance and efficiency of the HUAT model with those of the AST model based on traditional Transformer architecture. The experimental findings in Section 4.2.5 add to the body of evidence supporting the robustness of our feature fusion approach.

5.1. Datasets and training parameters

ShipsEar is a ship-radiated noise dataset with 90 recordings lasting from 15 s to 10 min in wav format, covering 11 vessel types and a class of background noise. Table 1 provides an overview of each distinct class created for the ship classification task, including specific details for each one. In this paper, we downsample the recordings to 32 kHz and split them with no overlap into 2137 5-s audio clips. To maintain the balance and fairness of the training data, we randomly split each class into five

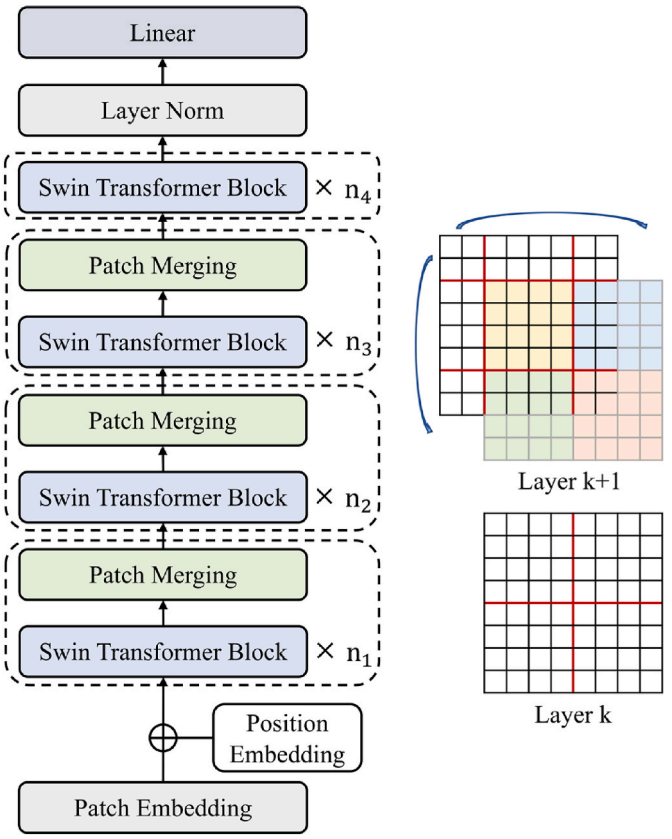


Fig. 5. The classifier architecture, n_1, n_2, n_3, n_4 mean the repeat time of the Swin Transformer Block in different modules, Layer k and Layer k+1 are two consecutive attention layers in Swin Transformer Block.

equal folds. During each experiment, four of these folds were selected for model training, while half of another fold was used for validation and the remaining half for testing.

The DeepShip dataset consists of 265 different ship signals belonging to four classes. There are a total of 47 h and 4 min of real-world underwater recordings with a sampling frequency of 32 kHz. The duration of each recording varies from about 6 to 1530 s. In this paper, we split them into 5-s segments and selected 2000 segments randomly from each category for model training and testing due to the limitation of computational resources. Then, using the same preprocessing method as ShipsEar, divide the training set, validation set, and test set. The details about the processed DeepShip dataset are shown in Table 2.

The Swin Transformer includes four versions of models with similar architecture, each of which varies in size and computational complexity. The structure of the Swin Transformer module of the HUAT model proposed in this paper is consistent with the smallest size of the Swin-T model. The cycle number of Swin Transformer Block layers in different modules shown in Fig. 5 is $n_1 = 2, n_2 = 2, n_3 = 6, n_4 = 2$, respectively. We give the number of parameters of each layer of the HUAT model in Table 3. We use a batch size of 64, the AdamW optimizer (Loshchilov and Hutter, 2019), and cross-entropy loss to train the HUAT model for 100 epochs on ShipsEar and 80 on DeepShip. The model uses a warm-up technique to adjust the learning rate from 0.02 to 0.1 over the first three epochs and then halves the learning rate every 10 epochs until it reaches 0.02. This approach follows the methodology of the Swin transformer (Z. Liu et al., 2021) and HTS-AT (Chen et al., 2022). We use a 5-fold cross-validation method to ensure the robustness of our experimental results. Additionally, this paper evaluates the classification performance using accuracy, precision, recall, and F1-score metrics.

Table 3
Number of parameters of each layer of the HUAT model, and the input and output of each module.

Modules	Layers	Number of Parameters
Feature Extractor 1	Attention layer	1579008
	Fully connected layer	524544
Feature Extractor 2	CNN1 (2 × 2)	235
	CNN2 (2 × 2)	17955
Classifier	Position embedding	393216
	Swin Transformer block 1	200480
	Patch merging layer 1	74496
	Swin Transformer block 2	893312
	Patch merging layer 2	296448
	Swin Transformer block 3	10668384
	Patch merging layer 3	1182720
	Swin Transformer block 4	14190144
	Layer Norm	1536
	Fully connected layer	3845
Total number of parameters		30026323

5.2. Results and analysis

5.2.1. Classification performance

We evaluated the classification performance of the proposed HUAT model on the ShipsEar and DeepShip datasets. The test results for each class on these two datasets are shown in Table 4 and Table 5, respectively. Fig. 6 presents the training curves of the HUAT model.

As shown in Table 4, the HUAT model obtains good classification results on the ShipsEar dataset. It achieves an accuracy of 98.62% for the classification of five classes of ships, and the F1 score, a comprehensive metric of precision and recall for each class, reaches more than 96%. Furthermore, it is worth noting that the F1 score on the ShipsEar dataset shows a positive correlation with the size of the data in each class. However, despite having the smallest number of samples, class E achieves 100% on all classification metrics. That is because class E only consists of background noise with relatively fewer features, making the label easier to predict. The training curves of the HUAT model on the ShipsEar dataset are shown in Fig. 6 (a) and (b). These curves demonstrate that the model converges normally.

