Archives of Acoustics Vol. 50, No. 2, pp. 161–171 (2025), https://doi.org/10.24425/aoa.2025.153662

Research Paper

CAPSE-ViT: A Lightweight Framework for Underwater Acoustic Vessel Classification Using Coherent Spectral Estimation and Modified Vision Transformer

NAJAMUDDIN, Usman Ullah SHEIKH, Ahmad Zuri SHA'AMERI

Faculty of Electrical Engineering, Universiti Teknologi Malaysia, UTM Skudai
Johor, Malaysia

*Corresponding Author e-mail: najamuddin@graduate.utm.my

Received February 24, 2025; revised April 10, 2025; accepted April 10, 2025; published online May 29, 2025.

Underwater acoustic target classification has become a key area of research for marine vessel classification, where machine learning (ML) models are leveraged to identify targets automatically. The major challenge is inserting area-specific understanding into ML frameworks to extract features that effectively distinguish between different vessel types. In this study, we propose a model that uses the coherently averaged power spectral estimation (CAPSE) algorithm. Vessel frequency spectra is first computed through the CAPSE analysis, capturing key machinery characteristics. Further, the features are processed via a vision transformer (ViT) network. This method enables the model to learn more complex relationships and patterns within the data, thereby improving the classification performance. This is accomplished by using self-attention mechanisms to capture global dependencies between features, enabling the model to focus on relationships throughout the entire input. The results, evaluated on standard DeepShip and ShipsEar datasets, show that the proposed model achieved a classification accuracy of 97.98 % and 99.19 % while utilizing just 1.90 million parameters, outperforming other models such as ResNet18 and UATR-Transformer in terms of both accuracy and computational efficiency. This work offers an improvement to the development of efficient marine vessel classification systems for underwater acoustics applications, demonstrating that high performance can be achieved with reduced computational complexity.

Keywords: underwater acoustic targets; CAPSE; vision transformer; CNN; LOFAR gram.

1. Introduction

Accurate classification of underwater acoustic targets is crucial in naval defense, underwater surveillance, and environmental monitoring (Bjørnø, 2017; Domingos *et al.*, 2022; Thomas *et al.*, 2020). The ability to distinguish between different types of marine vessels based on their acoustic signatures is essential for operations such as threat detection and marine traffic management (McKenna *et al.*, 2024). However, the underwater acoustic environment poses unique challenges due to complex propagation effects, ambient noise, and interference from various sources, making this task particularly difficult (Aslam *et al.*, 2024). Traditional classification methods, while effective in controlled conditions, become less efficient with high levels of noise and randomness present in real-world underwater scenarios. This creates a need for improved methods that can enhance the quality of the target signature and take advantage of deep learning to achieve more accurate and reliable classification (Luo *et al.*, 2023).

Conventional underwater acoustic target classification has relied on signal processing techniques such as Fourier transforms, wavelet analysis, and mel-frequency cepstral coefficients (MFCC), which are effective for identifying specific features in clean signals (Müller *et al.*, 2024). However, these methods face difficulties when dealing with highly noisy or distorted signals. With the rise of deep learning, convolutional

neural networks (CNNs) have been employed to classify acoustic signals by first transforming them into spectrograms and then treating the problem as an image classification task (Zeng *et al.*, 2019). CNNs leverage spatial hierarchies to capture local features from these spectrograms, but their reliance on local convolutions limits their ability to capture global dependencies in the data (Yang *et al.*, 2024). This shortcoming is particularly problematic for underwater acoustic signals, where the temporal and spectral relationships within the signal are essential for accurate target classification. Local features in a spectrogram refer to specific, small-scale patterns over short time or frequency ranges, such as individual machinery noises (Feng, Zhu, 2022). Global features, in contrast, represent broader patterns across time and frequency, capturing the overall acoustic signature of the source or vessel. The vision transformer (ViT), a recently developed deep learning model, offers an alternative approach by employing a self-attention mechanism that captures both local and global dependencies, making it more suited for tasks that require holistic data analysis (Dosovitskiy *et al.*, 2020).

Another major challenge in underwater acoustics is the low signal-to-noise ratio (SNR), making target detection and classification challenging. Earlier research indicates that targets become undetectable when SNR falls below critical levels of −14.4 dB, and with nearly 90 % of vessels receiving SNR below 0 dB in ambient noise conditions (Siddagangaiah *et al.*, 2016). Another significant challenge in underwater acoustics is the low SNR of acoustic signatures resulting from environmental noise, surface reflections, and interference (Lampert, O'Keefe, 2013). Preprocessing techniques, such as the coherently averaged power spectral estimation (CAPSE), are designed to enhance the quality of acoustic signals by averaging power spectra across multiple observations, thus reducing noise and improving the clarity of key signal features (Lan *et al.*, 2020).

The integration of CAPSE and ViT forms the core of this study. We use CAPSE as a preprocessing step to improve tonal signals and minimize noise, highlighting target-specific features through coherent spectral averaging. The processed signals are then converted into low frequency analysis and recording (LOFAR) grams, which are fed into the ViT model for classification. The ViT ability to capture both local and global machinery features dependencies using its attention mechanism is exploited. The proposed method was assessed using DeepShip (Irfan *et al.*, 2021) and ShipsEar (Santos-Domínguez *et al.*, 2016), a publicly available dataset, where it outperformed other methods reported in the literature, delivering higher accuracy and enhanced generalization. This approach highlights the potential for incorporating CAPSE and the modified ViT deep learning method for improving classification performance in noisy underwater.

The rest of the article is arranged as follows: Sec. 2 is an overview of existing studies in literature; Sec. 3 highlights the proposed methodology, dataset preprocessing techniques, and the model parameters employed in the experiments; Sec. 4 showcases the results, emphasizing the advantages of the proposed methodology; finally, Sec. 5 provides a conclusion, summarizing the main contributions and proposed future work.

