Expert Systems With Applications 183 (2021) 115270

Contents lists available at ScienceDirect

Expert Systems With Applications

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

DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification

Muhammad Irfan [a,*], Zheng Jiangbin [a], Shahid Ali [b], Muhammad Iqbal [c], Zafar Masood [a], Umar Hamid [d]

[a] School of Software, Northwestern Polytechnical University, Xian, China
[b] CESAT, Islamabad, Pakistan
[c] Faculty of Computer and Information Science, Higher Colleges of Technology, Fujairah, United Arab Emirates
[d] Comsats University, Islamabad, Pakistan

ARTICLE INFO

Keywords:
Underwater acoustics
Ship classification
Underwater dataset
Deep convolutional network

ABSTRACT

Underwater acoustic classification is a challenging problem because of presence of high background noise and complex sound propagation patterns in the sea environment. Various algorithms proposed in last few years used own privately collected datasets for design and validation. Such data is not publicly available. To conduct research in this field, there is a dire need of publicly available dataset. To bridge this gap, we construct and present an underwater acoustic dataset, named DeepShip, which consists of 47 h and 4 min of real world underwater recordings of 265 different ships belong to four classes. The proposed dataset includes recording from throughout the year with different sea states and noise levels. The presented dataset will not only help to evaluate the performance of existing algorithms but it shall also benefit the research community in future. Using the proposed dataset, we also conducted a comprehensive study of various machine learning and deep learning algorithms on six time–frequency based extracted features. In addition, we propose a novel separable convolution based autoencoder network for better classification accuracy. Experiments results, which are compared based on classification accuracy, precision, recall, f1-score, and analyzed by using paired sampled statistical t-test, show that the proposed network achieves classification accuracy of 77.53% using CQT feature, which is better than as achieved by other methods.

1. Introduction

Underwater acoustic classification attracted lot of attention in recent years because of its application to classification and detection of marine vessels, gauge environmental impact of sound of these vessels, quitter vessel design and marine life classification (Erbe et al., 2019; Malfante, Mars, Dalla Mura, & Gervaise, 2018). Factors, such as complex underwater environment, background noise, frequency dependent absorption and scattering of sound data, make it a challenging field (Erbe et al., 2019). Moreover, improvements in the design of propeller, engine and stealth hull technology make this field more challenging (Khishe & Mosavi, 2020).

Collection of real world underwater acoustic data is a very costly investment with respect to human resources, time, equipment and logistics. Moreover, the quality of recorded signal heavily depends upon factors such as mode of operation, recording equipment, area and underwater environmental conditions (Hovem, 2010). In addition, classified nature of signatures of military ships hinders possibility of publishing of such datasets. In last two decades, various studies applied machine learning as well as deep learning algorithms along with classical signal processing techniques to classify marine vessels through the radiated noise (Miglianti et al., 2020). Despite of lot of research work, accuracy achieved by various techniques remains unsatisfactory due to unavailability of appropriate size real-world dataset.

In order to fill the gap of underwater acoustic dataset and for the development of more accurate underwater acoustic classification techniques, we construct and propose a real world large scale dataset, named DeepShip. DeepShip offers a unique opportunity to train and evaluate the performance of different algorithms, and to identify their strengths and weaknesses. The advantage of this dataset is that it is recorded in the

* Corresponding author.
E-mail addresses: mirfan@mail.nwpu.edu.cn (M. Irfan), zhengjb@nwpu.edu.cn (Z. Jiangbin), miqbal1@hct.ac.ae (M. Iqbal), masoodzafar@mail.nwpu.edu.cn (Z. Masood).

https://doi.org/10.1016/j.eswa.2021.115270
Received 30 January 2021; Received in revised form 15 April 2021; Accepted 21 May 2021
Available online 5 June 2021

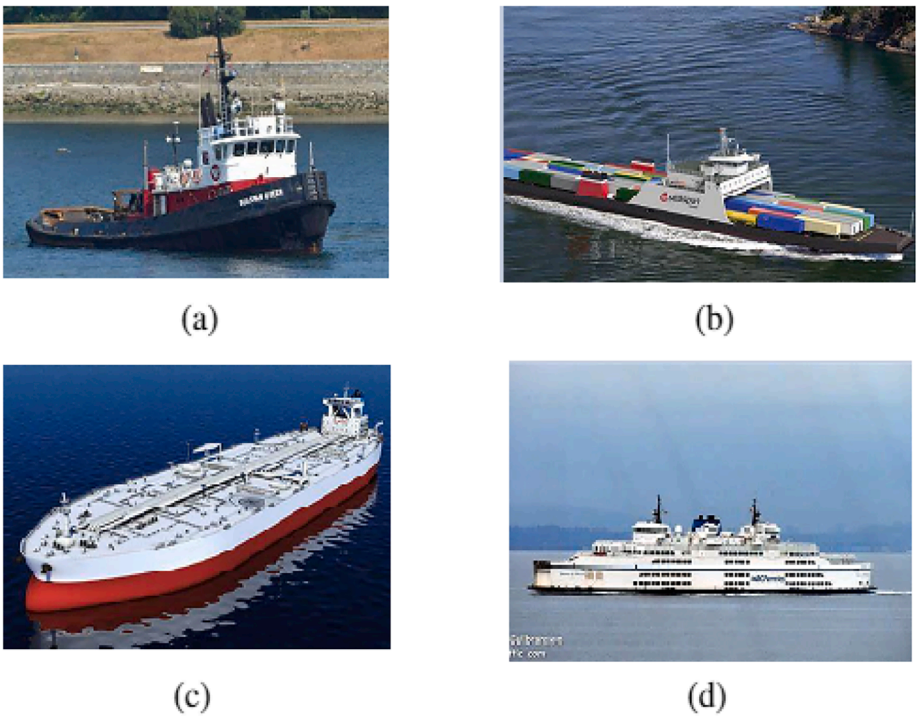

Fig. 1. Four ship classes (a) Tug (b) Cargo (c) Oil Tanker (d) Passenger ship.

real world sea environment in different seasons and sea conditions. Along with vessels signals, recorded signals also have natural background noise, marine mammal noise and noise of any other human initiated activity. This dataset consists of 47 h and 04 min of real world underwater recordings of 265 different ships which belong to four classes. Data for four commercial ship classes is provided. These classes include oil tanker, tug, passenger ship and cargo ship, as shown in Fig. 1. Moreover, the mechanism of collection of data and its labeling is also discussed. We hope that availability of a good benchmark data set will promote and accelerate the research and development in this field.

In addition to presentation of a new real world large size dataset, we also provide baseline evaluation results, to researchers, by conducting extensive experiments with several machine learning and deep learning based algorithms. These experiments offer better insight into performance of such algorithms and also pave the way for future research. We hope, with the availability of DeepShip, which consists of hundreds of recordings, deep learning based algorithms can be trained to improve the classification accuracy performance. For experiments, we extracted and utilized six features such as mel frequency cepstral coefficient (MFCC), mel-spectrogram, wavelet packets, gammatone frequency cepstral coefficients (GFCC), constant Q transform (CQT) and cepstrum.

Methods proposed in the literature, for classification of underwater signals, can be organized into classical machine learning based methods and deep learning based methods. In classical machine learning based studies, feature extraction and classifier are separately designed (Azimi-Sadjadi, Yao, Jamshidi, & Dobeck, 2002; Filho, Seixas, & Moura, 2011; Wang & Zeng, 2014; Wu, Li, & Wang, 2018; Karakos, Silovský, Schwartz, Hartmann, & Makhoul, 2018; Choi, Choo, & Lee, 2019). This has an inherent limitation that designed features may not be used as it is for broad range of classifiers. Moreover, classical machine learning models perform well for datasets of small size and may not achieve promising accuracy for large size datasets with diversified features space. Most of deep learning based methods, as reported in literature, use time–frequency based features for underwater acoustic classification (Yue, Zhang, Wang, Wang, & Lu, 2017; Yang et al., 2018; Luo & Feng, 2020; Cao, Togneri, Zhang, & Yu, 2019; Shen et al., 2020; Zheng, Gong, & Zhang, 2021). It is noted that such methods employ only one feature, and classification performance by utilizing other well-known features is still unknown. Based on the fact, that deep learning based methods achieve state of art by using hand crafted features, in this study, we propose a novel deep learning based method, which utilizes six time–frequency based features for better classification accuracy and offers a better insight into performance of deep learning based systems for broad category of features.

Inspired by the capability of convolution autoencoder (Irfan, Jiangbin, Iqbal, & Arif, 2021; Irfan, Zheng, Iqbal, & Arif, 2020) to extract better features and capability of separable convolution based network (Chollet, 2017; Zhang, Liang, & Ding, 2020) to extract features in an efficient way, we propose a novel separable convolution based autoencoder for training and classification of DeepShip. The proposed method offers an insight for effectiveness of such convolutional blocks for acoustic data classification by utilizing six time-frequecy based features. The proposed model can be used to detect and classify the source of noise in underwater environment, which can be maritime vessel or background noise. It can be used for military purposes as well as for commercial proposes such as maritime traffic management, fishing and protection of marine environment. Moreover, it also motivates the development of deep learning based models for underwater acoustic classification. Experiments results demonstrate that the proposed network performs better than other machine learning as well as deep learning based methods. Experimental results are also evaluated using paired statistical t-test, to have better insight into performance of all compared methods. Moreover, t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten & Hinton, 2008) is used to show 2D feature maps to assess their overlapping.

