Archives of Acoustics Vol. ***, No. ***, https://doi.org/10.24423/archacoust.2026.4434

Online First July 22, 2026

Technical Note

From Speech to Underwater Acoustics: A Transfer Learning Framework for Real-Time Passive Diver Detection Using Keyword Spotting Models

Osama Deeb[(1)], Saier Mahmoud[(2)*], Louay Saleh[(2)], Assef Jafar[(3)], Oumayma Al Dakkak[(1)], Ibrahim Chouaib[(2)]

[(1)] *Department of Telecommunications*
[(2)] *Department of Electronic and Mechanical Systems*
[(3)] *Department of Informatics*

Higher Institute for Applied Sciences and Technology
Damascus, Syria

*Corresponding Author: saier.mahmoud@gmail.com

Received February 9, 2026; revised April 9, 2026; accepted April 21, 2026; available online May 11, 2026; version of record July 22, 2026; published issue XXXX.

Passive acoustic detection of divers faces challenges such as low signal-to-noise ratios (SNRs), data scarcity, and latency in conventional methods. This paper proposes keyword spotting for diver detection (KWS-DD) – a transfer learning framework that repurposes speech-oriented KWS models for data-efficient diver detection. Diver inhalation signatures are treated as acoustic 'keywords,' enabling adaptation of the transformer-based HuBERT architecture (pre-trained on speech) to identify quasi-periodic respiratory events in underwater audio. The core innovation of this work lies in adapting the state-of-the-art speech model HuBERT for accurate diver detection via non-speech inhalation acoustics. This approach eliminates the need for accumulation during the respiratory cycle, enabling real-time detection using a minimal amount of domain-specific data (120 inhalation samples). The solution, deployed in diverse marine conditions, achieved 94.4 % accuracy and 94.6 % $F1$-score for inhalation sounds. This represents a more than 50 % range extension over conventional methods, which proved unreliable at distances greater than 10 meters in low-SNR environments. The framework reduces false alarms caused by boat noise and generalizes to external datasets, validating cross-domain transferability. This work bridges AI-based speech processing and passive sonar signal processing, offering a resource-efficient solution for real-time underwater surveillance.

Keywords: passive diver detection, keyword spotting, transfer learning, real-time detection, passive sonar, hydrophone.

1. Introduction

The acoustic signal generated by a scuba diver breathing through a regulator serves as the basis for passive sonar detection. This signal exhibits a broadband frequency spectrum (Johansson *et al.*, 2010), ranging from several hundreds of hertz to 75 kHz (Tu *et al.*, 2020), and characterized by quasi-periodicity (Gorovoy *et al.*, 2015). The repetition interval varies with an activity level, ranging from approximately 7.09 s at rest to 2.44 s during heavy flapping (Donskoy *et al.*, 2008), directly reflecting respiratory rates from 0.14 Hz to 0.41 Hz. These rates and the associated emitted power vary based on the diver's experience, exertion level, and equipment (Donskoy *et al.*, 2008). The respiratory cycle acoustically bifurcates: inhalation manifests predominantly above 2 kHz, whereas bubble-generating exhalation manifests predominantly below 2 kHz (Sun *et al.*, 2022). Both phases are useful for detection (Sun *et al.*, 2022). However, inhalation signals are often considered preferable due to their pulse characteristics (Tu *et al.*, 2020).

Detectability depends on both the signal power within relevant frequency bands and the repetition rate (Stolkin *et al.*, 2006). Research focus diverges between the high-frequency band (>2 kHz) (Tu *et al.*, 2020; Jin, Xu, 2020) and the low-frequency band (<2 kHz) (Mahmoud *et al.*, 2025; Radford *et al.*, 2005; Hari *et al.*, 2015), with notable challenges in detecting a closed-circuit scuba (rebreather) which emits significantly lower acoustic energy than an open-circuit scuba (Gorovoy *et al.*, 2015; Radford *et al.*, 2005).

Two primary methodologies are employed for detecting diver-generated acoustic signals: traditional signal processing techniques and artificial intelligence AI approaches. The first method follows a sequential processing chain. Initial band-pass filtering enhances the signal-to-noise ratio (SNR) by focusing on frequency bands relevant to the diver's spectral signature, which varies with the diving apparatus (Chung *et al.*, 2007; Hari *et al.*, 2015). This results in the use of different bands, such as 25 kHz to 75 kHz (Hari *et al.*, 2015), 200 Hz to 500 Hz (Gorovoy *et al.*, 2014), or 13 kHz to 18 kHz (Tu *et al.*, 2020). Subsequently, periodicity detection leverages energy estimation (Gorovoy *et al.*, 2014), envelope detection (Tu *et al.*, 2020), or matched filtering (Gorovoy *et al.*, 2014; Chen *et al.*, 2006). Finally, the fast Fourier transform (FFT) analysis quantifies power around the breathing rate and a threshold-based trigger is launched when results exceed a predetermined level (Stolkin *et al.*, 2006; Lennartsson *et al.*, 2009). These approaches have been augmented through adaptive noise subtraction for range extension (20 m to 40 m) (Tu *et al.*, 2020), background noise whitening to suppress transient interference (Johansson *et al.*, 2010), and complementary cross-correlation techniques applied to dual-hydrophone signals for enhanced detection ranges (Chung *et al.*, 2007; Korenbaum *et al.*, 2020).

Traditional signal processing techniques have limited capabilities when it comes to detecting divers and, more broadly, detecting underwater objects, particularly in detection accuracy and latency. These limitations have motivated the adoption of artificial intelligence approaches (the second method), with machine learning (ML) and deep learning emerging as promising solutions (Feng *et al.*, 2024). Like most AI-based solutions, the ML pipeline for this application typically comprises four critical stages as shown in Fig. 1:

1. Data acquisition, where generative adversarial networks (GANs) are used for both synthetic data generation and augmentation (Fuchs *et al.*, 2019; Jin *et al.*, 2020; Karjalainen *et al.*, 2019; Phung *et al.*, 2019; Zhao *et al.*, 2023).
2. Preprocessing, where autoencoder architectures have been successfully applied to critical preprocessing tasks including noise reduction, resampling, and signal enhancement (Dong *et al.*, 2022; Huang *et al.*, 2020; Vincent *et al.*, 2010; Wang *et al.*, 2014).
3. Feature extraction, where learned features derived through autoencoders and self-supervised learning techniques demonstrate the ability to automatically extract high-level representations by exploiting inherent data structures (Chen, Shang, 2019; Wang *et al.*, 2022).
4. Classification which has evolved through successive generations of ML approaches including but not limited to support vector machines (SVMs) (Zhang *et al.*, 2003), K-nearest neighbors (KNN) (Alvaro *et al.*, 2021), recurrent neural networks (RNNs), convolutional neural networks (CNNs), attention-based mechanisms and transformers (Hewamalage *et al.*, 2021; Wang *et al.*, 2021).

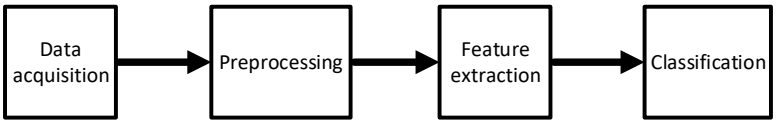

Fig. 1. General ML-based pipeline for underwater acoustic detection task.