Table 5 shows that the predicted results of HUAT on DeepShip have achieved 99.01% across all four indicators. Moreover, as shown in Fig. 6 (c) and (d), the accuracy curve of the model steadily increases while the loss curve steadily decreases. It indicates that the HUAT model learns the potential features of the DeepShip dataset well. The predicted values and training curves demonstrate that the powerful feature extraction capability of the transformer architecture is used to predict ship-radiated noise datasets. And the proposed HUAT model is robust on ship classification, which obtains better prediction performance on the DeepShip dataset with larger samples and fewer vessel types.

5.2.2. Feature visualization

To visualize the distribution of samples in the space before and after the model prediction intuitively, we use the t-SNE to project the high-dimensional features of the data into a two-dimensional space. Fig. 7 (a) is obtained by reducing each sample in the ShipsEar test set from

Table 4
Test results of HUAT model on ShipsEar dataset. Accuracy, precision, recall, and F1-score are the evaluation metrics (%). Support is the number of test samples for each class.

Label	Accuracy	Precision	Recall	F1-score	Support
Class A	/	98.4	97.61	98	35
Class B	/	95.4	98.57	96.96	30
Class C	/	98.3	99.5	98.9	81
Class D	/	98.8	98.42	98.61	51
Class E	/	100	100	100	20
Average	98.62	98.18	98.82	98.5	217

Table 5
Test results of HUAT model on DeepShip dataset.

Label	Accuracy	Precision	Recall	F1-score	Support
Cargo	/	99.50	98.02	98.75	200
Passengership	/	99.00	99.00	99.00	200
Tanker	/	98.53	99.01	98.77	200
Tug	/	99.01	100	99.50	200
Average	99.01	99.01	99.01	99.01	800

160,000 dimensions to 2 dimensions and mapping these two dimensions as the x-axis and y-axis, respectively, into a two-dimensional coordinates system. Fig. 7 (b) shows the distribution of the ShipsEar test samples after model prediction. The predicted features are reduced to 2 dimensions using the same method. Fig. 7 (c) and (d) are the results of the DeepShip test set.

As shown in Fig. 7 (a) and (c), the original signals on the test set are distributed randomly in 2-dimensional space. Additionally, the output features of the last layer of the model are reduced to two dimensions and visualized. As displayed in Fig. 7(b) and (d), the output features are categorized and aggregated in a two-dimensional space, with only a few points mixed with other categories. This observation suggests that the HUAT model has successfully learned discriminative information from the fusion features. This conclusion is consistent with the results presented in Section 4.2.1.

To find out the reason that some points mixed with other categories shown in Fig. 7(b) and (d). We analyze several acoustic signal segments where the predicted labels are inconsistent with the ground truth. Fig. 8 gives some examples, including their waveform, DEMON spectrum, and Mel spectrogram. It was found that the target information in these samples was usually hard to hear, or only a brief target sound could be heard at the beginning or end. As shown in the second column in Fig. 8, the DEMOM spectra of the signals with false predicted labels have no distinct line spectra, i.e., no target features are extracted. The brighter the color of the Mel spectrogram, the more concentrated the signal energy is at the corresponding frequency. The Mel spectrograms in Fig. 8. (c) and (d) show that the signal energy is weak in the lower frequency bands, suggesting that the segment contains too little target information. Moreover, when the model gives incorrect prediction results, the model is usually not very sure. For example, the prediction result distribution of signals in Fig. 8(a), (d) are $[0.1996, 0.1945, 0.1929, 0.2238, 0.1893]$ and $[0.2180, 0.2811, 0.3287, 0.1722]$, respectively. The index of the maximum value in the prediction result is the index corresponding to the predicted label. It presents that the probability of the model predicting the signal as each label is very close, which further illustrates the difficulty of the model in extracting useful classification features from these signals.

5.2.3. *Classification performance using different features*

This section compared the classification performance of our model with one kind of feature and fusion feature as input, which illustrates the effectiveness of the proposed novel feature fusion strategy that involves the DEMON spectrum and Mel spectrogram. Where DEMON spectrum and Mel Spectrogram in Table 6 means the input of our model is only one kind of feature.