The main contributions of this paper are summarized as follows:

- We construct and propose a real world large scale dataset, named DeepShip, which is recorded in the real world sea environment in different seasons and sea conditions. This dataset consists of 47 h and 04 min of real world underwater recordings of 265 different ships which belong to four classes. DeepShip offers an opportunity to train and evaluate the performance of different algorithms, and to identify their strengths and weaknesses.

Table 1
Comparison of Recording of ShipsEar Dataset with DeepShip for all Classes. (Duration in Seconds)

		DeepShip			
Class	Cargo	Passenger	Tug	Tanker	
Duration	38,400	44,520	40,620	45,900	
		ShipsEar			
Class	A	B	C	D	Background
Duration	1729	1228	1098	2041	923

- With the constructed DeepShip dataset, we perform comprehensive evaluations and analyses by conducting extensive experiments, with several machine learning and deep learning based algorithms, which offers better insights into their performance and also paves the way for future research.
- We propose a novel separable convolution based autoencoder network for training and classification of DeepShip. The proposed network offers an insight for effectiveness of such convolutional blocks for acoustic data classification by utilizing six time–frequency based features, and exhibits the improvement in the classification accuracy for classification of marine vessels using underwater acoustic signals.

The rest of this document is organized as: In Section 2, we review the existing datasets and existing classification methodologies. In Section 3, we present details of the proposed DeepShip dataset. Section 4 is related to the proposed classification methodology, which includes feature extraction sub-section and proposed method section. Experimental evaluation is done in Section 5, which includes details about experimental setup, compared methods, results and discussion respectively. Conclusion is drawn in Section 6.

2. Literature review

2.1. Existing datasets

Most of the studies reported in the literature, collected and used own private data and did not publish the used data. Such studies include: Bao, Li, Wang, Wang, and Du (2010) used data of recordings of 6 boats for training, McKenna, Ross, Wiggins, and Hildebrand (2012) recorded data of 29 freighters, Roth, Schmidt, Hildebrand, and Wiggins (2013) recorded and analyzed hours of data of icebreaker emitted noise, Das, Kumar, and Bahl (2013) used synthetically generated sound data of 6 boats along with their own recorded data, Jiang, Shi, Huang, and Xiao (2020) used data collected in an inland lake. The data used in above mentioned studies for experiments is not publicly available.

Domínguez, Guijarro, López, and Giménez (2016) proposed a database of underwater audio recordings of ships and boats, named ShipsEar, which is publicly available. This database consists of 90 different recordings of 11 vessel types with total recording duration of 6189 s. Up to our knowledge, this database is the largest publicly available database. Table 1 describes the comparison of ShipsEar and our proposed dataset DeepShip with respect to recording duration of each class and number of recordings. In Table 1 ShipsEar dataset is arranged into four classes as per the guidelines of the author for experiments. Further details of our proposed dataset are given in Section III. It is to mention that number of recordings in our proposed dataset are almost seven times greater than number of recordings in ShipsEar dataset. Moreover, total duration of recorded data of the proposed dataset is almost twenty-five (22) times greater than then biggest publicly available dataset. It can also be noted that duration of recorded data for each class is also far greater than the recording for each class of the previous database with almost same ratio. Keeping in view the duration of data of ShipsEar, it may be useful for acoustic classification by using signal processing and classical machine learning techniques, however, it may not be suitable for acoustic classification with promising accuracy by using deep learning methods, as deep learning methods may suffer from overfitting.

2.2. Previous classification methods

Methods proposed in literature, for classificatin of underwater signals, can be organized into classical machine learning based methods and deep learning based methods. Azimi-Sadjadi et al. (2002), proposed a K-nearest neighbor (KNN) based system memory system to find the similarities in the feature space. Pezeshki, Azimi-Sadjadi, and Scharf (2007) employed canonical correlation analysis (CCA), which is invariant to changes in aspect angle in a fixed bottom condition for multi aspect feature extraction for classfication. Filho et al. (2011) used spectrogram feature as an input to classification model. Wang and Zeng (2014) proposed Bark-wavelet analysis and Hilbert–Huang transform method to extract features and employed SVM as the classifier. Wu et al. (2018) used Wigner-Ville distribution (WVD) based features and support vector machine (SVM) as classifier. Karakos et al. (2018) used probabilistic linear discriminant analysis and i-vectors, as an input to classification model. Choi et al. (2019) used SVM, random forest (RF), convolutional neural network (CNN) and feed-forward neural network (FNN) to classify underwater and surface vessels by using features such as mode-space cross-spectral density matrix (mCSDM) and phone-space cross-spectral density matrix (pCSDM). The study concludes that the FNN performed better than other methods. In conventional machine learning based studies, feature extraction and classifier are separately designed. This has an inherent limitation that designed features may not be used as it is for broad range of classification model. Moreover, classical machine learning models perform well for datasets of small size and may not achieve promising accuracy for large size datasets with diversified features space.

Regarding deep learning methods, Yue et al. (2017) utilized CNN and deep belief network (DBN) for classification of ship radiated noise by using MFCC and low frequency analysis recording (LOFAR). The proposed network is tested for dataset consists of only 16 recordings. Yang et al. (2018) presented a competitive deep belief network (CDBN) through ensemble of DBN and competitive learning based algorithm. Unsupervised pre-training of DBN was conducted on large amount of unlabeled data, then fine tuning of DBN on small amount of own collected labeled data is done, and then verified for only two classes. The competitive learning mechanism is used to enhance deep features discriminating information. Reported results claim that CDBN achieved better accuracy as compared to conventional DBN. Luo and Feng (2020) used normalized frequency spectrum's such as MFCC and GFCC of the signal as input, and used a restricted Boltzmann machine to perform unsupervised automatic encoding of the data, and classified the acquired features through the Boltzmann machine based neural network. Although DBNs, Boltzmann machine based and restricted Boltzmann machine based networks are effective for training with unlabeled data, however, training such networks for supervised learning, is computationally more difficult specially for large size datasets. Cao et al. (2019) proposed a framework by using ensemble of CNN and second-order pooling (SOP) to extract the temporal correlations from the time frequency (T-F) representation CQT of the radiated acoustic signals, for input to CNN. Reported results show improved accuracy results. Some studies as follows, are proposed inspired by human auditory system. Yang, Junhao, Sheng, and Xu (2019) proposed network consisting of multi scale filter inspired by frequency component perception neural mechanism. Max pooling and fully connected layers are stacked at the end of the network and a fusion layer is used to merge features from each decomposed signal. The proposed method is used for classification of three ship types and background noise. Yin, Sun, Liu, Wang, and Tang (2020) employed LOFAR spectrogram and VGG based CNN for classification of underwater signals. Experiments were conducted on own

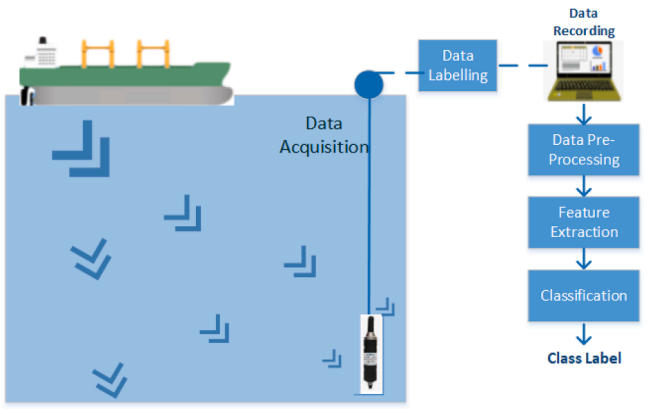

Fig. 2. Overall setup for underwater recordings of vessel audio by icListen AF hydrophone.

Fig. 3. icListen AF Hydrophone.

Table 2
DeepShip Benchmark Dataset Summary.

Ship Type	No. of Ships	Total Time (hr:min)	Total Recordings	Duration per Recording (sec)
Cargo Ship	69	10 h 40 min	110	180–610
Tug	17	11 h 17 min	70	180–1140
Passenger Ship	46	12 h 22 min	193	06–1530
Tanker	133	12 h 45 min	240	06–700

Table 3
Specifications of The icListen Smart Hydrophone.