Among all of these types, transformers, with their multi-head self-attention (MHSA) mechanisms, have shown particular promise despite computational challenges, prompting the development of specialized training protocols incorporating transfer learning and advanced data augmentation techniques to mitigate these constraints (Gong *et al.*, 2022; Li *et al.*, 2022).

Diver detection constitutes a specialized application of underwater acoustic signal processing, distinguished by the unique temporal and spectral characteristics of respiratory signatures. Of particular diagnostic value are

inhalation patterns, which exhibit distinctive pulsed waveforms (Tu *et al.*, 2020). SVM classifier, employing energy characteristics across eight distinct non-overlapping sub-bands (49 kHz to 51 kHz) as discriminative features, demonstrated superior capability in processing irregular respiratory patterns that challenge conventional matched filter approaches (Zhao *et al.*, 2016). However, it should be noted that these promising results were exclusively validated in controlled pool environments, potentially limiting generalizability to real-world settings. Subsequent research demonstrates the superiority of frequency-domain multi-sub-band energy (FMSE) features over both Mel frequency cepstral coefficient filtering (MFCC) and gamma tone frequency cepstral coefficient filtering (GFCC) within the 2 kHz to 8 kHz band when processed through SVM classifiers (Sun *et al.*, 2024), with validation extended to diverse environments including pools, lakes, and shallow sea. While these approaches operate on 1D acoustic data, CNNs that process 2D spectrogram, demonstrate robust detection capabilities in marine environments, achieving 93 % correct classification rates at ranges up to 12 m (Cole, 2019).

The aforementioned methods (both traditional and AI-based) rely heavily on respiratory periodicity – a feature consistently leveraged in existing literature. This dependence introduces inherent latency of up to 10 s due to the requisite observation window for capturing multiple respiratory cycles (Jin, Xu, 2020) which is considered a long time for some applications especially those related to protecting strategic maritime facilities (military and civilian) such as aircraft carriers and oil platforms against terrorist attacks. While approaches using energy exclusively within dominant diver frequencies eliminate this delay, they exhibit proportionally heightened vulnerability to noise interference, compromising detection reliability (Stolkin *et al.*, 2006). Jin *et al.* (2020) addressed this limitation through a deep learning framework that uses time-frequency features and treats the diver's signal as a keyword; however, their framework relies on a synthesized and augmented dataset (Jin, Xu, 2020).

Recent breakthroughs in speech processing using deep learning present significant opportunities for cross-domain applications. Keyword spotting (KWS) systems exemplify this potential, achieving >99 % accuracy in detecting target keywords within continuous speech. This success suggests promising applicability to diver detection by treating characteristic acoustic signatures (e.g., inhalation/exhalation patterns) as analogous to spoken keywords. Extending the KWS architecture of (Deeb *et al.*, 2025), which leverages a pretrained hidden unit bidirectional encoder representations from the transformer (HuBERT) model to achieve 99.7 % classification accuracy with only 15 samples per keyword, this work adapts the framework for diver detection via respiratory acoustic pattern recognition. Although HuBERT was originally trained on speech data, its transformer-based architecture, employing multi-head self-attention, exhibits robust representation learning capabilities. Through fine-tuning, the model can discern fundamental sound units within divers' non-speech vocal signals and align them with latent representations acquired during pretraining. This adaptation facilitates effective detection and discrimination of acoustic patterns in diver respiratory signals. This extension delivers three principal advantages:

1. Enhanced detection accuracy through physiologically distinctive acoustic signatures.
2. Real-time processing capability eliminating detection latency.
3. Data-efficient through employing transfer learning from the speech domain that makes it suitable for resource-constrained scenarios.

Collectively, this approach directly addresses the critical challenge of limited training data availability inherent in diver detection applications.

Beyond this introductory section, the paper is divided into eight parts. Section 2 details the characteristics of acoustic signals produced by divers. Section 3 describes the experimental setup, including location, conditions, and instrumentation used for data acquisition. Section 4 reviews the traditional signal processing algorithm employed for diver detection using acoustic signatures. Section 5 introduces the novel transfer learning algorithm designed for keyword spotting, and its adaptation for detecting divers based on inhalation signatures. Section 6 presents the experimental tests performed, providing a comparative analysis of the traditional method versus the proposed AI-based approach. Section 7 analyzes the obtained experimental results and examines the shortcomings of the proposed algorithm. Section 8 concludes the paper, summarizing the system's principal advantages and outlining potential avenues for future enhancement.

2. Acoustic signals emitted by diver

Diver breathing generates two distinct acoustic signatures: an inhalation signal characterized by energy predominantly above 2 kHz, and an exhalation signal concentrated below 2 kHz. Each respiratory cycle exhibits a duration of approximately 3 s, as illustrated in Fig. 2. This figure presents the time-domain waveform and the corresponding spectrogram of acoustic data recorded from a diver swimming within a ~3-meter radius of the hydrophone used in this research. The spectrogram was computed using a 1024-sample Hamming window, 50 % overlap, and 1024-point FFT.

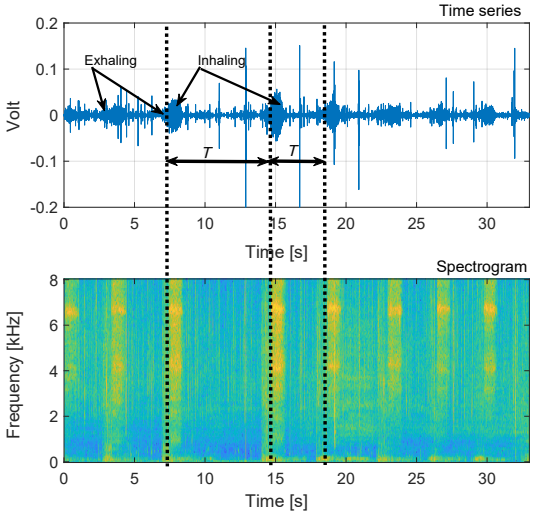

Fig. 2. Time-domain waveform and spectrogram of acoustic emissions from a diver breathing while swimming within a 3-meter radius of the hydrophone. The spectrogram was computed using a 1024-sample Hamming window, 50 % overlap, and 1024-point FFT.

An analysis of Fig. 2 reveals that inhalation manifests as a series of quasi-periodic transient events, visible as vertical striations in the spectrogram, though the periodicity is not strictly constant. The exhalation signal is less prominent in the time-domain waveform but appears in the spectrogram as low-frequency energy band preceding each inhalation event.

