Applied Acoustics 181 (2021) 108141

Contents lists available at ScienceDirect

Applied Acoustics

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

Diverse ocean noise classification using deep learning

B. Mishachandar [a], S. Vairamuthu [a,*]

[a] School of Computer Science and Engineering, Vellore Institute of Technology, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 9 October 2020
Received in revised form 6 April 2021
Accepted 22 April 2021
Available online 9 May 2021

Keywords:
Deep learning
Convolutional Neural Networks
Recognition
Classification
Natural acoustic systems
Artificial acoustic systems

ABSTRACT

The alarming rise in the ocean noises due to the ramping up of anthropogenic activities had adversely impacted the marine fauna and altered the soundscape of the oceans. In recent times, Deep learning has gained high significance in assessing this pervasive underwater noise data. Studying the underwater soundscape is crucial in conserving the ocean health and in achieving a "quiet ocean". As an effort to it, in this paper, we present a deep neural network architecture, Convolutional Neural Network-based ocean noise classification cum recognition system capable of classifying vocalization of cetaceans, fishes, marine invertebrates, anthropogenic sounds, natural sounds, and the unidentified ocean sounds from passive acoustic ocean noise recordings. The challenge is to classify these noises amidst the highly non-stationary sound spectrum of short low grunts to long high peaked vocals, limited sensor receiving range, and the need for a large annotated training data. The proposed method can self-learn the features from the training audio data with no need for feature extraction proving better adaptability to complex audio acoustic signals. Experimental results prove that the proposed system is befitting for classifying ocean noises with 96.1% accuracy. Classification helps in distinguishing natural acoustic systems from artificial acoustic systems.

© 2021 Elsevier Ltd. All rights reserved.

1. Introduction

The research on the ocean-dwelling marine mammals like dolphins, whales, and porpoises is constantly increasing for the awe and fascination humans have for these lovely creatures. Life in oceans especially for air-breathers is highly challenging and deserves our respect and protection. Like marine mammals, other marine species like fishes and invertebrates also needs conservatory measures, which is neglected by most researchers. Research for the protection and conservation of these marine species is scientifically progressing and still demands the intervention of automated recognition and classification systems combined with the traditional Passive Acoustic Monitoring. Laboratory-based marine life conservatory research is time-consuming, labor, and equipment demanding, and the results are mostly approximated as it involves manually observing the surface behavior of the cetaceans. Adding to this, like most marine species, the marine mammal's constant mobility and stringent ocean environment makes this process more infeasible. The root cause of this pitiful state is found to be the intrusion of anthropogenic activities in the ocean that has drastically changed the soundscape of the oceans [1]. The sounds produced from anthropogenic activities like shipping, offshore ocean exploration, sonar operations harm the entire marine fauna and not just the marine mammals [2]. It causes adverse irreversible physical, biological, and physiological damage as severe as masking in mammals, which leads to difficulty in identifying crucial sounds from the interfering background noise and destroys fish eggs and larvae affecting its breeding [3]. The overall impact is viewed from the perspective of "noise" contribution by the diverse sources in the ocean. Deep Neural Network has recently advanced in the field of audio signal processing with numerous applications like speech detection [4–5], audio enhancement, [6–7] music information retrieval [8], and source separation [9–10]. Numerous time-critical, cost-incurring, and human interpretation infeasible conservatory needs can be accomplished with this method [11]. Drawn by the promising results achieved in the field of audio signal processing, Convolutional Neural Networks has inspired marine oncologists and bio acousticians to investigate and imply it to the problems in their field of study. One such crucial problem is the analysis of ocean acoustic data produced by the diverse ocean noise sources. CNN is an extensively preferred choice of a deep learning model to perform acoustic tasks like classifying events or patterns found in acoustic recordings with precision. Unlike other threshold techniques, Deep CNN is found to be highly learnable and adaptive and the best results in terms of accuracy can be achieved through proper training [12].

* Corresponding author.
E-mail address: vairamuthu@vit.ac.in (S. Vairamuthu).

https://doi.org/10.1016/j.apacoust.2021.108141
0003-682X/© 2021 Elsevier Ltd. All rights reserved.

Various sources contribute to the overall ocean noise, namely: (a) Natural physical processes like cracking icebergs, underwater earthquakes (b) Marine life that comprises largely marine mammals (dolphins, porpoises, seals) and imperceptibly marine fishes (Arctic char, Red drum) and invertebrates (Snapping shrimp, lobsters) (c) Anthropogenic sources like ships, sonars and UANs [13]. The complexities involved in analyzing ocean noises are high compared to other acoustic scene classification due to both the structured and unstructured nature of the diverse ocean noises. High peaks or strong vocal patterns from marine mammals calls to randomly dispersed ambient noise challenges the audio processing system. However, ocean ambient sounds may include strong Spectro-temporal signatures. Thus, it is important to consider non-stationary aspects of signal and capture its variation in both time and frequency domains. Acoustic ocean noises are recorded using Active Acoustic Monitoring (AAM) and Passive Acoustic Monitoring (PAM) techniques of which PAM technique is mostly used to capture sounds produced by biological sources using underwater microphones called hydrophones. The key idea of PAM is "acoustics" which is based on the cetacean's sound-producing ability and their strong auditory systems. It is a non-invasive method unlike GPS tagging for recording sounds from the sound-producing source. Active acoustic detection doesn't require the animal to produce sound instead relies on their reflected echoes. This practice of detection can do serious harm to the cetaceans hence is deployed for non-biological sources. Passive acoustic detection, on the contrary, is based on listening to the acoustic signals from the animals with no intervention to the mammal's behavior. The conventional method of surveying cetaceans visually through sighting methods is slowly being replaced with hearing through sound techniques as their dwelling is mostly misjudged. Adding to it conditions like highly mobile nature and unpredictable behavioral pattern hinders precise detection, classification, and localization of marine mammals in their environment. Despite the fact, that PAM being a non-invasive method of recording, it still suffers few drawbacks. The cost incurred in deploying mobile PAM is comparatively high than static PAM leading to its deployment underwater for a relatively long period. The long-term deployment leads to the recording of enormous volumes of data that becomes manually infeasible for classification [14]. This underlying problem has created the need for an automated recognition and classification system to effectively recognize and classify the diverse ocean noises. Traditionally acoustic ocean sounds were manually analyzed by visually inspecting the spectrograms generated from the audio data as it was considered to be a faster alternative than listening to the entire audio recordings. Classified noise sources can aid in resolving various interlinked research problems in Underwater Acoustic Sensor Networks (UASN) [15] like efficient channel allocation [16], power adjustment [17], energy-efficient designing [18], localizing marine mammals [19]. This paper aims to cover a deep learning- based automatic ocean noise classification system using CNN trained using spectrogram representation of acoustic audio recordings. Unlike other work in this area, we emphasize and consider all the possible marine species like invertebrates and fishes and not just limit it to marine mammals as the impact of anthropogenic noises on them is equally saddening. Similarly, no work ever done so far have considered a category of ocean noise called the "unidentified sources". This unidentified source contributes about 2–5% of the total ocean noise. This paper, claims that effort.