Characteristic	Value
Bandwidth	1 Hz to 12 kHz
Dynamic Range	120 dB
Sensitivity	−170 dBV re. μ Pa
Sampling frequency	32 kHz
Power	Input 12–24 Vdc, 0.8 W
Output Interface	Ethernet
Internal Storage	32 GB
Case Material and Depth	Engineered Plastic – 200 m Titanium - 3500 m

collected private dataset, and model was used to classify only two classes. Shen et al. (2020) proposed a model which consists of a cochlea network and an auditory center network. In cochlea network, decomposition of audio signal is done by convolutional layer. In the auditory center network, spectro-temporal patterns are extracted by deep network. The proposed model was used for classification of four ship types and ocean background noise. Zheng et al. (2021) utilized short time Fourier transform to capture the time–frequency feature of the audio signals and used it as an input to the deep neural network model GoogleNet. Above mentioned deep learning based methods use time–frequency based features for underwater acoustic classification. However, it can be observed that above discussed methods employee only one feature, and classification performance results by utilizing other well-known features is still unknown. Based on the fact, that deep learning based methods achieve state of art by using hand crafted features, in this study, we propose a deep learning based novel method, which utilizes six time–frequency based features for better classification accuracy and offer a broad insight into performance of deep learning based systems for broad category of features.

3. Proposed deepship dataset

Data recording is done in duration between 02 May 2016 to 04 October 2018 in strait of Georgia delta node. Fig. 2 illustrates the over all setup for recording of vessel audio. As data recording site is situated in one of the busiest shipping routes of the Pacific Northwest coast, and near to the busiest port in Canada, i.e. Vancouver, the background has great influence of river discharge and strong tidal currents with presence of vibrant biological marine environment. The seabed composition in this area is comprised of silt and sand sediment. The central location of strait of Georgia is strongly influenced by semi-diurnal tidal currents, which vary over the spring-neap cycle with speeds ranging between 1 and 3 knots (0.5–1.5 m/s). Further, in the summer and early fall, deep dense gravity currents periodically sweep along the bottom. In this area main species include cetaceans (whales and dolphins) and salmon. The southern area is acutely influenced by human activity such as recreational boating, shipping, and industry affecting the acoustic environment of the ocean. Based on above mentioned reasons it is to mention that the recorded signals may contain background noise both from human and marine life activities.

The data is acquired by using an ocean sonics icListen AF hydrophone, as shown in Fig. 3. Main specifications of the used hydrophone are shown in Table 3. The icListen smart hydrophone is a broadband digital ultra-quit hydrophone, with bandwidth 1 Hz-12 kHz, dynamic range 120 dB, sensitivity −170 dBV re. μ Pa, power & input 12–24 Vdc, 0.8 W, case material is engineered plastic and titanium with depth capacity of 200 m & 3500 m respectively. It is a compact, all-in-one instrument designed and manufactured by replacing separate pre-amplifier, filters, converters and data link units with a compact unit, capable of processing real time acquired data with direct digital output.

The total recording duration is divided into three periods. From 02 May 2016 to 24 June 2017, it was placed at Longitude −123.338713333 and Latitude 49.080926666 at depth of 141 meter (Ocean Networks Canada Society, 2017a). From 24 June 2017 to 03 November 2017, it was placed at Latitude 49.08082191 Longitude −123.33923008 at depth of 147 m below sea level (Ocean Networks Canada Society, 2017b), and from 04 November 2017 to 04 October 2018 it was placed at Latitude 49.080811 Longitude −123.3390596 at depth of 144 m below sea level (Ocean Networks Canada Society, 2017c). It shows that data of almost 29 months comprising of all weathers is collected.

In order to label the recorded data, automatic identification system (AIS) data is used for getting location and timestamp of any particular ship pass by the deployed sensor. AIS data is stored using NMEA (National Marine Electronics Association) format. NMEA is a standard specification for communication between marine electronics such as sonars, gyrocompass, anemometer, echo sounder, GPS receivers, autopilot and various other such devices. We parsed NMEA files for AIS receiver of our interest for complete 29 months. Mainly message number 03 and message number 05 were considered for data extraction, as these

Table 4
DeepShip Recordings Details (Ten sample recordings per class displayed)

ID	class ID	Recording ID	Ship Name	Date & Time	Duration(sec)	Distances(m)
Cargo						
1	70	1	Sakizaya Justice	20171104:023756	357	1883-1285-583
2	70	2	Seaspan swift	20171104:203623	458	1538-378-1981
3	70	3	Istra ace	20171104:223241	441	1535-691-1779
4	70	4	Samos warrior	20171105:084406	297	641-712-1520
5	70	5	NYK Remus	20171106:115234	331	1579-1697-1906
6	76	6	Seaspan reliant	20171107:000452	208	476-685-1241
7	79	7	Princess superior	20171107:075012	487	1600–310-1905
8	70	8	United spirit	20171110:121636	534	1261-692-1905
9	74	10	Ital universo	20171111:201843	300	160-13-1942
10	70	12	Eminent ace	20171111:234213	441	1138–762-1704
Tug						
1	52	1	Seaspan queen	20171104:200959	636	1315-1905-1922
2	52	2	Seaspan raven	20171105:052000	1148	1999-275-1993
3	52	4	Glendale	20171106:144640	1032	1996-195-1987
4	52	6	Jose narvaez	20171107:095309	387	1985-1857-1954
5	52	7	Seaspan commander	20171111:025236	389	1205–1606
6	52	9	Millennium star	20171115:133938	203	803-565-10
7	52	10	Seaspan osprey	20171116:065145	534	1268-630-1998
8	52	11	Ocean betty	20171116:221717	746	1999-958-1670
9	52	16	Sea imp	20171118:234607	889	1982-1250-1748
10	52	20	North arm prowler	20171122:124609	718	1806-724-1520
Passenger Ship						
1	60	1	Queen of new west	20160505,125658	37	1592–1667-1729
2	60	2	Celebrity solstice	20160506,081240	334	1928-717-801
3	60	3	Mayne queen	20160506,233804	556	1988-694-1972
4	60	4	Malaspina	20160507,043512	204	1454-1847-1880
5	60	5	Ruby princess	20160507,115325	14	1089–1186
6	60	6	Coastal inspiration	20160507,202625	12	1430-1493-1665
7	60	7	Coral prince	20160514,113853	350	1974–166
8	60	8	Kennicott	20160515,062337	12	1998-1939
9	60	9	Carnival legend	20160516,112116	11	1111-1169
10	60	10	Crystal serenity	20160522,182048	481	1994-696-1212
Tanker						
1	89	1	Kirkeholmen	20160509:123109	278	1488–417-465
2	80	2	Cherry galaxy	20160511:030034	150	949-952-1560
3	89	3	Champion ebony	20160515:020119	383	1174-1282-1363
4	80	4	Eser K	20160516:134837	277	157–1821
5	80	5	Lynda victory	20160519:123958	47	1251-1111
6	80	7	Chembulk new orleans	20160531:142749	503	1993-1698-1998
7	80	9	Songa ruby	20160602:203639	18	1849-1930-1980
8	80	10	Nave orbit	20160602:225434	30	1956-1819-1750
9	80	11	Champion cornelia	20160604:171238	371	1290-901
10	80	13	High endurance	20160613:061037	157	987-109-95

messages provide all required information of ship for dataset preparation. Message number 03 provides dynamic information of ship, such as longitude, latitude, navigational status, true heading, timestamp, speed and maritime mobile service identity (MMSI) etc. Message number 05 provides static information about any ship such as MMSI, name, type, max draught, length, breadth and dimension etc. As per AIS standards, ship type IDs from 70 to 79 represent cargo ship, IDs from 60 to 69 represent passenger ship, IDs from 80 to 89 represent tanker, and ID 52 represents tugs. Information from message number 03 and message number 05 is combined together to get complete dynamic profile of a ship.

For our dataset, we consider only signals emitted by a ship when only a single vessel appears within a radius of 2 km of the hydrophone. Whenever a ship get out of range of 2 km from the hydrophone data is stopped. For this purpose, information extracted from message number 03 and message number 05 is used. We used wav file format and all of our data files are available in wav file format, as this format is supported by most of platforms/ libraries used by machine learning/ deep learning community. Data is recorded and saved at sampling rate of 32 kHz.

Table 2 provides overall summary of the proposed dataset, which includes ship type, number of ships, number of recordings, duration of each recording (seconds), and total recording time for each type. Data for four commercial ship classes is provided. These classes include oil tanker, tug, passenger ship and cargo ship. It can be observed that total recording duration for four ship classes is around 47 h and 04 min. Recordings of different 265 ships of four categories are included and there are total 613 number of recordings. Duration of the each recording varies from about 06 s to 1530 s, depending on position of ship with respect to sensor and navigational speed. In Table 4 we provide ten sample records for each class to offer better insight about the dataset. In order to balance the duration of recording for each class, we have tried to keep recording duration of classes almost equal. Because of excessive cargo activity in this area, there are excessive recordings for the cargo and the tug classes, only those recordings of these two classes are included in the dataset, which consists of at least three 03 min of continuous recording. For passenger ships and tankers, all recordings which consists of at least six 06 s of continuous recording are included, as there is relatively less activity related to these ship classes. For all classes, one event of continuous detection of any vessel is kept as a single wave file.

