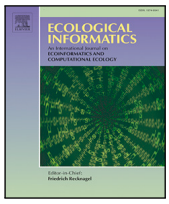
Ecological Informatics 95 (2026) 103709

Contents lists available at ScienceDirect

Ecological Informatics

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



Linear probing enables Ship-Radiated Noise recognition with pretrained audio embeddings

Hilde I. Hummel [a,*], Sandjai Bhulai [b], Rob van der Mei [a,b], Burooj Ghani [c]

[a] Centre of Mathematics and Computer Science, Stochastics group, Science Park 123, Amsterdam, 1098 XG, Netherlands
[b] Vrije Universiteit, Department of Mathematics, De Boelelaan 1111, Amsterdam, 1081 HV, Netherlands
[c] Naturalis Biodiversity Center, Darwinweg 2, Leiden, 2333 CR, Netherlands

ARTICLE INFO

Dataset link: https://www.sciencedirect.com/science/article/pii/S0957417421007016, https://github.com/irfankamboh/DeepShip, https://www.sciencedirect.com/science/article/pii/S0003682X16301566, https://underwaternoise.atlanttic.uvigo.es/

Keywords:
Ship radiated noise
Pretrained audio models
UATR
PAM
Underwater acoustics

ABSTRACT

Even though the ocean covers the majority of the planet's surface, it remains the least explored ecosystem. As light and radio waves do not propagate through water, underwater acoustics is the main choice for various ocean applications ranging from marine biology to pollution monitoring. Increasing levels of anthropogenic noise from ships contribute significantly to underwater sound pollution, posing risks to marine ecosystems. This makes monitoring crucial to understand and quantify the impact of the ship radiated noise. Passive Acoustic Monitoring (PAM) systems are widely deployed for this purpose, generating years of underwater recordings across diverse soundscapes. Manual analysis of such large-scale data is impractical, motivating the need for automated approaches based on machine learning. Recent advances in automatic Underwater Acoustic Target Recognition (UATR) have largely relied on supervised learning, which is constrained by the scarcity of labeled data. Transfer Learning (TL) offers a promising alternative to mitigate this limitation. In this work, we conduct the first empirical comparative study of transfer learning for UATR, evaluating multiple pretrained audio models originating from diverse audio domains. The pretrained model weights are frozen, and the resulting embeddings are analyzed through classification, clustering, and similarity-based evaluations. The analysis shows that the geometrical structure of the embedding space is largely dominated by recording-specific characteristics. However, a simple linear probe can effectively suppress this recording-specific information and isolate ship-type features from these embeddings. As a result, linear probing enables effective automatic UATR using pretrained audio models at low computational cost, significantly reducing the need for a large amounts of high-quality labeled ship recordings.

1. Introduction

The ocean covers roughly 71% of the planet's surface, however, it is the least explored ecosystem (Ramirez-Llodra et al., 2010). As light and radio waves do not propagate well through water, exploration of the ocean is mainly driven by acoustic analysis (Z. Li et al., 2025). This leads to diverse applications of underwater acoustics ranging from climate change measurement and marine biology to pollution monitoring, such as sound pollution (Z. Li et al., 2025). The latter comes from various anthropogenic underwater noises, such as *Ship-Radiated Noise* (SRN). Due to the increase in activity of ships over the years, the increased levels of anthropogenic noise generated by ships pose a threat to the sustainability of marine ecosystems (Williams et al., 2015; Chahouri et al., 2022). The low-frequency sounds produced by the engines, propellers, and hull vibrations of ships (Hummel et al., 2024) interfere with the natural acoustic environment of the oceans, disrupting vital behaviors such as communication between marine species (Williams et al., 2015). This makes monitoring of ship sounds crucial to better understand and mitigate their ecological impact (Hummel et al., 2024).

For this, Passive Acoustic Monitoring (PAM) systems are deployed to study the underwater soundscape. PAM systems enable continuous, non-invasive observation of ocean environments by recording ambient sounds through hydrophones. Since PAM systems do not emit an acoustical signal but only listen, their monitoring does not disturb the marine ecosystem. This makes PAM particularly suitable for long-term ecological studies and for detecting anthropogenic noise in various habitats.

Although PAM supports research on marine mammal detection (Bittle and Duncan, 2013), fish biodiversity (Lindseth and Lobel, 2018), and even climate change (Affatati et al., 2022), this study focuses on SRN. Each individual ship has its own unique sound profile, which

* Corresponding author.
E-mail address: h.i.hummel@cwi.nl (H.I. Hummel).

https://doi.org/10.1016/j.ecoinf.2026.103709
Received 13 January 2026; Received in revised form 10 March 2026; Accepted 10 March 2026
Available online 25 March 2026

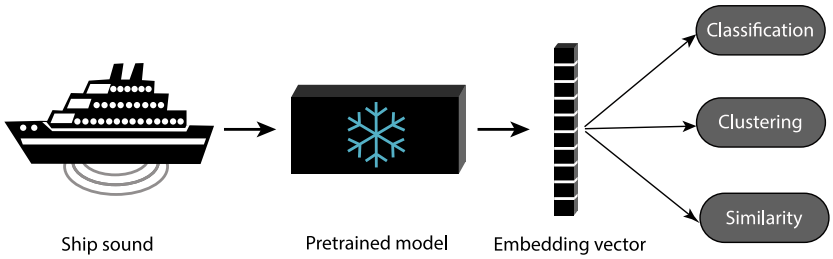


Fig. 1. Illustration of the evaluation method for various pretrained audio models.

can be used to identify its type and even the individual ship (Hummel et al., 2024). This process is referred to as *Underwater Acoustic Target Recognition* (UATR). Manual analysis of these recordings is time-consuming and labor-intensive. Given that PAM systems often produce years of unlabeled continuous recordings, the resulting data volume makes manual annotation and analysis impractical. This creates the need for the automation of this process.

To this end, a variety of machine learning (ML) methods have been suggested to automatically monitor ship sounds (Hummel et al., 2024). Most of these studies focus on supervised learning, which limits their use due to the scarcity of labeled data. The labeled datasets span a small region of the ocean (Santos-Domínguez et al., 2016; Irfan et al., 2021), which limits the environmental diversity within the recordings. In addition, due to the small size of the dataset, it is not feasible to train large ML models. Therefore, these methods may not be generalizable to newly seen ships or other ocean environments (Hummel et al., 2026). To cope with the limited amount of labeled data, *Transfer Learning* (TL) has emerged as a promising approach, as demonstrated in fields such as computer vision (Kornblith et al., 2019; Mensink et al., 2021), general audio processing (Tsalera et al., 2021), and bioacoustics (Çoban et al., 2020; Ghani et al., 2025).

In TL, the backbone of the model is first pretrained using a large dataset of a related domain. By doing so, it has learned some general features that could be generalizable to other domains. Next, the pretrained model can be repurposed as a feature extractor to extract information from another dataset. These features can then be applied to perform a new task, where only a small classifier needs to be trained instead of the whole model (Ghani et al., 2025). This approach reduces the number of trainable parameters and, therefore, a small labeled dataset is sufficient, eliminating the need for a large qualitative labeled dataset (Van Merriënboer et al., 2024). However, the success of TL depends on the choice of the backbone of the pretrained model (Schwinger et al., 2025). Numerous pretrained models based on large-scale audio data have recently been proposed (Zaman et al., 2023), and several studies have shown their effectiveness in various downstream tasks (Turian et al., 2022; Saeed et al., 2021).