The main contributions of the paper are:

1. Deep learning-based CNN that can classify all the six possible sources of ocean noises, namely: marine mammals vocalizations, sounds produced by invertebrates and fishes, anthropogenic noises, natural noise, and noise produced by unidentified sources.
2. An adaptable classifier that can classify sounds ranging from very low (invertebrates) to high (Baleen whales).
3. The proposed system performance is evaluated using two types of DNN architectures CNN and RNN to provide a comparative analysis.
4. The highest range that can be captured is 8000 Hz hence can fit in a wide variety of high- frequency audio signals.
5. Data augmentation techniques are applied to avoid the overfitting problem in the training process and in achieving better accuracy.

This paper aims to propose an application using annotated data that were created for environmental conservatory purposes. The data collected are raw ocean noise data from diverse sources with no prior cleaning or manipulation done. Preliminary denoising and soundscape source separation can help improve the classification accuracy [20–21]. The separated sources from the soundscape can be precisely identified using this classification system. Unlike sound source identification using a supervised learning method using clustering techniques, this method fetches precise results with limited annotated data. This work is a section in the underwater cognitive acoustic network strategy for efficient spectrum utilization framework proposed in [22].

The remaining sections of the paper are organized as follows, Table 1 briefs the various ocean noise sources with their impact, In Section 2, the previous related work on CNN-based audio classification is reviewed. Section 3 provides insights about the experimental setup and analysis covering data collection and feature extraction. The proposed classification model is detailed in Section 4. The corresponding results are analyzed in Section 5. Finally, the conclusion and the intended future work are presented in Section 6.

2. Related work

Numerous classification techniques to classify audio-based data have been developed extensively over the years. Based on the implementation pattern, the techniques are broadly categorized as statistical-based and threshold-based techniques [23]. In which the most widely used are the Gaussian Mixture Models (GMM) [24], Hidden Markov Models (HMM) [25], Support Vector Machine (SVM) [26], Neural Networks (NN) [27], Spectrogram Cross- Correlation (SPCC) [28], Dynamic Time Wrapping (DTW) and Matched Filtering (MF) [29]. The choice and performance of the technique depends on a few influencing factors like source, region, size, and features of the collected data.

Authors Shawn and Sourish in [30] have performed CNN-based video classification and audio event detection on a massively large data set. The model training is explored with different sized subsets with initialization using Gammatone filter bank, a human ear-inspired model of initializing convolutional layers to study its effect on the accuracy of the system. Classification of audio data using terrestrial acoustic sensors is both flexible and cost-effective but that is not the case in an ocean environment. Drawbacks like expensive underwater sensors, difficulty in deployment and recharging and the strident marine environment challenges data collection underwater [31]. Environmental sound classification (ESC) is a commonly experimented idea in acoustic scene classification as in [32] where the authors have presented an end-to-end approach using 1D CNN that directly learns representation from the audio signals. The proposed approach was found to outperform the 2D approach and significantly reduce the size of data

Table 1
Different sources of ocean noises with their impact.

Natural Noise Sources	Category	Sub-category	Frequency range
Marine Mammals	Baleen Whales	Blue Whale	15–40 Hz
		Bowhead Whale	25–900 Hz
		Fin Whale	30 to 15 Hz
		Grey Whale	20 and 200 Hz
	Toothed Whales	Common Dolphin	200Hz- 24kHz(Communication) 200Hz-150kHz (Click echolocation)
		Dall's Porpoise	Short with high frequencyclicks 117 to 160 kHz
		False Killer Whale	0.1 kHz to about 40 kHz
	Pinnipeds	Bearded Seal	0.02 – 11 kHz
		Ribbon Seal	200 Hz to 5 k Hz
		Harbor Seal	Peak sounds at 1.2 kHz
Marine Invertebrates		Mantis Shrimp	167 Hz
		Snapping Shrimp	2 kHz to 200 kHz
		Sea Urchin (Kina)	700–2000 Hz
		Spiny Lobster	–
Fishes		Arctic Char	Peaks at 1114 Hz
		Red Drum	140–160 Hz
		Black Drum	60–1100 Hz
		Clown Fish	95–240 Hz

Ambient Noise Sources	**Causes of Noise**	**Frequency**	**Duration**
Ocean floorEarthquakes	Movement of the ocean floor	5 to 100Hz	20s
Hydrothermal vents	Hydrothermal vents are analogous to hotsprings on land	5–15 Hz	—
Ice Cracking	Popping and cracking of ocean ice	100–500 Hz	40-60s
Iceberg Collisions	Iceberg breaking off from an ice shelf or a glacier	1–10 Hz for several hours	Several Hours
Lightning	Lightning that strikes the ocean surface at a rate of 2stikes/km in a year	2 strikes per square kilometre per year	10 times longer than that in land
Rainfall	Raindrops hitting the ocean surface	1000Hz-50,000Hz	Depends on various factors
Volcanic Eruptions	Eruption sound from the ocean floor cracks	100–500 Hz	Hours to years
Tsunami	Gushing waves	0.14 to 8 MHz	5–90 min
Waves	Breaking of bubbles and waves	20–20,000Hz	~20 min
Anthropogenic Noise Sources	Purpose	Frequency Range	Impact
Airgun	*Ocean layer study *Study of Earth's history *Gas and oil excavation	10–500 Hz	*Irrecoverable Hearing damage *Aberrations in the digestive tracts and stomach.
Bubble Curtain	Mitigate impacts caused by noise from pile driving.	——	Positive impact- Ocean noise reduced by 30 dB
Dredging	Ocean floor excavations	Broadband signals Peaks at 1kHz	Affects fish eggs and larvae
Explosive Sound Sources	*Study of seafloor structure * Oil and gas exploration.	Strong intensity signals	*Mass strandings *Causes cetaceans death
Icebreaker	*Aids in escorting ships trapped in icebergs. * Helps in reaching out inaccessiblepolar locations.	Continuous and broadband, with highest frequencies of approximately 5kHz	Contributes to the local underwater soundscape.
Ocean Acoustic Tomography Transmission	To measure the temperature of the ocean over large areas.	High frequencies up to 250Hz	No biological effects
Outboard Motor	Used in small boats to power up propellers	High frequencies like 6300 Hz.	Makes small fishes an easy prey
Pile Driving	*Helps in laying foundations for docks, bridges, wind turbines, and offshore oil and gasplatforms.	Low frequency 20-40 Hz	Induces emigration
Personal Watercraft	Recreational vehicle	100 Hz to 1kHz	Shortens cetacean's lifespan
Ships	Transportation	High frequency of above 20,000 Hz	Decreases the mammal population
Sonar	Ocean searching	Mostly above 20,000 Hz	*Adversely affects sound- sensitive mammals such as whales *Mass standings.
SURTASSLFA Sonar Sound	*Used as Antisubmarine	100 to 500 Hz	*Beaching
	warfare in Us Navy*Surveillance		*Bleeding from eyes and ears (sometimes)
Wind Turbine Sounds	For generating windpower	Low frequencies 1kHz and 700 Hz	Leads to emigration
Underwater Breathing Apparatus	Recreational, scientific, commercial, and military purposes	Bubbles: 100 to400 Hz	No significant biological impact