As the diver's distance from the hydrophone increases, the signal attenuation increases significantly, particularly for the higher-frequency inhalation components due to greater absorption losses in the water column (Tu *et al.*, 2020). Consequently, robust detection of the diver presence using conventional signal processing techniques presents significant challenges. Figure 3 illustrates this difficulty, showing the time-domain waveform and

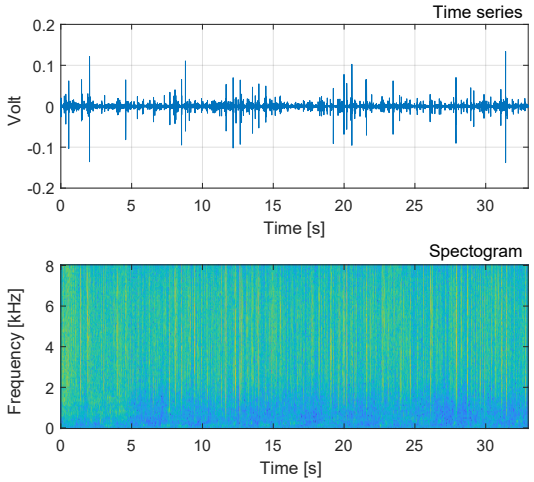

Fig. 3. Time-domain waveform and spectrogram of diver acoustic emissions at extended range (10 m). Spectrogram computed using a 96 kHz, 1024-sample Hamming window, 50 % overlap, and 1024-point FFT.

the corresponding spectrogram of acoustic data recorded from a diver swimming within a ~10-meter radius of the hydrophone. The spectrogram was computed using a 1024-sample Hamming window, 50 % overlap, and 1024-point FFT. Critically, the effect of inhalation is not clearly discernible either in the time-domain signal or in the spectrogram representation.

3. Data acquisition

Acoustic signatures were acquired during three collection rounds (August 2–3 and October 2, 2024) using two omnidirectional hydrophones mounted 70 cm apart on a rigid stationary platform positioned 60 cm above the seabed. An acoustic vector sensor (AVS) was co-located at the platform center but remained unused in this study, as shown in Fig. 4. Critically, each hydrophone channel was processed in strict isolation with zero cross-sensor techniques applied – including beamforming, time-difference-of-arrival estimation, coherence analysis, or other array-based methods. This dual-hydrophone configuration provided exclusively time-synchronized redundant measurements.

Fig. 4. Experimental setup showing diver, sensor platform (hydrophones + AVS), and support boat.

Experiments were conducted in a large seawater basin (350 m × 250 m × 5 m depth), where two divers, equipped with an open-circuit SCUBA, sequentially executed controlled maneuvers: circular traversals at 1 m to 30 m radial distances, radial approaches/moves away, and swimming-to-stationary state transitions. Activities spanned azimuthal bearings from 0° to 360° relative to the sensor platform to capture representative single-channel signatures across diverse source–receiver geometries and motion dynamics. Environmental conditions were deliberately varied: days 1–2 (August) featured calm, sunny conditions, while the third day (of October) was marked by sustained winds that generated surface wave disturbance. This deliberate diversity in diver activities, motion profiles, and environmental conditions yields a rich source for building a diverse dataset that closely mirrors real-world operational variability, making it particularly valuable for training and validating robust AI-based detection models. All data was recorded at 44.1 kHz with 24-bit resolution per channel.

To rigorously evaluate both traditional diver detection methods and the proposed transfer learning-based model, a diverse set of acoustic signals was curated. These signals encompass varying environmental conditions, SNRs, motion patterns, and interference sources. The test suite includes both locally recorded data and externally sourced signals to assess robustness and generalization capability. Table 1 describes these signals.

These signals were reserved for testing purposes and never used for model training.

Table 1. Evaluation signal specifications.

ID	Description
Diver_1	Diver moves within 3 m (high SNR (~10 dB))
Diver_2	Stationary diver at 10 m (low SNR <−8 dB)
Diver_3	Moving diver: starts at 25 m, approaches up to 15 m at $t = 22$ s, stays in place for 25 s, moves away at $t = 47$ s
Noise	Ambient noise (no divers/boats)
Boat	Moving boat (0 m to 150 m) approaching and departing
Diver_4	Externally sourced diver signal (Davi-sh, 2022)
Sonar	Externally sourced sonar signal (VoiceBosch, 2024)

4. Conventional diver detection algorithm

The conventional passive diver detection algorithm follows the processing chain illustrated in Fig. 5, consisting of these sequential stages (Mahmoud *et al.*, 2025):

1. Acoustic signal acquisition: raw hydrophone signals are captured for processing.
2. Bandpass filtering: signals are filtered within the dominant breathing frequency band to enhance the SNR by attenuating out-of-band 2 kHz to 7 kHz.
3. Periodicity enhancement: an envelope detector highlights breathing periodicity by a computing signal power within 100-ms sliding windows with 50 % overlap.
4. Spectral analysis: the FFT is applied to the envelope signal.
5. Diver index calculation: the detection energy is quantified by integrating the FFT power within the 0.14 Hz to 0.42 Hz band, corresponding to expected breathing rates.
6. Detection decision: the computed diver index is compared against a predefined threshold to determine the diver presence.

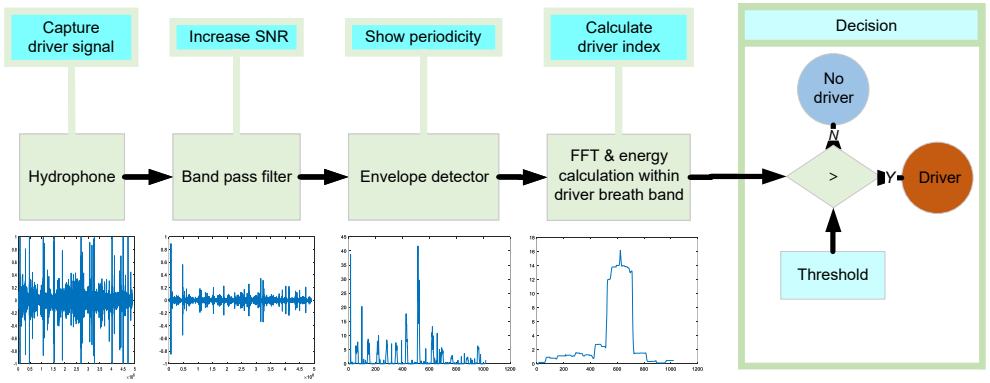

Fig. 5. Diver detection algorithm.

A critical implementation challenge involves the non-stationary nature of ambient noise, which necessitates adaptive threshold determination. This may be achieved through either:

- a reference hydrophone at a suitable distance from the primary sensor, or
- real-time noise power estimation algorithms applied to the acquired signal.

Figure 6 demonstrates the performance of the algorithm when applied to the signals described in Sec. 3 (Table 1). The analysis uses a 10-second window, with power estimation windows (100-ms duration, advanced in 50-ms increments). Key observations include:

- Diver_1: the diver index significantly exceeds noise levels more than 100 times, enabling reliable detection.
- Diver_2: the reduced diver index amplitude degrades detection reliability.
- Boat: the diver index for boat noise reaches observed levels for distant divers, highlighting susceptibility to false alarms from high-energy sources.

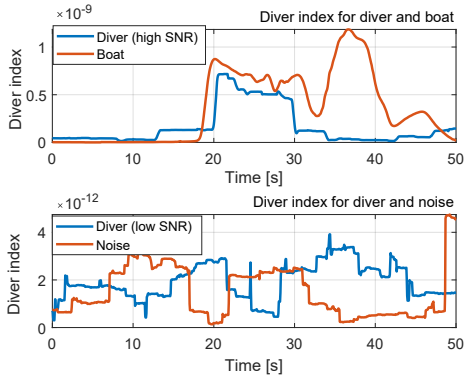

Fig. 6. Diver index estimation for a close-proximity diver (<3 m), distant diver (>10 m), boat noise, and ambient noise.