This dataset is available for download at https://github.com/irfan-kamboh/DeepShip repository. (Keeping in view the limitation of file size

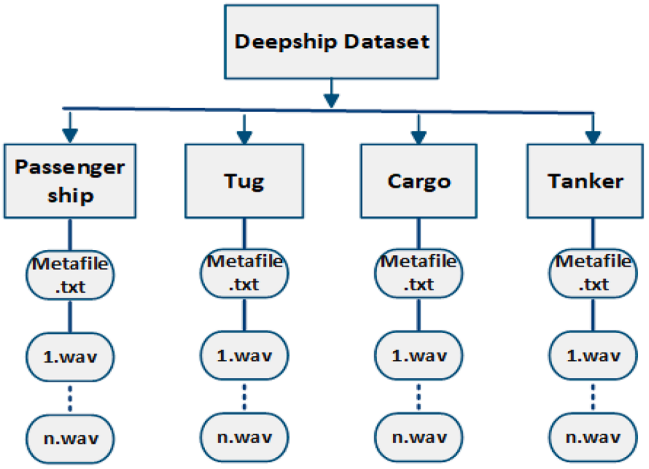

Fig. 4. DeepShip Dataset Structure.

and space on github directory, only a part of dataset is uploaded on this repository. The complete dataset can be downloaded by following the information provided on the same repository web page). The dataset is arrange in four folders named after every ship class as shown in Fig. 4. In each folder there are two types of files, first are.wav files which contain audio data, second type of file is classname-metafile such as cargo-metafile, tug-metafile, tanker-metafile and passenger-metafile. This metafile contains information about.wav files in the folder. Each meta file contains information such as (i) class id, (ii) recording id, (iii) ship name, (iii) date and time of recording, (iv) duration of each recording in seconds, (v) distances of ship from the sensor. Table 4 shows ten sample records for each class. Similar data arrangement of records in form of columns is done in meta files. It is to mention that we provide three different distances, at three different times, of ship from sensor for each recording to offer better understanding of the trajectory followed by the ship. Middle distance is recorded almost at the middle of recorded sortie.

4. Classification methodology

4.1. Feature extraction

In this study we extracted six types of features, which include cepstrum, mel spectrogram, mfcc, cqt, gfcc and wavelet packets. These features are extracted based on the explanation given below and these features data in time–frequency domain is stored in form of images (Xie & Zhu, 2019). These image are given as input to the proposed model as well as machine learning and deep learning models used for comparison in this study.

4.1.1. Cepstrum

Cepstrum is computed by taking the log magnit,ude of the spectrum followed by inverse Fourier transform as:

$$C = |F^{-1}(log(|F(x(t))|^2))|^2 \tag{1}$$

Here, F represents the Fourier transform and F^{-1} is inverse Fourier transform. The resulted signal is in quefrency. In the cepstrum, the low quefrencies represent data of the log spectrum features.

4.1.2. Mel spectrogram

A mel spectrogram is a non-linear transformation of the frequency scale, where frequencies in the spectrum are converted to the mel scale. In order to compute it, hamming window of length N is applied to the signal $x(n)$. Then FFT is taken of the signal. After that power spectrum is computed and then mel filterbank are applied as:

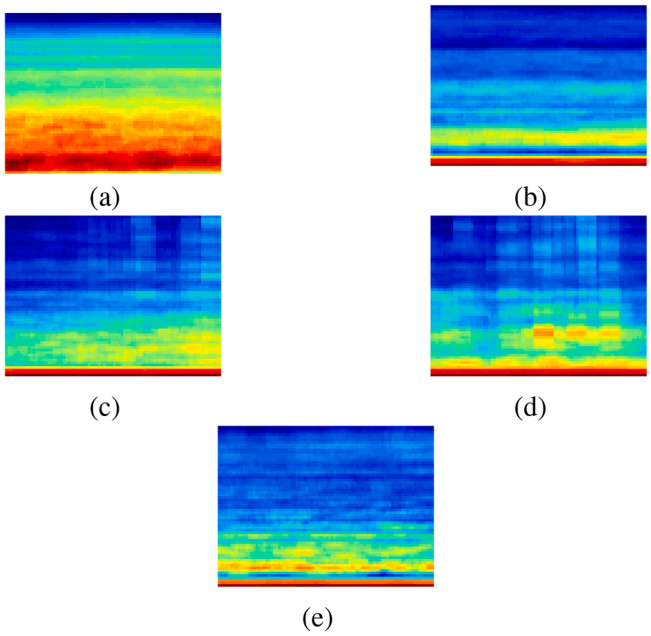

Fig. 5. GFCC display of (a) Background Noise; (b) Cargo; (c) Passenger ship; (d) Tanker; (e) Tug.

$$M_m = \sum_{k=0}^{N-1} \omega_m(k)|X(k)|^2 0 \leqslant m \leqslant M-1 \tag{2}$$

$\omega_m(k)$ corresponds to the *mth* window function and $X(k)$ is the FFT of input signal $x(t)$. After that, frequencies scale in output is converted to the mel scale.

4.1.3. MFCC

MFCC offers better representation of sound as frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely (Zhang, Wu, Han, & Zhu, 2016). MFCC is computed by applying the log function to the magnitude of signal spectrum M_m. Then MFCC for the *kth* frame is calculated by applying DCT on the logarithm of M_m.

$$C_m = \sum_{m=0}^{M-1} log_{10}(M_m) cos(\pi n(m-0.5)) \Big/ M) 0 \leqslant n \leqslant C-1 \tag{3}$$

Here, C is number of MFCCs.

4.1.4. CQT

Underwater radiated signals carry very useful signal information in low frequency subbands, which are associated to the propeller's signature. The CQT is a T-F based representation which offers better frequency resolution for low frequencies sub bands as compared to STFT (Schörkhuber, 2010). It is computed as:

$$X(m,n) = \sum_{i=n-\lfloor N_m/2 \rfloor}^{n+\lfloor N_m/2 \rfloor} x(i)(a_m^*)((i-n+N_m \Big/ 2)) \tag{4}$$

Here, $m = 1,2,\ldots,M$ denotes number of frequency bins of CQT, $\lfloor . \rfloor$ is rounding towards negative infinity, N_m is the window length, and a_m^* is complex conjugate of basis function.

4.1.5. GFCC

GFCC is computed by applying Gammatone filter bank to the spectrum, followed by loudness compression and then DCT is applied for dimensionality reduction and its components de-correlation (Xu, Lin, Sun, & Jin, 2012). The impulse response of gammatone filter is defined

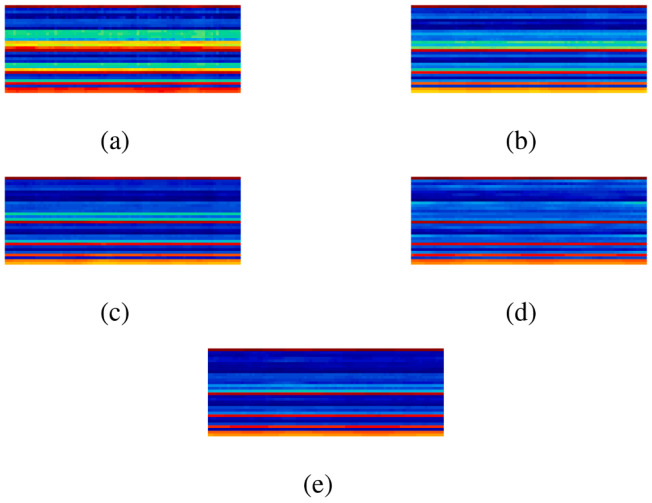

Fig. 6. Wavelets Display of (a) Background Noise; (b) Cargo; (c) Passenger ship; (d) Tanker; (e) Tug.

as:

$$g_i(t) = t^{n-1} exp(-2\pi b_i t) cos(2\pi f_i + \alpha_i) U(t) \tag{5}$$

Where, N represents the number of filters, n denotes filter order, i is the order number of filter arranged by frequency from 1 to N, α_i denotes the phase of *ith* ordinal filter, $U(t)$ denotes the unit step function. Fig. 5 shows sample images formed by GFCC of four classes.

4.1.6. Wavelet packets

It is a generalization of wavelet decomposition. In this, wave forms are indexed by three naturally interpreted parameters like frequency, position and scale. Wavelet packets are superior at time–frequency analysis as compared to Discrete Wavelet Transform (DWT), as it offers the energy-preserving property of the DWT with superior frequency resolution and better computational efficiency. For a given function $f(t)$ the wavelet transform coefficients (WTCs) at *xth* point and *jth* level are calculated by recursion relations as:

$$W_j^{2i}(x) = \int f(t)\psi_{j,k}^{2i}(t)dt = \sqrt{2}\sum_{k=0}^{2N-1} h(n) W_{j-1}^{i}(2x-k) \tag{6}$$

$$W_j^{2i+1}(x) = \int f(t)\psi_{j,k}^{2i+1}(t)dt = \sqrt{2}\sum_{k=0}^{2N-1} g(n) W_{j-1}^{i}(2x-k) \tag{7}$$

Where, $W_0(x)=\phi(x)$ and $W_1(x)=\psi(x)$ are scaling and wavelet functions, $\psi_{j,k}^{2i}$ and $\psi_{j,k}^{2i+1}$ are wavelet basis, i is band number, j defines the decomposition level, $g(n)$ and $h(n)$ are quadrature mirror filters of length $2N$ correspond to wavelet. Fig. 6 shows sample images formed by wavelet packets of four classes.