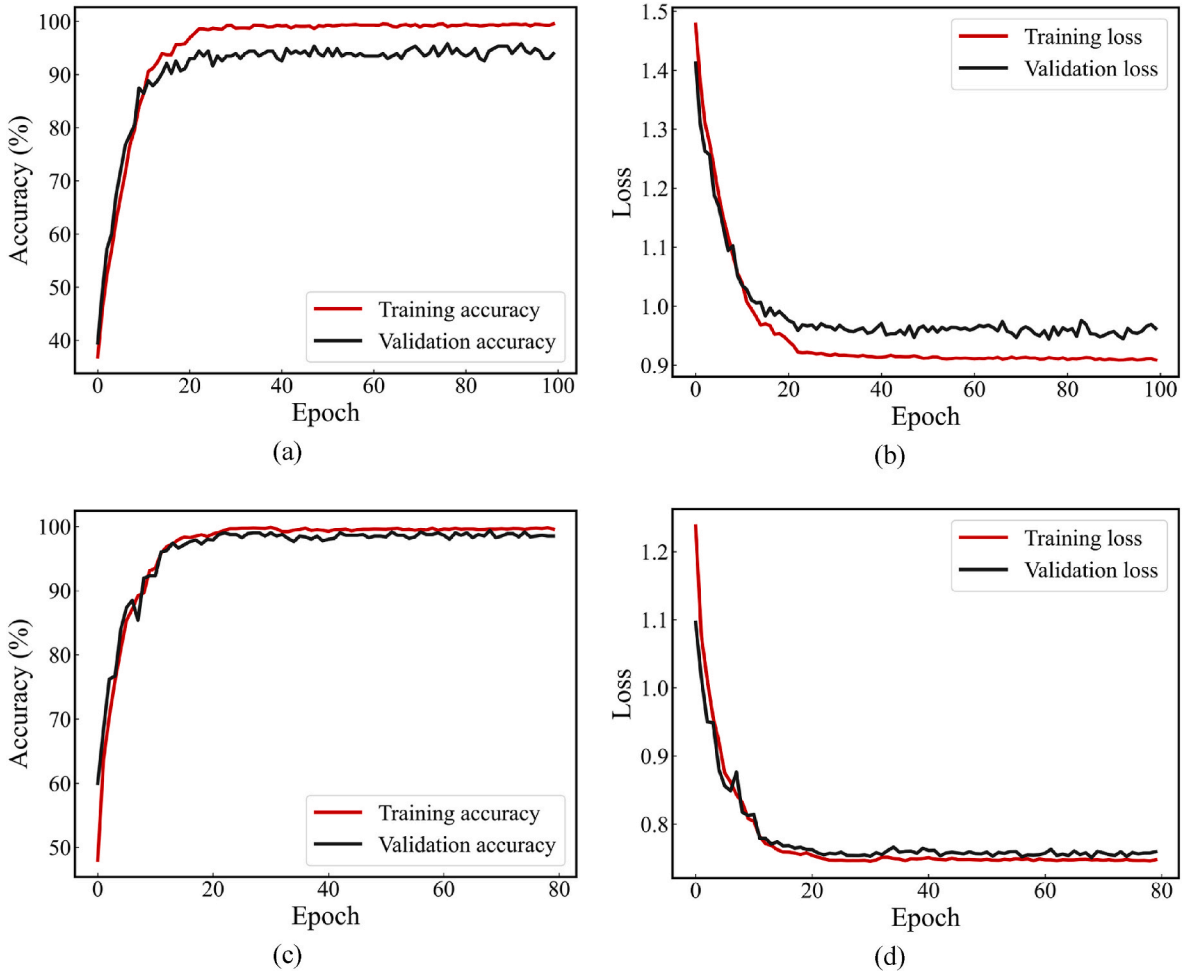


Fig. 6. Training curves on training and validation set: (a) accuracy curves on ShipsEar dataset, (b) loss curves on ShipsEar dataset, (c) accuracy curves on DeepShip dataset, and (d) loss curves on DeepShip dataset.

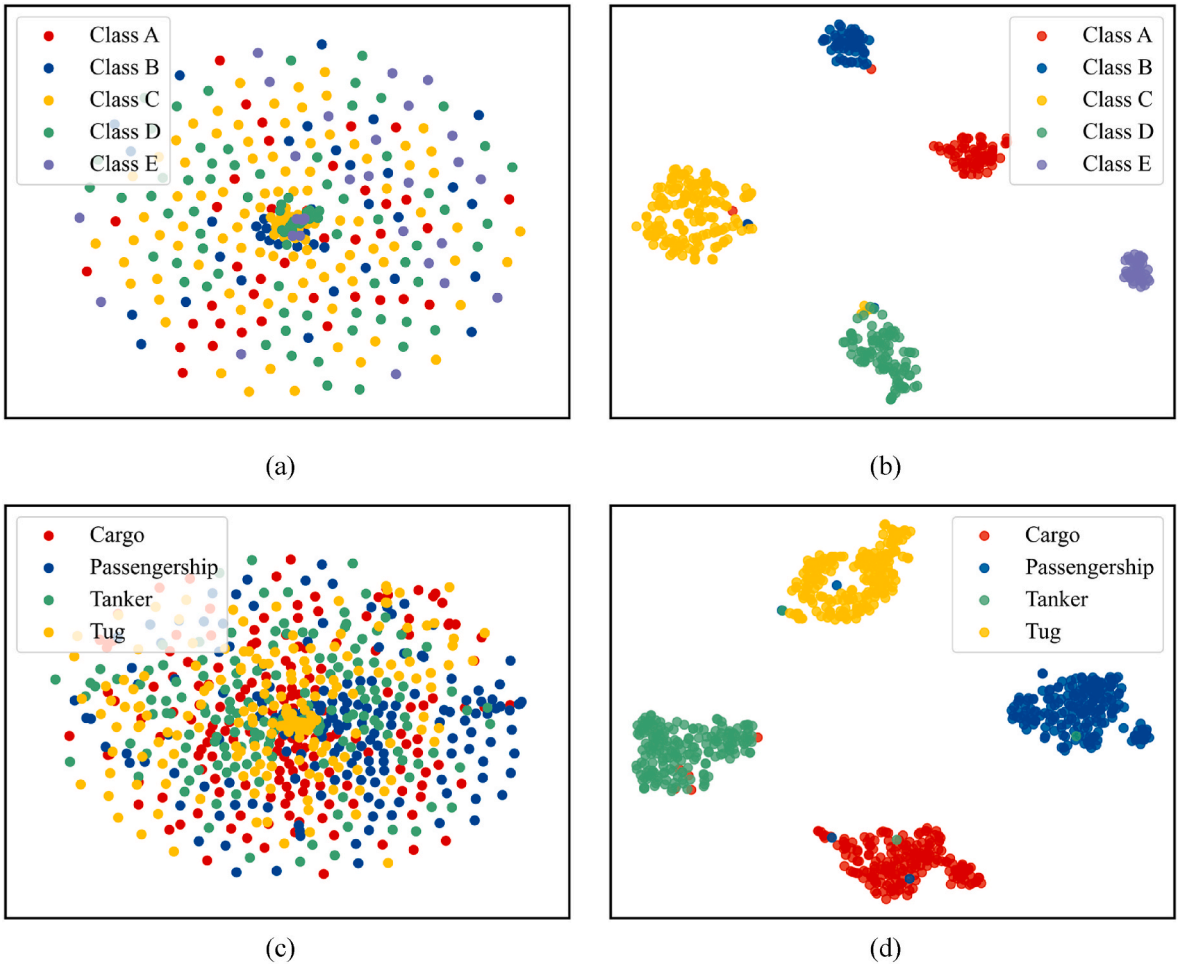


Fig. 7. Feature visualization of test set: (a) original signals of ShipsEar, (b) high dimensional features of ShipsEar, (c) original signals of DeepShip, and (d) high dimensional features of DeepShip.