5. AI-based diver detection

This section introduces the proposed method that leverages transfer learning to adopt an AI-based KWS model to solve the diver detection task.

5.1. Adapting keyword spotting for diver detection

The primary challenge in applying KWS systems to diver detection lies in the substantial training data requirements. Conventional KWS frameworks typically demand approximately 4000 positive and negative samples per keyword to achieve sufficient accuracy (Lin *et al.*, 2020). This requirement proves particularly problematic for diver detection given the inherent difficulties in acquiring sufficient sonar acoustic data, which has limited the adoption of advanced ML architectures such as transformers. While data augmentation and synthesis techniques offer partial solutions, transfer learning presents a more fundamental approach by enabling knowledge transfer between domains. This paradigm has demonstrated particular effectiveness in KWS applications, where models pre-trained on English speech could be adapted to other languages with minimal target-language samples through fine-tuning (Seo *et al.*, 2021). Building on this foundation, a transfer learning framework is proposed, where a general speech representation model – using transformers in its architecture – is fine-tuned to detect diver respiratory sounds (particularly inhalation patterns), significantly reducing the required training data while maintaining the performance advantages of transformer architectures, as demonstrated in underwater object detection tasks (Feng *et al.*, 2024; Domingos *et al.*, 2022).

5.2. Model architecture

The proposed framework adapts and modifies a previously established KWS architecture (Deeb *et al.*, 2025) for the diver detection downstream task through fine-tuning on underwater acoustic recordings of diver sounds. As illustrated in Fig. 7, the architecture comprises two fundamental components: (1) a HuBERT encoder block (Hsu *et al.*, 2021) and (2) a classifier block (this architecture is denoted as KWS-DD). The HuBERT block employs a multi-head self-attention mechanism and has demonstrated state-of-the-art performance in speech processing since its introduction in 2021. Pretrained using masked prediction self-supervised learning, this component generates discriminative contextual representations of input signals that effectively separate distinct acoustic patterns in the embedding space.

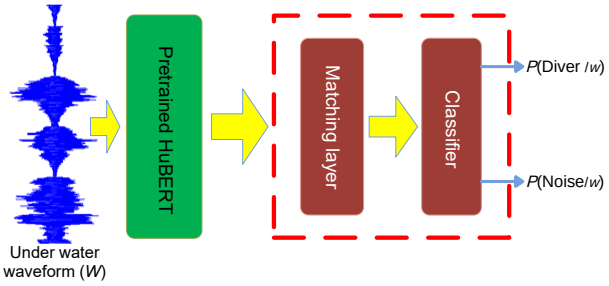

Fig. 7. KWS-based diver detection model.

The classifier block subsequently learns to map these representations to predefined target classes during fine-tuning. Notably, the HuBERT architecture processes raw waveform inputs without requiring manual feature engineering, as its integrated feature extractor – comprising 7-layer convolutional network – automatically extracts optimal acoustic features. This end-to-end design preserves signal integrity while allowing feature extraction parameters to be jointly optimized during training, yielding superior representational capacity compared to conventional preprocessing pipelines.

Upon receiving a raw underwater waveform lasting approximately one second, the pretrained HuBERT model segments it into 20 ms frames, each frame is represented as a (1×1024)-dimensional tensor at the output of HuBERT. These tensors are fed to the matching layer which computes the average of the tensor across the time,

resulting in a condensed (1×1024)-tensor for the whole waveform, which is subsequently fed into the classifier. The classifier is designed to categorize the output tensors into noise or diver classes. A softmax layer at the final stage of the classifier generates the posterior probabilities $P(c_i/w) : i \in [\text{Diver, Noise}]$, which quantify the likelihood of observing the class c_i at the output of the model when a tensor representing the input word w is applied to its input. The final classification decision is made based on these posteriors following the formula provided in Eq. (1):

$$\text{ChosenClass} = \underset{c_i}{\text{Argmax}} \left(P(c_i/w) \right). \tag{1}$$

The framework employs HuBERT-Large, a pretrained version of HuBERT with 317 million parameters, to extract contextual embeddings from raw acoustic waveforms, followed by the KWS head. The final diver detection model contains (317.02 million parameter), it was trained by fine-tuning the whole model's parameters to adapt the system for diver detection. The fine-tuning process was performed using a specialized dataset containing: (1) positive samples of authentic diver-generated acoustic signatures (inhaling signals), and (2) negative samples encompassing various underwater acoustic interference sources. This approach enables effective transfer learning while maintaining the model's ability to discriminate the subtle acoustic patterns characteristic of a diver presence.

5.3. Upgrading model for continuous waveform

In operational scenarios, the exact timing of diver inhalation events is inherently unpredictable, necessitating continuous analysis of the hydrophone's time-domain acoustic signal. To reconcile this streaming data requirement with the model's fixed-length input specification (1.0 ±0.2 s segments), a sliding window algorithm with 100 ms temporal increments was implemented (this temporal resolution was carefully selected to optimize the trade-off between computational efficiency and signal capture reliability, ensuring complete coverage of critical acoustic features while enabling a near-real-time operation). Each 1-second window undergoes probabilistic classification, generating posterior estimates $P(c_i/w) : i \in [\text{Diver, Noise}]$ through the trained model. As the window advances across the continuous waveform, this process produces discrete-time probability functions (Eq. (2)) that quantitatively characterize the evolving likelihood of diver presence versus ambient noise (Fig. 8), enabling real-time detection without compromising accuracy;

$$f_{c_i}(k) = P(c_i/w_k) \quad @k^{\text{th}}\text{window}. \tag{2}$$

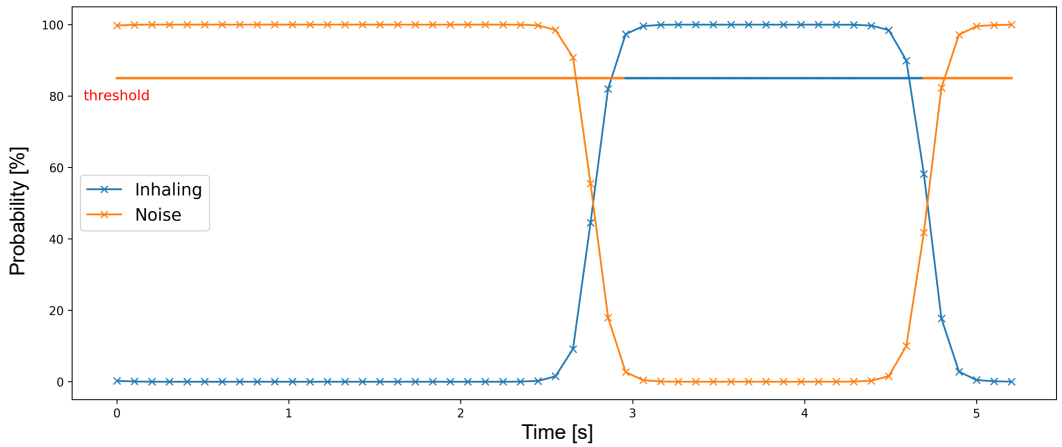

Fig. 8. Applying the sliding window approach on a waveform containing a single inhaling signal, the blue curve indicates the discrete probability function of diver's inhaling signal, the orange curve represents the discrete probability function of the noise.