A common approach involves freezing the pretrained model weights and using the resulting embedding vectors as input features for a simple classifier (Rauch et al., 2025; Kather et al., 2025). To the best of our knowledge, no comprehensive study has examined the transferability of these pretrained audio models to UATR. Motivated by this, in this work, a range of pretrained audio models from various audio domains are systematically evaluated to assess their generalizability to UATR. The performance of the models is evaluated using three complementary methods in two benchmark datasets: (1) ship-type classification through linear probing, (2) clustering-based embedding analysis, and (3) similarity-based embedding evaluation. Together, these methods show classification performance and the representational quality of the embedding vectors generated. An illustration of this is given in Fig. 1. All data, code, and trained models used in this study are openly available under an Open Access license at this GitHubpage[1] to ensure transparency and reproducibility. In summary, the contributions of this study are as follows:

- A systematic review of a number of pretrained models in various audio domains;
- An extensive empirical study on the transfer of these models to automatic UATR;
- A combination of evaluation metrics to evaluate and compare the performance of pretrained audio models for UATR.

2. Related work

This section describes recent work on automated UATR and briefly describes comparison studies for other audio domains.

2.1. Classification of ships

Previous work on automated UATR has primarily focused on *Supervised Learning* (SL) (Hummel et al., 2024). These methods have largely relied on *Convolutional Neural Networks* (CNNs), although more recent work has shifted the focus to Transformer-based models. However, limited attention has been paid to the generalizability of these models (Müller et al., 2024). This gap has been addressed in Hummel et al. (2026), where a small Conformer model is proposed that is optimized using *Self-Supervised Learning* (SSL). The model is trained on a large amount of unlabeled data extracted from a single hydrophone. After pretraining, the model weights were frozen, and a linear classifier was added for downstream classification. The model was shown to be generalizable in both the classification of ship type and the classification of marine mammal calls. Recently, SSL frameworks have received increasing attention for UATR.

For example, Xu et al. (2023) proposed an SSL framework in which the pretraining of the framework has not been done on underwater acoustic data, but they utilized a general dataset called AudioSet (Gemmeke et al., 2017). The framework employs a so-called Swin Transformer encoder that processes masked and patched Mel-spectrograms, with one decoder reconstructing the masked patches and the other reconstructing the gammatone spectrogram of the original audio. Here, the first decoder captures local information, and the second focuses on the global information of the recording. They evaluated their method on the benchmark dataset Deepship, reaching 80% accuracy (Gemmeke et al., 2017). Similarly, W. Li et al. (2025) also performed the pretraining of their learning framework on AudioSet. Their framework consists of two modules, a combination of masked learning and contrastive learning using a Swin Transformer during pretraining. The knowledge captured within the Transformer model is then transferred to a light-weight CNN model. This reduces the number of parameters, facilitating the possibility of using the model on edge devices. Their model achieved over 87% accuracy on the Deepship dataset. Although these studies introduce novel architectures and training paradigms, none of the well-established SSL methods have been proven to be effective in other audio domains and have yet been systematically applied to UATR.

[1] https://github.com/hildeingvildhummel/Decodable_not_Structured

2.2. Previous comparison studies

For speech-based models, a standardized benchmark known as the Speech Processing Universal PERformance Benchmark (SUPERB) (Yang et al., 2021) has been proposed to fairly evaluate the generalizability of models across a wide range of speech-related tasks. This approach is extended in Yang et al. (2024), where they evaluated 33 speech foundation models in 15 speech-related tasks. These tasks range from speaker verification to speech enhancement to automatic speech recognition and keyword recognition. Each layer of every model was evaluated, motivated by the understanding that different layers of *Bidirectional Encoder Representations from Transformers* (BERT) capture distinct types of information. Their findings demonstrated that SSL models have superior performance in the generalizability to these various speech-related tasks. However, these speech models have thus far been tested only on speech-related downstream tasks and do not take any other audio domain into account.

In parallel, various studies have been conducted to explore the potential of TL in the bioacoustic domain. Kather et al. (2025) compared 15 bioacoustic models by classifying either birds or frogs using these pretrained bioacoustic models. In addition, they evaluated the generated embedding spaces by applying clustering and Adjusted Mutual Information (Kather et al., 2025). In this study, they concluded that SL models still outperform SSL ones overall, though SSL models exhibit better generalization to new tasks such as frog classification.

Schwinger et al. (2025) compared bioacoustic and general audio models, providing an extensive review of their architectures and evaluating their performance on two bioacoustic datasets using both linear and attentive probing. Here, they showed that Transformer models benefit from attentive probing, achieving performance comparable to or exceeding that of SL models. Similarly, Miron et al. (2025) compared various general audio models with several bioacoustic models. They extended the research (Kather et al., 2025) by using clustering to evaluate the embedding space and adding a retrieval evaluation. During retrieval evaluation, the embedding space is directly evaluated using the cosine similarity of each test sample compared to the full test set. They found that pretraining on a mixture of bioacoustic and general audio data, followed by supervised post-training on the same mixture, yields the best in- and out-of-distribution performance.

3. Methods

This section outlines the datasets used for pretraining and benchmarking, followed by descriptions of the selected models and evaluation methodologies.

3.1. Datasets

The models employed in this study were pretrained on various domain-specific datasets prior to further evaluation. Subsequently, the performance of the models are evaluated using two benchmark datasets. This section provides a detailed description of both the pretraining and benchmark datasets.

3.1.1. Datasets: pretraining datasets

Depending on the audio domain, various pretraining datasets are in place. A well-known dataset for general audio purposes is AudioSet (Gemmeke et al., 2017), which contains more than 2 million annotated 10-second audio files sourced from YouTube. This makes the dataset highly diverse, containing more than 600 classes, although the content is still dominated by music and speech. It can be difficult to obtain the complete AudioSet, because YouTube videos may be removed over time. To address this, another dataset, FSDK50 (Fonseca et al., 2021) has been introduced. This dataset contains over 51k audio clips that were manually annotated and derived from the AudioSet ontology. Similarly, VGGSound (Chen et al., 2020) extracted audio and visuals from YouTube videos, resulting in a dataset of paired audio-visual samples. In the speech domain, Librispeech (Panayotov et al., 2015) is a large dataset considering English-only speech. This dataset was explicitly designed to train automatic speech recognition systems. Unlike general audio datasets, this dataset does not contain classes, but transcriptions of the speech that have been recorded. Beyond speech and general audio, the availability of bioacoustic datasets has grown to monitor biodiversity in a non-invasive manner. Bird sound datasets, in particular, have reached a substantial scale. In this domain, Xeno-Canto (Vellinga and Planqué, 2015) plays a central role in this domain. This dataset offers globally observed bird songs that have been recorded by various devices, spanning more than 10,000 classes. Based on Xeno-Canto, another large-scale labeled dataset BirdSet is introduced (Rauch et al., 2024). This dataset is a collection of several bioacoustic datasets, with the focals derived from the Xeno-Canto dataset.

In addition to bird recordings, other species are also represented in recent bioacoustic efforts. The Meerkat data set (Schäfer-Zimmermann et al., 2024) comprises more than 1000 h of annotated meerkat vocalizations labeled by type of social interaction. In total, the recordings are 10 s long, completing a total of 1068 h of recordings. The final bioacoustic dataset taken into account is the iNatSounds dataset (Chasmai et al., 2024). This dataset differentiates itself from the others by incorporating sounds from a larger diversity of animals than only birds or meerkats, including mammals, insects, reptiles, and amphibians, resulting in more than 5500 species. Unfortunately, all of these recordings are weakly labeled. This means that it contains a single positive label for the entire recording. Therefore, it is possible that the majority of the recording contains background noise or is even interrupted by the sound of another species. Aquatic ecosystems are also being studied through passive acoustic monitoring. To automatically monitor biodiversity, especially within coral reefs, a dataset ReefSet (Williams et al., 2024) is introduced. This dataset is a combination of 16 individual datasets that have been recorded across 12 countries. Additionally, *National Oceanic and Atmospheric Administration* (NOAA) (Passive Acoustic Data Collection, 2017) has an extensive amount of unlabeled data that can be utilized for underwater biodiversity monitoring. In Harrell (2024), they annotated part of the NOAA data themselves to train the model. A complete overview of all the pretraining datasets is given in Table 1.