Table 6 presents the test results of the HUAT model for various inputs, with the best classification achieved using the fusion features. While the input features are DEMON spectra, the branch corresponding to the Mel spectrogram and its automatic feature extractor is removed, and vice versa. But the patch embedding dimension of the individual feature inputs is the same as that of fusion features. The results shown in Fig. 9 and Table 6 mean that while the worst classification results are obtained using the DEMON spectrum as input only, its fusion with the Mel spectrogram still contributed to the training and prediction of the model. Since the DEMON spectrum only contains part of the ship-radiated information. And the limited valuable information is highlighted during the automatic feature extraction stage. Experimental results suggest that embedding domain-specific knowledge into deep learning models is beneficial to prediction results.

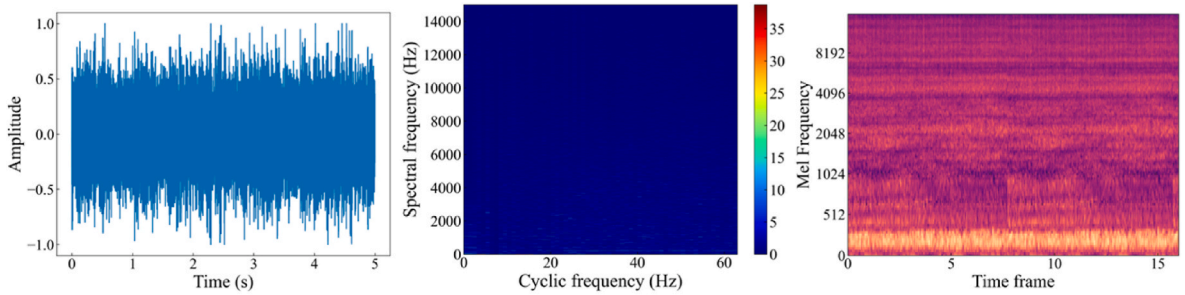
5.2.4. Comparison with previous works

Most previous ship classification research based on deep learning techniques typically applies the CNN model. And several of them use self-constructed datasets to evaluate the classification performance of the proposed methods. But these datasets cannot be reconstructed, which makes it difficult to compare the performance of different models in the same dimension. In this paper, we compare the classification accuracy of HUAT with the previous deep learning-based works that use the ShipsEar or DeepShip dataset. The results displayed in Fig. 10 show that HUAT outperformed the baseline models. The accuracy of HUAT is 1.72 percentage points higher than that of the highest UATR-Transformer on the ShipsEar dataset and 2.64 percentage points higher than that of the highest baseline model on the DeepShip dataset. The dataset processing methods of baselines are similar to this paper, so the comparison is fair.

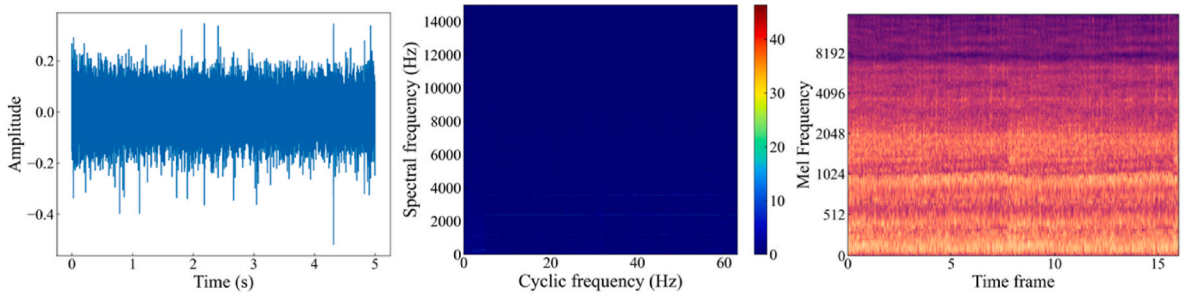
5.2.5. Prediction performance comparison of the HUAT and the AST model

Table 7 lists the details of the model performance of the HUAT and AST (Gong et al., 2021) models on two public datasets. The AST model uses Mel spectrograms as the input in the original sound recognition task, and this paper also employs Mel spectrograms of the ship-radiated noise to train the model. To verify the robustness of our feature fusion strategy, we also train the AST model using fusion features as input. Where *_Mel* and *_Fuison* in Table 7 means the input of the classifier is Mel spectrograms or fusion features, respectively. The results of the second and third columns demonstrate that the feature fusion method suggested in Section 4.1 can be conducted with various classifiers to improve classification performance.

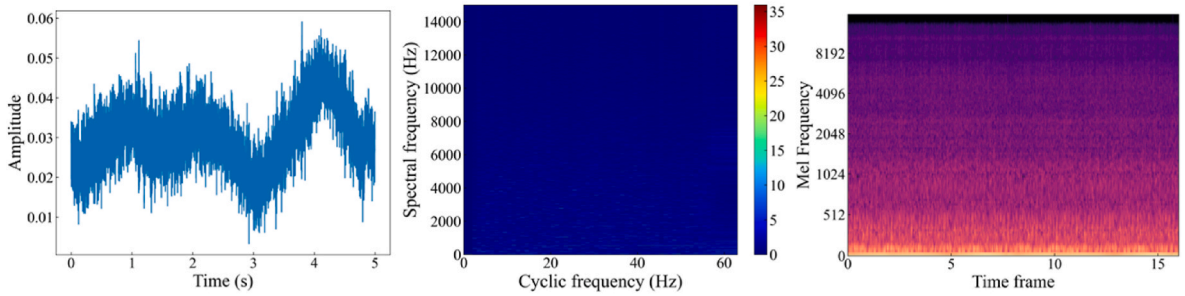
The last column of Table 7 shows the number of model parameters. The HUAT model has 30.3 M parameters, which is about one-third of the 88.19 M parameters of the AST model based on the traditional Transformer architecture. We train them on the same device (an Intel(R) Core (TM) i7-9700K CPU @ 3.60 GHz), with the same settings, and using the same dataset. Experimental results show that the prediction time of the HUAT model saves more than tenfold, and the classification accuracy on the ShipsEar dataset of the HUAT model is higher than AST. That's because the main module of the HUAT model is the Swin Transformer, which has linear complexity. Additionally, the prediction accuracy of HUAT_Mel is close to AST_Mel, which means the Swin Transformer