2. Related works

Research on marine vessel classification using acoustic noise has explored a range of signal processing techniques and machine learning models (Bianco *et al.*, 2019). Early methods relied heavily on manual interpretation of acoustic signatures by sonar operators, depending entirely on their expertise (Domingos *et al.*, 2022). However, with advancements in computational power and deep learning techniques, automated classification has become an area of increasing interest.

One of the most established techniques is the fast Fourier transform (FFT), which converts time-domain signals into the frequency domain, enabling the identification of spectral components, for understanding the underlying patterns (Feng *et al.*, 2021). However, it is ineffective for representing underwater acoustic signals due to their non-stationary nature. The wavelet transform offers both time and frequency information, providing variable resolution that makes it effective for analyzing non-stationary signals with varying patterns, but it is sensitive to ambient noise (Kim *et al.*, 2021). MFCCs are frequently applied in sound analysis due to their ability to capture the perceptual characteristics of audio signals (Lim *et al.*, 2007). They are also designed to mimic the human ear's sensitivity to different frequencies, making them particularly useful in tasks such as speech and sound classification (Sharma *et al.*, 2020). The constant-Q transform offers constant resolution across octaves, making it well-suited for logarithmic frequency analysis (Singh *et al.*, 2021). In addition to these general techniques, LOFAR is a method focused on detecting long-term spectral patterns, which is particularly useful for identifying sustained sounds such as engine noise or other mechanical signals (Li *et al.*, 2023). On the other hand, the detection of envelope modulation on the noise (DEMON) technique is specifically designed to detect modulation spectrum caused by rotating components such as propellers and blades (Park, Jung, 2021).

Numerous multi-modal recognition techniques have been investigated for marine vessel classification. In (Yuan *et al.*, 2019), a method was developed that combines both optical images and radiated noise from vessels as input data, allowing for a more comprehensive classification approach. Luo *et al.* (2021) applied a multi-window spectral analysis method to capture a range of in-band frequency features, provid-

ing a more detailed and accurate representation of the acoustic environment. Additionally, Song *et al.* (2021) significantly improved underwater noise classification by extracting the one-third octave noise spectrum, power spectral density, and MFCC features. These various approaches aim to increase classification accuracy by integrating and leveraging multiple feature sets, enhancing the robustness of the recognition process.

Machine learning techniques, such as support vector machines (SVM) and shallow neural networks (SNN), have long been used in underwater acoustic classification. These techniques rely on efficient feature extraction methods to transform raw acoustic signals into feature vectors, which are subsequently input into the network (De Moura, De Seixas, 2016). For example, Sherin and Supriya (2015), used enhanced SVM classifiers to differentiate types of vessel noise. With advances in deep learning, research has increasingly focused on more complex neural networks. Khishe and Mohammadi (2019) applied MFCC as inputs to a neural network optimized by the salp swarm algorithm and achieved an accuracy of 97.1 % (Hegazy *et al.*, 2020). However, these fully connected networks still face challenges in capturing deep, complex features in multiple-class scenarios due to their relatively simple architecture.

To overcome these limitations, CNNs have been used to map raw waveforms or time-frequency representations directly to vessel types (Hu *et al.*, 2021; Luo *et al.*, 2021). CNNs have demonstrated good performance in classifying vessels using acoustic signals. For instance, Cao *et al.* (2019), introduced the CNN combining second-order pooling (SOP) and the constant-Q transform for feature extraction, outperforming traditional classifiers such as VGG-Net and deep belief networks by achieving an accuracy of 96.3 %. Custom CNN architectures, such as VesselNet, have also been proposed to enhance the classification of LOFAR spectrograms. Cinelli *et al.* (2018) designed VesselNet specifically for spectrogram classification, using the two-pass split-window filter with resulting in a precision of 88.1 % on proprietary dataset (de Carvalho *et al.*, 2021).

The transformer architecture has been extensively applied in fields such as natural language processing (NLP) (Raffel *et al.*, 2020), computer vision (CV) (Dosovitskiy, 2020), and audio classification (Noumida, Rajan, 2022), consistently demonstrating superior performance. Recently, Chen *et al.* (2024) introduced Swin transformer for ship-radiated noise classification, combining DEMON spectra and mel-spectrograms through feature fusion and attention mechanisms. The achieved performance on standard dataset was 98.62 % and 99.01 %. However, its performance with weak acoustic signals due to masking by both self-generated broadband noise, an increase in distance from the receiver, and ambient noise from natural sources is unknown. This masking effect degrades the clarity and detectability of the vessel's tonal components (Ikpekha *et al.*, 2018). Similarly, the large size of the deep learning network makes it unsuitable for real-time applications. This paper introduces the ViT, with self-attention, as the classifier. This lightweight transformer architecture significantly reduces training time and resource requirements (Chen *et al.*, 2024).

3. Methodology

This section details the methodology for underwater acoustic target classification, using CAPSE for signal enhancement and ViT for classification. CAPSE improves spectral clarity by reducing noise, while ViT leverages self-attention mechanisms to capture patterns in the enhanced spectrograms and improve classification accuracy.

3.1. CAPSE

CAPSE is a signal processing technique designed to enhance the detection of sinusoids in noisy environments. Unlike traditional methods such as the periodogram and Welch's method, CAPSE preserves phase coherence across multiple signal segments, resulting in a substantial improvement in SNR (Feng *et al.*, 2021).