The discrete probability functions enable robust diver detection through temporal analysis of the posterior probability evolution. A detection event is triggered when $P(\text{Diver}|w)$ exceeds a predefined threshold θ for a minimum duration Δt, ensuring protection against transient false positives. While not strictly required, observed

periodicity in detection events (typically 0.14 Hz to 0.42 Hz for human respiration) provides secondary validation of a diver presence. In the following, the continuous-processing variant of KWS-DD is denoted as CKWS-DD.

6. Experiments and results

6.1. Data preparation

The most distinctive acoustic signatures associated with divers and their equipment are generated by respiratory activity, particularly inhalation and exhalation processes. An air bubble movement during exhalation produces unique signatures that aid detection. However, since similar acoustic patterns may be produced by marine creatures or equipment, the inhalation signal was selected as the primary discriminative feature due to its high specificity to human divers.

For model fine-tuning, representative acoustic samples were required for each target class (inhalation signatures and ambient noise). Sample extraction involved segmentation of hydrophone recordings with different SNRs into 1-second intervals, corresponding to the mean inhalation duration (1.0 ±0.2 s). Two qualified experts independently performed the segmentation through aural and visual inspection of spectrographic representations. Inter-rater reliability was enforced, with only concordantly verified samples retained for the final dataset. The curated collection ultimately comprised four distinct marine acoustic categories relevant for coastal surveillance:

1. Inhalation signatures: isolated inhalation events.
2. Exhalation signatures: isolated exhalation events.
3. Vessel signatures: propeller cavitation and engine harmonics.
4. Ambient noise: site-specific hydrodynamic and biological noise.

While the present study employs binary classification, this categorical taxonomy enables future transition to multiclass detection frameworks. For the current objective of diver detection, samples were aggregated into two classes: the diver class encompasses verified inhalation events and the noise class encompassing composite non-target acoustics (exhalation, vessels, ambient noise).

The acquired acoustic samples underwent standard preprocessing, including down sampling to 16 kHz to ensure compatibility with the HuBERT architecture's input specifications. The curated dataset consisted of 120 validated inhalation exemplars (the diver class) and 344 interference samples (the noise class). The dataset was partitioned into three subsets while preserving class distributions: a training set (70 %), validation set (15 %), and test set (15 %). The dataset's limited scale, particularly the test portion containing fewer than 70 samples, restricts its ability to provide a comprehensive evaluation of model performance. While this constrained benchmark may not fully represent the model's generalization capabilities, it serves as an initial performance indicator. Instead, the definitive evaluation utilizes extended, continuous audio samples incorporating diverse acoustic conditions, as presented in the experiments section.

6.2. Implementation and training details

The proposed architecture was implemented in Python using PyTorch, incorporating the pre-trained Hubert-large model (facebook/hubert-large-ls960-ft) as described in Sec. 5. During fine-tuning, the 7-layer convolutional feature extractor parameters remained frozen to preserve learned acoustic representations. The model optimization was performed on a workstation equipped with the Intel Core i7-11370H processor and NVIDIA GTX 1650 GPU (16GB RAM) using the training partition of the dataset. Training ran for 30 epochs, converging in ~110 min. Figure 9 illustrates the training dynamics through the evolution of both training and validation loss curves throughout the optimization process.

The trained model underwent comprehensive performance assessment using the reserved test dataset subset. Quantitative evaluation employed four standard classification metrics: accuracy, precision, recall, and $F1$-score, providing complementary perspectives on detection capability. These metrics collectively characterize: (1) overall prediction correctness (accuracy), (2) positive prediction reliability (precision), (3) target detection sensitivity

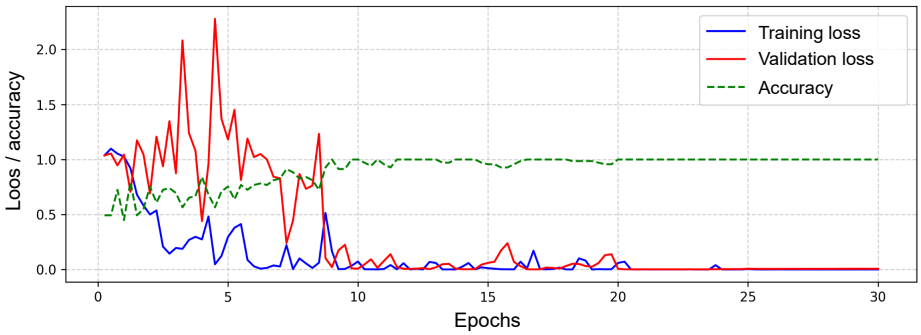

FIG. 9. Training metrics of the KWS-DD model through the fine-tuning phase.

(recall), and (4) their harmonic balance ($F1$-score), enabling robust assessment of the system's operational viability. Results are shown in Table 2.

TABLE 2. Testing result of the KWS-DD model using the curated dataset.

Precision	Recall	$F1$-score	Accuracy
0.955	0.944	0.946	0.944

6.3. RESULTS OF CONTINUOUS SIGNALS

Following the evaluation of the model on the curated dataset containing isolated samples of various underwater sounds, a comprehensive evaluation is conducted on evaluation signals presented in Table 1. These results are compared with the traditional detection methods presented in Sec. 4.

6.3.1. DIVER WITH HIGH SNR 'DIVER_1'

Figure 10 shows the detection results of the CKWS-DD model when applying 'Diver_1' to its input. The model detected nearly all inhalations with high probabilistic value, and could successfully distinguish it from the ambient noise. The periodicity is also clear, and it enhances the detection decision. Both the CKWS-DD model and the traditional method successfully detect the diver, demonstrating robust performance under high-SNR conditions.

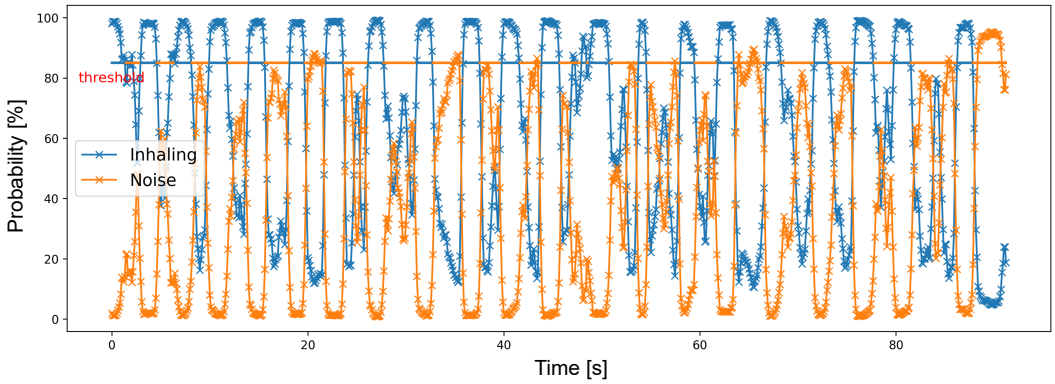

FIG. 10. Model prediction for diver signal (high SNR).