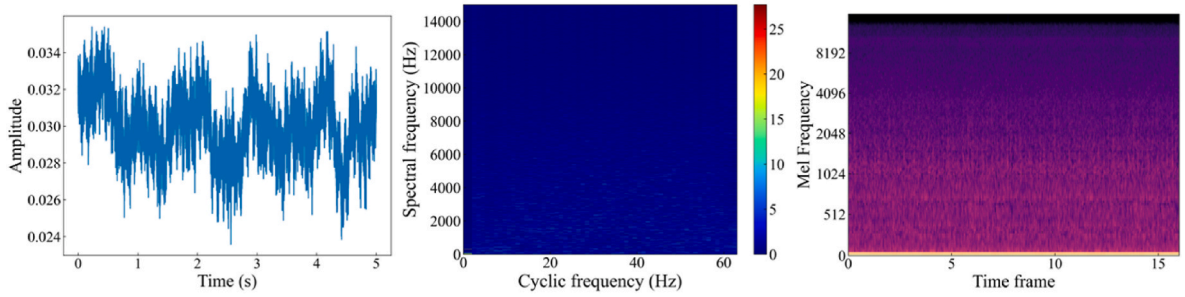
(a) Failure example 1 from ShipsEar. Ground truth: Class C, Predicted label: Class D



(b) Failure example 2 from ShipsEar. Ground truth: Class A, Predicted label: Class D.



(c) Failure example 1 from DeepShip. Ground truth: Tanker, Predicted label: Passengership.



(d) Failure example 2 from DeepShip. Ground Truth: Cargo, Predicted label: Tanker.

Fig. 8. The waveform, DEMON spectrum, and Mel spectrogram of failure examples.

Table 6
Test results of HUAT model using different features.

Dataset	Feature	Accuracy	Precision	Recall	F1-score
ShipsEar	DEMON spectrum	75.81	78.21	73.29	75.09
	Mel Spectrogram	96.63	96.66	97.13	96.83
	Fusion feature	**98.62**	**98.18**	**98.82**	**98.5**
DeepShip	DEMON spectrum	87.48	87.73	87.49	87.48
	Mel Spectrogram	97.94	98.00	97.92	97.96
	Fusion feature	**99.01**	**99.01**	**99.01**	**99.01**

reduces the number of model parameters and improves efficiency without diminishing the model prediction performance.

6. Conclusion

A new deep learning-based ship classification framework, HUAT, is presented to improve the classification performance of ship-radiated noise. The HUAT model combines the fusion features of DEMON spectra and Mel spectrograms, as well as Swin Transformer based on shifted window self-attention. Firstly, the cyclostationarity of ship signals is analyzed to achieve the DEMON spectra, which reveal the