For a sinusoidal signal S_0 embedded in noise, the Fourier transform for each segment, S_k can be expressed as

$$S_k(\omega) = S_0(\omega)e^{(j\phi_k)}, \tag{1}$$

where $\phi_k = \omega_0 kD$ represents the phase difference between the Fourier transforms of the k-th and the 1st segments at frequency ω_0. CAPSE aims to coherently average the signal across multiple segments K, thus enhancing SNR:

$$\overline{X}(\omega) = (1K)\sum_{k=0}^{K-1} X_k(\omega)e^{(-j\phi_k)}. \tag{2}$$

This offset introduces a phase variation across segments, which can be corrected by applying an additional DFT along the segment indices given in Eq. (3), yielding:

$$\widehat{X}(\omega_l, \dot{\omega}_m) = (1/K)\sum_{k=0}^{K-1} X_k(\omega_l)e^{(-j\omega_v k)}, \tag{3}$$

$$\omega_l = \arg\max_{\omega_m} |X(\omega_l, \omega_v)|^2. \tag{4}$$

CAPSE spectrum is then defined in Eq. (5), where ω_l and ω_m are the indices of angular frequencies, measured in radians per second, and K is the number of segments (Lan *et al.*, 2020):

$$P_{x_{xx}}^{\mathrm{CAPSE}}(\omega) = (1/\mathrm{UM})\left|\widehat{X}(\omega_l, \omega_{\delta l})\right|^2. \tag{5}$$

By maximizing the energy component, CAPSE preserves the most significant spectral information, making it a robust method for tonal detection in noisy environments. Details of the algorithm can be found in (Feng *et al.*, 2021).

3.2. Vision transformer

ViT is employed as the classification model, leveraging its self-attention mechanism to capture both local and global dependencies on data features (Dosovitskiy *et al.*, 2020). The acoustic signals preprocessed with CAPSE are transformed into LOFAR grams and treated as 2D images, displaying frequency components along the horizontal axis and temporal progression along the vertical axis. Let LOFAR gram be denoted as $x \in \mathbb{R}^{W\times H}$, where W is the number of frequency bins and H is the number of time steps. The LOFAR gram is divided into non-overlapping patches, each of size $Q \times P$, where Q and P represent the patch dimensions in frequency and time domains, respectively. These patches are then flattened into 1D vectors, creating a sequence of patch embeddings, $X_p = [x_1, x_2, ..., x_N]$, where $N = 400$ is the total number of patches. Each patch acts as an independent token for the transformer input. Positional encodings $E \in R^{N\times D}$ are added to the patch embeddings to retain the relative positional information, creating the input sequence, z_0 for the transformer layers,

$$z_0 = [x_1 E, x_2, ..., x_N E] + E. \tag{6}$$

The core of the ViT model is its multi-head self-attention mechanism, which allows the model to compute attention weights between different patches in the sequence. For each attention head, the input sequence z_0 is transformed into a query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) matrices:

$$\mathbf{Q} = z_0\mathbf{W_Q}, \qquad \mathbf{K} = z_0\mathbf{W_K}, \qquad \mathbf{V} = z_0\mathbf{W_V}, \tag{7}$$

where $\mathbf{W_Q}, \mathbf{W_K}, \mathbf{W_V} \in \mathbb{R}^{\mathrm{D}\times\mathrm{D}}$ are the learnable weight matrices. The attention score for each patch is computed as (Pang *et al.*, 2023)

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{D}}\right)\mathbf{V}. \tag{8}$$

Multiple attention heads are applied in parallel, enabling the model to focus on different regions of the LOFAR gram simultaneously. The outputs from the attention heads are concatenated and then passed through a feed-forward network for additional processing.

Following the attention module, a series of transformer encoder layers are applied, each containing a multi-head self-attention block and a position-wise feed-forward network. These layers help in progressively learning higher-level representations from the patch sequence. Each encoder layer includes residual connections and layer normalization to stabilize training:

$$z_l' = \mathrm{LayerNorm}(z_{l-1} + \mathrm{MultiHeadAttention}(z_{l-1})), \tag{9}$$

$$z_l = \mathrm{LayerNorm}(z_l' + \mathrm{FeedForward}(z_l')), \tag{10}$$

where l denotes the current transformer layer.

After passing through several transformer layers, the final sequence representation z_L is obtained from the last encoder block. A class token is appended to the patch sequence during input, and this token's representation at the final layer z_L^{class} is extracted and passed to a classification head. The classification head consists of a fully connected layer followed by a softmax activation function, which produces the class probabilities for the acoustic target:

$$y = \mathrm{softmax}\left(\mathbf{W}_{\mathrm{class}} z_L^{\mathrm{class}}\right), \tag{11}$$

where $\mathbf{W}_{\mathrm{class}}$ is the weight matrix of the classification layer.

Figure 1 presents the process flow, where the classifier takes $50 \times 4000 \times 1$ grayscale LOFAR grams as

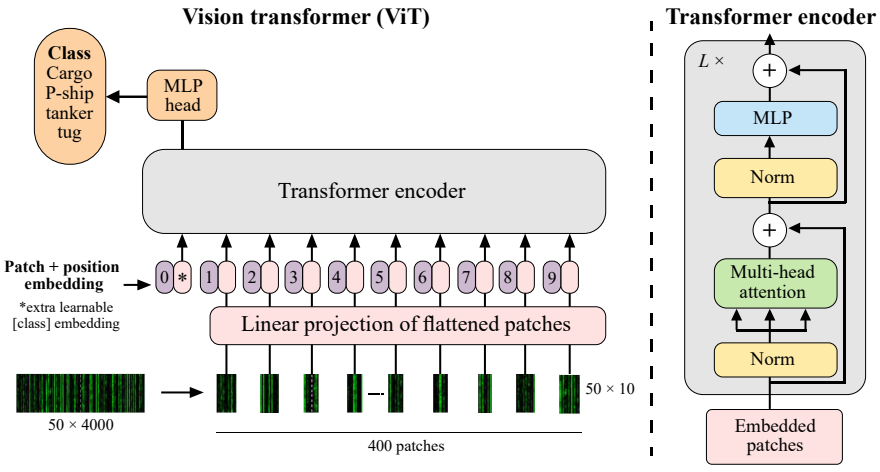

Fig. 1. Modified ViT network architecture with custom size LOFAR gram images (Dosovitskiy *et al.*, 2020).

input and is divided into 400 patches of size 50×10 to ensure that each patch spans the full temporal resolution while maintaining fine spectral resolution, allowing the model to preserve tonal shifts caused by Doppler effects or environmental variability within a single patch. The ViT model was trained using stochastic gradient descent with momentum as the optimization method, with an initial learning rate set to 0.001. Training was conducted over a maximum of 10 epochs with a mini-batch size of 64, utilizing a GPU to accelerate the process. A modified ViT network model with description and learnable parameters for each layer is presented in Table 1. The training was conducted on a system featuring an AMD Ryzen 5 3600 processor (6-core), 32 GB RAM, 500 GB SSD storage, and an NVIDIA GTX 1660 SUPER GPU with 6 GB of memory.

