KSOE

Journal of Ocean Engineering and Technology 39(1), 92−102, February, 2025 eISSN 2287-6715
https://doi.org/10.26748/KSOE.2024.069

Original Research Article

Real-Time Acoustic Signal Classification Using RNN for Underwater Cutting Process Monitoring and Situational Awareness

Seunghyun Pyo[1,5], Tae-Kyeong Yeu[2], Yeongjun Lee[3], Jong-Boo Han[4,7], Yujin Cho[8] and Daegil Park[4,6]

[1]*Graduate student, Department of Ocean Engineering, University of Science & Technology (UST), Daejeon, Korea*
[2]*Principal researcher, Korea Research Institute of Ships & Ocean Engineering (KRISO), Daejeon, Korea*
[3]*Senior Engineer, Korea Research Institute of Ships & Ocean Engineering (KRISO), Daejeon, Korea*
[4]*Senior researcher, Korea Research Institute of Ships & Ocean Engineering (KRISO), Daejeon, Korea*
[5]*UST collaborative researcher, Korea Research Institute of Ships & Ocean Engineering (KRISO), Daejeon, Korea*
[6]*Associate professor, Department of Ocean Engineering, University of Science & Technology (UST), Daejeon, Korea*
[7]*Assistant professor, Department of Ocean Engineering, University of Science & Technology (UST), Daejeon, Korea*
[8]*Associate engineer, Department of Naval & Special Ship Design Coordination, Hyundai Heavy Industries, Ulsan, Korea*

KEYWORDS: Audio classification, Underwater operation, Acoustic feature extraction, Recurrent neural network (RNN), Long short term memory (LSTM)

ABSTRACT: *Seabed crushing, a critical underwater operation for mineral resource extraction and infrastructure construction, necessitates acoustic monitoring due to the limited visibility caused by debris generated during the process. This study proposes an acoustic classification method to enable real-time monitoring of underwater robotic operations. Given the acoustic characteristics of crushing operations, which predominantly manifest in the low-frequency band, acoustic features were extracted and processed using a deep learning model to classify the operational states into four categories: idling, cutting, hard cutting, and base. The model was developed using a recurrent neural network (RNN), which is particularly suited for real-time time-series data processing. The classification performance of long short-term memory (LSTM) networks and standard RNN models was systematically evaluated. Training on a dataset collected from crushing operations conducted on land, the LSTM model achieved an accuracy of 89%, outperforming the RNN, which achieved 84%. Furthermore, real-time operational state prediction was performed at a speed of 10 Hz, demonstrating high accuracy. These findings indicate that the proposed method effectively enables real-time classification of seabed crushing operations, thereby enhancing the safety and efficiency of remote underwater operations.*

1. Introduction

With the growing interest in marine resources and their strategic importance for military applications, underwater infrastructure construction and resource extraction technologies have garnered significant attention, leading to an increased demand for remote underwater operations. Among these, seabed crushing operations, which are essential for underwater mineral resource extraction and infrastructure development, involve flattening the seabed and constructing foundational structures by operating manipulators mounted on underwater robots.

However, remote control of seabed crushing presents considerable technological challenges. Operators must rely solely on data from cameras and sound navigation and ranging (SONAR) systems mounted on the robots, as visibility in underwater environments is often compromised by turbidity. During operations such as seabed crushing and thruster maneuvering, suspended sediments frequently render visibility completely unachievable, necessitating interruptions until the sediments settle. These extreme remote working conditions not only slow the pace of operations but also contribute to frequent accidents, including machinery damage caused by excessive operational demands.

To address these challenges, this study proposed a cyber-physical operation system (CPOS) (Yeu et al., 2023). This system is designed to ensure robust performance under varying environmental and operational conditions by employing real-time analysis of physical models based on dynamics and inverse dynamics. It integrates information about the robot

Received 19 August 2024, revised 11 October 2024, accepted 25 November 2024
Corresponding author Daegil Park: +82-42-866-3865, daegilpark@kriso.re.kr

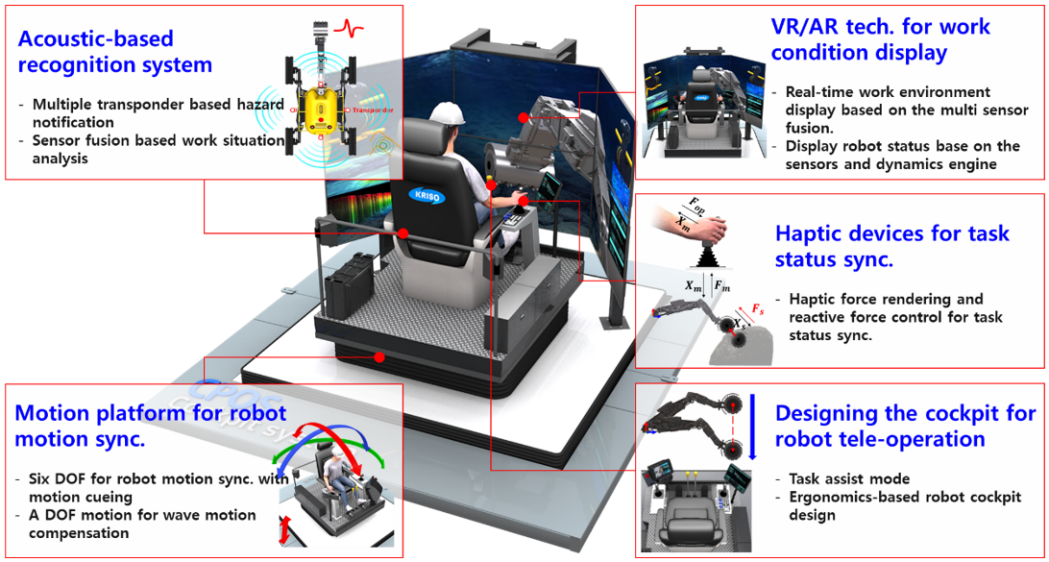


Fig. 1 Development of a haptic-based robot environment awareness and control framework for synchronization task environments and situations

and its surrounding environment, collected from onboard sensors, into a digital environment. Additionally, a human-in-the-loop simulation (HILS) environment was developed to effectively convey the operational status and work environment of the robot to the operator, creating a setting in which operators can control robots as though they were physically present at the worksite. Specifically, an HILS environment tailored for underwater vehicles (hereafter referred to as "rovers"), a representative category of remotely operated underwater robots, was developed, and a cognitive and control framework for this environment was designed (Fig. 1).