6.3.2. DIVER WITH LOW SNR 'DIVER_2', 'DIVER_3'

In this case, 'Diver_2' was applied to the CKWS-DD model, and results are shown in Fig. 11, the traditional method failed to distinguish the diver's signals from the ambient noise in this signal, as depicted in Sec. 4 (Fig. 6). In contrast, the CKWS-DD model exhibited reliable detection. To demonstrate the model's performance

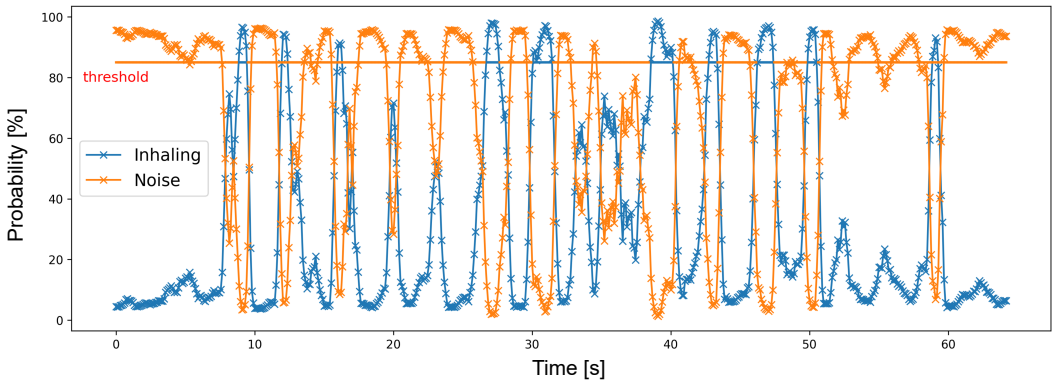

FIG. 11. Model prediction for diver signal 'Diver_2.'

in a moving target, 'Diver_3' was applied to the system input. Figure 12 shows the result. It can be seen that the detection becomes unstable when the distance between the diver and the sensor exceeds 15 m, regardless of whether the diver is approaching or moving away from the sensor.

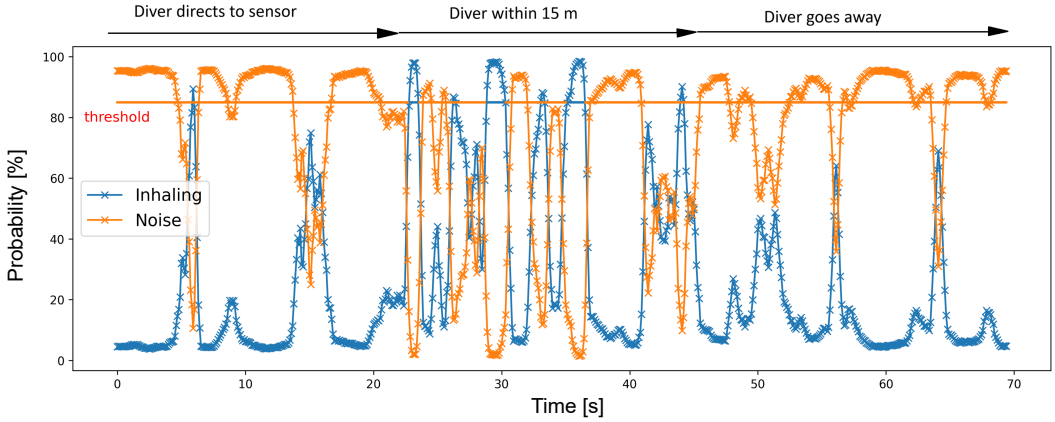

FIG. 12. Model prediction for diver signal 'Diver_3.'

6.3.3. AMBIENT NOISE DETECTION 'NOISE'

The model accurately identifies the ambient noise with a minimal number of false alarms, confirming strong noise-rejection capabilities, as shown in Fig. 13, obtained after applying 'Noise' to the model.

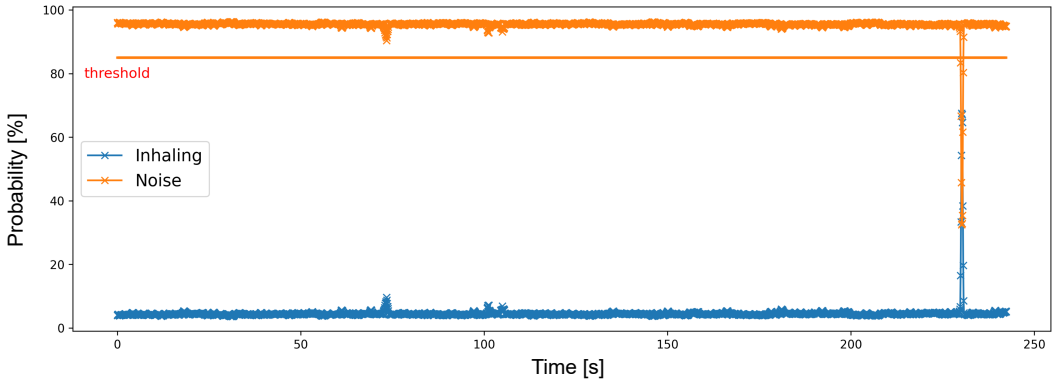

FIG. 13. Model prediction for ambient noise 'Noise.'

6.3.4. BOAT SIGNAL DETECTION 'BOAT'

Boat signals are classified as noise but exhibit sporadic false alarms (Fig. 14, 'Boat'). This stems from non-stationary motor noise and dynamic electrical status variations.

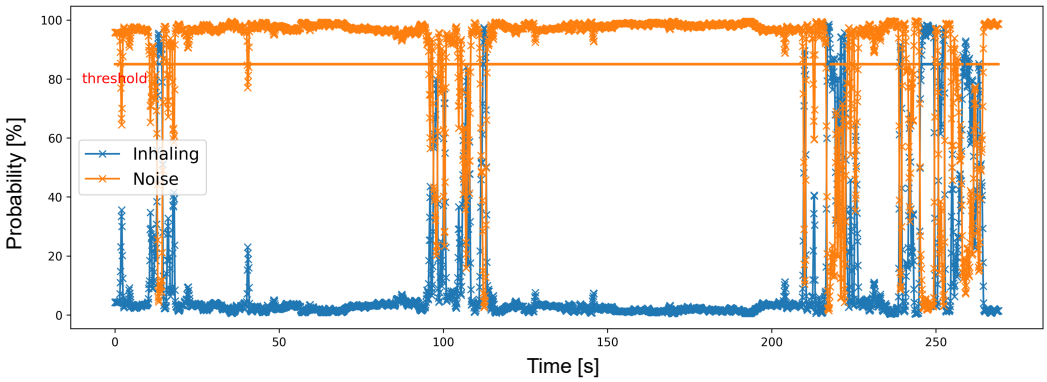

FIG. 14. Model prediction for boat noise 'Boat.'