required for training. In a similar work [33] deep CNN is used to learn discriminative spectrogram patterns in an environmental sound. Audio-based augmentation techniques are used to cope with data scarcity during the model training and thereby avoid model overfitting. The combined performance of the deep CNN high-capacity model with data augmentation had highly reflected in the accuracy of the classified results. In [34], ESC is performed using spectrogram images of sound generated from environmental sounds to train CNN and tensor deep stacking networks (TDSN). The results showed a decent spike in the accuracy proving it to a better sound recognition and classification system. In a similar audio processing application, the authors Rigas et al in [35] aimed

at performing spoken language classification in broadcasting content is a strategical pattern to surpass other state of art models. Initially, the voices are separated using hierarchical discrimination schemes and the deep CNN– based classification is performed using semi-automatically annotated audio data. Maccagno and Mastropietro in [36] have developed an application to classify five classes of construction vehicles and tools using audio emitted in a construcion site using a stack of convolutional layers.

Acoustic-based audio detection and classification have drawn the attention of ecosystem managers to aid in conserving animals that mainly use acoustic means of communication notably birds, reptiles, amphibians, and both terrestrial and marine mammals. Going with the motivation of this paper the remaining part of the supporting literature will largely speak about animal life protection and precisely marine fauna conservatory-based audio recognition and classification. Supporting it authors in [37] have performed marine mammal species detection and classification from acoustic recording using CNN. It is designed to classify the vocalizations of three different whale species, non-biological sources of noise, and ambient noise. In addition to the classification of marine mammals, the authors have proposed a novel representation of acoustic signals as spectrogram representation using Short-Time Fourier Transform (STFT) parameters. They intend to apply transfer learning to species regarded as endangered or less annotated to include additional species as part of their future work. In a similar effort authors, Dan and Micheal in [38] have attempted acoustic detection and classification of bird sounds using deep learning to observe the presence and abundance of birds. A very high retrieval rate is achieved with no need for manual calibration and pretraining of the target species detector. In [39], the authors have approached the detection of three endangered varieties of baleen whales using Region-based Convolutional Neural Networks from spectrogram representations of acoustic vocalization recordings. The detection is performed on both time and frequency domain with a background of anthropogenic and ambient noise. Transfer learning is applied to extent the detection process to more species with limited training data due to less population. The authors intend to apply active learning to correct partial labelling as their future work. Unlike other efforts in the literature, detecting marine mammals calls in acoustic vocalizations is approached in [40] to assess the abundance of fish varieties in an underwater video using CNN. The idea of this work is to accurately perform fish identification in a complex, deformed, low-light video recording. An end-to-end deep learning architecture is introduced to process the video stream and extract important data required for the fish assessment.

Acoustic audio detection and classification of marine mammal vocalizations in [41] were performed using two open-source software ISHMAEL and PAMGAURD. On comparing both, PAMGUARD was found to perform better than ISHMAEL by effectively detecting blue whale vocalizations, and both the software were tested on MATLAB. The detection and classification of fishes are performed using traditional Faster R-CNN with three distinct classification networks on data obtained from sea trial experiments. A novel deep CNN-based detection and classification technique is proposed in [42] to classify whistle sounds produced by False killer whales and long-finned pilot whales. Feature extraction was performed directly on the training data eliminating the need for a separate feature extractor. STFT feature extraction was used to create contours in the spectrogram to facilitate whistle detection. A separate graphical user interface (GUI) was developed to aid in the visualization of the results. The challenge of pre- processing and the lack of a large training dataset in applying conventional machine learning and automated algorithms in classifying sound sources from the acoustic recordings is overcome by the authors in [43] using supervised learning Empirical Mode Decomposition (EMD). EMD is completely data-driven and eliminates the need for human analysis. The application of Hilbert transform on the generated IMF was avoided and facilitated quick sound detection from all the sound sources in a recording without the need for prior knowledge.

3. Experimental setup and analysis

3.1. Data collection and pre-processing

The ocean acoustic recordings used for training and testing the classifier are obtained from various open-source databases. The marine mammal vocalizations were collected from the Watkins marine mammal sound database[44]. The acoustic data is observed to have been collected from various locations in the US for decades

Fig. 1. Geographic locations of the collected data [44].

till the 2000 s for more than 60 marine mammal species during specific seasons of a year. The acoustic recordings have been recorded using the fixed PAM technique in the presence of ambient noise and the collected non-continuous audio data has been manually annotated. The fish vocalization and the sounds produced by the invertebrates, anthropogenic noise, ambient noise are open-source data from "Discovery of sound in the sea" [45] funded by the University of Rhode Island. The marine mammal vocalization had been collected using digital acoustic recording tags (DTAG3) and mooring devices for better coverage. Owing to the non-invasive nature of PAM, all the data used in this work were ensured to be collected only using the PAM technique as a little harm can