3.1.2. Benchmark datasets: The evaluation sets

For the evaluation of the proposed models, two benchmark datasets were selected. The first, Deepship (Irfan et al., 2021), contains more than 45 h of single-ship recordings collected in the water near Vancouver. The annotations were made by defining an inclusion zone of a two km range around the hydrophone, such that each single ship within this range was defined as recorded. The dataset comprises four ship classes.

The second benchmark dataset, ShipsEar (Santos-Domínguez et al., 2016), is smaller compared to Deepship, totaling eight hours of single-ship recordings. The audio was recorded on the Spanish Atlantic coast in northwest Spain and consists of five classes. Both Deepship and ShipsEar are commonly used datasets for the evaluation of automatic UATR (Hummel et al., 2024). Both datasets were divided into a training set and a test set. For Deepship, a time-wise split was made as described in Hummel et al. (2026). This split ensures that there is no data leakage in the test set and that the test set represents the future relative to the training set. For ShipsEar, this information is not present to make such a split. Therefore, 80% of the individual recordings were utilized for training, and the remaining recordings for testing. Both datasets represent distinct acoustic environments and vessel populations, making cross-dataset evaluation a useful measure of model generalizability.

Table 1
Overview of pretraining datasets.

Type	Dataset	Duration	Labeled	# Classes	Descriptor	Citation
General Audio	AudioSet	5790 h	Yes	631	AS	Gemmeke et al. (2017)
	FDSK50	100+ h	Yes	200	FDSK50	Fonseca et al. (2021)
	VGGSound	550 h	Yes	309	VGG	Chen et al. (2020)
Speech	Librispeech	960 h	Yes	Transcriptions	LS	Panayotov et al. (2015)
Bioacoustics	Xeno-Canto	± 1100 h	Yes	10,000+	XC	Vellinga and Planqué (2015)
	BirdSet	6877 h	Yes	±10,000	BS	Rauch et al. (2024)
	iNatSounds	± 1200 h	Yes	5,500+	iN	Chasmai et al. (2024)
	MeerKAT	1068 h	Partially	8	MK	Schäfer-Zimmermann et al. (2024)
Marine Life sounds	Reefset	±1,800 h	Yes	37	RS	Williams et al. (2024)
	National Oceanic and Atmospheric Administration	–	No	–	NOAA	Passive Acoustic Data Collection (2017)

3.2. Models: Pre-trained models used in this study

The models in this comparative study were selected based on the audio domain in which they were initially trained. This could either be "*general audio*", "*speech*", "*bioacoustics*", or "*marine life sounds*". Models trained on a large amount of data and possessing the potential to generalize well to other related audio domains were selected. In addition, these models are not pretrained on either of the benchmark datasets, and they need to have a clearly defined embedding space for evaluation purposes. This is important, since it is then clear where the information bottleneck is defined within the architecture, and therefore, it is clear which layer should be used to extract the features. An overview of the selected models is given in Table 2. For some of these models, multiple sizes are available. If this were the case, the base models were chosen for evaluation.

3.2.1. General audio models

For the general audio domain, three individual pretrained models were selected: (1) Audio Masked AutoEncoder (AudioMAE) (Huang et al., 2022), (2) Bidirectional Encoder representation from Audio Transformers (BEATS) (Chen et al., 2022b), and (3) *Hidden-Unit BERT AudioSet* (HuBERT AS) (La Quatra et al., 2024). Each model employs a distinct self-supervised learning strategy. AudioMAE first converts the raw audio into a Mel-spectrogram. These spectrograms are then divided into patches, where some of these patches are masked before being passed to the model. The masked patches are encoded using a Transformer-based encoder, and the resulting representations are processed by a Transformer-based decoder with local attention mechanisms. The decoder tries to reconstruct the masked patches of the input signal and is trained using a simple Mean Squared Error (MSE). This reconstruction objective primarily focuses on the correctness of low-level time-frequency features, but does not take into account the semantic abstraction of high-level audio (Ramesh et al., 2021; Bao et al., 2021). To include this in the training process, BEATS introduced discrete label prediction in the general audio domain. The model employs an acoustic tokenizer to generate discrete labels from unlabeled audio. The SSL framework is then optimized jointly using both a masked reconstruction loss and a discrete label prediction loss, enabling BEATS to capture both acoustic and semantic information. Similarly, HuBERT (Hsu et al., 2021) uses pseudo-labels to train its representations. During the first iteration, these labels are generated by clustering MFCC features using a simple K-Means algorithm. In the subsequent iterations, a random layer from the encoder is selected, and its outputs are clustered to form progressively more informative pseudo-labels. Although originally designed for speech representation learning, for this work, a HuBERT model trained on the full AudioSet is selected (La Quatra et al., 2024). This version is referred to as HuBERT AS.

3.2.2. Speech models

In recent years, several pre-trained speech models have been proposed (Malik et al., 2021). In this study, four representative models were selected: Wav2Vec 2.0 (Baevski et al., 2020), Data2Vec (Baevski et al., 2022), WavLM (Chen et al., 2022a), and HuBERT (Hsu et al., 2021). Wav2Vec2.0 (Baevski et al., 2020) consists of a CNN encoder that inputs the raw audio. The output of this encoder is then fed to a Transformer to extract the representation. In addition, the output of the CNN encoder is discretized using a quantization module. During pretraining, a contrastive learning objective is employed, portions of the encoder output are masked before being passed to the Transformer, while the full latent sequence is quantized. The model learns to identify the correct quantized latent audio representation. This setup has shown great performance in speech-related tasks (Kunešová et al., 2024). Therefore, a similar approach is utilized in Data2Vec (Baevski et al., 2022). Here, they presented a general framework for SSL that can be applied to vision, text, and speech. For speech, Data2Vec adopts a similar architecture to Wav2Vec 2.0 but replaces the contrastive loss with a latent reconstruction objective based on knowledge distillation. Another approach is presented in WavLM (Chen et al., 2022a). Again, this model consists of a CNN encoder coupled to a Transformer. However, they suggested Gated Relative Position Bias (Chi et al., 2021) in the Transformer encoder. The model is pretrained using a masked speech denoising and prediction objective. They suggest that the combination of this objective improves robustness for complex environments (Chen et al., 2022a), making WavLM a promising candidate for transfer learning to other audio domains. Finally, HuBERT (Hsu et al., 2021), as described in Section 3.2.1 above, learns through pseudo-label prediction. It generates labels by clustering features extracted from the raw audio and iteratively refines these labels across training stages. Although originally designed for speech representation learning, its structure allows it to generalize well to other audio-based applications.

3.2.3. Bioacoustic models

Beyond the general audio and speech domain, the bioacoustic domain presents an equally valuable field for evaluation. This domain has access to an enormous amount of labeled data (Vellinga and Planqué, 2015; Chasmai et al., 2024), allowing supervised training of large models. Such large models are expected to generalize to similar audio applications. Moreover, the implementation of various pretrained bioacoustic models is facilitated in the repository called **bacpipe** developed by Kather et al. (2025). This repository provides standardized tools for model evaluation and comparison. In this study, several representative bioacoustic models were selected for evaluation.

The first, BirdMAE (Rauch et al., 2025), is inspired by the original AudioMAE (Huang et al., 2022) and trained on a large number of bird songs. During pre-training, they suggested some small changes; for instance, they replaced the original Swin Transformer with a Vision Transformer. In addition to BirdMAE, another popular bird model, called BirdNet (Kahl et al., 2021), has been selected. This CNN-based

Table 2
Information about the pretrained models chosen for this comparison study.