4.2. The proposed method

The proposed novel separable convolution autoencoder (SCAE) based network is comprised of the encoder network and the decoder network. Inspired by Xception net, in this study, we utilized its modified version as a back bone to the encoder module. The main idea of the proposed separable convolution based encoder network is based on the idea that the mapping of cross channels correlations and spatial correlations in the feature maps are decoupled. Usage of this strategy helps to make learning of the network more efficient and also enhances performance in terms of classification accuracy. Separable convolution is based on the idea of depth wise separable convolutions, which is a spatial convolution performed independently over each channel of an input, followed by a point wise convolution, which is a 1 × 1 convolution, mapping the channels output by the depth wise convolution to the new channel space. In addition to depth wise separable convolution, shortcut connections between convolution blocks similar to ResNet are also utilized.

Table 5
Xception Block.

Layer Type	Configuration
SeperableConv Block	
SeparableConv2D + BatchNorm ReLU	Filters: N × (3 × 3)
Xception Block	
SeperableConv Block SeparableConv2D + BatchNorm MaxPooling (layer3) Convolution + BatchNorm (layer4) ADD	Filters: N Filters: N × (3 × 3) Pool size: 3 × 3 Filters: N × (1 × 1) outputof(layer3,layer4)
Middle Block	
ReLU 3 × SeperableConv Block (layer2) ADD	 Filters:256 [input, output of layer2]

Table 6
The Seperable Convolution Based Autoencoder Architecture.

Layer Type	Configuration
Encoder	
Convolution + BatchNorm	Filters: 64 × (3 × 3), ReLU
Convolution + BatchNorm	Filters: 64 × (3 × 3), ReLU
Xception Block	Filter: 128
Xception Block	Filters: 256
4 × Middle Block ReLU	
Xception Block	Filters: 256
SeperableConv Block	Filters: 256
SeperableConv Block globalAvgPooling	Filters: 256
Fully connected	10 units; Softmax classifier
Decoder	
Upsampling	Factor: 4 × 4
Convolution + BatchNorm	Filters: 128 × (3 × 3), ReLU
Upsampling	Factor: 3 × 4
Convolution + BatchNorm	Filters: 64 × (3 × 3), ReLU
Upsampling	Factor: 2 × 2
Convolution + BatchNorm	Filters: 32 × (3 × 3), ReLU
Upsampling	Factor: 2 × 2
Convolution	Filters: 1 × (3 × 3), ReLU

Table 5 shows separableConv block, Xception block and middle block, which are main building blocks of the encoder network. In Table 5 names of given blocks are defined by ourselves for better elaboration of these blocks used in the encoder network. Separableconv block is an ensemble of separableconv2D layer and ReLU. Xception block is consist of five layers. This block uses the seperable conv block for its construction. Middle block is consist of ReLU and 3 × separable conv blocks. Several separable conv blocks, Xception blocks and middle blocks are stacked together to form a deeper neural network, which help to learn variant feature representation.

Table 6 shows the proposed separable convolution block based autoencoder architecture. For each block and layer number of filters and size of each convolution filter is also mentioned. The encoder network consists of three xception blocks, ensemble with one middle block and two separable conv blocks. Two convolution layers are stacked in the start. globalAvgPooling layer is stacked after the network. Then one fully connected layer is stacked at the end of the network. Softmax classifier is used in the last layer for classification. Output of the encoder module is taken as an input to the decoder module. The decoder network

is consist of four upsampling layers and three convolution layers. Deconvolution operation is performed by combining the upsampling and convolution layer to regenerate the input.

Input image is convolved with the learned convolution filters of the encoder module (Irfan, Jiangbin, Iqbal, & Arif, 2021). Non-trivial features from input images are extracted. Let z represents the input image, the feature extraction process by the encoder can be defined as

$$e_i = \sigma(z_i * w^i + b) \tag{8}$$

Where e_i represents the extracted features of the input z_i in a compressed space, b is bias, * represents convolution in 2D space, w^i is 2D filter for convolution and σ represents non linear function for activation i.e. ReLU in this case.

Let L represent the total loss of the proposed network. In order to achieve maximum classification accuracy, L is adjusted to be minimum in each training iteration. L is calculated by weighted sum of L_1 and L_2 losses weighted by the loss weights coefficients c_1 and c_2 respectively as:

$$L = c_1 L_1 + c_2 L_2 \tag{9}$$

As in this study our main concern is classification accuracy, weights coefficients values are set as c_1=0.5 and c_2=1.0. L_1 represents mean square error (MSE) loss, which is used to generate image D_{ki} at the last convolution layer of the decoder as close as possible to the encoder input as:

$$L_1 = 1 \Big/ n \sum_{i=1}^{m} (D_{ki} - I_{ki})^2 \tag{10}$$

L_2 represents categorical cross entropy loss. The feature maps generated by convolutional blocks, are flattened as one dimensional in fully connected layer. The softmax is used as classifier, which uses input from fully connected layers. The weights of the network are updated by back propagation on the base of adaptive delta. It is used to reduce categorical cross entropy loss L_2, which is used for classification as:

$$L_2 = -\sum_{j=1}^{N} \sum_{k=1}^{K} t_{j_k} log(y_{i_k}) \tag{11}$$

Where, N represents number of observations, t_{j_k} is the target, y_{j_k} is computed output probability, K represents number of classes, the input image may belong to.

5. Experimental evaluation

5.1. Experimental setup

As mentioned in the previous section, we extracted and utilized six feature types for this study. Classification experiments are performed on the four ship categories and background noise. As the site of deployment of the hydrophone i.e. strait of Georgia delta is very busy, so for background noise we downloaded data from internet from various websites for our experiments. This downloaded background noise data is of almost 7 h and 27 min. Its divided into 183 recordings, where duration of each recording is from 03 s to 300 s.

For the experiments, data is re-sampled at sampling frequency of 16 kHz. Each of audio file are sliced into multiple segments to process for the input of machine/ deep learning algorithms. The specific length of the segment is decided based on the nature of acquired signal and input dimensions of algorithms being used in the study. Keeping in view, properties of recorded signals, computational resources and classification accuracy, we divided each recording in segments of 3 s. Each segment is treated by the system for classified independently. As each audio file is divided in many segments and each segment corresponds to a single image, images generated by a single file are kept either in training or testing dataset to avoid any bias. Each segment signal is divided/ windowed into short frames of 250 ms hopped 64 ms before extracting features. Based on specified frame and hopped size, we get 43 frames for each segment signal. Above mentioned features are computed from these 43 frames and features saved as time–frequency in image forms with dimensions as specified in Table 7. In addition, resized dimensions used for input to deep CNN methods are also mentioned in the Table 7.

Table 7
Features and Dimensions.

Feature	Dimension	Resized (for DL methods input)
Mel-Spectrogram	40 × 43	48 × 48
MFCC	13 × 43	16 × 48
Cepstrum	256 × 43	256 × 48
Gammatone	64 × 43	64 × 48
Wavelets	32 × 43	32 × 48
CQT	64 × 43	64 × 48

For all experiments, all algorithms are trained and tested on raw data without any pre-processing and without any data augmentation technique, with 70% segments of data randomly allocated for training and remaining 30% of segments for testing. The performance of all algorithms against all features used in this study is evaluated by parameters such as accuracy, precision, recall and F1-score.

Accuracy is computed as by following expression:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

Where, *TP* is true positive, *TN* is true negative, *FP* is false positive, and *FN* is false negative.

Precision is computed as:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

Recall is computed as:

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

F1-Score is computed as:

$$F1 - Score = 2 * \frac{precision * recall}{precision + recall} \tag{15}$$

5.2. Compared methods

In this study, we selected four machine learning based models and three deep learning based model, based on their popularity and frequent usage in underwater acoustic classification problem, and compared results of our proposed method with them.

5.2.1. Machine learning based methods

These models include SVM, KNN, Naive Bayes and RF (Strain & Olszewska, 2020; Carbonera & Olszewska, 2019). Various variants of these models have been reported in literature for classification of underwater acoustic data.

5.2.2. Deep learning based methods

We compared classification results with variants of a deep neural network (DNN), ResNet, VGG and Inception networks. Description of these methods is given in coming sub sections.