Table 1. Number of parameters of each layer of the CNN architecture.

Modules	Layers	Number of parameters
Input processing	Image input	–
	Patch embedding	160 400
	Embedding concatenation	400
	Position embedding	200 400
	Layer norm	800
Feature extractor	Self-attention	641 600
	Layer norm	800
	Encoder block 1	480 400
	Encoder block 2	480 400
	Layer norm	800
Classifier	Head	1604
Total parameters		1 967 604

3.3. Data preprocessing

The generation of LOFAR grams involves a systematic analysis of publicly available DeepShip and ShipsEar datasets. The process starts with loading audio files and configuring the key parameters of the CAPSE algorithm. This includes setting a window size of 16 000 samples with a 50 % overlap and a sampling rate of 8 kHz.

For the LOFAR gram, the algorithm processes the audio signal in segments, applying a Hanning window to reduce spectral leakage. A real FFT is performed on each window, normalizing the power in each frequency bin.

The first half of the bins is preserved, followed by applying an FFT to each column of the spectrum. After squaring the magnitudes, the maximum value in each column is stored. The resulting spectrum is saved as a row vector in a (50×4000) matrix in logarithmic scale, and the matrix is saved as PNG images. Figure 2 shows samples of zoomed LOFAR gram (0 Hz–1600 Hz) images generated for different classes in DeepShip dataset (Irfan *et al.*, 2021). Here, the spectral components due to machinery are visible against normalized broadband noise.

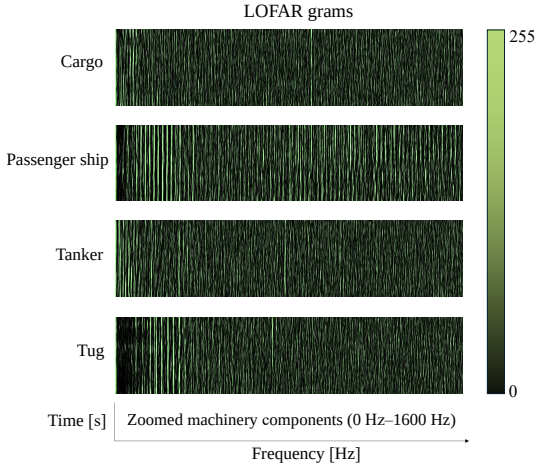

Fig. 2. Sample of LOFAR grams of DeepShip dataset developed using CAPSE algorithm.

Following preprocessing, the dataset was randomly divided into three distinct subsets: 70 % of the data was allocated for training, where the model learns patterns and features within the data; 15 % was kept for validation, which is used to adjust the weights of the neural networks of the model. Early stopping was used to prevent overfitting by evaluating its performance on unseen data during training; the remaining 15 % of the dataset was used for testing the trained model to evaluate its performance metrics.

4. Results and discussion

This section presents and analyzes the results of our proposed method for underwater acoustic target classification. The performance of the model is evaluated on a benchmark dataset, with a focus on classification accuracy, and the advantages of CAPSE enhanced spectral representations and ViT. Comparative results with existing methods are also discussed. Figures 3 and 4 show the accuracy and loss curves for the

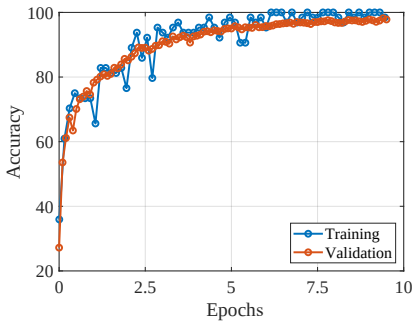

Fig. 3. Accuracy curves for training and validation on DeepShip dataset.

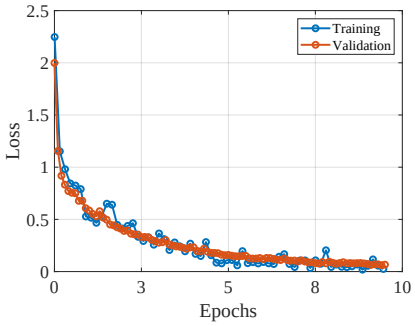

Fig. 4. Loss curves for training and validation on DeepShip dataset.

training and validation process on DeepShip dataset. The network shows rapid accuracy improvement and convergence with minimal overfitting. The loss steadily decreases, indicating stable and effective training.

4.1. Classification performance

Tables 2 and 3 provide an overview of DeepShip and ShipsEar datasets used for evaluating the classification performance, showing the number of samples for each vessel class. Despite the variation in sample sizes, the model demonstrated effective generalization across all classes, maintaining high performance even for classes with fewer samples, such as the tug class in DeepShip or class B form ShipsEar, shown in the confusion matrix in Figs. 5a and 5b.

Table 2. Class description of DeepShip dataset (Irfan *et al.*, 2021).

Class label	Number of samples
Cargo	4242
Passenger ship	4641
Tanker	4454
Tug	4054

Table 3. Class description of ShipsEar dataset (Santos-Domínguez *et al.*, 2016).