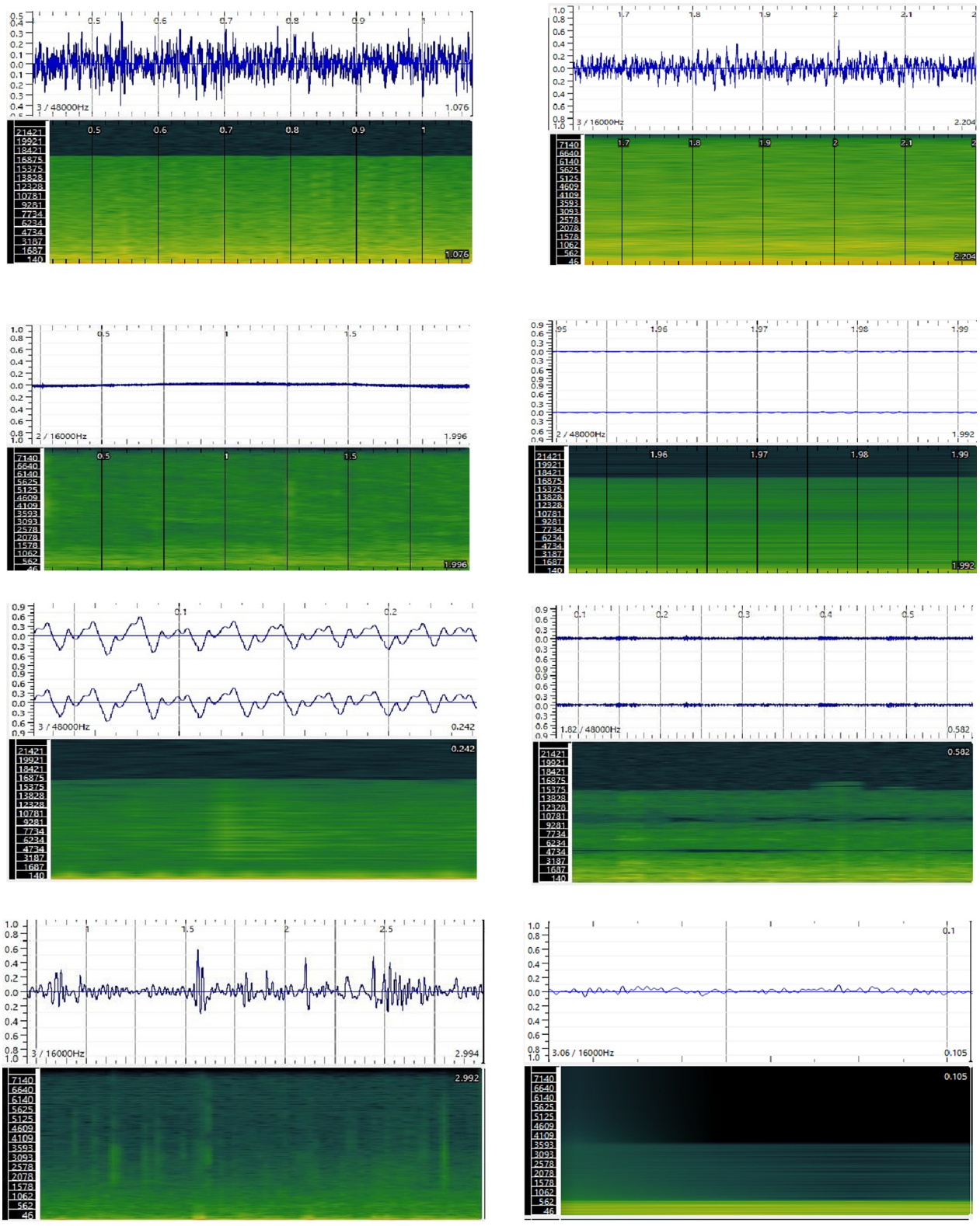

Fig. 2. Example waveforms and spectrograms pre- and post-cleaning. By the model, from the top in pairs (a) High peaked marine mammal vocalization, (b) Low grunts of marine invertebrates, (c) Noise generated by anthropogenic sources, (d) Intermitted ambient noise.

cause a huge impact on the aquatic life. Six categories of ten different classes of ocean noises consisting of vocalizations of four categories of forty-three varieties of cetaceans, twenty-seven fish species, four marine invertebrates, eighteen varied sources of anthropogenic sounds, nine varieties of natural sounds, and the two unidentified ocean sounds with an average duration of 3 s 16-bit audio are considered for classification. The acoustic recordings were sampled at 44.1KHz to capture both low frequencies of invertebrates and high-frequency vocalizations of marine mammals. Fig. 1 shows the recording sites of the Watkins marine mammal sounds data.

The pre-processing of the audio is performed using Digital signal processing (DSP). DSP aids in converting the real-world digitalized audio data and manipulates it mathematically that aids in distinguishing the audio signal of interest from the other signals. The time-domain signal before being fed to the classifier is first cleaned to remove empty dead spaces with fading low magnitude to enhance the quality of the audio. Noise threshold detection is employed to remove frequencies lower than the minimum range that an underwater microphone can detect. The highest frequency that can be captured from an underwater microphone recording is 8 kHz. Post denoising the audio is annotated. In our case, the audio data used had been manually annotated by experts via visual inspection of the spectrograms. Logarithmic scaling is applied to the features to compensate for the diverse dynamic ranges in the ocean noises. A major challenge in most audio classification systems is the overshadowing of acoustic scenes by background noises that overlap in time and frequency. In that case, the source separation technique using CNN can be adopted to recover the source signals of interest mixed with interfering signals and to remove irrelevant noise components from the acoustic scene background and the recording device. The pre-processing of the audio data used in this work is done by calculating and cleaning the periodical noise, removing the dead spots, and down sampling to 16 KHz was sufficient in reaching the desired results. Fig. 2 shows sample wave form and spectrogram representations of the marine species calls pre and post denoising.

3.2. Spectrogram extraction

Mel scaled spectrogram is preferred over the linear spectrogram to aid the spatially invariant nature of CNN, where CNN is incapable of interpreting frequencies expressed in a linear scale. A spectrogram is an audio waveform that is encoded as a visual representation before being fed into the network as training data. The frames were sampled at 16 KHz with a window length of 25 ms for 400 samples with a step size of 10 ms for 160 samples. The bandwidth of the data used in this work is bounded from very low to high, hence Mel scaled spectrogram generation eases feature extraction. The audio signals in the time domain are converted into the frequency domain using Fast Fourier Transform (FFT). Underwater acoustic signals are highly non-stationary due to the presence of interfering signals from various ocean noise sources, which causes hindrance and reduces the overall performance of the classifier. Owing to its easier implementation and providing time-localized frequency information of the inputted signal, Short-Time Fourier Transform (STFT) is performed by stacking multiple FFTs computed overtime to produce spectrograms. A logarithmic spectrogram is considered for visualizing more profound counters in the spectrogram representation that aids in better classification. Filter banks and Mel Frequency Cepstrum Coefficient (MFCC) feature extraction techniques are employed to extract the linguistic features of the spectrograms and to get rid of the background noise in the acoustic recordings. However, CNN eliminates