DNN: Different variant of DNN are being reported and used for underwater acoustic classification tasks. For the comparison we used DNN with number of nodes: 2752–1024-256–64-5 for gammatone and CQT, 11008–3072-768–192-5 for cepstrum, 1720-768–192-48-5 for mel-spectrogram, 602-256–64-16-5 for MFCC. ReLU is used as activation function and adam is used as an optimizer.

CNN: A sequential CNN architecture is designed by a stacking convolutional layers, pooling layers and fully connected layers such that the

Table 8
Accuracy Comparison (%) for Mel-spectrogram.

Method	Accuracy	Precision	Recall	F1-Score
SCAE	70.18	70	70	69
CNN	69.76	70	70	69
Residual	69.29	69	69	69
Inception	68.11	68	68	67
DNN	63.87	65	64	65
RF	63.84	64	64	64
SVM	64.83	65	65	65
KNN	58.11	59	58	58
Naive Bayes	48.00	48	48	46

Table 9
Accuracy Comparison (%) for CQT.

Method	Accuracy	Precision	Recall	F1-Score
SCAE	77.53	78	77	77
Residual	76.98	77	77	77
CNN	76.35	76	76	76
Inception	76.16	76	76	76
DNN	73.11	73	73	73
SVM	72.24	72	72	72
RF	69.71	70	70	69
KNN	62.71	64	63	63
Naive Bayes	53.97	57	53	52

output of one layer is the input of the next layer. we utilized modified variant of Simonyan and Zisserman (2015) with different layer configuration for comparison. we used five convolution layer with filter size varying from 62 to 512 filters in the last convolution layer, used max-pooling layer and used 3 fully connected layers with nodes 512, 128 and 5.

Residual: ResNet (He, Zhang, Ren, & Sun, 2016) introduced the idea of residual blocks and the skip connection, where output x_{l-1} of the previous layer is added with the output $g(x_{l-1}, k_l)$ of the subsequent next layer, to form input to the current *lth* layer. The architecture used for comparison consists of three residual blocks. Each residual block is comprised of two convolutional layers. Each layer has 64 filter of 3×3. Each residual block is followed by a convolution layer plus batch normalization layer and a max pooling layer.

Inception: Inception net (Szegedy et al., 2015) introduced the idea of inception block, which consists of multi-scale convolutional transformations using split, transform and merge. It aimed to learn diverse types of feature variations present in the input images with the help of convolution filters of various sizes such as 1×1, 3×3, and 5×5. For comparison we used network architecture consisting of three inception blocks. Each inception block is comprised of three types of convolution. Each inception block is followed by a convolution layer plus batch normalization layer and then max pooling layer.

5.3. Results and discussion

Accuracy results for all methods for mel-spectrogram are elaborated in Table 8. Results show that the proposed SCAE network achieved accuracy of 70.18% and outperformed all other methods. It achieved scores of 70 %, 70% and 69% for precision, recall and f1-score respectively. CNN based sequential network remained on second number and performed better than other deep CNN methods and achieved accuracy of 69.76%, 70%, 70% and 69% for classification accuracy, precision, recall and f1-score respectively. Among machine learning methods SVM performed better than other methods and achieved scores of 64.83%, 65%, 65% and 65% for classification accuracy, precision, recall and f1-score. Among other machine learning methods random forest remained on second number with score of 63.84%. Results show that the proposed method performed better than other deep learning and machine learning based methods.

Table 10
Accuracy Comparison (%) for MFCC.

Method	Accuracy	Precision	Recall	F1-Score
Residual	59.58	59	58	58
SCAE	58.04	58	57	57
Inception	57.91	56	56	56
CNN	57.35	56	55	55
DNN	52.76	54	53	53
RF	56.96	57	56	56
SVM	58.33	59	58	58
KNN	54.96	56	55	55
Naive Bayes	46.12	45	46	45

Table 11
Accuracy Comparison (%) for Wavelets.

Method	Accuracy	Precision	Recall	F1-Score
Inception	59.85	60	60	60
CNN	58.48	59	58	58
Residual	58.68	59	58	58
SCAE	57.57	58	57	57
DNN	51.33	51	51	51
SVM	55.47	56	55	56
RF	53.73	54	54	54
KNN	50.04	50	50	50
Naive Bayes	45.09	46	45	43

Results for classification accuracy for all methods for classification of CQT feature are depicted in Table 9. The proposed SCAE method achieved best accuracy score of 77.53%, and outperformed all other methods. It achieved scores of 78%, 77% and 77% for precision, recall and f1-score. Residual based network remained at the second position by achieving scores of 76.98%, 77%, 77% and 77% for classification accuracy, precision, recall and f1-score. Whereas, CNN remained on third with score of 76.35% and Inception remained on fourth position with classification accuracy score of 76.16% respectively. Among machine learning methods SVM remained on top and it outperformed other machine learning based methods and achieved accuracy score of 72.24%. Among other machine learning based methods RF remained on the second position by achieving classification accuracy score of 69.71%. It is observed that proposed SCAE method outperformed all other methods.

Table 10 shows the results of classification accuracy of all methods for MFCC. Results show residual outperformed all other methods with score of 59.58 %, 59 %, 58 % and 58 % for classification accuracy, precision, recall and f1-score respectively. SCAE performed better than other deep CNN methods by achieving classification accuracy score of 58.04%. Among machine learning approaches, SVM outperformed all other machine learning methods and achieved score of 58.33 %, 59 %, 58 % and 58 % in classification accuracy, precision, recall and f-score. It is observed that SVM remained on second number among all methods, and it performed slightly better than three deep CNN based methods.

Table 12
Accuracy Comparison (%) for Cepstrum.

Method	Accuracy	Precision	Recall	F1-Score
SCAE	73.10	73	73	72
Inception	72.27	72	70	70
CNN	71.46	72	71	71
Residual	70.80	71	71	71
DNN	67.97	68	68	68
SVM	71.74	72	72	72
RF	63.88	65	64	64
Naive Bayes	53.50	52	54	52
KNN	49.75	51	50	50

Table 13
Accuracy Comparison (%) for Gammatone.

Method	Accuracy	Precision	Recall	F1-Score
SCAE	73.59	74	73	73
CNN	73.49	73	73	73
Residual	71.67	72	72	72
Inception	70.85	70	69	69
DNN	67.61	68	66	66
SVM	68.86	70	69	69
RF	68.55	70	69	69
KNN	61.93	63	62	62
Naive Bayes	55.09	57	55	54

Table 11 shows the results of classification accuracy of all methods for Wavelet feature. Results show that inception outperformed all other methods with score of 59.85 %, 60 %, 60 % and 60 % for classification accuracy, precision, recall and f1-score respectively. Residual performed better than other deep CNN methods by achieving classification accuracy score of 58.68 %. Among machine learning approaches, SVM outperformed all other machine learning methods and achieved score of 55.47 %, 56 %, 55 % and 56 % in classification accuracy, precision, recall and f-score. Among other machine learning based methods RF remained on second position by achieving score of 53.73%, whereas, Naive Bayes method remained at bottom and achieve classification accuracy score of 45.09%.

Classification accuracy of all methods for Cepstrum are elaborated in Table 12. Results show that the proposed SCAE network achieved accuracy of 73.10% and outperformed all other methods. It achieved scores of 73 %, 73% and 72% for precision, recall and f1-score respectively. Inception based network remained on second number and performed better than other deep CNN methods and achieved accuracy of 72.27%, 72%, 70% and 70% for classification accuracy, precision, recall and f1-score respectively. Among machine learning methods SVM performed better than other machine learning based methods and achieved scores of 71.74%, 72%, 72% and 72% for classification accuracy, precision, recall and f1-score.

Results for accuracy for all methods for classification of gammatone feature are depicted in Table 13. SCAE remained at the top position by achieving scores of 73.59%, 74%, 73% and 73% for classification accuracy, precision, recall and f1-score. The CNN method achieved best accuracy score of 73.49%, and outperformed all other methods. It achieved scores of 73%, 73% and 73% for precision, recall and f1-score. Whereas, Residual and inception remained on third and fourth position respectively. Among machine learning methods SVM remained on top and it outperformed other machine learning based methods and achieved classification accuracy score of 68.86%. Among other machine learning based methods RF remained on the second position by achieving classification accuracy score of 68.55%.

From overall results it can be observed that overall accuracy results remained better for CQT feature as compared to other five features, with score 77.53%, 78%, 77%, and 77% for accuracy, precision, recall and f1-score. Whereas, results for classification accuracy for gammatone and cepstrum remained at second and third position. It can be inferred that CQT features proved to be more effective than other features and proved to be better option to be used for input to deep learning methods for classification.

The paired t-test is performed to show the statistical significance of the classification results achieved by methods used in the study. Table 14 demonstrates the results of t-test for the proposed SCAE against deep learning methods and other machine learning methods for GFCC, CQT, MFCC, wavelet, cepstrum and mel-spectrogram features. It can be observed that for gammatone, CQT, cepstrum and mel-spectrogram propsed SCAE method exhibit statistically significant as compared to other methods, even for under the significance level of $p < 0.001$. For MFCC, SVM and inception performed significantly better than SCAE. Whereas, the difference with residual is not significant. For wavelet, CNN and inception performed significantly better than SCAE.