Domain	Model	Architecture	Num. params	Dataset	Input type	Sample rate	Window size	Embedding size	Training paradigm	Open access	Citation
General Audio	AudioMAE	Transformer	86M	AS	Mel-spectrogram	16 kHz	10 sec	768	SSL	Yes	Huang et al. (2022)
	BEATS	Transformer	90M	AS	Spectrogram	16 kHz	10 secs	768	SSL	Yes	Chen et al. (2022b)
Speech	Wav2Vec2.0	Transformer	95M	LS	Raw audio	16 kHz	10 sec	768	SSL	Yes	Baevski et al. (2020)
	Data2Vec	Transformer	95M	LS	Raw audio	16 kHz	10 secs	768	SSL	Yes	Baevski et al. (2022)
	WavLM	CNN and Transformer	94.70M	LS	Raw audio	16 kHz	10 secs	768	SSL	Yes	Chen et al. (2022a)
	HuBERT Base	CNN and Transformer	95M	LS	Raw audio	16 kHz	10 secs	768	SSL	Yes	Hsu et al. (2021)
Bioacoustics	BirdMAE	Transformer	86M	BirdSet	Spectrogram	32 kHz	10 sec	1280	SSL	Yes	Rauch et al. (2025)
	BirdNet	ResNet	27M	XC, eBird, Macauley Library	Spectrogram	48 kHz	3 secs	1024	SL	Yes	Kahl et al. (2021)
	AvesEcho	PaSST	85M	–	Mel-spectrogram	16 kHz	1 sec	768	SL	Yes	Ghani et al. (2024)
	Animal2vec MK	Transformer	315M	MK	Raw audio	8 kHz	10 secs	1024	SSL	Yes	Schäfer-Zimmermann et al. (2024)
	Perch	EfficientNet	7.8M	XC	Raw audio	32 kHz	5 secs	1280	SL	Yes	Ghani et al. (2023)
	Perch 2.0	EfficientNet	12M	XC, iN, Tierstimmenarchiv, FSDK50	log Mel-spectrogram	32 kHz	5 sec	1536	SL	Yes	van Merriënboer et al. (2025)
	AVES	CNN and Transformer	95M	FSD50K, AS 20k, VGG	Raw audio	16 kHz	1 sec	768	SSL	Yes	Hagiwara (2023)
Marine life sounds	Google Whale	EfficientNet		NOAA	Spectrogram	24kHz	5 sec	1280	SL	Yes	Harrell (2024)
	SurfPerch	EfficientNet		RS	Spectrogram	32 kHz	5 sec	1280	SL	Yes	Williams et al. (2025)

model is optimized in a supervised manner and is able to classify almost 10,000 different bird species. As a follow up on BirdNET, AvesEcho is developed (Ghani et al., 2024), where a student model is optimized with BirdNET as the teacher using knowledge distillation.

Beyond avian models, the Animal2Vec model (Schäfer-Zimmermann et al., 2024) was developed for meerkat vocalizations. This method uses a self-distillation method to reconstruct masked embeddings, which is similar to the Data2Vec framework. The architecture starts with a feature extractor that is constructed of multiple SincNet-style filterbanks followed by a stack of 1D CNN layers. This is a promising method for UATR, as previous research has shown that decomposition-based features perform effectively in underwater acoustic classification tasks (Hummel et al., 2024). On top of the feature extractor, the Data2Vec2.0 framework is used, and the complete model is trained in a self-supervised manner. Although the aforementioned models focus on specific animal groups, other bioacoustic models take advantage of more diverse pre-training datasets that encompass a wide range of species. One such model is Perch (Ghani et al., 2023), a lightweight model based on the EfficientNet network (Tan and Le, 2019). Recently, an updated version of this model was released Perch 2.0 (van Merriënboer et al., 2025). Both models are designed to be computationally efficient while maintaining strong generalization performance. The authors of Perch 2.0 emphasize that the model can effectively adapt to multiple domains through linear probing, making it a promising candidate for transfer learning in ship type classification.

Finally, the *Animal Vocalization Encoder based on Self-Supervision* (AVES) (Hagiwara, 2023) was included for evaluation. As stated in the name of the model, AVES is optimized using SSL on a broad range of animal vocalizations. The model adopts the HuBERT architecture and training paradigm, extending its application from speech to the bioacoustic domain.

3.2.4. Marine life sound models

For the marine life sound models, two models were selected: Google Whale (Harrell, 2024) and Surfperch (Williams et al., 2025). Both models are based on the EfficientNet architecture (Tan and Le, 2019) and are trained in a supervised manner. The Google Whale has been trained with data derived from the NOAA and manually annotated these data to construct a labeled dataset. Their dataset comprises 12 classes, each representing a different type of whale call. The resulting model is designed to detect and classify these calls from PAM recordings. The SurfPerch model (Williams et al., 2024) focuses on monitoring the biodiversity within coral reefs. This model is trained on various underwater sound recordings made within coral reefs in a supervised manner. Both models are relatively lightweight and optimized for efficient inference, making them promising candidates for transfer learning in underwater acoustic classification tasks such as UATR.

3.2.5. Baseline

The performance of the pretrained audio models were compared to a relatively simple benchmark. This comparison was made to detect whether these complex models were able to outperform a relatively simple and basic method. This benchmark employs a logistic regression using log Mel-spectrograms as input features for classification. These Mel-spectrograms were constructed using 128 mel bins, a Fast Fourier size of 1024, and a hop length of 512. Subsequently, mean pooling was applied over the temporal dimension to obtain a fixed-length feature vector for each audio clip. These representations were then used to perform the model performance evaluation.

3.3. Evaluation set-up

For the evaluation of the models, the embedding spaces were evaluated using both unsupervised evaluation methods and a supervised evaluation method. During the unsupervised evaluations, the embeddings were clustered and the similarity of the embeddings were measured. Next, during the supervised evaluation the embeddings are utilized to perform ship type classification.

3.3.1. Unsupervised: Clustering with normalized mutual information

First, the embedding space was evaluated using a clustering-based approach. Following the methodology described in Miron et al. (2025) and Kather et al. (2025), the embeddings are clustered using K-Means. The number of training clusters was set equal to the number of classes within the benchmark dataset. In this case, four classes for the Deepship embeddings and five classes for the ShipsEar embeddings. The similarity between the clusters and the ground truth labels was quantified using the Normalized Mutual Information (NMI) score. The score is given by:

$$NMI = \frac{2 \times I(Y;C)}{[H(Y) + H(C)]}, \tag{1}$$

where Y corresponds to the class label, C to the cluster label, function $H(.)$ to the entropy and $I(Y;C)$ to the mutual information between the cluster labels and the class labels. This score ranges between 0, indicating no mutual information, and 1, indicating a perfect correlation.

3.3.2. Unsupervised: Similarity scores with ranking

In addition to the clustering evaluation, the similarity between neighboring data samples in the test set was analyzed. Each test sample is treated as a query, and the cosine similarity score is calculated on all other test samples, as described in Schwinger et al. (2025) and van Merriënboer et al. (2025). The samples were then ranked from most similar to least similar. From this ranking, the ROC-AUC score is calculated, where samples belonging to the same class as the query were considered positive, and all others were treated as negatives.

3.3.3. Supervised: Ship type classification

To finalize the evaluation, automated UATR was introduced as a down-stream task. The performance for ship-type classification was measured by freezing the backbone of each pretrained audio model and extracting the embeddings. Then, a linear probe was trained using these embeddings to predict the type of ships. This approach eliminates the need to fine-tune the entire model, which reduces the number of trainable parameters and minimizes the amount of labeled data required for training (van Merriënboer et al., 2025).

4. Results

This section presents the results of both the unsupervised and supervised evaluation methods across the various pretrained models. The section is concluded with a comparison of the performance of the HuBERT pretrained in various domains.

4.1. Unsupervised evaluation

First, the embeddings are evaluated in an unsupervised manner. During this evaluation, the geometric separability of the complete embedding space is evaluated by comparing this separability with the label information. This separability is defined by the clusters optimized by a K-Means algorithm or by defining neighboring data samples by the cosine similarity.