Class label	Vessel type	Number of samples
Class A	Mussel boats, dredgers, fishing boats, trawlers, and tugboats	389
Class B	Sailboats, motorboats, and pilot boats	313
Class C	Passenger ferries	842
Class D	Ro-ro vessels and ocean liners	492
Class E	Background noise recordings	229

In terms of classification performance on DeepShip dataset in Fig. 5a, as presented in Table 4, the model achieved excellent results across all vessel types. The cargo class achieved the highest accuracy at 98.90 %,

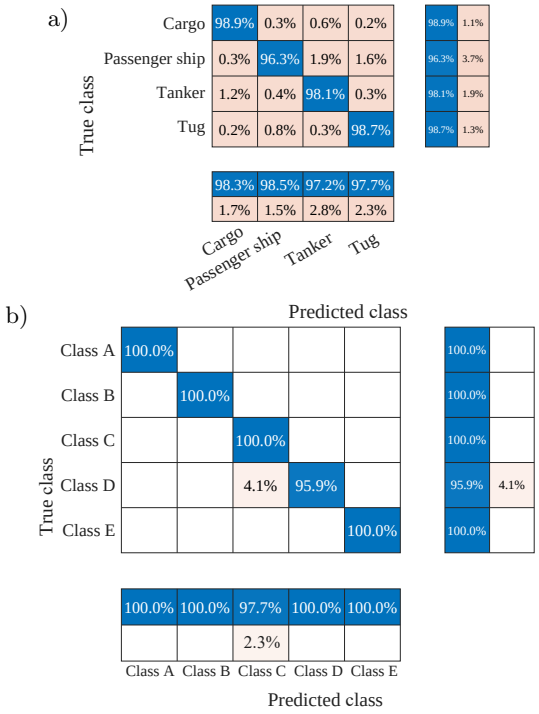

Fig. 5. Confusion matrix for modified ViT network: a) DeepShip dataset; b) ShipsEar dataset.

Table 4. Classification performance on DeepShip dataset.

Label	Accuracy [%]	Precision [%]	Recall [%]	$F1$-score [%]
Cargo	98.90	98.36	98.90	98.63
Passenger ship	96.26	98.38	96.26	97.31
Tanker	98.05	97.20	98.05	97.62
Tug	98.68	97.98	98.68	98.33
Average	97.98	97.98	97.98	97.97
Std Dev	1.1971	0.5516	1.1971	0.6119

with the tug class following at 98.68 %. The passenger ship and tanker classes show slightly lower accuracies of 96.26 % and 98.05 %, respectively. These small differences indicate that the model is consistent in identifying all vessel types, regardless of their sample size.

The precision, recall, and $F1$-score metrics further demonstrate the model robustness. The $F1$-scores, which balance precision and recall, are consistently high for all classes, ranging from 97.31 % for the passenger ship class to 98.63 % for the cargo class, highlighting the model ability to maintain high classification performance across diverse acoustic characteristics. The model average $F1$-score of 97.97 % across all classes reflects its ability to generalize well to unseen data, making it a reliable tool for underwater acoustic target classification tasks. Although there are some minor performance variations, particularly for the passenger ship class, the overall results confirm the model's effective classification capability, demonstrating robust generalization across all vessel types.

Similarly, the confusion matrix for the ShipsEar dataset in Fig. 5b, shows high classification performance across all vessel classes as shown in Table 5, with an accuracy of (100 %) achieved for classes A, B, and E, indicating that the model can reliably distinguish mussel boats, sailboats, and background noise recordings. Class C (passenger ferries) and class D (Ro-ro vessels and ocean liners) exhibit minor confusion, with 2.3 % of class C misclassified as class D and 4.1 % of class D misclassified as class C. This overlap suggests that these vessel types share similar acoustic characteristics, likely due to comparable propulsion systems or operational behaviours. However, the model maintains over 95 % accuracy for all classes, demonstrating strong generalization.

Table 5. Classification performance on ShipsEar dataset.

Label	Accuracy [%]	Precision [%]	Recall [%]	$F1$-score [%]
Class A	100	100	100	100
Class B	100	100	100	100
Class C	100	96.10	100	98.01
Class D	95.95	100	95.95	97.93
Class E	100	100	100	100
Average	99.19	99.22	99.19	99.19
Std Dev	1.8130	1.7424	1.8130	1.1111

4.2. Features visualization

The t-SNE method is employed to visually analyze the model feature extraction process. High-dimensional features of vessel radiated noise data are projected into a two-dimensional space to observe how well the model separates different vessel classes. The first visualization in Fig. 6 shows the t-SNE plot for the input layer before the ViT model initialization, where the samples are scattered with no clear patterns or groupings. After the model processes the data, the second visualization shown in Fig. 7 presents a clear separation of classes, with most samples correctly clustered into distinct groups according to their labels. Very few instances remain misclassified, potentially

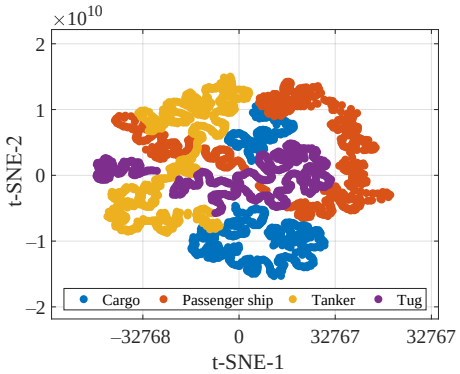

Fig. 6. t-SNE high dimensional features visualization of untrained network.