In order to calculate the total computation time for a single epoch of the network, the time required for a single batch for forward and backward pass is calculated as:

$$t_b = \sum_{i=0}^{n} b(L_i) \tag{16}$$

Where, n is the number of layers in the network, bL_i is the batch execution time of layer i and L_i is of the type of layer i. The total execution time for the network is computed as:

$$t = omt_b \tag{17}$$

Where, m is number of batches required to process the data and o is number of epochs required to train the network.

Table 15 presents the computational efficiency with respect to number of floating point operations (FLOPs) and computation time taken by the model per epoch, of all deep methods used in this study. Results in Table 15 are taken by using system with i7-7700HQ CPU @2.8 GHz, NVIDIA GeForce GTX 1050 Ti. It can be observed that the number of FLOPs of the proposed method SCAE are far less than CNN, Residual and Inception based models. However, number of FLOPs of the proposed model are greater than DNN, but accuracy of the proposed model is far greater than DNN. In case of computation time, time taken by the proposed method is less than Residual and Inception based models, however, it is greater than computation time taken by DNN and CNN.

Table 15
Computational Efficiency.

Method	FLOPs (million)	Computation Time (sec/ epoch)
SCAE	6163	65
CNN	20317	44
Residual	52110	104
Inception	84396	143
DNN	345	15

Table 14
Results of paired t-test for classification accuracy (* means $p < 0.05$, ** means $p < 0.001$).

Method1	Method2	Gammatone (t-value)	CQT (t-value)	MFCC (t-value)	Wavelet (t-value)	Cepstrum (t-value)	Mel-spectrogram (t-value)
SCAE	KNN	59.80**	87.38**	15.36**	38.22**	130.95**	65.88**
	SVM	23.65**	27.73**	−1.48**	10.29**	3.39**	27.33**
	RF	25.26**	43.64**	5.45*	19.96**	52.35**	34.57**
	Naive Bayes	93.46**	134.85**	60.39**	63.35**	100.74**	113.17**
	DNN	30.05**	24.54**	27.77**	11.25**	26.07**	36.81**
	Inception	14.25**	77.78**	−7.72**	−11.96**	8.77**	11.47**
	Residual	9.6**	3.46*	0.65	−5.43**	4.88**	4.72**
	CNN	0.41	6.16**	3.51*	−4.63**	12.57**	2.33*

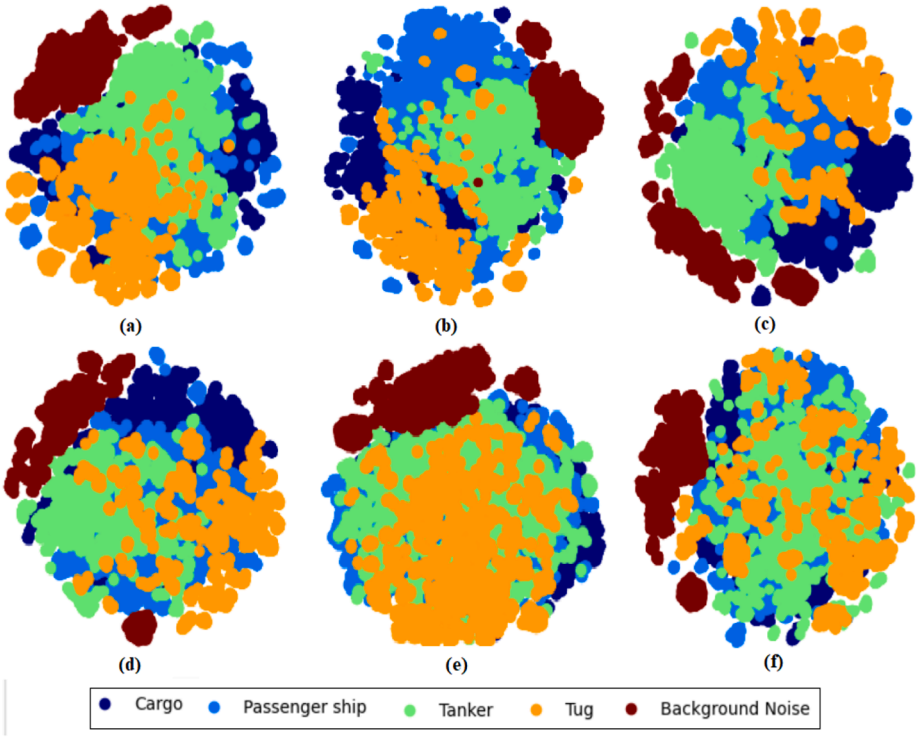

Fig. 7. Feature vectors 2D display (a) Cepstrum (b) CQT (c) Gammatone (d) Melspectrogram (e) MFCC (f) Wavelet.

5.4. Features visualization

The data visualization method t-SNE is used to visually analyze features extracted from avg. pooling layer of SCAE in a two dimensional feature space as shown in Fig. 7. Testing data is used to construct it. There are five classes represented by different colors. Six features used in this study, are used to display 2D maps represented by (a) to (f) respectively. It can be observed that samples from different classes are more overlapped generally. It can also be observed that features from background noise and cargo are relatively more discriminating and lying in separate areas as compared to features from other classes. It is also be inferred from the results that though the features are not very discriminating as shown in 7, our proposed deep convolutional networks attempt to classify these classes with better classification accuracy.

6. Conclusion

In this paper, we have constructed an underwater acoustic benchmark dataset, which offer large scale real world underwater audios of different vessel classes with different levels of background noise. The data is recorded throughout the year. It offers unmatched opportunities to researcher community working in this area for training and evaluation of algorithms. Up to our knowledge, its the biggest publicly available underwater acoustic dataset. A number of famous machine learning and deep learning algorithms have also been evaluated on the proposed dataset by using six time–frequency domain extracted features. In addition, we proposed a separable convolution based autoencoder network for better classification. In the future work, we want to include signals from more vessel classes and include background noise recording as well. Moreover, we want to propose more deep learning based models with lifelong learning capability, which can learn features automatically, retain knowledge and utilize learnt knowledge for multiple tasks, for better accuracy.

CRediT authorship contribution statement

Muhammad Irfan: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Zheng Jiangbin:** Supervision, Conceptualization, Writing - review & editing. **Shahid Ali:** Supervision, Writing - review & editing. **Muhammad Iqbal:** Supervision, Writing - review & editing. **Zafar Masood:** Writing - review & editing, Methodology. **Umar Hamid:** Writing - review & editing, Methodology.

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.

Acknowledgement

Data used in this work is collected by using infrastructure of Ocean Networks Canada. The authors would like to gratefully acknowledge the support from Ocean Networks Canada.

References

Azimi-Sadjadi, M. R., Yao, D., Jamshidi, A. A., & Dobeck, G. J. (2002). Underwater target classification in changing environments using an adaptive feature mapping. *IEEE Transactions on Neural Networks and Learning Systems, 13*(5), 1099–1111.

Bao, F., Li, C., Wang, X., Wang, Q., & Du, S. (2010). Ship classification using nonlinear features of radiated sound: An approach based on empirical mode decomposition. *The Journal of the Acoustical Society of America, 128*(1), 206–214. https://doi.org/10.1121/1.3436543

Cao, X., Togneri, R., Zhang, X., & Yu, Y. (2019, 4 15). Convolutional neural network with second-order pooling for underwater target classification. IEEE Sensors Journal, 19 (8), 3058–3066. doi: 10.1109/JSEN.2018.2886368.

Carbonera, J.L., & Olszewska, J.I. (2019). Local-set based-on instance selection approach for autonomous object modelling. International Journal of Advanced Computer Science and Applications, 10(12). Retrieved from https://doi.org/10.14569/IJACSA.2019.0101201 doi: 10.14569/IJACSA.2019.0101201.

Choi, J., Choo, Y., & Lee, K. (2019, Aug). Acoustic classification of surface and underwater vessels in the ocean using supervised machine learning. Sensors, 19(16), 3492. Retrieved from https://doi.org/ 10.3390/s19163492 doi: 10.3390/s19163492.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *2017 ieee conference on computer vision and pattern recognition (cvpr)* (pp. 1800–1807). https://doi.org/10.1109/CVPR.2017.195

Das, A., Kumar, A., & Bahl, R. (2013). Marine vessel classification based on passive sonar data: the cepstrum-based approach. *IET Radar, Sonar Navigation, 7*(1), 87–93.

Domínguez, D. S., Guijarro, S. T., López, A. C., & Giménez, A. P. (2016). Shipsear: An underwater vessel noise database. *Applied Acoustics, 113*, 64–69.

Erbe, C., Marley, S.A., Schoeman, R.P., Smith, J.N., Trigg, L.E., & Embling, C.B. (2019). The effects of ship noise on marine mammals–a review. Frontiers in Marine Science, 6, 606. Retrieved from https://www.frontiersin.org/article/10.3389/fmars.2019.00606 doi: 10.3389/fmars.2019.00606.