As a core component of the HILS framework, this study proposed a real-time operational judgment technology based on recurrent neural networks (RNNs). This technology classifies and communicates work-related sounds to the operator in real time, aiming to prevent work-related accidents and ensure consistent operational performance (Fig. 2). The proposed system categorizes work states into four classes: no load (idling), working (cutting), excessive damage risk (hard cutting), and other conditions (base). Acoustic characteristics associated with these states, which frequently recur during crushing operations, were analyzed and modeled. Mel frequency cepstral coefficients (MFCCs) were employed to enhance the acoustic feature extraction process, while deep learning models based on long short-term memory (LSTM) networks—a prominent technique for processing time-series acoustic data—were used for learning and classification. A comparative analysis with standard RNN models was also conducted to evaluate the real-time judgment performance of the proposed algorithm.

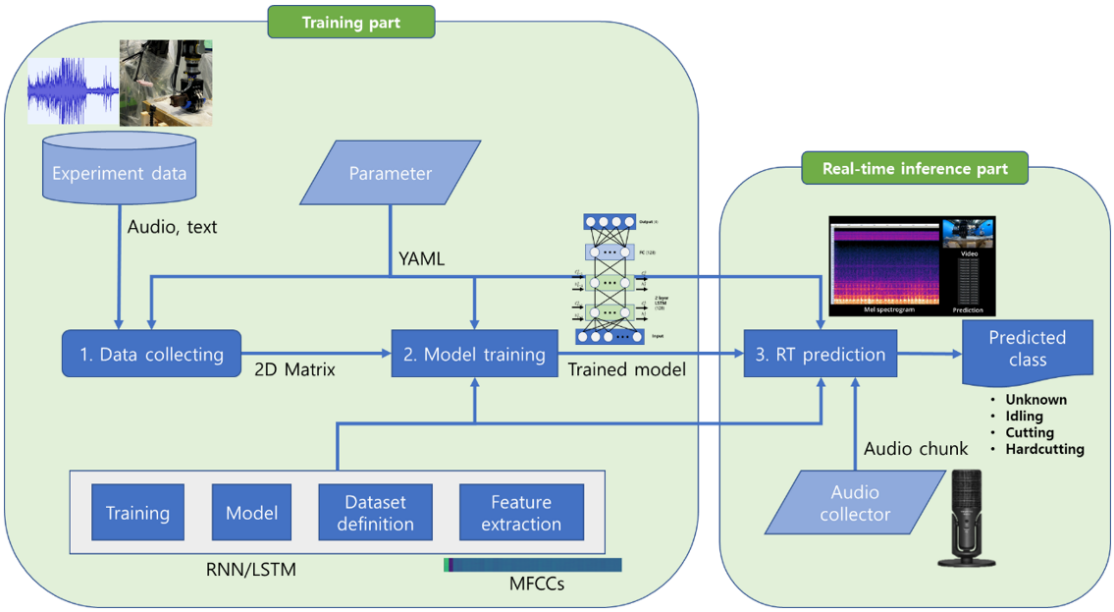


Fig. 2 Schemes diagram of deep learning-based audio classification for underwater cutting situational awareness of CPOS

The structure of this paper is as follows: Chapter 2 reviews related research on audio classification using deep learning techniques. Chapter 3 details the proposed deep learning-based real-time audio classification methodology using MFCCs and LSTM networks. Chapter 4 describes the experimental setup for validating the proposed algorithm, while Chapter 5 presents the experimental results and discussion.

2. Related Works

2.1 Deep Learning-Based Audio Classification

Research on the application of deep learning to audio classification has predominantly focused on convolutional neural networks (CNNs) since 1995, when Yann LeCun achieved success in voice recognition using CNNs (LeCun and Bengio, 1995). CNN-based audio classification typically involves converting audio data into spectrograms, which represent the data in the frequency domain and along the time axis. These spectrograms are input into CNNs, which extract features with high accuracy and speed. However, CNNs have limitations in capturing the temporal characteristics inherent in audio data. In 2006, Graves et al. introduced a method for learning temporal characteristics using recurrent RNNs for voice recognition and classification (Graves et al., 2006). While RNNs are capable of processing time-series data, their ability to learn temporal characteristics diminishes when dealing with longer sequences, such as those encountered in voice recognition tasks. To address these limitations, improved RNN-based models were developed, including LSTM networks and Gated Recurrent Units (GRUs). These models demonstrated enhanced capability in processing longer time-series data. For instance, Yu et al. (2020) utilized bidirectional RNNs, which process hidden states in both forward and backward directions, in combination with attention mechanisms to classify music genres. Similarly, Gan (2021) employed bidirectional LSTMs for music feature classification. While CNNs and the RNN family of models are primarily supervised learning approaches, where correct answers are explicitly provided during training, unsupervised learning methods have also been explored. For example, autoencoders, including those connected to variational autoencoders or multi-layer perceptrons, have been applied to learn and classify data features without labeled data. These approaches have been employed to address dataset imbalance issues through data augmentation (Saldanha et al., 2022) or to generate spectrograms for robust feature extraction (Huang et al., 2022).

2.2 Construction Site Audio Classification

Among the various applications of deep learning-based audio classification, construction sites present acoustic environments similar to the crushing operations examined in this study. Scarpiniti et al. (2021) conducted research that employed CNNs and LSTM networks, an improvement over traditional RNNs, for the classification of heavy equipment sounds at construction sites. This study utilized a variety of acoustic feature extraction methods and achieved an average accuracy of 95% in distinguishing between different manufacturers of the same type of heavy equipment. The methodology involved collecting features extracted through diverse approaches and inputting them into CNNs as images and into RNNs as sequences (Scarpiniti et al., 2021).

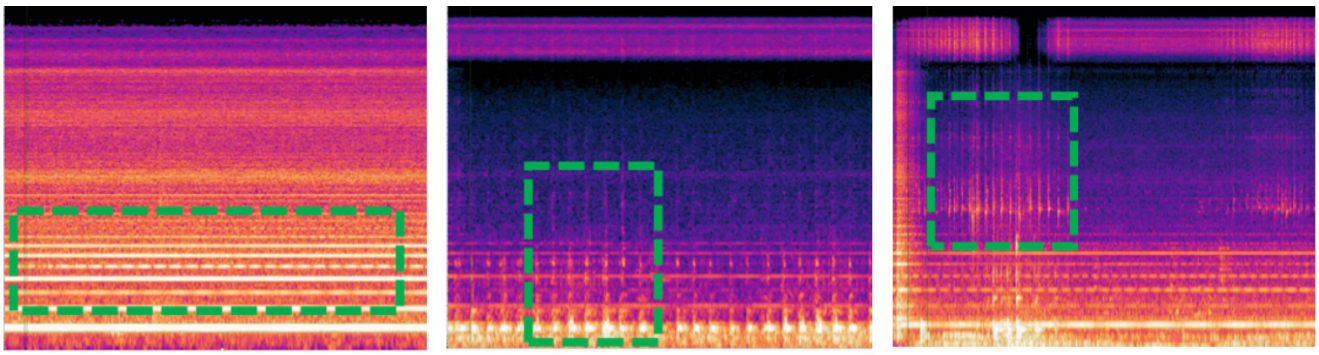
The present research builds on audio classification techniques used in construction site contexts to classify repetitive work noise. Its objective was to enable situational assessment of remote crushing operations conducted in underwater environments. The study focused on robust classification of sounds from work sites, which vary in real time, using deep learning-based audio classification techniques. To extend the applicability of audio classification, a real-time pipeline was designed to process crushing sounds, classify operational scenarios, and validate the approach through ground-based crushing experiments.