4.1.1. Clustering with normalized mutual information

This evaluation assesses whether the clusters defined by a simple K-Means algorithm contain informative structure with respect to the ground-truth labels of the benchmark datasets. As shown in Table 3, most models achieve low NMI scores that are comparable to the performance of the baseline model. Unlike the classification evaluation, where most of the models outperformed the baseline, the clustering evaluation does not show this. The best performing model in this evaluation for ShipsEar is AvesEcho, achieving an NMI of 0.39. For Deepship, the best performing model is Data2Vec, which reaches an NMI score of 0.18. Interestingly, Data2Vec and Wav2Vec2.0 exhibit

Table 3
NMI values of cluster performance on benchmark datasets. The overall best performance in bold and the best performance per domain are underlined.

Domain	Model	Deepship	ShipsEar
General Audio	AudioMAE	0.10	0.18
	BEATS	0.13	0.22
	HuBERT AS	0.10	0.23
Speech	Wav2Vec2.0	0.06	0.13
	Data2Vec	**0.18**	0.24
	WavLM	0.04	0.17
	HuBERT	0.04	0.15
Bioacoustics	BirdMAE	0.14	0.26
	BirdNet	0.05	0.30
	AvesEcho	0.15	**0.39**
	Animal2Vec MK	0.03	0.14
	Perch	0.10	0.15
	Perch 2.0	0.14	0.28
	AVES	0.09	0.19
Marine life sounds	Google Whale	0.06	0.16
	SurfPerch	0.10	0.16
Baseline	Mel Spectrogram	0.00	0.16

an opposite trend. Here, the accuracy scores of Data2Vec were lower compared to Wav2Vec2.0 in the classification evaluation. However, in the clustering evaluation, Data2Vec outperforms Wav2Vec2.0. Overall, these findings indicate that the clusters produced by the K-Means algorithm do not appear to align well with the true ship-type classes, suggesting limited semantic separability in the embedding spaces.

4.1.2. Similarity scores with ranking

The next analysis of the embedding space evaluates whether neighboring data samples within the embedding space are actually derived from the same ship-type. At first glance, the results in Table 4 show that none of the pretrained models substantially outperform the baseline model. Although BEATS is the strongest performer among the pretrained models reaching a ROC-score of 0.61 on Deepship and 0.72 on ShipsEar, it performs similarly to the baseline model reaching 0.62 on Deepship and 0.71 on ShipsEar. These results align with the clustering evaluation, suggesting that the neighboring vectors in the embedding space are often not of the same ship type. In other words, the embeddings do not appear to encode consistent or meaningful characteristics of the ship sound, indicating that the geometric separability of the complete embedding space is not primarily driven by ship-type characteristics.

4.1.3. Are the embeddings defined by record ID information?

As can be seen from the results, the BEATS model achieves the best performance or a competitive performance in both evaluation metrics. To further investigate this, the embeddings of the ShipsEar test set generated by BEATS is visualized in Fig. 2. The plots show a clear separation between the classes. However, it can also be seen that the classes form small subclusters within the embedding space. This is especially visible for the background class. In particular, the background class splits into three distinct subgroups, while the test set also contains three individual recordings. This suggests that the geometric separability of the embedding space generated by the pretrained model is mainly driven by individual recording information.

To determine whether the pretrained BEATS embedding space learned to differentiate recordings, the evaluation was repeated using record IDs instead of ship-type labels. Since individual recordings do not appear in both the training set and the test set, the classification of the recording ID is excluded from this evaluation. For cluster evaluation, the number of clusters for the K-Means algorithm was set to the number of individual recordings in the test set. For ShipsEar,

Table 4
Mean ROC-AUC values from ranking based on the cosine similarity. The overall best performance is in bold, and the best performance per domain are underlined.

Domain	Model	Deepship	ShipsEar
General Audio	AudioMAE	0.60	0.63
	BEATS	0.61	**0.72**
	HuBERT AS	0.59	0.67
Speech	Wav2Vec2.0	0.54	0.55
	Data2Vec	0.60	0.64
	WavLM	0.52	0.62
	HuBERT	0.53	0.61
Bioacoustics	BirdMAE	0.61	0.70
	BirdNet	0.59	0.67
	AvesEcho	0.58	0.71
	Animal2Vec MK	0.57	0.64
	Perch	0.59	0.66
	Perch 2.0	0.60	0.69
	AVES	0.60	0.67
Marine life sounds	Google Whale	0.55	0.62
	SurfPerch	0.59	0.66
Baseline	Mel Spectrogram	**0.62**	0.71

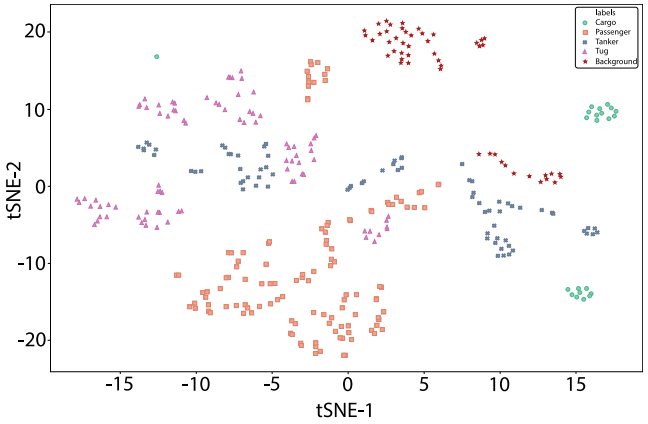


Fig. 2. t-SNE plot of the embeddings generated by BEATS using the ShipsEar test set.

Table 5
Evaluation of the recording ID information in the BEATS embeddings.

Metric	Deepship	ShipsEar
NMI	0.62	0.66
ROC-AUC	0.90	0.93

this was set to 21, and for Deepship to 237. Table 5 shows that both the resulting NMI and ROC-AUC scores are substantially higher compared to those obtained with label-based evaluation. In particular, the ShipsEar NMI increased from 0.22 to 0.66, while the ROC-AUC increased from 0.72 to 0.93. This shows that the placement of data samples within the complete embedding space is mainly driven by record ID information. This analysis has been repeated for the other pretrained models, resulting in the same phenomenon.

4.2. Supervised evaluation

To finalize the evaluation, the embedding spaces were evaluated by introducing a downstream task. In this study, this task was defined as classifying various ship-types from two distinct benchmark datasets. This analysis checks whether the pretrained model is able to translate

to a newly seen task by training a classifier in a supervised manner for a task that the model was initially not pretrained on.

4.2.1. Ship type classification: linear probing

By applying a linear probe, the majority of the models substantially outperform the baseline model (see Table 6). This indicates that, even though the geometric separability of the embedding space did not indicate containing any ship-type characteristics, a low-cost linear probe was still able to effectively differentiate between various ship-types.

In general, the results in Table 6 show that the accuracy scores in the ShipsEar dataset are higher compared to those of the Deepship dataset in most cases. This difference was expected because of the way the datasets were split into a training set and a test set. For Deepship, a time-wise split was defined to differentiate between the training set and the test set. This ensured no data leakage, but introduced previously unseen ships in the test set. In addition, seasonality differences can also induce variability between the recordings in the training set and the test set. Together, this explains the overall lower performance on Deepship compared to ShipsEar.

When examining the performance of individual models, Wav2Vec2.0 has the best performance in the ShipsEar dataset reaching 78% accuracy. Additionally, BEATS performed the best in the Deepship dataset reaching 65.4% accuracy, and is also the second best performer in ShipsEar, reaching 74% accuracy. In contrast, WavLM underperformed relative to the baseline model in both ShipsEar and Deepship. This could be explained by the denoising objective that is used for pretraining, which may suppress relevant noise characteristics of the ship, and thus remove critical information required for accurate classification. Another notable result is the high performance of Animal2Vec, despite its use of the lowest sampling rate (8 kHz) among all models. This suggests that this lower temporal resolution did not affect the model's ability to distinguish between ship types. In contrast, marine life sound models perform poorly. Although both models were optimized for various underwater sounds, they failed to make the translation to ship noises. Their supervised training in marine life sounds likely limited their ability to generalize to other acoustic domains such as ships. Similarly, BirdNET, Perch, and Perch2.0 were also pretrained in a supervised manner. However, these models benefited from multiple labeled datasets, introducing more varied data diversity, and consequently improved their generalization capability. In addition to linear probing, a more advanced probing method, attentive probing, has also been evaluated. Overall, these results did not compete with the linear probe, making the linear probe the best performer while also being computationally more efficient.