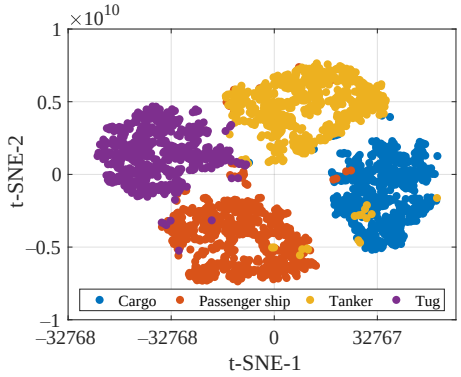

Fig. 7. t-SNE high dimensional visualization after the network is fully trained.

due to a weak or ambiguous target signature. By highlighting these outliers, the model limitations and areas for improvement become apparent. Overall, this visualization confirms that the model effectively learns discriminative features, resulting in well-formed class clusters in the feature space.

4.3. Performance under varying SNR

In real-world maritime scenarios, the underwater acoustic environment is subject to varying levels of environmental noise originating from natural sources such as wind, wave activity, and marine life. Such noise can complicate the accurate classification and interpretation of acoustic signals, highlighting the need to evaluate the performance of classification models under adverse conditions. To systematically examine the robustness and generalization abilities of the classifiers, this study simulated diverse noise environments by injecting Gaussian white noise at multiple SNR levels into the original acoustic signals. Using DeepShip dataset, the power of the signal was computed for each case, and zero-mean noise with a specified power level was generated and added to achieve a targeted SNR value. The objective was to mimic real-world situations where the clarity of received signals is degraded by external noise sources.

The performance of the model was assessed, as depicted in Fig. 8. As the SNR decreases from 20 dB to 0 dB, the overall accuracy declines, demonstrating the negative impact of increasing noise on the model's performance. Notably, the average classification accuracy remains above 50 % at 5 dB, indicating a moderate level of robustness in noisy conditions. Among the vessel classes, the cargo class consistently achieves higher accuracy, which can be attributed to its stronger and more distinguishable acoustic signature. In contrast, the model exhibits reduced performance for certain classes such as tankers and passenger ships, whose acoustic characteristics are more susceptible to noise interference. To address this limitation, future work will focus on exploring alternative classifier configura-

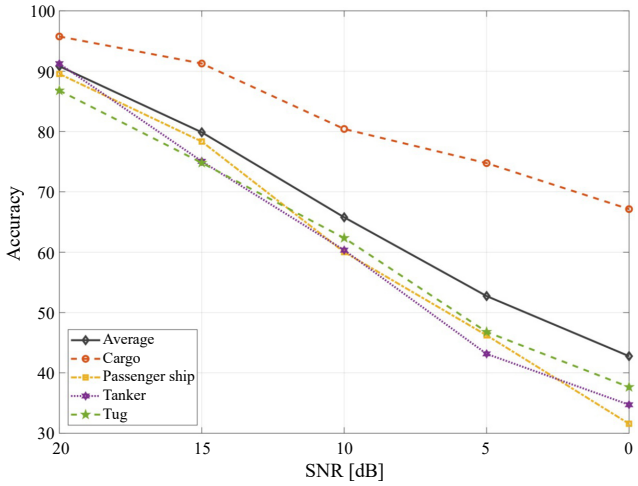

Fig. 8. Performance of classifier under varying SNR conditions.

tions to further enhance robustness and improve class-specific accuracy under challenging acoustic environments.

4.4. Comparison with earlier research

Most of the prior vessel classification research using deep learning techniques commonly employs CNN models, with a majority of these studies relying on non-standard datasets to evaluate the classification performance of their proposed methods. This reliance on diverse datasets makes it challenging to consistently compare the performance of different models.

Table 6 compares the classification accuracy of our model with earlier deep learning-based studies that utilize both the DeepShip and ShipsEar datasets. The comparative analysis demonstrates that our proposed model achieves an accuracy of 97.98 % on the DeepShip dataset, which is competitive with the state-of-the-art. It surpasses models such as UATR-transformer, which achieved 95.30 %, and significantly outperforms DRA-CNN, which lagged at 89.20 %. On the ShipsEar dataset, our model achieves an impressive accuracy of 99.19 %, further solidifying its competitiveness. Although the HAUT Fusion model slightly outperforms our method with an accuracy of 99.01 % on the DeepShip dataset and 98.62 % on the ShipsEar dataset, it does so at a considerable computational cost. The HAUT Fusion model utilizes 30.33 million parameters, compared to the 1.90 million parameters used by our classifier model. This highlights that our model had a better balance between accuracy and computational efficiency, making it a favorable choice for practical applications.

Moreover, the ResNet18 model, despite achieving commendable accuracies of 96.37 % on the DeepShip dataset and 94.30 % on the ShipsEar dataset, operates with a significantly larger parameter count of 11.70 million. In contrast, our model maintains high performance while using only 1.90 million parameters, underscoring its efficiency in terms of model complexity and memory requirements. Similarly, DRA-CNN, while achieving better accuracies of 97.10 % on the ShipsEar dataset, uses only 0.26 million parameters, but its performance on the DeepShip dataset is considerably lower (89.20 %).

These findings suggest that the proposed model provides an effective solution with lower computational demand, making it suitable for deployment in environments with limited resources without sacrificing accuracy. The consistent high performance across both datasets (DeepShip and ShipsEar) further validates the generalizability of our approach.

5. Conclusions and future work

A new framework leveraging deep learning, based on CAPSE as preprocessing and ViT as classifier is designed to enhance the performance of classification of marine vessel based on their radiated noise. The model demonstrates a robust performance, achieving an accuracy of 97.98 % while maintaining a significantly lower parameter of 1.9 million compared to other state-of-the-art models. The results highlight the model's efficiency in extracting discriminative features with minimal computational complexity, making it suitable for real-time or resource-constrained environments. Despite marginally lower accuracy compared to HAUT Fusion, our model's efficiency in terms of parameter usage offers a compelling advantage. These findings emphasize the effectiveness of the proposed approach in balancing accuracy and computational cost for passive underwater acoustic target classification tasks.