3. Methods

3.1 Audio Feature Extraction with Cutting Operation's Audio Characteristics

Understanding the features of the data being processed is a critical aspect of problem-solving using deep learning. In the context of audio classification for this study, comprehending the characteristics of the collected sounds is essential, as it directly influences the extraction of appropriate acoustic feature vectors and the training of models. The acoustic features of the crushing operation, which serve as the target for classifying operational states, can be categorized into noise generated during idle operation and noise produced during active crushing. The idle-state noise primarily includes sounds originating from the hydraulic power unit (HPU) and the crushing tools, while the active crushing noise is limited to the interaction between the crushing tools and the work material, which in this study is gypsum. The HPU functions as a hydraulic supply unit, powering the hydraulic devices utilized by the robots in the CPOS. Since HPUs are commonly employed in heavy construction equipment, and the crushing sound closely resembles noises typically observed at construction sites, the methodology for this study was primarily informed by techniques used in construction site audio classification. Examination of the noise in the robot's idle state (Fig. 3(a)) through the Mel spectrogram revealed continuous signals in the 500, 700, and 1,100 Hz frequency bands. In contrast, the noise generated during the crushing operation (Fig. 3(b)) exhibited a distinct signal at 900 Hz, recurring at intervals of 53 to 57 ms. When strong vibrations occurred during crushing, signals were observed across a broad frequency range, particularly above 2,000 Hz, as illustrated in Fig. 3(c).

Acoustic signals, as one-dimensional time-series data, require the extraction of feature vectors to enable input into deep learning model classifiers. This process involves transforming the signals to reflect the acoustic features described above. The feature vector extraction methodology, illustrated in Fig. 4, is primarily divided into two stages. In the first stage, the amplitudes of the input acoustic signals are normalized and standardized. This step addresses variations in the distance and direction of the sound source relative to the sound collection device, ensuring consistency in the input data. In the second stage, the audio samples are converted into feature vectors using an acoustic



(a) During the idle state (b) Cutting state (c) Hard cutting with large oscillation that confirmed in visual check

Fig. 3 Mel spectrogram of audio data from the cutting experiment

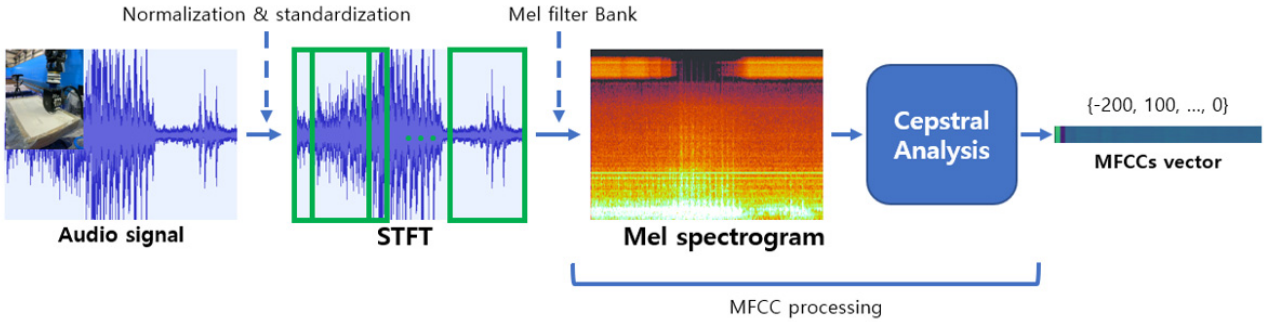


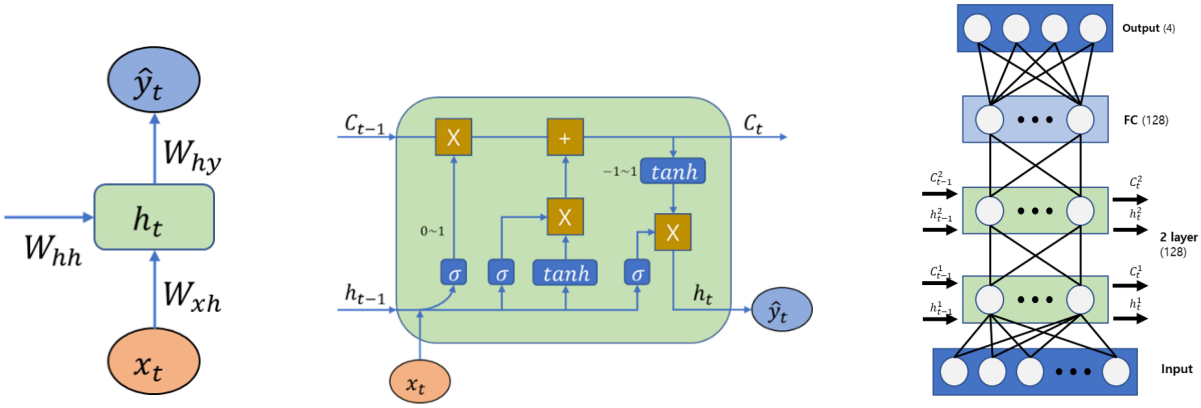
Fig. 4 Diagram of process for extracting acoustic feature vector from audio signal

feature vector extraction algorithm. Given that the acoustic features of the crushing operation are distinguishable by the human ear, MFCCs were employed for feature extraction (Majeed et al., 2015). MFCCs leverage the Mel scale (Eq. (1)), which enhances features in the audible frequency range by representing low frequencies on a linear scale and high frequencies on a logarithmic scale. In Eq. (1), m denotes the mel frequency and f is represents frequency. For this study, the parameters were configured to optimize frequency resolution and capture signals generated during the crushing operation. Specifically, the number of cepstral coefficients was set to 40, the window length for the fast Fourier transform (FFT) was defined as 2048, and the hop length was set to 512. Based on this, the frequency resolution was increased to include the signals generated during crushing.