Table 6
Accuracy values of performance on benchmark datasets using linear probing. The overall best performance is in bold, and the best performance per domain are underlined.

Domain	Model	Deepship	ShipsEar
General Audio	AudioMAE	61.9%	67.2%
	BEATS	**65.4%**	74.0%
	HuBERT AS	54.2%	42.7%
Speech	Wav2Vec2.0	56.8%	**78.0%**
	Data2Vec	56.8%	53.1%
	WavLM	49.7%	45.0%
	HuBERT	50.1%	32.8%
Bioacoustics	BirdMAE	63.4%	67.2%
	BirdNet	58.6%	57.5%
	AvesEcho	57.22%	64.26%
	Animal2Vec MK	64.0%	61.0%
	Perch	55.0%	64.3%
	Perch 2.0	61.4%	73.2%
	AVES	56.4%	59.1%
Marine life sounds	Google Whale	45.9%	50.6%
	SurfPerch	57.8%	58.9%
Baseline	Logistic Regression	56.4%	48.6%

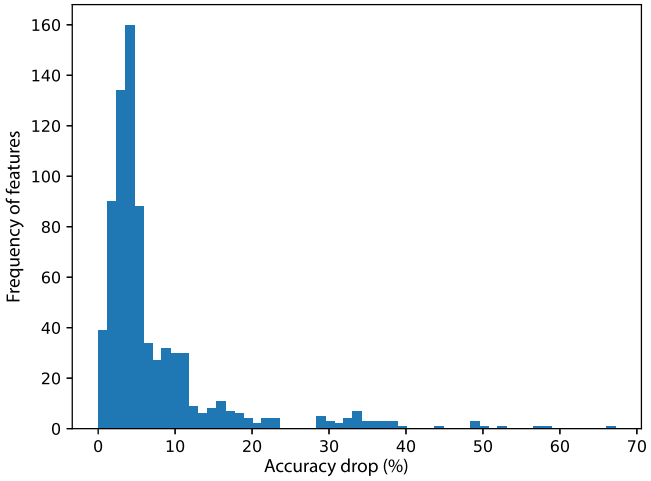


Fig. 3. Histogram of the accuracy drop of removing each feature from the BEATS embeddings individually on the classification of ShipsEar.

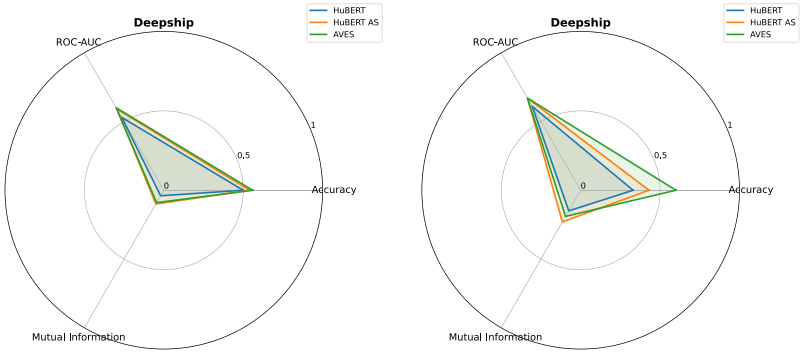
4.2.2. Is the classification performance driven by record ID information?

To assess whether the observed classification performance is primarily driven by recording-specific artifacts rather than ship-type information, a label-shuffling control experiment was conducted on the BEATS embeddings. Specifically, the ship-type labels were randomly permuted across recordings and retrained the linear probe on these shuffled labels. If the embeddings mainly encoded recording identity or other spurious cues that correlate with ship type, the classifier would still achieve relatively high accuracy even after shuffling. Instead, performance dropped sharply to 44.6% accuracy on ShipsEar and 22.9% on Deepship. This decrease in performance indicates that once the semantic link between embeddings and ship types is broken, the model cannot generalize meaningfully. In other words, the original classification performance relies on genuine ship-type structure present in the embedding space, not solely on recording-specific information.

To further probe this observation, the clustering analysis was repeated using the logits of the trained linear probe rather than the raw embeddings generated by BEATS. Clustering in logit space yields substantially higher alignment with ship types, with NMI scores of 0.42 for ShipsEar and 0.36 for DeepShip, approximately twice the NMI obtained from the full embedding space. This suggests that while the embedding space contains mixed factors of variation, the linear probe successfully isolates ship-type-relevant dimensions, suppressing recording-specific effects.

To verify that this result is not an artifact of dimensionality reduction, the analysis is repeated on the reduced embedding space of BEATS by performing a *Principal Component Analysis* (PCA). The reduced embedding space resulted in a slightly higher NMI score from 0.13 in Deepship and 0.22 on ShipsEar originally, to 0.14 on Deepship and 0.27 on ShipsEar. This shows that indeed the increase in the NMI score on the logits is not driven by the dimensionality reduction, but rather by the ship characteristic information.

The same experiments have been conducted on the other pretrained audio embeddings, showing similar results. All together, these results suggest that the linear probe can select a small subset of the complete embedding space to extract ship type characteristics even though the complete embedding space is overruled by record ID information. To support this claim, a feature importance analysis was performed on the original BEATS classifier. During this analysis, each feature is set to 0 at the time while the original classifier tries to classify the various ship types. The results in Fig. 3 show a histogram of the drop in



(a) Performance on Deepship (b) Performance on ShipsEar

Fig. 4. Spider plot of the performances of HuBERT-based models trained on datasets from different domains.

accuracy per feature. Here, the figure shows that the majority of the elimination of the features did not influence the performance of the model. Only a small portion of the features had a severe impact on the performance of the model, underscoring the hypothesis that the linear layer selects a small subset of the embedding space to perform satisfactory classification.

4.3. Influence on data domain: Comparison of HuBERT

Among the evaluated models, three models can be directly compared: HuBERT, HuBERT (AS), and AVES. All of these models share the same architecture and training paradigm, but differ in their pretraining dataset. By comparing the performance of these three models, the influence of the pretraining domain can be visualized. The results summarized in Fig. 4 show that all three models have limited performance on evaluation tasks, suggesting that the HuBERT pseudo-label-based learning objective may not effectively capture features relevant to the acoustics of the ship. However, some domain-related differences can be seen. The model pretrained in speech is underperforming the other models for all three evaluation methods, indicating a poor transfer from speech to UATR. Interestingly, the bioacoustic model (AVES) even slightly outperforms the general audio model. This phenomenon can also be seen in other models, where the majority of the bioacoustic models slightly outperform the general audio models. Especially the models that have been trained on bird data show good performance, which could be explained by the diversity of bird data. Bird-based datasets, in particular, contain recordings from varied environments and devices worldwide, exposing models to a wide range of natural and anthropogenic sounds. The model may benefit from this diversity and, therefore, be able to generalize slightly more to other domains.

5. Discussion

This work presents an extensive empirical study on the transfer learning of pretrained audio models from various domains to UATR. When comparing performance between different pretrained audio models, it is important to consider variations in preprocessing techniques specific to each model. For example, differences in sample rates may influence the results, where higher sample rates retain more information than lower sample rates, potentially benefiting downstream performance (Schwinger et al., 2025). The same principle applies to the embedding size; larger embedding sizes may retain more detailed information, but can also increase the risk of overfitting to the pretraining data (Hummel et al., 2026). Nevertheless, this variation enables the examination of model diversity, providing deeper insight into how design choices impact downstream performance.