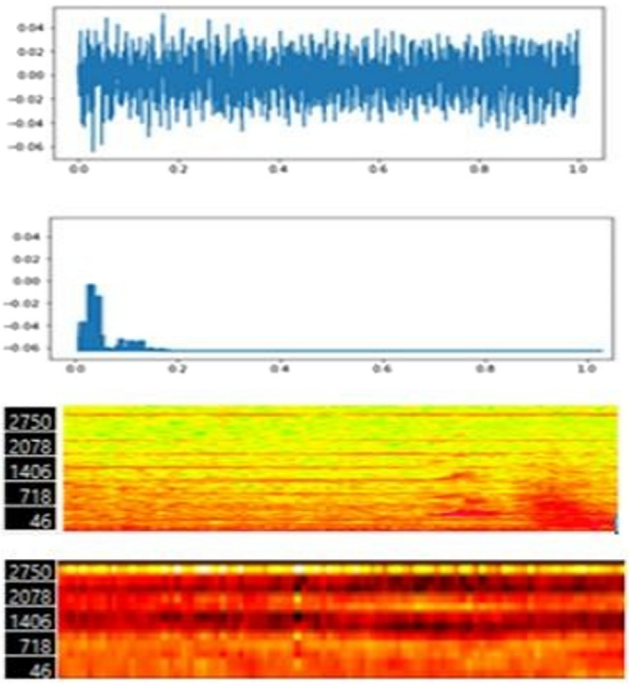

Fig. 3. Sample of the spectrogram extraction. By the model, from the top (a) Time series (time-domain) (b) Short-Term Fourier Transform (STFT-frequency domain) (c) Filter bank coefficients (d) Mel Frequency Cepstrum Coefficient (MFFCs).

the need for feature extraction, in this work it is used to create visible contours in the annotated spectrograms. Feature extraction using MFCCs favors low computational complexity and suits best for characterizing high-frequency broadband signals and amplitude-modulated sounds. Fig. 3 represents a sample of the spectrogram extraction.

The filter banks and the MFCCs are computed by passing the signal through a pre-emphasis filter where it gets sliced into frames. A window function is applied to the sliced frames. individually and a Short-Time Fourier Transform is performed on each frame to extract features. The power spectrum is calculated from the resulting frames and finally, the filter banks are computed. A Discrete Cosine Transform (DCT) is applied to the filter banks only with coefficients to compute MFCCs. This procedure converts the input audio signal from a time domain to a frequency domain periodogram. Since the audio signals tend to constantly vary with time, we assume that in a fixed time frame, the audio signals are stationary. Another recent trend in spectrogram extraction is the long durational false colour spectrograms that aids experts visualize and identify the species of interest from long duration of audio data in a single long durational spectrogram [46,47]. A sample spectrogram representation of a humpback whale vocalization at a low SNR is presented in Fig. 4.

4. Methodology

Convolutional Neural Networks (CNN) is a deep learning architecture that is commonly employed to analyze and classify features in an image for computer vision. It is a class of deep feed-forward artificial neural networks designed to perform numerous audio-related tasks with successful practical applications in speech and music processing. In this paper, a CNN-based ocean noise recognition and classification model is designed. The model can aid in distinguishing the natural acoustic systems (marine mammals, fishes,

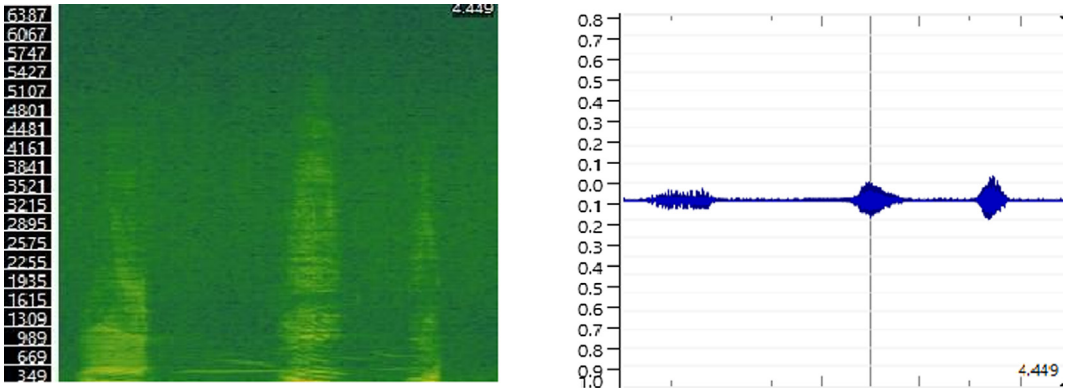

Fig. 4. Sample of a 4 s spectrogram of audio data with humpback whale vocalization at a low SNR.

and invertebrates) from the artificial acoustic systems (UANs, sonar, ships) that can help in addressing a challenging research problem in Underwater Acoustic Sensor Networks (UASN), Spectrum allocation. Efficient spectrum allocation will save marine life significantly by avoiding frequency clashes with the artificial acoustic systems that cause fatal effects on the marine fauna. Characteristics like sparse interactions, parameter sharing, and equivariant representations of CNN assists in perceiving the two-dimensional features of the input images. Unlike other traditional machine learning-based classification methods, the proposed method can self-learn features from the generated training spectrograms with no prior need for feature extraction, showing better adaptability to complicated signals like marine acoustic sounds. Firstly, inputted raw audio data is cleaned as in Section 3.1. A denoising mechanism is adopted to remove fading dead spots in the audio. Then automatic recognition and the classification system are trained using the annotated acoustic spectrograms generated from the ocean noise recordings obtained in section 3.2. In the recognition and the classification process, the trained recognition model learns few prominent features like intensity, spread, and overlapping in the spectrograms needed for performing the classification. Furthermore, from the recognition process, the model learns unique distinguishing features of each source individually. The recognized contours are then forwarded to the trained classification model. The CNN model then classifies the spectrograms to their corresponding sources based on the maximum prediction percentage value. Added to it the classifier predicts the non-generic features of the spectrograms and distinguishes the audio noise data as a biological or a non-biological source.

The structuring of the CNN model is done in a way to learn more different features from the contours in the MFCCs than following the regular stacking up of convolutional layers followed by max-pooling that pulls down the data leading to loss in prominent features. The CNN model is designed in a way wherein multiple 2D convolutional layers are followed by a 2D max-pooling layer then a dropout of 0.5, a flatten layer, and then finally few fully connected layers are stacked to improve the efficiency of the classifier. The size of the filters is progressively increased layer after layer and the tank activation function is used for the first layer, ReLu activation function for the remaining layers, and a softmax activation for the final dense layer to favor categorical cross-entropy. The padding same is used for the convolutional layers to preserve the dimensions of the input matrix and a stride of 1 for the first layer and 2 for the remaining layers is followed. The idea behind this model design is to minimize the complexity of the architecture that has a positive effect on the overfitting problem. Through this model, CNN can efficiently learn from the log Mel spectrograms and extract more deep features suitable for performing the classification. The techniques used in the paper are implemented with Spyder 4 IDE and google collab with python 3.7. The visualization of the graphs is done using IPython console in Spyder.