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \quad (1)$$

3.2 Deep Learning Model Configuration for Real-Time Audio Classification

Given the distinct frequency-band features of the crushing sound described above, audio classification can theoretically be performed using traditional machine learning models such as support vector machines. However, the acoustic features mentioned were derived from sounds collected in terrestrial environments and cannot adequately account for variations introduced by underwater environmental factors, such as water temperature and pressure (Wenz, 1972). To address this limitation and construct models and pipelines capable of robustly classifying sounds under underwater conditions, deep learning models were employed in this study. Among machine learning approaches designed to process data-driven information, RNNs and transformer models are commonly used for time-series data. In this study, an RNN model was selected due to its lower computational cost, which aligns with the requirements of real-time operations in underwater crushing machinery. The RNN model, illustrated in Fig. 5(a), incorporates a hidden layer between input and output, enabling it to retain and reflect prior outputs in subsequent inputs. This characteristic makes it well-suited for capturing temporal contexts in real-time audio classification tasks. However, RNNs suffer from the issue of gradient vanishing, which prevents the effective propagation of initial input information to later stages in the sequence. To mitigate this problem, LSTM networks, depicted in Fig. 5(b), were employed. LSTM networks address this limitation by incorporating a cell structure to store long-term memory and gate structures to manage short-term memory. In typical speech processing applications using RNNs, signals are segmented into 20–40 ms intervals for analysis (Paliwal et al., 2011). However, in this study, longer sound segments were used, with audio divided into 100-ms intervals to ensure that signals with an average duration of 55 ms were captured. In this study, the RNN and LSTM models were trained on crushing sound data, and their classification performances were analyzed to evaluate whether LSTM provided superior accuracy compared to RNN for this specific application. The deep learning models were designed to classify task situations using audio data as input, as shown in Fig. 5(c). To enhance the classification



(a) Basic RNN (b) LSTM (c) Total model architecture

Fig. 5 Schemes diagram of deep learning models which process time series data

of similar task situations, such as standard cutting and hard cutting, two RNN layers, each comprising 128 hidden units, were stacked. Similarly, a second model with two LSTM layers was constructed. For each model, a classification layer followed the RNN or LSTM layers. This classification layer was designed to predict task situations using a fully connected layer and a Softmax output layer.

3.3 Audio Classification Pipeline

In this study, a three-stage real-time prediction pipeline was developed. The first stage involved generating a dataset comprising audio files collected during the experiment and their corresponding class labels. In the second stage, an untrained deep learning model, along with learning functions and feature extraction functions, was imported, and training was conducted using the previously generated dataset. In the final stage, the audio stream, formatted in Pulse Code Modulation (PCM), was stored as a buffer within the sound collection device. The device's sampling rate was set to 44,100 Hz, and audio samples spanning 100 ms were input into the trained model to determine the predicted class. During this process, components such as deep learning models, learning methods, and feature extraction techniques were designed as modular elements to allow for future research expansion (Fig. 6). The pipeline's learning functions were implemented using PyTorch, a Python-based deep learning library. Additionally, the PyAudio library was utilized for acoustic feature extraction and the real-time collection of audio data.

4. Experiment

4.1 Experiment Setup

The crushing experiment conducted to construct the dataset utilized a gypsum plate measuring 1.06 × 0.66 m^2. During the operation of the crusher at 100 rpm, the vertical cutting operation was configured to a depth of 5 mm, while the horizontal cutting operation was executed such that the manipulator's endpoint moved at a rate of 0.1 to 0.2 mm/s. The generated crushing sound was recorded in the MP3 format at a sampling rate of 44,100 Hz using a GoPro™ 8 camera, as a dedicated sound collection device could not be installed during the experiment. The recorded audio was subsequently decoded into the PCM format for use in model training. The experiments were conducted nine times, with horizontal cutting (Fig. 7(d)) performed after vertical cutting (Fig. 7(c)). Each experiment lasted approximately 6 min, resulting in a total of 57.42 min of audio data. The collected audio data were divided into 34,452 segments, each with a duration of 100 ms, to align with the audio frame required for real-time prediction. A two-dimensional dataset was generated by associating these segments with their respective situational

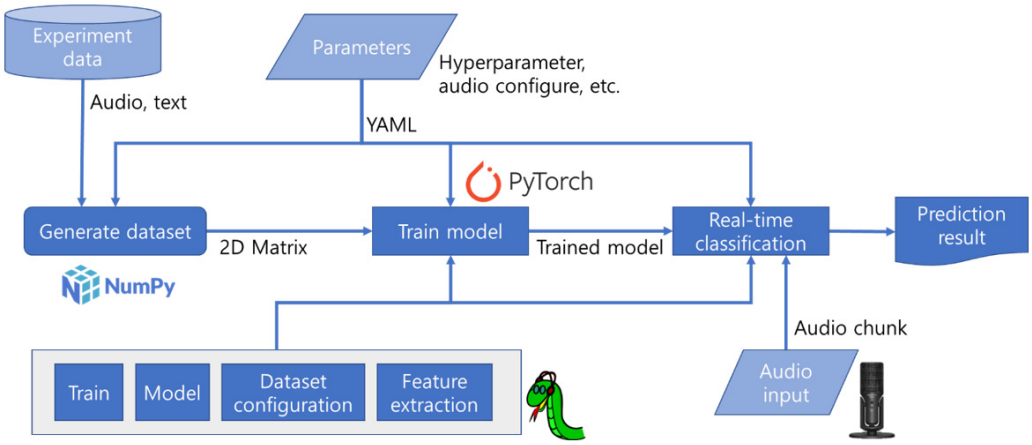
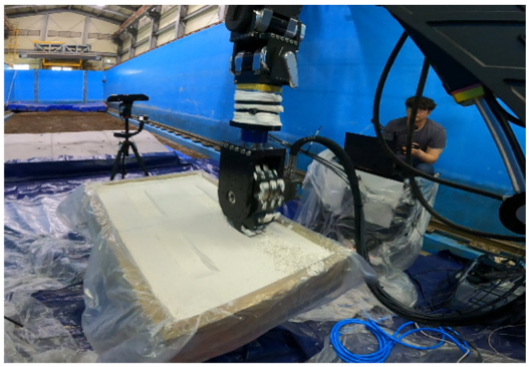
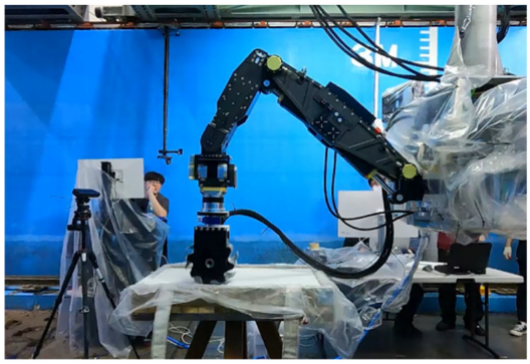