Initially, the geometric separability of the embedding space was evaluated by unsupervised metrics. Both metrics showed a lower or comparable performance compared to the baseline model. This suggests that the geometric separability of the pretrained embeddings was similar to the baseline, despite outperforming the baseline in the classification evaluation. In other words, embeddings that were close in the embedding space were not necessarily derived from the same ship class. To explore this, the embedding spaces were analyzed in more detail. Visualization of BEATS embeddings indicated that embeddings from the same recording clustered closely together, and similarity scores were largely driven by recording-specific properties. Environmental factors, hydrophone types, and recording configurations likely introduce greater variance between the recordings than the various ship-types.

Despite this, the performance of the pretrained models was evaluated using linear probing to classify ship types from the labeled benchmark datasets. At first glance, most of these models achieved satisfactory accuracy scores, given that the ship-type classification is a multi-class problem. Surprisingly, the models optimized for marine life showed poor performance. This was unexpected, as these models were presumed to have learned some general underwater acoustic properties. Other research has similarly reported that the generalizability of GoogleWhale to other related tasks is limited (Kather et al., 2025). In contrast, models pretrained on bioacoustic or general audio data demonstrated stronger transfer performance to UATR. Among them, BEATS achieved particularly strong results in both ShipsEar and DeepShip datasets. Its robustness in DeepShip is especially notable, given that the test set was constructed using a time-based split, which means that the majority of individual ships in the test data were absent from the training set (Hummel et al., 2026). This finding indicates that BEATS was able to extract meaningful ship-type characteristics directly from raw audio that generalize to previously unseen vessels.

Overall, these results indicate that, while ship-type information is linearly decodable from the embeddings, the global structure of the embedding space is dominated by recording-specific characteristics. This suggests that recording artifacts and session-dependent properties exert a stronger influence on the learned representations than ship-type semantics. The linear probe effectively learned to downweight embedding dimensions dominated by recording-specific information, focusing instead on the small subset that carried ship-identifying features. This acts much like a feature selection mechanism, highlighting that even when the embedding space is noisy and entangled, simple models can extract meaningful signals. However, supervision is needed to do a meaningful selection for the downstream task. These insights underline both the promise and limitations of pretrained audio models in transferring to complex underwater acoustic tasks like UATR.

6. Conclusion

In conclusion, pretrained audio models effectively transfer to UATR tasks. Especially audio models pretrained on general audio tasks were

able to transfer successfully to ship sounds, where the BEATS model outperformed the other models. An in depth analysis of the BEATS embedding space showed that the embeddings were mainly driven by recording specific characteristics. However, a low cost linear probe was able to extract a small subset of the embedding space to accurately classify ship types in two different benchmark datasets. This means that pretrained audio models can effectively be reused for UATR tasks by only optimizing a low cost linear probe. This greatly reduces the need for a large scale labeled dataset. This may enable pretrained audio models to transfer to other related underwater acoustic applications, such as climate change measurements and marine life tracking. Together, this study gives the potential for automated exploration of the ocean ecosystem without the need for large-scale annotations and heavy computations.

CRediT authorship contribution statement

Hilde I. Hummel: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Sandjai Bhulai:** Writing – review & editing, Supervision. **Rob van der Mei:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Burooj Ghani:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

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

Data availability

The data used in this study can be accessed by contacting the authors of the original manuscript of the benchmark datasets. For Deepship it can be found at https://www.sciencedirect.com/science/article/pii/S0957417421007016 along with https://github.com/irfankamboh/DeepShip and for ShipsEar the data can be accessed via https://www.sciencedirect.com/science/article/pii/S0003682X16301566 and https://underwaternoise.atlanttic.uvigo.es/.

References

Affatati, A., Scaini, C., Salon, S., 2022. Ocean sound propagation in a changing climate: Global sound speed changes and identification of acoustic hotspots. Earth's Futur. 10 (3), e2021EF002099.

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., Auli, M., 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In: International Conference on Machine Learning. PMLR, pp. 1298–1312.

Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. 33, 12449–12460.

Bao, H., Dong, L., Piao, S., Wei, F., 2021. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254.

Bittle, M., Duncan, A., 2013. A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring. In: Proceedings of Acoustics. 2013, Australian Acoustical Society Victor Harbor, SA, pp. 1–8.

Chahouri, A., Elouahmani, N., Ouchene, H., 2022. Recent progress in marine noise pollution: A thorough review. Chemosphere 291, 132983.

Chasmai, M., Shepard, A., Maji, S., Van Horn, G., 2024. The iNaturalist sounds dataset. Adv. Neural Inf. Process. Syst. 37, 132524–132544.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al., 2022a. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE J. Sel. Top. Signal Process. 16 (6), 1505–1518.

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Wei, F., 2022b. Beats: Audio pre-training with acoustic tokenizers. arXiv preprint arXiv:2212.09058.

Chen, H., Xie, W., Vedaldi, A., Zisserman, A., 2020. Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 721–725.

Chi, Z., Huang, S., Dong, L., Ma, S., Zheng, B., Singhal, S., Bajaj, P., Song, X., Mao, X.-L., Huang, H., et al., 2021. XLM-E: Cross-lingual language model pre-training via ELECTRA. arXiv preprint arXiv:2106.16138.

Çoban, E.B., Pir, D., So, R., Mandel, M.I., 2020. Transfer learning from youtube soundtracks to tag arctic ecoacoustic recordings. In: ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 726–730.

Fonseca, E., Favory, X., Pons, J., Font, F., Serra, X., 2021. Fsd50k: an open dataset of human-labeled sound events. IEEE/ACM Trans. Audio, Speech, Lang. Process. 30, 829–852.

Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M., 2017. Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 776–780.

Ghani, B., Denton, T., Kahl, S., Klinck, H., 2023. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. Sci. Rep. 13 (1), 22876.

Ghani, B., Kalkman, V.J., Planqué, B., Vellinga, W.-P., Gill, L., Stowell, D., 2024. Generalization in birdsong classification: impact of transfer learning methods and dataset characteristics. arXiv preprint arXiv:2409.15383.

Ghani, B., Kalkman, V.J., Planqué, B., Vellinga, W.-P., Gill, L., Stowell, D., 2025. Impact of transfer learning methods and dataset characteristics on generalization in birdsong classification. Sci. Rep. 15 (1), 16273.

Hagiwara, M., 2023. AVES: animal vocalization encoder based on self-supervision. arxiv.

Harrell, L., 2024. Whistles, songs, boings, and biotwangs: Recognizing whale vocalizations with AI. https://research.google/blog/whistles-songs-boings-and-biotwangs-recognizing-whale-vocalizations-with-ai/. [Online; accessed 7-Oct-2025].

Hsu, W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio, Speech, Lang. Process. 29, 3451–3460.

Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., Feichtenhofer, C., 2022. Masked autoencoders that listen. Adv. Neural Inf. Process. Syst. 35, 28708–28720.

Hummel, H.I., Gansekoele, A., Bhulai, S., van der Mei, R.D., 2026. The computation of generalized embeddings for underwater acoustic target recognition using contrastive learning. Appl. Acoust. 243, 111103. http://dx.doi.org/10.1016/j.apacoust.2025.111103, URL https://www.sciencedirect.com/science/article/pii/S0003682X25005754.

Hummel, H.I., van der Mei, R., Bhulai, S., 2024. A survey on machine learning in ship radiated noise. Ocean Eng. 298, 117252.

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 Syst. Appl. 183, 115270.

Kahl, S., Wood, C.M., Eibl, M., Klinck, H., 2021. BirdNET: A deep learning solution for avian diversity monitoring. Ecol. Informatics 61, 101236.

Kather, V.S., Ghani, B., Stowell, D., 2025. Clustering and novel class recognition: evaluating bioacoustic deep learning feature extractors. arXiv preprint arXiv:2504.06710.