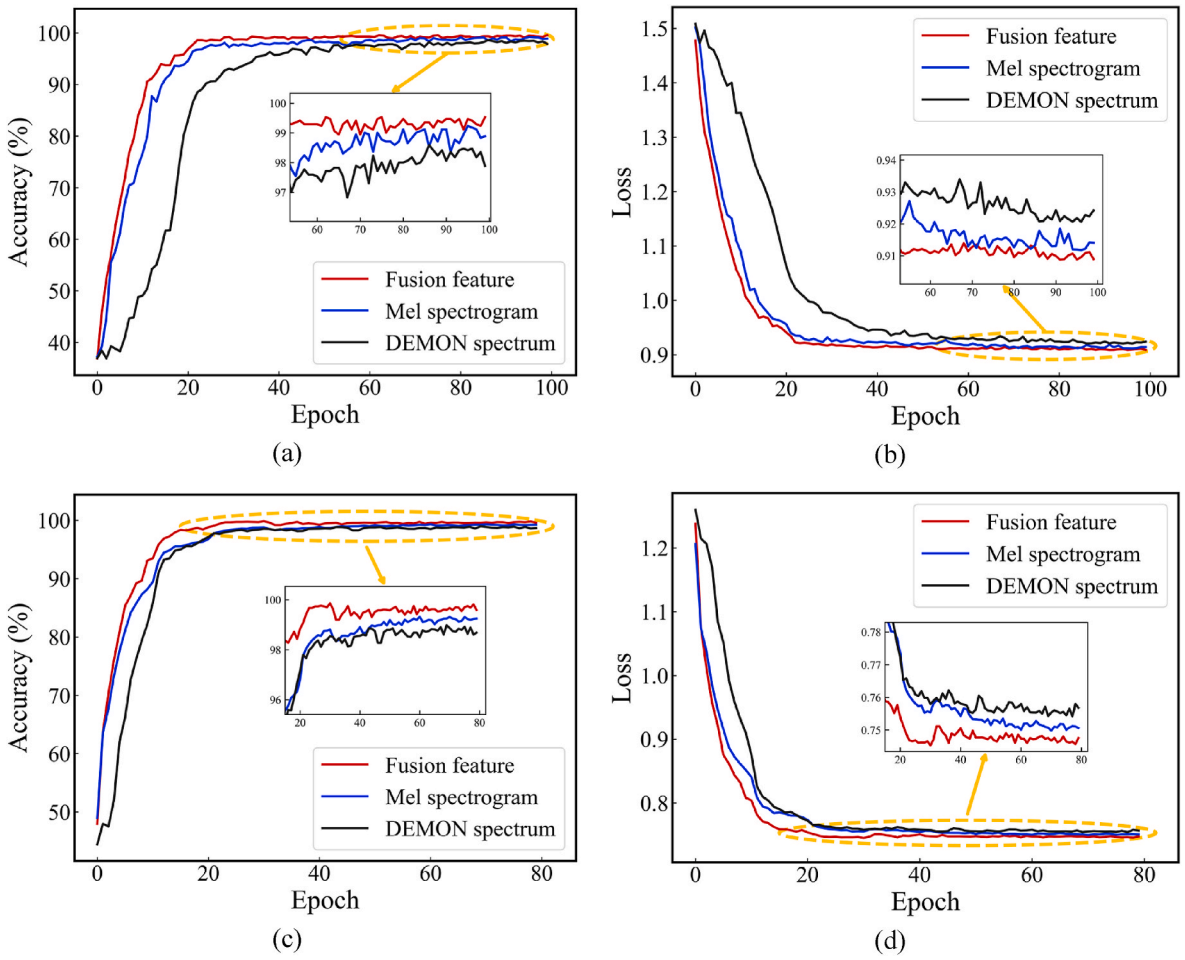


Fig. 9. Training curves of the model using different features: (a) accuracy curves on ShipsEar dataset, (b) loss curves on ShipsEar dataset, (c) accuracy curves on DeepShip dataset and (d) loss curves on DeepShip dataset.

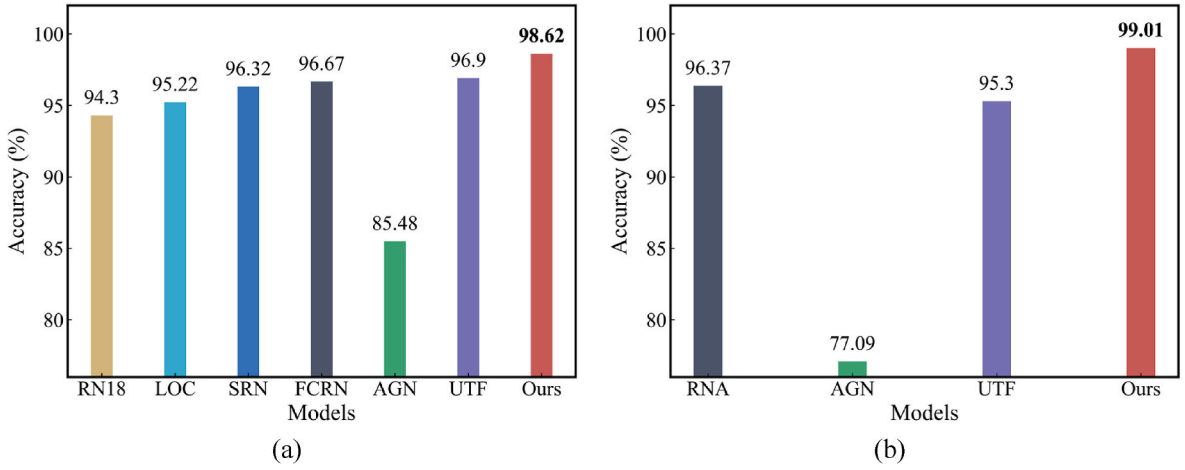


Fig. 10. Test performance comparison of baselines and the HUAT model: (a) test accuracy on ShipsEar dataset, and (b) test accuracy on DeepShip dataset. RN18: ResNet18 (Hong et al., 2021), LOC: LOFAR-CNN (Chen et al., 2021), SRN: S-ResNet (Luo et al., 2021), FCRN (Zhang et al., 2022), AGN: AGNet (Xie et al., 2022), UTF: UATR-Transformer (Feng and Zhu, 2022), RNA: ResNet18_Aug (Yao et al., 2023).

propeller nature of different vessels. Then a novel feature fusion strategy is constructed for valuable feature learning and domain knowledge embedding. Finally, the Swin Transformer is used to classify ships. Experimental results show that the HUAT model acquires excellent classification performance on the ShipsEar and DeepShip datasets. The classification accuracy of the HUAT model is 1.72 and 2.64 percentage points higher than that of the strongest baseline on ShipsEar and DeepShip datasets, respectively. The HUAT model outperforms the AST model, based on the traditional Transformer architecture, in terms of classification performance and prediction efficiency. The HUAT model

Table 7
Model performance evaluation of HUAT and the traditional Transformer-based AST model. Acc_SE and ACC_DS are the test accuracy on the ShipsEar and DeepShip datasets, respectively (%). Time_SE and Time_DS mean the average prediction time for predicting one sample for one time (s). Parameters represents the number of model parameters (million, M).

Models	Acc_SE	Acc_DS	Time_SE	Time_DS	Paramaters
AST_Mel	95.26	98.27	2.85	2.77	87.26
HUAT_Mel	96.63	97.94	0.23	0.23	**29.1**
AST_Fusion	98.57	**99.15**	3.17	3.24	88.19
HUAT_Fusion	**98.62**	99.01	**0.35**	**0.36**	30.03

achieved very competitive prediction results compared to the AST model, with a tenfold saving in prediction time. Most importantly, our feature fusion method is effective in different datasets and models.

7. Limitations and future work

There are many types of ships in the real marine environment. However, collecting ship-radiated noise signals is a high-cost task. So, we use public datasets to verify the classification performance of our method. In addition, the research on ship-radiated noise classification was removed from the method based on manual features to the machine learning method just in a few of recent years. The number of public datasets and the types of ships covered are limited. Therefore, it is hard to verify the robustness of the model. We will further explore the ship-radiated noise in-depth and generate signals in different scenarios with the help of generative modeling and reduce the cost of data acquisition.