Filho, W. S., Seixas, J. M. D., & Moura, N. N. D. (2011). Preprocessing passive sonar signals for neural classification. *IET Radar, Sonar Navigation, 5*(6), 605–612.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 ieee conference on computer vision and pattern recognition (cvpr)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Hovem, J. M. (2010). *Marine acoustics: the physics of sound in underwater environments*. Los Altos Hills, California: Peninsula Publishing.

Irfan, M., Jiangbin, Z., Iqbal, M., & Arif, M.H. (2021b). A novel lifelong learning model based on cross domain knowledge extraction and transfer to classify underwater images. Information Sciences, 552, 80–101. Retrieved from http://www.sciencedirect.com/science/ article/pii/S0020025520311464 doi: 10.1016/j.ins.2020.11.048.

Irfan, M., Jiangbin, Z., Iqbal, M., & Arif, M.H. (2021a, 03). Enhancing learning classifier systems through convolutional autoencoder to classify underwater images. Soft Computing. Retrieved from https://doi.org/10.1007/s00500-021-05738-w doi: 10.1007/ s00500-021-05738-w.

Irfan, M., Zheng, J., Iqbal, M., & Arif, M. H. (2020). A novel feature extraction model to enhance underwater image classification. In *Intelligent computing systems* (pp. 78–91). Cham: Springer International Publishing.

Jiang, J., Shi, T., Huang, M., & Xiao, Z. (2020). Multi-scale spectral feature extraction for underwater acoustic target recognition. Measurement, 166, 108227. Retrieved from http://www.sciencedirect.com/science/article/pii/S0263224120307661 doi: https://doi.org/ 10.1016/j.measurement.2020.108227.

Karakos, D., Silovský, J., Schwartz, R., Hartmann, W., & Makhoul, J. (2018, 04). Individual ship detection using underwater acoustics. In (p. 2121-2125). doi: 10.1109/ICASSP.2018.8462193.

Khishe, M., & Mosavi, M. (2020). Classification of underwater acoustical dataset using neural network trained by chimp optimization algorithm. *Applied Acoustics, 157*, Article 107005. https://doi.org/10.1016/j.apacoust.2019.107005

Luo, X., & Feng, Y. (2020). An underwater acoustic target recognition method based on restricted boltzmann machine. *Sensors, 20*(18). https://doi.org/10.3390/s20185399

Malfante, M., Mars, J.I., Dalla Mura, M.,& Gervaise, C. (2018). Automatic fish sounds classification. The Journal of the Acoustical Society of America, 143(5), 2834-2846. Retrieved from https://doi.org/ 10.1121/1.5036628 doi: 10.1121/1.5036628.

McKenna, M., Ross, D., Wiggins, S., & Hildebrand, J. (2012). Underwater radiated noise from modern commercial ships. *The Journal of the Acoustical Society of America, 131*, 92–103. https://doi.org/10.1121/1.3664100

Miglianti, L., Cipollini, F., Oneto, L., Tani, G., Gaggero, S., Coraddu, A., & Viviani, M. (2020). Predicting the cavitating marine propeller noise at design stage: A deep learning based approach. Ocean Engineering, 209, 107481. Retrieved from http://www.sciencedirect.com/science/article/pii/S0029801820304960 doi: https://doi.org/ 10.1016/j.oceaneng.2020.107481.

Ocean Networks Canada Society. (2017a). Fraser river delta lower slope hydrophone deployed 2016–05-02. Ocean Networks Canada Society. Retrieved from https://data.oceannetworks.ca/DatasetLandingPage?doidataset=10.34943/ 5418bf93-45be-43d4-ab40-f76fef4d9a15 doi: 10.34943/5418BF93-45BE-43D4-AB40-F76FEF4D9A15.

Ocean Networks Canada Society. (2017b). Fraser river delta lower slope hydrophone deployed 2017–06-24. Ocean Networks Canada Society. Retrieved from https://data.oceannetworks.ca/DatasetLandingPage?doidataset=10.34943/ b521061b-43e8-49d7-8831-e34d4612521d doi: 10.34943/B521061B-43E8-49D7-8831-E34D4612521D.

Ocean Networks Canada Society. (2017c). Fraser river delta lower slope hydrophone deployed 2017–11-04. Ocean Networks Canada Society. Retrieved from https://data.oceannetworks.ca/DatasetLandingPage?doidataset=10.21383/ 650d90fa-ce87-473c-b932-278519062ab5 doi: 10.21383/650D90FA-CE87-473C-B932-278519062AB5.

Pezeshki, A., Azimi-Sadjadi, M.R.,& Scharf, L.L. (2007). Undersea target classification using canonical correlation analysis. IEEE Journal of Oceanic Engineering, 32(4), 948-955.

Roth, E. H., Schmidt, V. E., Hildebrand, J. A., & Wiggins, S. M. (2013). Underwater radiated noise levels of a research icebreaker in the central arctic ocean. *The Journal of the Acoustical Society of America, 133*(4), 1971–1980.

Schörkhuber, C. (2010). Constant-q transform toolbox for music processing.

Shen, S., Yang, H., Yao, X., Li, J., Xu, G., & Sheng, M. (2020). Ship type classification by convolutional neural networks with auditory-like mechanisms. *Sensors, 20*(1), 253.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In 3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7–9, 2015, conference track proceedings. Retrieved from http:// arxiv.org/abs/1409.1556.

Strain., N.,& Olszewska., J. (2020). Naive bayesian network for automated, fashion personal stylist. In Proceedings of the 12th international conference on agents and artificial intelligence - volume 2: Icaart, (p. 814–821). SciTePress. doi: 10.5220/0009123808140821.

Szegedy, C., Liu, Wei, Jia, Yangqing, Sermanet, P., Reed, S., Anguelov, D., & Rabinovich, A. (2015). Going deeper with convolutions. In *2015 ieee conference on computer vision and pattern recognition (cvpr)* (pp. 1–9). https://doi.org/10.1109/CVPR.2015.7298594

van der Maaten, L., & Hinton, G. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research, 9*, 2579–2605.

Wang, S., & Zeng, X. (2014). Robust underwater noise targets classification using auditory inspired time–frequency analysis. Applied Acoustics, 78, 68–76. Retrieved from http://www.sciencedirect.com/science/article/pii/S0003682X13002624 doi: https://doi.org/ 10.1016/j.apacoust.2013.11.003.

Wu, Y., Li, X., & Wang, Y. (2018). Extraction and classification of acoustic scattering from underwater target based on wigner-ville distribution. *Applied Acoustics, 138*, 52–59.

Xie, J., & Zhu, M. (2019). Investigation of acoustic and visual features for acoustic scene classification. *Expert Systems With Applications, 126*, 20–29.

Xu, H., Lin, L., Sun, X., & Jin, H. (2012). A new algorithm for auditory feature extraction. In *2012 international conference on communication systems and network technologies* (pp. 229–232).

Honghui Yang, X.Y.M.S.C.W., Sheng Shen. (2018, Mar). Competitive deep-belief networks for underwater acoustic target recognition. Sensors, 18(4), 952. Retrieved from https://doi.org/ 10.3390/s18040952 doi: 10.3390/s18040952.

Yang, H., Junhao, L., Sheng, S., & Xu, G. (2019, 03). A deep convolutional neural network inspired by auditory perception for underwater acoustic target recognition. Sensors, 19, 1104. doi: 10.3390/s19051104.

Yin, X., Sun, X., Liu, P., Wang, L., & Tang, R. (2020). Underwater acoustic target classification based on lofar spectrum and convolutional neural network. In *Proceedings of the 2nd international conference on artificial intelligence and advanced manufacture* (pp. 59–63). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3421766.3421890.

Yue, H., Zhang, L., Wang, D., Wang, Y., & Lu, Z. (2017). The classification of underwater acoustic targets based on deep learning methods. In *Proceedings of the 2nd international conference on control, automation and artificial intelligence (caai 2017)* (pp. 526–529). Atlantis Press. https://doi.org/10.2991/caai-17.2017.118 doi.

Zhang, L., Wu, D., Han, X., & Zhu, Z. (2016). Feature extraction of underwater target signal using mel frequency cepstrum coefficients based on acoustic vector sensor. Journal of Sensors, 2016, 1–11. Retrieved from https://doi.org/10.1155%2F2016%2F7864213 doi: 10.1155/2016/7864213.

Zhang, T., Liang, J., & Ding, B. (2020). Acoustic scene classification using deep cnn with fine-resolution feature. Expert Systems with Applications, 143, 113067. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417419307845 doi: https://doi.org/ 10.1016/j.eswa.2019.113067.

Zheng, Y., Gong, Q., & Zhang, S. (2021). Time-frequency featurebased underwater target detection with deep neural network in shallow sea. *Journal of Physics: Conference Series, 1756*(1), Article 012006. https://doi.org/10.1088/1742-6596/1756/1/012006