4.1. Data augmentation

Data augmentation is a technique to generate syntectic data to tackle data inadequacy for training purposes and to achieve better accuracy without the need to collect new data. It is a vital practice in many states of art systems to perform recognition and classification without the problem of overfitting. Growing demand in incorporating neural networks in audio recognition systems that require large volumes of data for training purposes has created a need for data augmentation techniques, especially in audio-based data. Data augmentation in a traditional ASR is done in two possible ways; one is through augmenting the waveform of the inputted audio and the other is through augmenting the spectrograms generated from the audio data as in SpecAugment, a google brain audio augmentation initiative. The former method of augmenting the raw audio recording is used in this work and is generally performed by adopting various methods such as (a) Time shifting: The audio sample is shifted to left/right with respect to time in seconds, keeping the pitch of the audio unaltered. The audio is mildly shifted by -2824 frames in the starting few seconds and then the remaining portion is padded to its original length. (b) Pitch and speed shifting: The tempo of the audio sample is raised or lowered keeping the duration static. A speed rate of 1.155 is applied to the audio waveform to alter the phase of the audio. (c) Noise injection: The audio sample of interest is augmented by injecting random values into the data. The random value is mostly a white background noise that is mixed with the audio data with a random volume tuning. Augmenting the audio data helped in narrowing the gap between the training and the testing data. In addition to the audio augmentation methods, a dropout of 0.5 is added in the CNN model to perform regularization and the complexity of the CNN architecture is reduced to avoid the model overfitting problem. Fig. 5 presents the visualization of the augmented spectrograms.

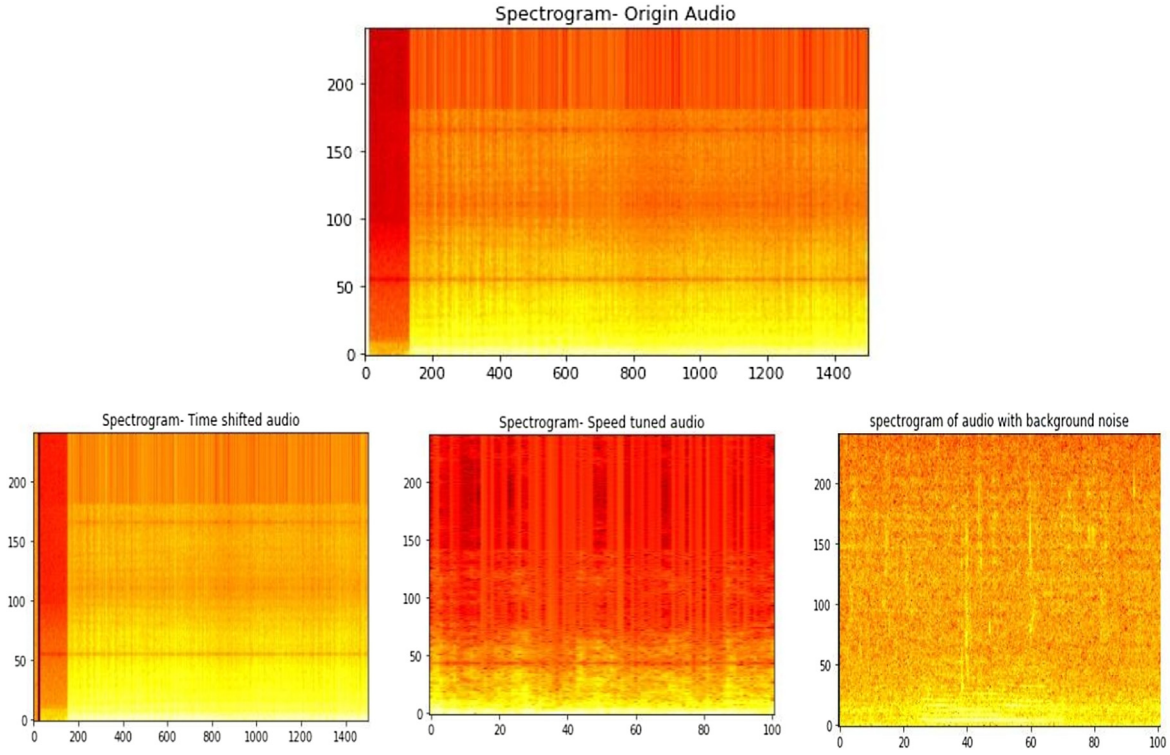

Fig. 5. Visualization of the spectrogram of the origin audio and the augmented spectrograms. By the model, (a) Time shifted (b) Speed tuned (c) Background noise added.

5. Experimental results

The Mel spectrograms generated in Section 3.2 are used in the training and the testing process of the classification model. The training was done with 185,056 samples and testing with 20,562 samples. The accuracy of the model increases with an increase in the number of epochs. The experiment was carried out with 10 and 25 epochs to observe the accuracy changes. The structuring of the proposed model with a combination of convolutional layers followed by flattening dense layers aids in minimizing the training and the validation loss of the models. The model is developed on Windows 10, 64bit PC with Intel(R) Core (TM) i5-7200U CPU, and NVIDIA GeForce 304 GTX 1080 GPU. The codes are scripted in python using the latest version of TensorFlow 2.2. TensorFlow is an open-source platform to integrate and develop large-scale AI and Deep Learning Models. The performance of the proposed model is evaluated using the accuracy metric. A threshold of 0.5 is used in the classification model and the results show almost similar performance in different CNN models. In this work, the performance of the CNN model is compared with a Recurrent Neural Network based Long Short-Term Memory Network (LSTM) model using the same data. According to the results, the 2D CNN model was able to fetch a training accuracy of 97.89% and 96.01% in the testing and validation. The RNN-LSTM model achieved a training accuracy of 89.46% and 92.36% of testing and validation accuracy. The graphs of the models show that both the models perform well but a little upper hand for the proposed CNN model with minimum overfitting. This proves that the proposed CNN-based ocean noise classification model with data augmentation performs better classification for highly challenging data like passive acoustic ocean noise data. Whereas manually validating such data by an expert analyst would have taken a paramount of time.

The graph in Fig. 6 shows the performance of the 2D CNN model during training and testing. The data annotation and data augmentation aids in the training and in avoiding model overfitting. Similarly, the performance of the RNN -LSTM model is shown in Fig. 7,

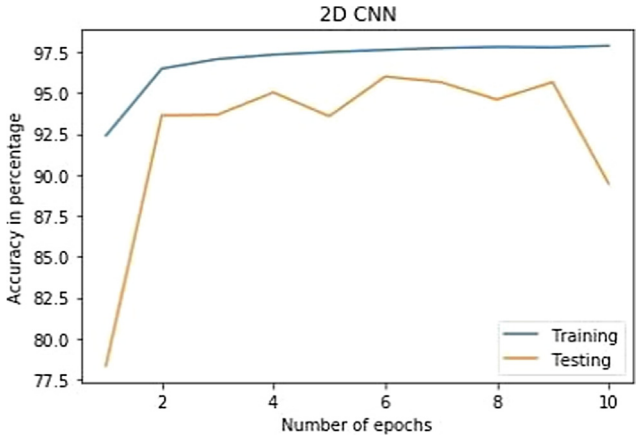

Fig. 6. Performance of the proposed 2D CNN model.