6.4. CROSS-PLATFORM GENERALIZABILITY ASSESSMENT

To evaluate model robustness across different acquisition systems, we conducted validation testing using externally-collected recordings comprising: diver respiratory and active sonar signal captured with alternative hydrophone. This independent verification protocol assesses the system's ability to maintain detection performance when deployed with varying sensor configurations and environmental conditions. The model generalizes effectively to real-world data:

– diver detection: robust inhaling signal identification when evaluating (Fig. 15, 'Diver_4').
– sonar rejection: sonar signals are correctly classified as noise when evaluating (Fig. 16, 'Sonar').

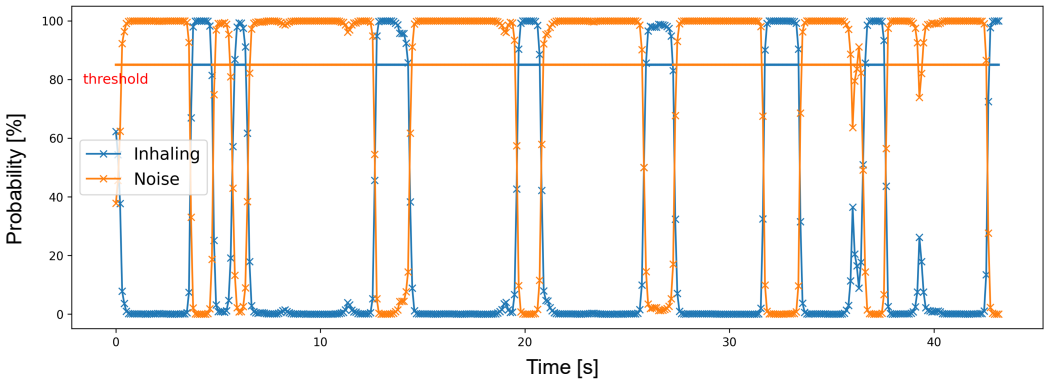

FIG. 15. External diver signal 'Diver_4.'

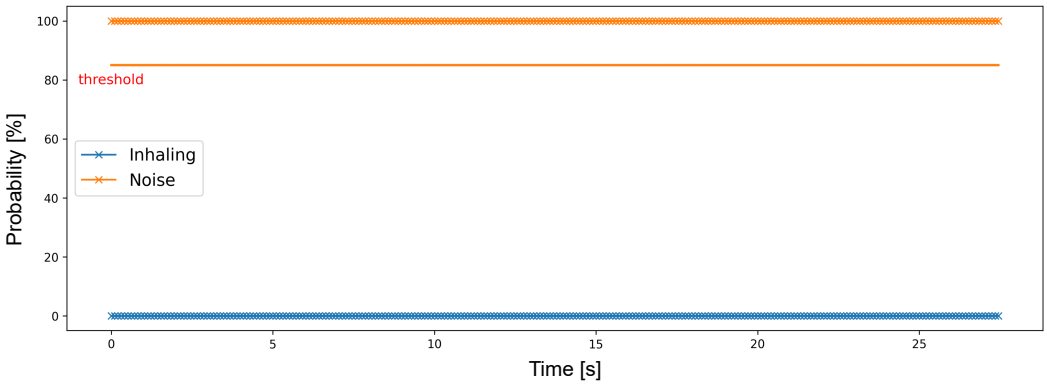

FIG. 16. Sonar signal misclassified as noise 'Sonar.'

As illustrated, the model discriminated the sonar signal from the diver's inhalation signal even though the sonar data were acquired using different equipment and under different environmental conditions – highlighting the trained model's high generalization ability.

7. Discussion

The results demonstrate that the CKWS-DD model represents a significant advancement in diver signal detection by addressing critical limitations of traditional methods. Most notably, the system extends a reliable diver detection range up to 15 m, outperforming conventional approaches that typically fail beyond 10 m in low-SNR conditions. This enhanced capability originates from the model's noise-resilient architecture, which effectively extracts faint inhaling signatures obscured by ambient noise.

The proposed system achieves an inference time of less than 0.45 s on a standard laptop computer (see Subsec. 6.2). Although the model requires a 1-second audio input window for reliable detection, the total decision latency from the start of a recording segment is approximately 1.45 s. This is substantially shorter than traditional periodicity-based methods, which often need more than 10 s to analyze multiple breathing cycles. By making frame-wise predictions on 1-second windows without waiting for a periodic structure to emerge, our model enables near-instantaneous detection.

Furthermore, transfer learning facilitates effective model training, even with limited data while maintains high detection accuracy. Validation using external real-world data, including publicly available diver and interference recordings, confirms robust generalization to unknown environments, underscoring practical deploy ability for underwater surveillance systems.

Despite these advances, two key limitations require consideration. First, false alarms during boat detection occur due to non-stationary motor noise and dynamic electrical variations related to the motor type and the operation mode, suggesting a need for additional dataset samples covering diverse boat noise profiles. Second, detection instability beyond 15 m appears linked to hydrophone sensitivity limitations.

The model depicts a clear performance dichotomy: while demonstrating performance comparable to that of traditional methods in high-SNR scenarios, it provides distinct advantages in noisy environments where conventional approaches prove ineffective.

While the current model demonstrates robust performance in distinguishing diver breathing sounds from various non-biological transient noises – including boat engine sounds and sonar pulses – we acknowledge that further enhancements are possible. In future work, we plan to enrich the training and testing dataset by incorporating a wider variety of marine animal sounds (e.g., snapping shrimp, fish vocalizations, and marine mammal calls) as well as additional artificial noise sources such as different types of boat engines, ship propellers, and other mechanical underwater equipment. We believe that exposing the model to a more diverse and challenging acoustic environment will further improve its generalization capability and reduce the risk of false positives in real-world deployment scenarios. This dataset expansion, along with corresponding model retraining and evaluation, constitutes a key direction for our ongoing research.

Another thing that should be mentioned is that water temperature and salinity gradients create sound speed profiles that can refract acoustic waves, thereby affecting the SNR of a diver's regulator sound over distance. For example, negative thermoclines, may bend sound downward and reduce the amount of surface-detected energy. Our experiments were conducted in a shallow, well-mixed environment where such gradients were minimal. For deeper or stratified waters, these effects could impact the detection range and should be considered in the system deployment.

8. Conclusion

This study establishes the viability of repurposing speech-focused KWS models for passive diver detection. Distinctive inhalation signatures are treated as acoustic 'keywords' in CKWS-DD, a transfer learning framework leveraging the HuBERT transformer architecture pre-trained on speech data. When fine-tuned with only 120 curated inhalation samples and 344 noise samples, the model achieves 94.4 % accuracy and 94.6 % $F1$-score, demonstrating significant data efficiency compared to conventional deep learning approaches (Jin, Xu, 2020). Key advantages include:

1. Real-time capability: detection latency reduces to <1 s through analysis of 1 s audio segments, eliminating the 10 s observation windows required by periodicity-based methods.