Future work may explore further optimization of the feature extraction process and the potential integration of additional domain-specific knowledge

Table 6. Classification performance on ShipsEar dataset.

Models	Accuracy [%] DeepShip	Accuracy [%] ShipsEar	Parameters (million)
ResNet18 (Hong *et al.*, 2021)	96.37	94.30	11.70
DRA-CNN (Chen *et al.*, 2021)	89.20	97.10	0.26
UATR-transformer (Feng, Zhu, 2022)	95.30	96.90	2.60
HAUT Fusion (Chen *et al.*, 2024)	99.01	98.62	30.33
Proposed method	**97.98**	**99.19**	**1.90**

to enhance performance. Furthermore, due to the limited availability of publicly available datasets and the inadequate class of vessel types that are recorded, it is difficult to fully assess the robustness of the model under various environmental scenarios. To address this issue, we plan to further investigate vessel radiated noise and synthetically generate signals for different scenarios using mathematical modelling.

Fundings

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

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

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' contribution

Study conception and design: Najamuddin, Usman Ullah Sheikh, and Ahmad Zuri Sha'ameri. Data collection, analysis and interpretation of results, and the draft manuscript preparation: Najamuddin. All authors reviewed the results and approved the final version of the manuscript.

Acknowledgments

The authors express their gratitude to Universiti Teknologi Malaysia (UTM) for their invaluable support and resources that have facilitated this research.

References

1. Aslam M.A. *et al.* (2024), Underwater sound classification using learning based methods: A review, *Expert Systems with Applications*, **255**(Part 1): 124498, https://doi.org/10.1016/j.eswa.2024.124498.

2. Bianco M.J. *et al.* (2019), Machine learning in acoustics: Theory and applications, *The Journal of the Acoustical Society of America*, **146**(5): 3590–3628. https://doi.org/10.1121/1.5133944.

3. Bjørnø L. (2017), Underwater acoustic measurements and their applications, [in:] *Applied Underwater Acoustics*, Neighbors T.H., III, Bradley D. [Eds.], pp. 889–947, Elsevier, https://doi.org/10.1016/B978-0-12-811240-3.00014-X.

4. Cao X., Togneri R., Zhang X., Yu Y. (2019), Convolutional neural network with second-order pooling for underwater target classification, *IEEE Sensors Journal*, **19**(8): 3058–3066, https://doi.org/10.1109/JSEN.2018.2886368.

5. 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**(10): 265, https://doi.org/10.3390/fi13100265.

6. Chen L., Luo X., Zhou H. (2024), A ship-radiated noise classification method based on domain knowledge embedding and attention mechanism, *Engineering Applications of Artificial Intelligence*, **127**(Part B): 107320, https://doi.org/10.1016/j.engappai.2023.107320.

7. Cinelli L.P., Chaves G.S., Lima M.V.S. (2018), Vessel classification through convolutional neural networks using passive sonar spectrogram images, [in:] *Proceedings of the Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT 2018)*, pp. 21–25, http://doi.org/10.14209/sbrt.2018.340.

8. de Carvalho H.T., Avila F.R., Biscainho L.W.P. (2021), Bayesian restoration of audio degraded by low-frequency pulses modeled via Gaussian process, *IEEE Journal of Selected Topics in Signal Processing*, **15**(1): 90–103, https://doi.org/10.1109/JSTSP.2020.3033410.

9. de Moura N.N., de Seixas J.M. (2016), Novelty detection in passive SONAR systems using support vector machines, *2015 Latin-America Congress on Computational Intelligence (LA-CCI)*, https://doi.org/10.1109/LA-CCI.2015.7435957.

10. Domingos L.C.F., Santos P.E., Skelton P.S.M., Brinkworth R.S.A., Sammut K. (2022), A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance, *Sensors*, **22**(6): 2181, https://doi.org/10.3390/s22062181.

11. Dosovitskiy A. *et al.* (2020). An image is worth 16x16 words: Transformers for image recognition at scale, arXiv, https://doi.org/10.48550/arXiv.2010.11929.

12. Feng S., Jiang K., Kong X. (2021), A line spectrum detector based on improved coherent power spectrum estimation, *Journal of Physics: Conference Series*, **1971**(1): 012006, https://doi.org/10.1088/1742-6596/1971/1/012006.

13. Feng S., Zhu X. (2022), A transformer-based deep learning network for underwater acoustic target recognition, *IEEE Geoscience and Remote Sensing Letters*, **19**: 1–5, https://doi.org/10.1109/LGRS.2022.3201396.

14. Hegazy A.E., Makhlouf M.A., El-Tawel G.S. (2020), Improved salp swarm algorithm for feature selection, *Journal of King Saud University – Computer and Information Sciences*, **32**(3): 335–344, https://doi.org/10.1016/j.jksuci.2018.06.003.

15. 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)*, pp. 1240–1244, https://doi.org/10.1109/ICET51757.2021.9451099.

16. Hu G., Wang K., Liu L. (2021), Underwater acoustic target recognition based on depthwise separable convolution neural networks, *Sensors*, **21**(4): 1429, https://doi.org/10.3390/s21041429.

17. Ikpekha O.W., Eltayeb A., Pandya A., Daniels S. (2018), Operational noise associated with underwater sound emitting vessels and potential effect of oceanographic conditions: A Dublin Bay port area study, *Journal of Marine Science and Technology*, **23**: 228–235, https://doi.org/10.1007/s00773-017-0468-4.

18. Irfan M., Jiangbin Z., Ali S., Iqbal M., Masood Z., Hamid U. (2021), DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification, *Expert Systems with Applications*, **183**: 115270, https://doi.org/10.1016/j.eswa.2021.115270.

19. Khishe M., Mohammadi H. (2019), Passive sonar target classification using multi-layer perceptron trained by salp swarm algorithm, *Ocean Engineering*, **181**: 98–108, https://doi.org/10.1016/j.oceaneng.2019.04.013.