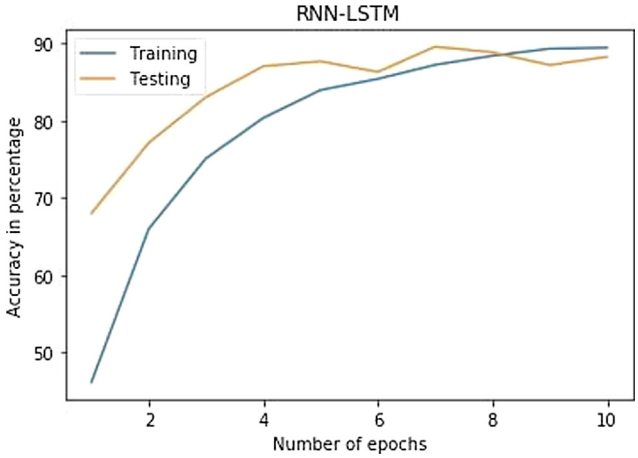

Fig. 7. Performance of RNN- LSTM based classification.

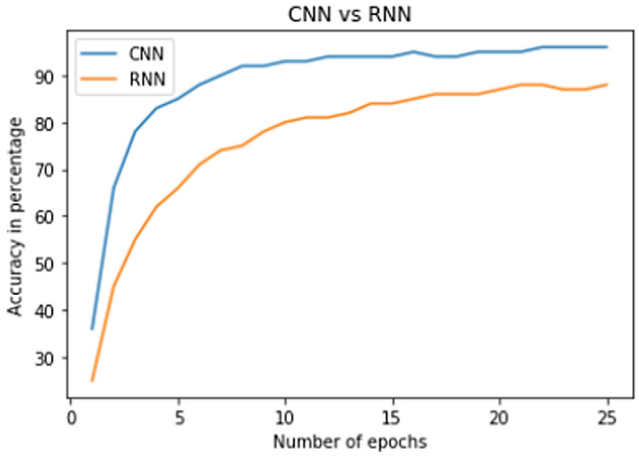

Fig. 8. Performance comparison of the proposed model with the RNN model.

where the model achieved better training and testing results but suffers model overfitting. The performance of the models is evaluated using the classification accuracy of the validation data. Fig. 8 provides a comparative analysis of the 2D CNN and RNN model performances. The graph depicts that the 2D CNN outperforms RNN in recognizing and classifying complex acoustic data and that CNN with data augmentation can produce better results in terms of accuracy than the RNN model that suits best for sequential time series audio data recorded continuously for long durations. The recognition and the classification model can be easily generalized to other audio acoustic sources of varied frequencies as the model showed better accuracy results in identifying and classifying acoustic sources from low to high frequencies efficiently. The model's performance can be significantly improved if the audio data used with marine species calls were source-separated from the raw ocean noise data rather than using a denoising algorithm for cleaning the data. The results prove that data as audio when effectively handled as images (spectrograms) can produce better accuracy with image classification models like CNN than in its time-domain representation.

6. Conclusion and future work

In this paper, a CNN-based recognition cum classification method is proposed that classifies almost all prominent categories of ocean noises in acoustic recordings and aids in distinguishes natural acoustic systems from artificial acoustic systems in the ocean. The proposed method effectively classifies four categories of forty-three varieties of cetaceans, twenty-seven varieties of fish species, four varieties of marine invertebrates, anthropogenic sounds, natural sounds, and the unidentified category of ocean noises from passive ocean acoustic recordings. The entire process of the proposed method, from data collection and pre-processing, spectrogram extraction, model creation, and training and validation is detailed upon in this paper. The experimental results prove that the proposed 2D CNN model adaptively learns the features of the audio data and classifies the ocean noises amidst the challenges with an accuracy of 96.1%. The problem of overfitting is observed to have been substantially reduced using the audio augmentation technique. The proposed methodology used in this work in successfully classifying ocean noises can be easily generalized to other passive acoustic sound classification problems. An efficient audio classification system like this can be of great help to marine oncologists in ensuring a marine habitat-friendly ocean noise-free environment. As part of the future work, we intend to try the audio augmentation on the spectrograms and apply transfer learning to include sources with limited data due to less population.

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.

References

[1] Williams R, Wright AJ, Ashe E, Blight LK, Bruintjes R, Canessa R, et al. Impacts of anthropogenic noise on marine life: Publication patterns, new discoveries, and future directions in research and management. Ocean Coast Manag 2015;115:17–24. https://doi.org/10.1016/j.ocecoaman.2015.05.021.
[2] De JK, Nesse T, Maria F, Amorim CP, Rieucau G, Slabbekoorn H, et al. Predicting the effects of anthropogenic noise on fish reproduction. Rev Fish Biol Fish 2020;30:245–68. https://doi.org/10.1007/s11160-020-09598-9.
[3] Lucke K, Erbe C, Reichmuth C, Cunningham K, Lucke K, Dooling R. Communication masking in marine mammals: A review and research strategy Communication masking in marine mammals: A review and research strategy. MPB 2015. https://doi.org/10.1016/j.marpolbul.2015.12.007.
[4] Hu W, Qian Y, Soong FK, Wang Y. ScienceDirect Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. Speech Commun 2015;67:154–66. https://doi.org/10.1016/j.specom.2014.12.008.
[5] Zhang X, Wang D. Boosting Contextual Information for Deep Neural Network Based Voice Activity Detection 2016;24:252–64.
[6] Pandey A, Member S, Wang D. A new framework for CNN-based speech enhancement in the time domain. IEEE/ACM Trans Audio Speech Lang Process 2019;27:1179–88. https://doi.org/10.1109/TASLP.2019.2913512.
[7] Hou J, Wang S, Lai Y, Tsao Y, Chang H, Wang H, et al. Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks n.d.
[8] Sandler M. Retrieval 2017.
[9] Heo W, Kim H, Kwon O. Source separation using dilated time-frequency densenet for music identification in broadcast contents. Appl Sci 2020.
[10] Benda-beckmann AM Von, Wensveen PJ, Samarra FIP, Beerens SP, Miller PJO, Benda-beckmann AM Von, et al. Correction : Separating underwater ambient noise from flow noise recorded on stereo acoustic tags attached to marine mammals Separating underwater ambient noise from flow noise recorded on stereo acoustic tags attached to marine mammals 2016:2271–5. https://doi.org/10.1242/jeb.148197.
[11] Canziani A, Culurciello E, Paszke A. An analysis of deep neural network models n.d.:1–7.
[12] Ibrahim AK, Zhuang H, Chérubin LM, Schärer-umpierre MT, Erdol N, Ibrahim AK, et al. networks Automatic classification of grouper species by their sounds using deep neural networks 2018;196. https://doi.org/10.1121/1.5054911.
[13] Rahmati M, Pompili D. UNISeC: Inspection, separation, and classification of underwater acoustic noise point sources. IEEE J Ocean Eng 2018;43:777–91. https://doi.org/10.1109/JOE.2017.2731061.
[14] Overview A. Fixed Passive Acoustic Observation Methods for Cetaceans An Overview of n.d.; 20:36–45.
[15] Akyildiz IF, Pompili D, Melodia T. Underwater acoustic sensor networks : research challenges 2005;3: 257–79. https://doi.org/10.1016/j.adhoc.2005.01.004.
[16] Yao G, Jin Z, Su Y. An environment-friendly spectrum decision strategy for underwater acoustic networks. J Netw Comput Appl 2016;73:82–93. https://doi.org/10.1016/j.jnca.2016.07.004.
[17] Sherlock B, Neasham JA, Tsimenidis CC. Spread-spectrum techniques for bio-friendly underwater acoustic communications. IEEE Access 2018;6:4506–20. https://doi.org/10.1109/ACCESS.2018.2790478.
[18] Wang K, Gao H, Xu X, Jiang J, Yue D. An energy-efficient reliable data transmission scheme for complex environmental monitoring in underwater acoustic sensor networks. IEEE Sens J 2016;16:4051–62. https://doi.org/10.1109/JSEN.2015.2428712.
[19] Han G, Zhang C, Shu L, Rodrigues JJPC. Impacts of deployment strategies on localization performance in underwater acoustic sensor networks. IEEE Trans Ind Electron 2015;62:1725–33. https://doi.org/10.1109/TIE.2014.2362731.
[20] Lin T. Source separation in ecoacoustics: A roadmap towards versatile soundscape information retrieval 2019:236–47. https://doi.org/10.1002/rse2.141.
[21] Lin TH. Improving acoustic monitoring of biodiversity using deep learning-based source separation algorithms n.d.
[22] Mishachandar B, Vairamuthu S. An underwater cognitive acoustic network strategy for efficient spectrum utilization. Appl Acoust 2021;175:. https://doi.org/10.1016/j.apacoust.2020.107861107861.
[23] Bittle M, Duncan A. A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring. 2013.
[24] Gillespie D, Caillat M, Gordon J, White P. Automatic detection and classification of odontocete whistles a) 2013;134:2427–37. https://doi.org/10.1121/1.4816555.