Kornblith, S., Shlens, J., Le, Q.V., 2019. Do better imagenet models transfer better? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2661–2671.

Kunešová, M., Zajíc, Z., Šmídl, L., Karafiát, M., 2024. Comparison of wav2vec 2.0 models on three speech processing tasks. Int. J. Speech Technol. 27 (4), 847–859.

La Quatra, M., Koudounas, A., Vaiani, L., Baralis, E., Cagliero, L., Garza, P., Siniscalchi, S.M., 2024. Benchmarking representations for speech, music, and acoustic events. In: 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops. ICASSPW, pp. 505–509. http://dx.doi.org/10.1109/ICASSPW62465.2024.10625960.

Li, W., Chen, Z., Wang, J., Liu, T., Chen, H., Xu, G., 2025. A cross-architecture masked contrastive learning framework for few-shot underwater acoustic target classification. Knowl.-Based Syst. 114252.

Li, Z., Chitre, M., Stojanovic, M., 2025. Underwater acoustic communications. Nat. Rev. Electr. Eng. 2 (2), 83–95.

Lindseth, A.V., Lobel, P.S., 2018. Underwater soundscape monitoring and fish bioacoustics: a review. Fishes 3 (3), 36.

Malik, M., Malik, M.K., Mehmood, K., Makhdoom, I., 2021. Automatic speech recognition: a survey. Multimedia Tools Appl. 80 (6), 9411–9457.

Mensink, T., Uijlings, J., Kuznetsova, A., Gygli, M., Ferrari, V., 2021. Factors of influence for transfer learning across diverse appearance domains and task types. IEEE Trans. Pattern Anal. Mach. Intell. 44 (12), 9298–9314.

van Merriënboer, B., Dumoulin, V., Hamer, J., Harrell, L., Burns, A., Denton, T., 2025. Perch 2.0: The bittern lesson for bioacoustics. arXiv preprint arXiv:2508.04665.

Miron, M., Robinson, D., Alizadeh, M., Gilsenan-McMahon, E., Narula, G., Pietquin, O., Geist, M., Chemla, E., Cusimano, M., Effenberger, F., et al., 2025. What matters for bioacoustic encoding. arXiv preprint arXiv:2508.11845.

Müller, N., Reermann, J., Meisen, T., 2024. Navigating the depths: A comprehensive survey of deep learning for passive underwater acoustic target recognition. IEEE Access 12, 154092–154118. http://dx.doi.org/10.1109/ACCESS.2024.3480788.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 5206–5210.

Passive Acoustic Data Collection, 2017. NOAA national centers for environmental information. NOAA National Centers for Environmental Information Asheville, NC, https://www.ncei.noaa.gov/products/passive-acoustic-data. [Accessed 1 May 2023].

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I., 2021. Zero-shot text-to-image generation. In: International Conference on Machine Learning. PMLR, pp. 8821–8831.

Ramirez-Llodra, E., Brandt, A., Danovaro, R., De Mol, B., Escobar, E., German, C.R., Levin, L.A., Martinez Arbizu, P., Menot, L., Buhl-Mortensen, P., et al., 2010. Deep, diverse and definitely different: unique attributes of the world's largest ecosystem. Biogeosciences 7 (9), 2851–2899.

Rauch, L., Heinrich, R., Moummad, I., Joly, A., Sick, B., Scholz, C., 2025. Can masked autoencoders also listen to birds?. arXiv preprint arXiv:2504.12880.

Rauch, L., Schwinger, R., Wirth, M., Heinrich, R., Huseljic, D., Herde, M., Lange, J., Kahl, S., Sick, B., Tomforde, S., et al., 2024. Birdset: A large-scale dataset for audio classification in avian bioacoustics. arXiv preprint arXiv:2403.10380.

Saeed, A., Grangier, D., Zeghidour, N., 2021. Contrastive learning of general-purpose audio representations. In: ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 3875–3879.

Santos-Domínguez, D., Torres-Guijarro, S., Cardenal-López, A., Pena-Gimenez, A., 2016. ShipsEar: An underwater vessel noise database. Appl. Acoust. 113, 64–69.

Schäfer-Zimmermann, J.C., Demartsev, V., Averly, B., Dhanjal-Adams, K., Duteil, M., Gall, G., Faiß, M., Johnson-Ulrich, L., Stowell, D., Manser, M.B., et al., 2024. Animal2vec and MeerKAT: A self-supervised transformer for rare-event raw audio input and a large-scale reference dataset for bioacoustics. arXiv preprint arXiv:2406.01253.

Schwinger, R., Zadeh, P.V., Rauch, L., Kurz, M., Hauschild, T., Lapp, S., Tomforde, S., 2025. Foundation models for bioacoustics–a comparative review. arXiv preprint arXiv:2508.01277.

Tan, M., Le, Q., 2019. Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tsalera, E., Papadakis, A., Samarakou, M., 2021. Comparison of pre-trained CNNs for audio classification using transfer learning. J. Sens. Actuator Networks 10 (4), 72.

Turian, J., Shier, J., Khan, H.R., Raj, B., Schuller, B.W., Steinmetz, C.J., Malloy, C., Tzanetakis, G., Velarde, G., McNally, K., et al., 2022. Hear: Holistic evaluation of audio representations. In: NeurIPS 2021 Competitions and Demonstrations Track. PMLR, pp. 125–145.

Van Merriënboer, B., Hamer, J., Dumoulin, V., Triantafillou, E., Denton, T., 2024. Birds, bats and beyond: Evaluating generalization in bioacoustics models. Front. Bird Sci. 3, 1369756.

Vellinga, W.-P., Planqué, R., 2015. The xeno-canto collection and its relation to sound recognition and classification. In: Conference and Labs of the Evalua- tion Forum (CLEF) (Working Notes). Vol. 1391.

Williams, B., van Merriënboer, B., Dumoulin, V., Hamer, J., Fleishman, A.B., McKown, M., Munger, J., Rice, A.N., Lillis, A., White, C., et al., 2025. Using tropical reef, bird and unrelated sounds for superior transfer learning in marine bioacoustics. Philos. Trans. B 380 (1928), 20240280.

Williams, B., van Merriënboer, B., Dumoulin, V., Hamer, J., Triantafillou, E., Fleishman, A., McKown, M., Munger, J., Rice, A., Lillis, A., et al., 2024. Leveraging tropical reef, bird and unrelated sounds for superior transfer learning in marine bioacoustics. arxiv.

Williams, R., Wright, A.J., Ashe, E., Blight, L.K., Bruintjes, R., Canessa, R., Clark, C.W., Cullis-Suzuki, S., Dakin, D., Erbe, C., et al., 2015. Impacts of anthropogenic noise on marine life: Publication patterns, new discoveries, and future directions in research and management. Ocean & Coastal Management 115, 17–24.

Xu, K., Xu, Q., You, K., Zhu, B., Feng, M., Feng, D., Liu, B., 2023. Self-supervised learning–based underwater acoustical signal classification via mask modeling. J. Acoust. Soc. Am. 154 (1), 5–15.

Yang, S.-w., Chang, H.-J., Huang, Z., Liu, A.T., Lai, C.-I., Wu, H., Shi, J., Chang, X., Tsai, H.-S., Huang, W.-C., et al., 2024. A large-scale evaluation of speech foundation models. IEEE/ACM Trans. Audio, Speech, Lang. Process. 32, 2884–2899.

Yang, S.-w., Chi, P.-H., Chuang, Y.-S., Lai, C.-I.J., Lakhotia, K., Lin, Y.Y., Liu, A.T., Shi, J., Chang, X., Lin, G.-T., et al., 2021. Superb: Speech processing universal performance benchmark. arXiv preprint arXiv:2105.01051.

Zaman, K., Sah, M., Direkoglu, C., Unoki, M., 2023. A survey of audio classification using deep learning. IEEE Access 11, 106620–106649.