Ship-radiated noise signals obtained from real marine environments contain various noises. The extraction of useful feature information from real signals, particularly those with low signal-to-noise ratios, and the development of an effective ship classification model remain areas that require further research and investigation.

Current research on the ML-based method for ship-radiated noise classification, including this paper, has to rely on traditional signal feature extraction methods to transform time-domain signals into spectral features. The quality of the spectrum will affect the final classification result. How to dynamically obtain higher quality spectra according to the characteristics utilizing the powerful learning ability of machine learning models of the signal itself is one of our future exploration directions.

CRediT authorship contribution statement

Lu Chen: Conceptualization, Investigation, Methodology, Software, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Xinwei Luo:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Funding acquisition. **Hanlu Zhou:** Investigation, Validation, Writing – review & editing.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, and there is no professional or other personal interest of any nature or kind in any product, service, or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, "**A ship-radiated noise classification method based on domain knowledge embedding and attention mechanism**".

Data availability

The authors do not have permission to share data.

Acknowledgements

This work was supported by the Stable Supporting Fund of National Key Laboratory of Underwater Acoustic Technology No. JCKYS2023604SSJS014, in part by the National Natural Science Foundation of China under Grant 12174053; and in part by the Fundamental Research Funds for Central Universities No.2242023K30003 and 2242023K30004.

References

Antoni, J., Xin, G., Hamzaoui, N., 2017. Fast computation of the spectral correlation. Mech. Syst. Signal Process. 92, 248–277. https://doi.org/10.1016/j.ymssp.2017.01.011.

Barros, R.E., de, B.A., Ebecken, N.F.F., 2022. Development of a ship classification method based on Convolutional neural network and Cyclostationarity Analysis. Mech. Syst. Signal Process. 170, 108778 https://doi.org/10.1016/j.ymssp.2021.108778.

Boustany, R., Antoni, J., 2005. Cyclic spectral analysis from the averaged cyclic periodogram. IFAC Proc. Vol. 38, 166–171. https://doi.org/10.3182/20050703-6-CZ-1902.00028.

Bynagari, N.B., 2020. The difficulty of learning long-term dependencies with gradient flow in recurrent nets. Eng. int. (Dhaka) 8, 127–138. https://doi.org/10.18034/ei.v8i2.570.

Chen, Z., 2020. A deep learning method for bearing fault diagnosis based on Cyclic Spectral Coherence and Convolutional Neural Networks. Mech. Syst. Signal Process. 140, 106683.

Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., Dubnov, S.. HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection. in: ICASSP 2022. https://doi.org/10.1109/ICASSP43922.2022.9746312.

Chen, J., Han, B., Ma, X., Zhang, J., 2021. Underwater target recognition based on multi-decision LOFAR spectrum enhancement: a deep-learning approach. Future Internet 13, 265. https://doi.org/10.3390/fi13100265.

Clark, P., Kirsteins, I., Atlas, L., 2010. Multiband analysis for colored amplitude-modulated ship noise. IEEE International Conference on Acoustics, Speech and Signal Processing. Presented at the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing 3970–3973. https://doi.org/10.1109/ICASSP.2010.5495776, 2010.

Cooley, J.W., Tukey, J.W., 1965. An algorithm for the machine calculation of complex fourier series. Math. Comput. 19, 297–301.

de Moura, N.N., de Seixas, J.M., 2015. Novelty detection in passive SONAR systems using support vector machines. 2015 Latin America Congress on Computational Intelligence (LA-CCI). Presented at the 2015 Latin America Congress on Computational Intelligence (LA-CCI) 1–6. https://doi.org/10.1109/LA-CCI.2015.7435957. IEEE, Curitiba.

Devlin, J., Chang, M.-W., Lee, K., Toutanova 2019., K.. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/arXiv.2010.11929.

Feng, S., Zhu, X., 2022. A transformer-based deep learning network for underwater acoustic target recognition. Geosci. Rem. Sens. Lett. IEEE 19, 1–5. https://doi.org/10.1109/LGRS.2022.3201396.

Fırat, U., Akgül, T., 2018. Compressive sensing for detecting ships with second-order cyclostationary signatures. IEEE J. Ocean. Eng. 43, 1086–1098. https://doi.org/10.1109/JOE.2017.2740698.

Gers, F.A., Schmidhuber, J., Cummins, F., 2000. Learning to forget: continual prediction with LSTM. Neural Comput. 12, 2451–2471. https://doi.org/10.1162/089976600300015015.

Gong, Y., Chung, Y.-A., Glass, J., 2021. AST: audio spectrogram transformer. Interspeech 2021. Presented at the Interspeech 2021 571–575. https://doi.org/10.21437/Interspeech.2021-698. ISCA.

Hanson, D., Antoni, J., Brown, G., Emslie, R., 2008. Cyclostationarity for passive underwater detection of propellor craft: A Development of DEMON Processing. *Proceedings of Acoustics 2008*, 24-26.

Hong, F., Liu, C., Guo, L., Chen, F., Feng, H., 2021. Underwater acoustic target recognition with ResNet18 on ShipsEar dataset. 2021 IEEE 4th International Conference on Electronics Technology (ICET). Presented at the 2021 IEEE 4th International Conference on Electronics Technology (ICET) 1240–1244. https://doi.org/10.1109/ICET51757.2021.9451099.

Hu, G., Wang, K., Liu, L., 2021. Underwater acoustic target recognition based on depthwise separable convolution neural networks. Sensors 21, 1429. https://doi.org/10.3390/s21041429.

Irfan, M., 2021. DeepShip: an underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification. Expert Syst. Appl. 12.

Kamal, S., Satheesh Chandran, C., Supriya, M.H., 2021. Passive sonar automated target classifier for shallow waters using end-to-end learnable deep convolutional LSTMs. Engineering Science and Technology, an International Journal 24, 860–871. https://doi.org/10.1016/j.jestch.2021.01.014.