[25] Putland RL, Ranjard L, Constantine R, Radford CA. A hidden Markov model approach to indicate Bryde ' s whale acoustics. Ecol Indic 2018;84:479–87. https://doi.org/10.1016/j.ecolind.2017.09.025.
[26] Jarvis, S., DiMarzio, N., Morrissey, R., & Moretti, D. (2008). A novel multi-class support vector machine classifier for automated classification of beaked whales and other small odontocetes. *Canadian Acoustics*, 36(1), 34–40.
[27] Jiang J, Bu L, Wang X, Li C, Sun Z. Clicks classification of sperm whale and long-finned pilot whale based on continuous wavelet transform and artificial neural network 2018;141:26–34. https://doi.org/10.1016/j.apacoust.2018.06.014.
[28] Mellinger DK, Clark CW. Recognizing transient low-frequency whale sounds by spectrogram correlation 2014;107:3518–29.
[29] Dreo R, Boudraa A, Denis S. Antarctic blue whale calls detection based on an improved version of the stochastic matched filter 2017:2319–23.
[30] Park H, Yoo CD, Member S. CNN-based learnable gammatone filterbank for environmental sound classification n.d.:1–5.
[31] Domingo MC. An overview of the internet of underwater things. J Netw Comput Appl 2012;35:1879–90. https://doi.org/10.1016/j.jnca.2012.07.012.
[32] Abdoli S, Cardinal P, Koerich AL. End-to-end environmental sound classification using a 1D convolutional neural network 2019: 1–24.
[33] Salamon J, Bello JP. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification 2017;9908:1–5. https://doi.org/10.1109/LSP.2017.2657381.
[34] Guzhov A, Raue F, Gmbh D. ESResNet: Environmental Sound Classification Based on Visual Domain Models n.d.
[35] Abbas M, Albadr A, Tiun S, Al-dhief FT, Mahmoud A, Sammour M. Spoken language identification based on the enhanced self-adjusting extreme learning machine approach 2018:1–27.
[36] Maccagno A, Mastropietro A, Mazziotta U. A CNN approach for audio classification in construction sites A CNN approach for audio classification in construction sites, 2019.
[37] Martin B, Kowarski K, Gaudet B. Marine mammal species classification using convolutional neural networks and a novel acoustic representation n. d.:1–16.
[38] Stowell D, Wood MD, Pamuła H, Stylianou Y, Glotin H, Stowell D. Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge 2019;2019:368–80. https://doi.org/10.1111/2041-210X.13103.
[39] Thomas M, Martin B, Kowarski K, Gaudet B, Matwin S. Detecting endangered baleen whales within acoustic recordings using R-CNNs 2019:1–5.
[40] Rathi D, Jain S, Indu S. Underwater fish species classification using convolutional neural network and deep learning. Ninth Int Conf Adv Pattern Recognit 2017;2017:1–6.
[41] Nanaware S, Shastri R, Joshi Y, Das A. Passive Acoustic Detection and Classification of Marine Mammal Vocalizations 2014:493–7. https://doi.org/10.1109/ICCSP.2014.6949891.
[42] This R, Attribution-noncommercial-noderivs CC, By-nc-nd CC, If T, Rose W. Whistle Detection and Classification for Whales Based on Convolutional Neural Networks 2019.
[43] Seger KD, Al-badrawi MH, Miksis-olds JL, Kirsch NJ, Lyons AP, Kirsch NJ, et al. marine mammal vocal signals An Empirical Mode Decomposition-based detection and classification approach for marine mammal vocal signals 2018; 3181. https://doi.org/10.1121/1.5067389.
[44] Sayigh L, Daher MA, Allen J, Gordon H, Joyce K, Stuhlmann C, et al. Fourth international conference on the effects of noise on aquatic life the watkins marine mammal sound database: An online , freely accessible resource 2019;040013. https://doi.org/10.1121/2.0000358.
[45] The University of Rhode Island. Discovery of Sound in the Sea (DOSITS) project, https://dosits.org/.
[46] Towsey M, Znidersic E, Broken-brow J, Indraswari K. Long-duration , false-colour spectrograms for detecting species in large audio data-sets 2018. https://doi.org/10.22261/JEA.IUSWUI.
[47] Towsey M, Zhang L, Cottman-fields M, Wimmer J. Visualization of long-duration acoustic recordings of the environment visualization of long - duration acoustic recordings of the environment. Proc Proc Comput Sci 2014;29:703–12. https://doi.org/10.1016/j.procs.2014.05.063.