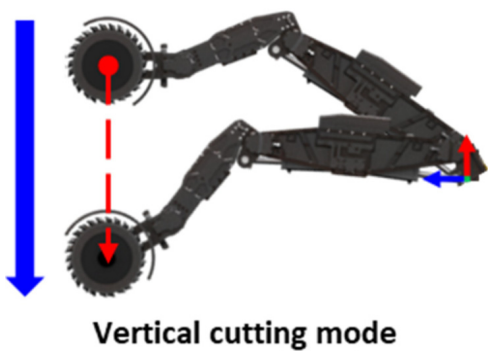
Fig. 6 Diagram of whole pipeline that collects dataset and classify situational awareness in real time



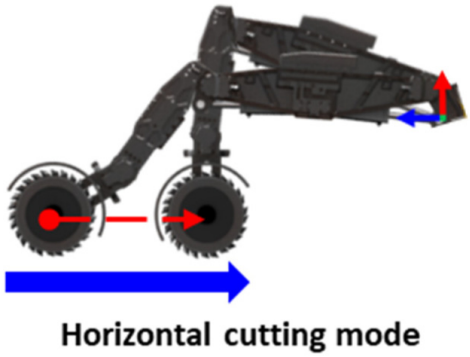
(a) X Gypsum plate view



(b) Side view of experiment



(c) Diagram of gesture and coordinate frame of vertical cutting operation



(d) Diagram of gesture and coordinate frame of horizontal cutting operation

Fig. 7 Ground experiment environment of gypsum crushing

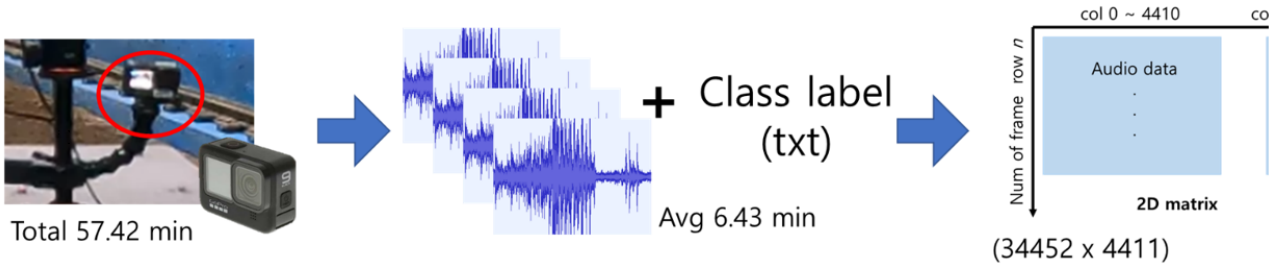
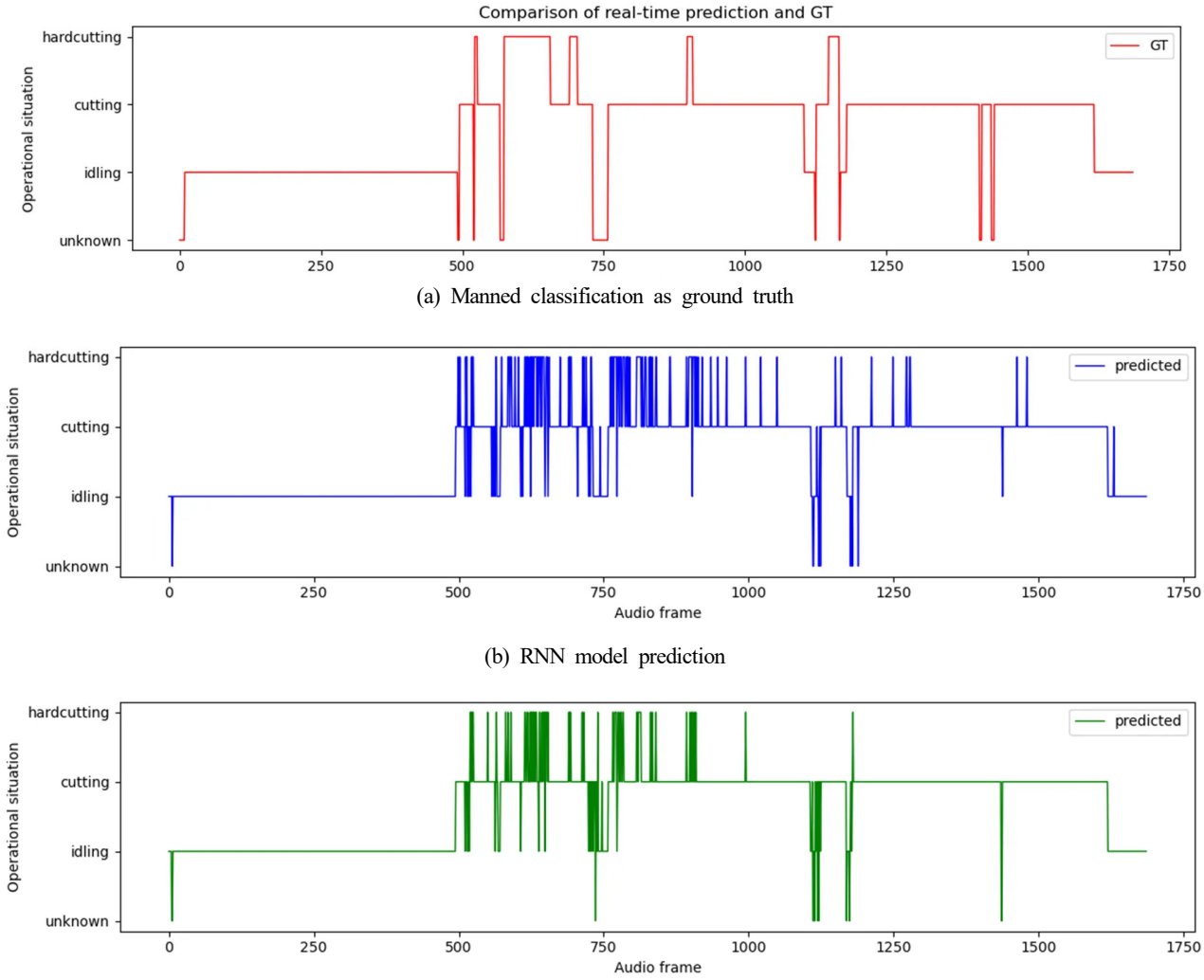


Fig. 8 Diagram of processing real cutting operation audio dataset with class label.

judgment class information (Fig. 8). The dataset was partitioned into training, validation, and testing sets in a 6:2:2 ratio for model training. The architecture of the deep learning models used for this study is depicted in Fig. 5(c). To enable a direct comparison between RNN and LSTM models, the structural parameters, including the number of hidden layers and nodes within these layers, were kept consistent. Two model variants were created by altering the node type to either RNN or LSTM. For a fair comparison of model performance, the batch size, representing the number of input data instances processed simultaneously during training, was set to 128, while the epoch, defining the number of iterations over the entire dataset, was set to 10 for both models. Given that the deep learning models must perform operations within the canister mounted on the CPOS rover, they cannot rely on graphics cards. Therefore, model training was conducted in an environment featuring an Intel 13th Gen i9-13900K™ CPU and 64 GB of RAM.