Kazemi, S., Ghorbani, A., Amindavar, H., Li, C., 2014. Cyclostationary approach to Doppler radar heart and respiration rates monitoring with body motion cancelation

using Radar Doppler System. Biomed. Signal Process Control 13, 79–88. https://doi.org/10.1016/j.bspc.2014.03.012.
Khishe, M., Mohammadi, H., 2019. Passive sonar target classification using multi-layer perceptron trained by salp swarm algorithm. Ocean Eng. 181, 98–108. https://doi.org/10.1016/j.oceaneng.2019.04.013.
Li, Q., 2012. Digital sonar design in underwater acoustics. Advanced Topics in Science and Technology in China. https://doi.org/10.1007/978-3-642-18290-7. Springer, Berlin, Heidelberg.
Liu, F., Shen, T., Luo, Z., Zhao, D., Guo, S., 2021. Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation. Appl. Acoust. 178, 107989 https://doi.org/10.1016/j.apacoust.2021.107989.
Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: hierarchical vision transformer using shifted windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV) 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986. IEEE, Montreal, QC, Canada.
Liu, X., Zhao, S., Su, K., Cen, Y., Qiu, J., Zhang, M., Wu, W., Dong, Y., Tang, J., 2022. Mask and reason: pre-training knowledge graph transformers for complex logical queries. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Presented at the KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 1120–1130. https://doi.org/10.1145/3534678.3539472. ACM, Washington DC USA.
Loshchilov, I., Hutter. Decoupled Weight Decay Regularization. https://doi.org/10.48550/arXiv.1711.05101.
Luo, X., Zhang, M., Liu, T., Huang, M., Xu, X., 2021. An underwater acoustic target recognition method based on spectrograms with different resolutions. JMSE 9, 1246. https://doi.org/10.3390/jmse9111246.
Napolitano, A., 2016. Cyclostationarity: new trends and applications. Signal Process. 120, 385–408. https://doi.org/10.1016/j.sigpro.2015.09.011.
Noumida, A., Rajan, R., 2022. Multi-label bird species classification from audio recordings using attention framework. Appl. Acoust. 197, 108901 https://doi.org/10.1016/j.apacoust.2022.108901.
Qiao, W., Khishe, M., Ravakhah, S., 2021. Underwater targets classification using local wavelet acoustic pattern and Multi-Layer Perceptron neural network optimized by modified Whale Optimization Algorithm. Ocean Eng. 219, 108415 https://doi.org/10.1016/j.oceaneng.2020.108415.
Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21, 67.
Robert, J.Urick, 1983. Principles of underwater sound, 3rd ed. McGraw-Hill Book Company.
Roberts, R.S., Brown, W.A., Loomis, H.H., 1991. Computationally efficient algorithms for cyclic spectral analysis. IEEE Signal Process. Mag. 8, 38–49. https://doi.org/10.1109/79.81008.
Santos-Domínguez, D., Torres-Guijarro, S., Cardenal-López, A., Pena-Gimenez, A., 2016. ShipsEar: an underwater vessel noise database. Appl. Acoust. 113, 64–69. https://doi.org/10.1016/j.apacoust.2016.06.008.
Sherin, B.M., Supriya, M.H., 2015. Selection and parameter optimization of SVM kernel function for underwater target classification. 2015 IEEE Underwater Technology (UT). Presented at the 2015 IEEE Underwater Technology (UT) 1–5. https://doi.org/10.1109/UT.2015.7108260. IEEE, Chennai, India.
Simic, D.C., Simic, J.R., 1999. The strip spectral correlation algorithm for spectral correlation estimation of digitally modulated signals. 4th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services. TELSIKS'99 (Cat. No.99EX365). Presented at the 4th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Services. TELSIKS'99. Papers 277–280. https://doi.org/10.1109/TELSKS.1999.804745. IEEE, Nis, Yugoslavia.
Simonyan, K., Zisserman, A.,2015. In: Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/arXiv.1409.1556.
van der Maaten, Laurens, Hinton, Geoffrey, 2008. t_SNE.pdf. J. Mach. Learn. Res. 9, 2579–2605.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc.
Wang, X., Liu, A., Zhang, Y., Xue, F., 2019. Underwater acoustic target recognition: a combination of multi-dimensional fusion features and modified deep neural network. Rem. Sens. 11, 1888. https://doi.org/10.3390/rs11161888.
Wenz, G.M., 1972. Review of underwater acoustics research: noise. J. Acoust. Soc. Am. 51, 1010–1024. https://doi.org/10.1121/1.1912921.
Xie, Y., Ren, J., Xu, J., 2022. Adaptive ship-radiated noise recognition with learnable fine-grained wavelet transform. Ocean Eng. 265, 112626 https://doi.org/10.1016/j.oceaneng.2022.112626.
Yao, Q., Wang, Y., Yang, Y., 2023. Underwater acoustic target recognition based on data augmentation and residual CNN. Electronics 12, 1206. https://doi.org/10.3390/electronics12051206.
Zhang, W., Lin, B., Yan, Y., Zhou, A., Ye, Y., Zhu, X., 2022. Multi-features fusion for underwater acoustic target recognition based on convolution recurrent neural networks. 2022 8th International Conference on Big Data and Information Analytics (BigDIA). Presented at the 2022 8th International Conference on Big Data and Information Analytics (BigDIA) 342–346. https://doi.org/10.1109/BigDIA56350.2022.9874151.
Zhou, X., Yang, K., 2020. A denoising representation framework for underwater acoustic signal recognition. J. Acoust. Soc. Am. 147, EL377–EL383. https://doi.org/10.1121/10.0001130.
Zhu, Z.K., Feng, Z.H., Kong, F.R., 2005. Cyclostationarity analysis for gearbox condition monitoring: approaches and effectiveness. Mech. Syst. Signal Process. 19, 467–482. https://doi.org/10.1016/j.ymssp.2004.02.007.