20. Kim K.-I., Pak M.-I., Chon B.-P., Ri C.-H. (2021), A method for underwater acoustic signal classification using convolutional neural network combined with discrete wavelet transform, *International Journal of Wavelets, Multiresolution and Information Processing*, **19**(04): 2050092, https://doi.org/10.1142/S0219691320500927.

21. Lampert T.A., O'Keefe S.E.M. (2013), On the detection of tracks in spectrogram images, *Pattern Recognition*, **46**(5): 1396–1408, https://doi.org/10.1016/j.patcog.2012.11.009.

22. Lan H., White P.R., Li N., Li J., Sun D. (2020), Coherently averaged power spectral estimate for signal detection, *Signal Processing*, **169**: 107414, https://doi.org/10.1016/j.sigpro.2019.107414.

23. Li X., Wang D., Tian Y., Kong X. (2023), A method for extracting interference striations in lofargram based on decomposition and clustering, *IET Image Processing*, **17**(6): 1951–1958, https://doi.org/10.1049/ipr2.12768.

24. Lim T., Bae K., Hwang C., Lee H. (2007), Classification of underwater transient signals using MFCC feature vector, *2007 9th International Symposium on Signal Processing and Its Applications, ISSPA 2007, Proceedings*, pp. 1–4, https://doi.org/10.1109/ISSPA.2007.4555521.

25. Luo X., Chen L., Zhou H., Cao H. (2023), A survey of underwater acoustic target recognition methods based on machine learning, *Journal of Marine Science and Engineering*, **11**(2): 384, https://doi.org/10.3390/jmse11020384.

26. Luo X., Zhang M., Liu T., Huang M., Xu X. (2021), An underwater acoustic target recognition method based on spectrograms with different resolutions, *Journal of Marine Science and Engineering*, **9**(11): 1246, https://doi.org/10.3390/jmse9111246.

27. McKenna M.F. *et al.* (2024), Understanding vessel noise across a network of marine protected areas, *Environmental Monitoring and Assessment*, **196**(4): 369, https://doi.org/10.1007/s10661-024-12497-2.

28. Müller N., Reermann J., Meisen T. (2024), Navigating the depths: A comprehensive survey of deep learning for passive underwater, *IEEE Access*, **12**: 154092–154118, https://doi.org/10.1109/ACCESS.2024.3480788.

29. Noumida A., Rajan R. (2022), Multi-label bird species classification from audio recordings using attention framework, *Applied Acoustics*, **197**: 108901, https://doi.org/10.1016/j.apacoust.2022.108901.

30. Pang D., Wang H., Ma J., Liang D. (2023), DCTN: A dense parallel network combining CNN and transformer for identifying plant disease in field, *Soft Computing*, **27**(21): 15549–15561, https://doi.org/10.1007/s00500-023-09071-2.

31. Park J., Jung D.-J. (2021), Deep convolutional neural network architectures for tonal frequency identification in a lofargram, *International Journal of Control, Automation and Systems*, **19**(2): 1103–1112, https://doi.org/10.1007/s12555-019-1014-4.

32. Raffel C. *et al.* (2020), Exploring the limits of transfer learning with a unified text-to-text transformer, *Journal of Machine Learning Research*, **21**(140): 1–67.

33. Santos-Domínguez D., Torres-Guijarro S., Cardenal-López A., Pena-Gimenez A. (2016), ShipsEar: An underwater vessel noise database, *Applied Acoustics,* **113**: 64–69, https://doi.org/10.1016/j.apacoust.2016.06.008.

34. Sharma G., Umapathy K., Krishnan S. (2020), Trends in audio signal feature extraction methods, *Applied Acoustics*, **158**: 107020, https://doi.org/10.1016/j.apacoust.2019.107020.

35. Sherin B.M., Supriya M.H. (2015), Selection and parameter optimization of SVM kernel function for underwater target classification, [in:] *2015 IEEE Underwater Technology (UT)*, pp. 1–5, https://doi.org/10.1109/UT.2015.7108260.

36. Siddagangaiah S., Li Y., Guo X., Chen X., Zhang Q., Yang K., Yang Y. (2016), A complexity-based approach for the detection of weak signals in ocean ambient noise, *Entropy*, **18**(3): 101, https://doi.org/10.3390/e18030101.

37. Singh P., Saha G., Sahidullah M. (2021), Non-linear frequency warping using constant-Q transformation for speech emotion recognition, [in:] *2021 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–6, https://doi.org/10.1109/ICCCI50826.2021.9402569.

38. Song G., Guo X., Wang W., Ren Q., Li J., Ma L. (2021), A machine learning-based underwater noise classification method, *Applied Acoustics*, **184**: 108333, https://doi.org/10.1016/j.apacoust.2021.108333.

39. Thomas M., Martin B., Kowarski K., Gaudet B., Matwin S. (2020), Marine mammal species classification using convolutional neural networks and a novel acoustic representation, [in:] *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Lecture Notes in Computer Science*, **11908**: 290–305, https://doi.org/10.1007/978-3-030-46133-1_18.

40. Yang Y., Yao Q., Wang Y. (2024), Underwater acoustic target recognition method based on feature fusion and residual CNN, *IEEE Sensors Journal*, **24**(22): 37342–37357, https://doi.org/10.1109/JSEN.2024.3464754.

41. Yuan F., Ke X., Cheng E. (2019), Joint representation and recognition for ship-radiated noise based on multimodal deep learning, *Journal of Marine Science and Engineering*, **7**(11): 380, https://doi.org/10.3390/jmse7110380.

42. Zeng Y., Zhang M., Han F., Gong Y., Zhang J. (2019), Spectrum analysis and convolutional neural network for automatic modulation recognition, [in:] *IEEE Wireless Communications Letters*, **8**(3): 929–932, https://doi.org/10.1109/LWC.2019.2900247.