4.2 Experiment Result

The developed algorithm was determined to operate in real time, as it could classify task situations at a rate of 10 Hz when 100 ms of acoustic data were input. Audio data, not used during the training process, were recorded using a Sennheiser™ Profile microphone. The results of classifying situations by both a human operator and the trained model were then compared. In this case, the microphone was configured to collect data in the PCM format at a sampling rate of 44,100 Hz, consistent with the settings used for collecting training data. These results are illustrated in the graph in Fig. 9. For the manual classification of operational situations (Fig. 9(a)), the results predicted by the trained RNN model (Fig. 9(b)) achieved an accuracy of 84%, while the results from the LSTM model (Fig. 9(c)) achieved an accuracy of 88%. The prediction process for both models required a total operation time of 5 ms, which included 4 ms for extracting acoustic features and 1 ms for



(a) Manned classification as ground truth

(b) RNN model prediction

(c) LSTM model prediction

Fig. 9 Comparison of real-time cutting operation situation

predicting the class. This confirmed that predictions could be performed at a maximum frequency of 200 Hz. The learning process for both the RNN and LSTM models was represented by the average loss (Figs. 10(a) and 10(c)) and the average accuracy for the test set (Figs. 10(b) and 10(d)) plotted over each epoch, showing the performance after one out of ten epochs. The performance metrics for the models' training are presented in Tables 1 and 2 for the RNN and LSTM models, respectively. In this study, data imbalance and the influence of the base class were prominently observed.First, the models exhibited the lowest performance for the hard cutting class, as it had the smallest amount of corresponding data. Thus, the models struggled to learn this class effectively. In contrast, for the cutting class, the precision, recall, and F1 score (the harmonic mean of precision and recall) all showed high values. This was due to the abundance of training data, with 4,333 frames available for learning. Second, for the base class, the precision and recall values were relatively lower, with scores of 78 and 66, respectively, for the LSTM model. This is likely because the base class did not belong to any of the three primary classes, leading to misclassifications where other classes were incorrectly predicted. The confusion matrix derived from the test set (Fig. 11) confirmed these predictions for the base class.

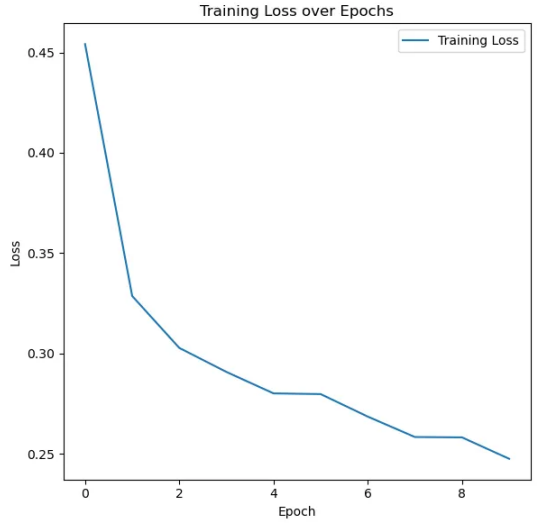
Table 1 RNN training result metrics with the collected dataset

Class	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Num
Base	88	76	49	60	938
Idling		78	94	85	1485
Cutting		95	97	96	4333
Hard cutting		67	29	40	135

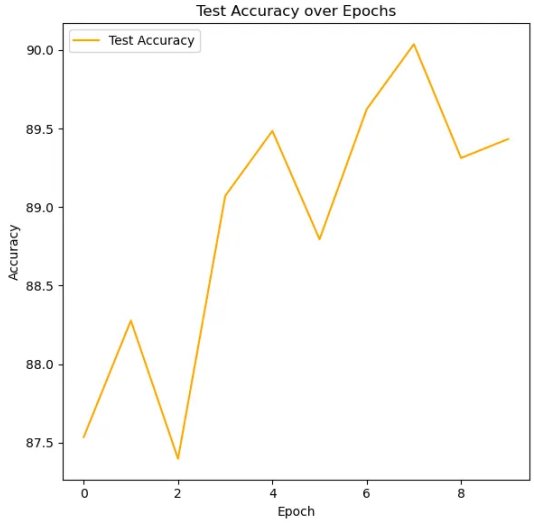
Table 2 LSTM training result metrics with the collected dataset

Class	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Num
Base	91	78	66	71	938
Idling		89	89	89	1485
Cutting		95	98	96	4333
Hard cutting		64	59	61	135

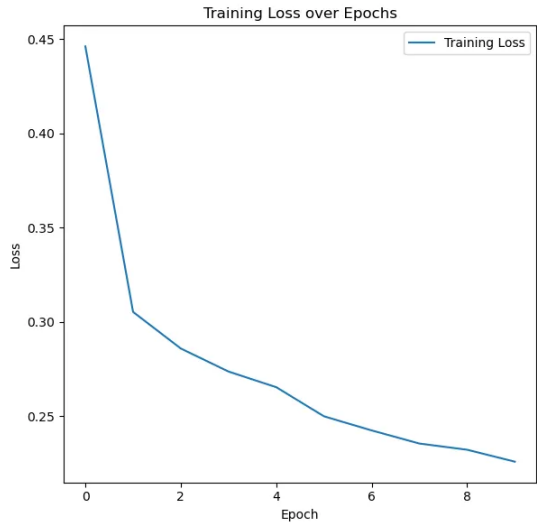
As previously described in section 3.2, the LSTM model is known to



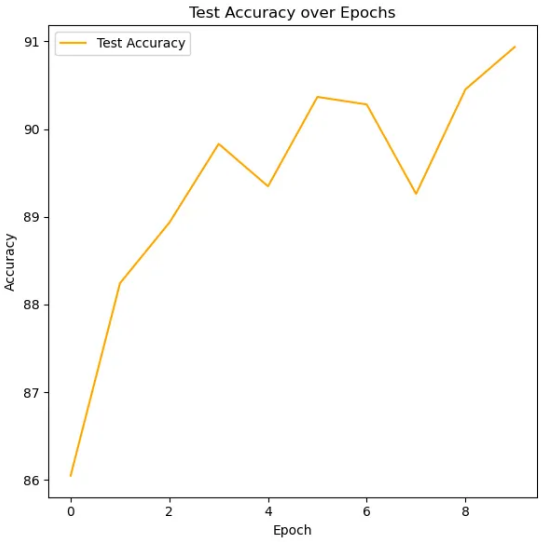
(a) Training loss over epochs of RNN



(b) Test accuracy over epochs of RNN

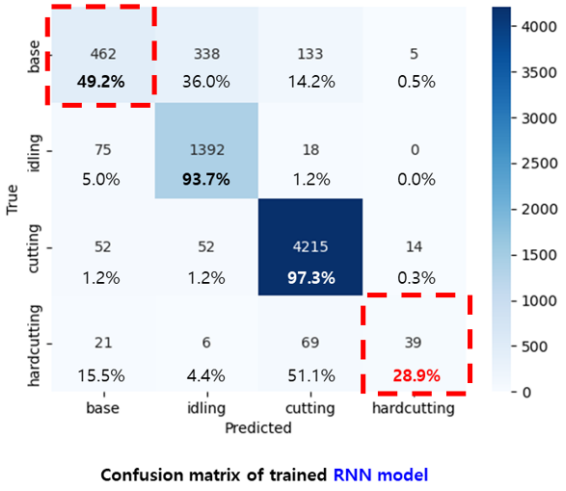


(c) Training loss over epochs of LSTM

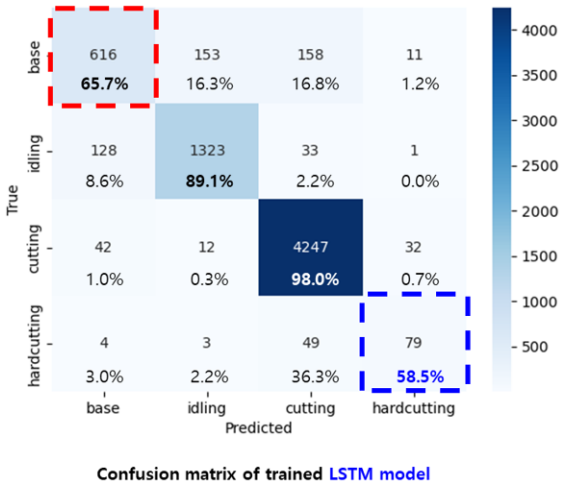


(d) Test accuracy over epochs of LSTM

Fig. 10 Comparison of RNN and LSTM model performance in training loss and test accuracy



(a) Trained RNN model



(b) Trained LSTM model

Fig. 11 Confusion matrix of RNN, LSTM model with test dataset

demonstrate superior performance compared to the conventional RNN. To verify whether this holds true for the actual learning results, we compared the performance of both models using the same hyperparameters, model structure, and data as outlined in section 4.1. The comparison was conducted through the confusion matrix (Fig. 11), and the precision, recall, and F1 score classification performance evaluation tables were analyzed for both the RNN (Fig. 11(a), Table 1) and LSTM (Fig. 11(b), Table 2). The RNN displayed distinct characteristics when compared to the LSTM model, and similar performance was observed for the idling and cutting classes, which had the largest amount of training data. However, for the hard cutting class, which is similar to the cutting class, the RNN demonstrated lower performance in both precision and recall than the LSTM, thereby confirming that the RNN is not suitable for the classification task in this study.

5. Discussions

This study successfully classified task situations with higher accuracy using LSTM; however, the constructed dataset contains insufficient data for the hard cutting class, as it primarily focuses on crushing and idling. Additionally, since the audio data were extracted from video acquired through a GoproTM 8 camera, the camera had already performed some audio preprocessing (e.g., noise cancellation), meaning that the actual on-site sounds were not fully represented for the four task situations. To address these limitations, future work will involve acquiring audio data through additional ground experiments aimed at mitigating the data imbalance, as well as installing a microphone at the crushing position to collect sound data directly. Furthermore, the audio classification performed in this study relied on supervised learning, in which humans directly define the task situations. This approach introduces an element of subjectivity, as classification is based on the judgment of the individual who provides the correct answer. For example, cutting and hard cutting are similar situational classes. Since excessive vibration during crushing operations can damage machinery, the definitions used in this study were based on observations of audio data in situations where excessive vibration occurred. However, as there are various crushing situations depending on the location and the target, the occurrence of excessive vibration is often based on the empirical judgment of the operator. In fact, the judgment of hard cutting differs between individuals who have performed the crushing operation and those who have not, making it challenging to establish objective class definitions and resulting in the presence of boundary conditions. This ambiguity in class definitions is further illustrated by the greater variability observed in the classification of cutting and hard cutting compared to other classes in the real-time audio classification results (Fig. 9). These findings also confirm the superiority of LSTM over RNN, particularly with respect to precision and recall, as discussed in section 4.2, in the context of the real-world experiment. Moreover, the performance and characteristics of the LSTM model present a promising breakthrough in audio classification tasks involving boundary conditions. However, due to the lack of audio data captured at the moment when machinery damage occurs (e.g., damage to crushing blades or failure of the crusher motor), it is clear that direct damage situations cannot be classified using audio data alone. To address this limitation, future research should explore multi-input learning (Fig. 12), which combines acoustic information

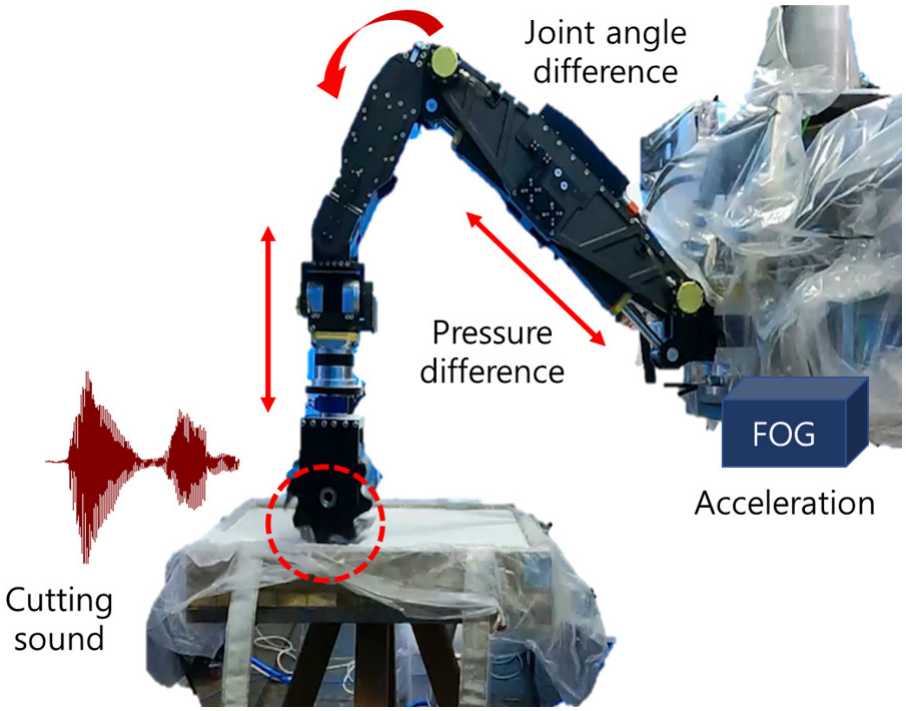


Fig. 12 The configuration of physical and acoustic information for a multi-input deep learning model designed to enhance cutting operational situation awareness.

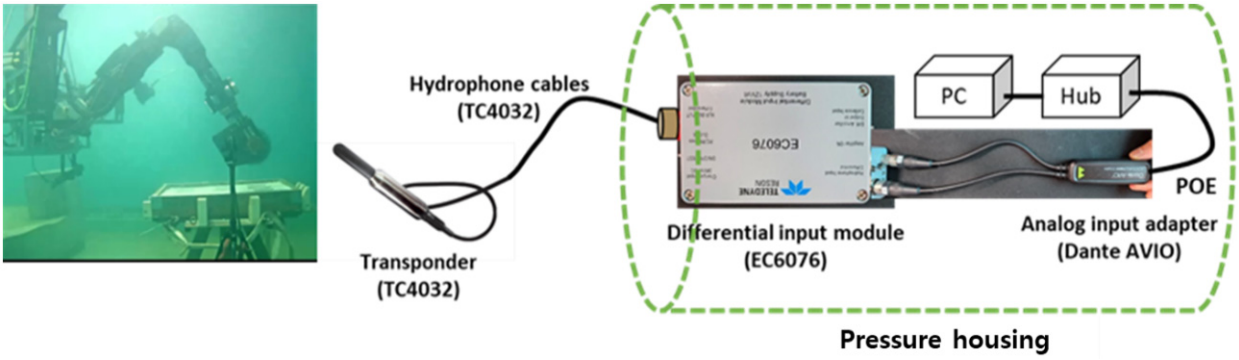


Fig. 13 Diagram of data collection in an underwater cutting operation.

with physical data, such as differential pressure from the hydraulic manipulator joint, joint angle, and acceleration data obtained from an inertial measurement unit (IMU).

6. Conclusions

This research focused on real-time audio classification to enable remote workers performing crushing operations in underwater environments to assess task situations under conditions of poor visibility. Acoustic feature vectors were extracted from the original data provided by an acoustic input device and were subsequently input into deep learning models to predict classification results for task situations. When RNN and LSTM models, which are commonly used for processing time-series data, were trained using the same dataset and their classification performance was compared, the results demonstrated that the LSTM model was more effective in classifying crushing operations.

The study was conducted using crushing operation experiments performed on land prior to underwater operations. As the effects of sound speed, which is influenced by air pressure and temperature on land compared to underwater conditions, are reflected in the data, it will be necessary to account for the characteristics of the underwater environment, including the experimental setup, when acquiring sounds using a hydrophone in future underwater experiments (Fig. 13).

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Funding

This work was supported by “the marine robot cyber physical operation system (CPOS) key technology development project” funded by the Korea Research Institute of Ships and Ocean Engineering (PES5200), and Korea Research Institute of Ships and Ocean engineering a grant from Endowment Project of “Digital platform development to support marine digital transformation” funded by Ministry of Oceans and Fisheries (2520000291, PES5581).

References

Demir, F., Turkoglu, M., Aslan, M., & Sengur, A. (2020). A new pyramidal concatenated CNN approach for environmental sound classification. *Applied Acoustics, 170*, 107520. https://doi.org/10.1016/j.apacoust.2020.107520

Gan, J. (2021). Music feature classification based on recurrent neural networks with channel attention mechanism. *Mobile Information Systems, 2021*(1), 7629994. https://doi.org/10.1155/2021/7629994

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)* (pp. 369–376). https://doi.org/10.1145/1143844.1143891

Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., & Feichtenhofer, C. (2022). Masked autoencoders that listen. In *the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*. https://doi.org/10.48550/arXiv.2207.06405

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time-series. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks*, *3361*. MIT Press.

Majeed, S. A., Husain, H., Samad, S. A., & Idbeaa, T. F. (2015). Mel frequency cepstral coefficients (Mfcc) feature extraction enhancement in the application of speech recognition: A comparison study. *Journal of Theoretical & applied Information Technology, 79*(1), 38–56.

Paliwal, K., Lyons, J., & Wojcicki, K. (2011). Preference for 20–40 ms window duration in speech analysis. *Proceedings of the 4th International Conference on Signal Processing and Communication Systems (ICSPCS'2010)*, 1–4. https://doi.org/10.1109/ICSPCS. 2010.5709770

Saldanha, J., Chakraborty, S., Patil, S., Kotecha, K., Kumar, S., & Nayyar, A. (2022). Data augmentation using Variational Autoencoders for improvement of respiratory disease classification. *PLoS One, 17*(8), e0266467. https://doi.org/10.1371/journal.pone.0266467

Sang, J., Park, S., & Lee, J. (2018). Convolutional recurrent neural networks for urban sound classification using raw

waveforms. In *Proceedings of the 26 th European Signal Processing Conference (EUSIPCO)* (pp. 2444–2448). https://doi.org/10.23919/EUSIPCO.2018.8553247

Scarpiniti, M., Comminiello, D., Uncici, A., & Lee, Y.-C. (2021). Deep fecurrent neural network for audio classification in construction sites. In *2020 8th European Signal Processing Conference (EUSIPCO)* (pp. 810–814). https://doi.org/10.23919/Eusipco47968. 2020.9287802

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *the 31st Conference on Neural Information Processing Systems (NIPS 2017).* https://doi.org/10.48550/arXiv.1706.03762

Wenz, G. M. (1972). Review of underwater acoustics research: Noise. The Journal of the Acoustical Society of America, 51, 1010–1024. https://doi.org/10.1121/1.1912921

Yeu, T. K., Han, J. B., Lee, Y. J., Park, D. G., Kim, S. S., & Hong, S. (2023). Preliminary study on develpment of CPOS (Cyber physical operation system) for underwater robots. *OCEANS 2023 - Limerick*, 1–4. https://doi.org/10.1109/OCEANSLimerick52467.2023 .10244254

Yu, Y., Luo, S., Liu, S., Qiao, H., Liu, Y., & Feng, L. (2020). Deep attention based music genre classification. *Neurocomputing, 372*, 84–91. https://doi.org/10.1016/j.neucom.2019.09.054

Zaman, K., Sah, M., Direkoglu, C., & Unoki, M. (2023). A survey of audio classification using deep learning. *IEEE Access, 11*, 106620–106649. https://doi.org/10.1109/ACCESS.2023.3318015

Author ORCIDs

Author name	ORCID
Pyo, Seunghyun	0009-0002-6830-8813
Yeu, Tae kyeong	0000-0003-2742-3284
Lee, Yeongjun	0000-0002-3808-8349
Han, Jong-Boo	0000-0002-5670-538X
Cho, Yujin	0009-0007-9583-310X
Park, Daegil	0000-0001-5724-1794