2. Noise resilience: reliable detection is maintained at SNRs as low as −8 dB (at ranges ≤15 m), outperforming traditional methods that fail beyond 10 m.
3. Interference rejection: false alarms from boat noise decrease compared to spectral energy thresholding approaches.

Current limitations included performance degradation beyond 15 m and occasional false positives from non-stationary motor noise. Cross-validation using externally sourced data confirms strong generalizability. Future research directions include dataset expansion with diverse interference signals, incorporation of marine animal sounds, and investigation of multi-diver scenarios. The CKWS-DD framework provides a scalable solution for resource-constrained underwater monitoring systems by enabling accurate, low-latency diver identification with minimal training data requirements.

Fundings

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

Conflict of interest

The authors declare that there are no known competing financial interests or personal relationships that could have influenced the work described in this paper.

Authors' contributions

Osama Deeb and Saier Mahmoud: writing – original draft, writing – review and editing, visualization, validation, resources, software, methodology, investigation, data curation, conceptualization. Louay Saleh and Assef Jafar: validation, methodology, supervision, project administration, conceptualization. Oumayma Al-Dakkak and Ibrahim Chouaib: validation, methodology, supervision, conceptualization. All authors reviewed and approved the final manuscript.

Data availability

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

References

1. Alvaro A., Schwock F., Ragland J., Abadi S. (2021), Ship detection from passive underwater acoustic recordings using machine learning, *The Journal of the Acoustical Society of America*, **150**(4_Supplement): A124, https://doi.org/10.1121/10.0007848.
2. Chen X., Wang R., Tureli U. (2006), Passive acoustic detection of divers under strong interference, [in:] *OCEANS 2006*, https://doi.org/10.1109/OCEANS.2006.306869.
3. Chen Y., Shang J. (2019), Underwater target recognition method based on convolution autoencoder, [in:] *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, https://doi.org/10.1109/ICSIDP47821.2019.9173362.
4. Chung K.W., Li H., Sutin A. (2007), A frequency-domain multi-band matched-filter approach to passive diver detection, [in:] *2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers*, https://doi.org/10.1109/ACSSC.2007.4487426.
5. Cole A.M. (2019), *Automated open circuit scuba diver detection with low cost passive sonar and machine learning*, MA Thesis, Massachusetts Institute of Technology, https://hdl.handle.net/1721.1/122269.
6. Davi-sh (2022), Underwater scuba diver, Pixabay, https://pixabay.com/sound-effects/underwater-scuba-diver-28467/ (access: 21.06.2025).

7. Deeb O., Jafar A., Al Dakkak O. (2025), A deep learning framework for Arabic continuous speech keyword spotting in low-resource settings using isolated-word keyword spotting and posterior probability functions, *Advances in Artificial Intelligence and Machine Learning*, **5**(3): 4074–4093, https://doi.org/10.54364/AAIML.2025.53229.
8. Domingos L.C.F., Santos P.E., Skelton P.S.M., Brinkworth R.S.A., Sammut K. (2022), A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance, *Sensors*, **22**(6): 2181, https://doi.org/10.3390/s22062181.
9. Dong Y., Shen X., Wang H. (2022), Bidirectional denoising autoencoders-based robust representation learning for underwater acoustic target signal denoising, *IEEE Transactions on Instrumentation and Measurement*, **71**: 1–8, https://doi.org/10.1109/TIM.2022.3210979.
10. Donskoy D.M., Sedunov N.A., Sedunov A.N., Tsionskiy M.A. (2008), Variability of SCUBA diver's acoustic emission, [in:] *Proceedings of SPIE – The International Society for Optical Engineering*, **6945**, https://doi.org/10.1117/12.783500.
11. Feng S., Ma S., Zhu X., Yan M. (2024), Artificial intelligence-based underwater acoustic target recognition: a survey, *Remote Sensing*, **16**(17): 3333, https://doi.org/10.3390/rs16173333.
12. Fuchs L.R., Larsson C., Gällström A. (2019), Deep learning based technique for enhanced sonar imaging, [in:] *Underwater Acoustic Conference and Exhibition Series*.
13. Gong Y., Lai C.-I., Chung Y.-A., Glass J. (2022), SSAST: self-supervised audio spectrogram transformer, [in:] *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**(10): 10699–10709, https://doi.org/10.1609/AAAI.V36I10.21315.
14. Gorovoy S. *et al.* (2014), A possibility to use respiratory noises for diver detection and monitoring physiologic status, *The Journal of the Acoustical Society of America*, **135** (4_Supplement): 2303, https://doi.org/10.1121/1.4877584.
15. Gorovoy S. *et al.* (2015), Detecting respiratory noises of diver equipped with rebreather in water, [in:] *Proceedings of Meetings on Acoustics*, **24**(1): 070020, https://doi.org/10.1121/2.0000171.
16. Hari V.N., Chitre M., Too Y.M., Pallayil V. (2015), Robust passive diver detection in shallow ocean, [in:] *OCEANS 2015 – Genova*, https://doi.org/10.1109/OCEANS-Genova.2015.7271656.
17. Hewamalage H., Bergmeir C., Bandara K. (2021), Recurrent neural networks for time series forecasting: current status and future directions, *International Journal of Forecasting*, **37**(1): 388–427, https://doi.org/10.1016/J.IJFORECAST.2020.06.008.
18. Hsu W.-N., Bolte B., Tsai Y.-H.H., Lakhotia K., Salakhutdinov R., Mohamed A. (2021), HuBERT: self-supervised speech representation learning by masked prediction of hidden units, *IEEE/ACM Transactions on Audio Speech and Language Processing*, **29**: 3451–3460, https://doi.org/10.1109/TASLP.2021.3122291.
19. Huang F., Zhang J., Zhou C., Wang Y., Huang J., Zhu L. (2020), A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction, *Landslides*, **17**(1): 217–229, https://doi.org/10.1007/s10346-019-01274-9.
20. Jin B., Xu G. (2020), A passive detection method of divers based on deep learning, [in:] *2020 IEEE 3rd International Conference on Electronics Technology (ICET)*, https://doi.org/10.1109/ICET49382.2020.9119556.
21. Jin G., Liu F., Wu H., Song Q. (2020), Deep learning-based framework for expansion, recognition and classification of underwater acoustic signal, *Journal of Experimental and Theoretical Artificial Intelligence*, **32**(2): 205–218, https://doi.org/10.1080/0952813X.2019.1647560.
22. Johansson A.T., Lennartsson R.K., Nolander E., Petrović S. (2010), Improved passive acoustic detection of divers in harbor environments using pre-whitening, [in:] *OCEANS 2010 MTS/IEEE SEATTLE*, https://doi.org/10.1109/OCEANS.2010.5664549.
23. Karjalainen A.I., Mitchell R., Vazquez J. (2019), Training and validation of automatic target recognition systems using generative adversarial networks, [in:] *2019 Sensor Signal Processing for Defence Conference (SSPD 2019)*, https://doi.org/10.1109/SSPD.2019.8751666.
24. Korenbaum V., Kostiv A., Gorovoy S., Dorozhko V., Shiryaev A. (2020), Underwater noises of open-circuit scuba diver, *Archives of Acoustics*, **45**(2): 349–357, https://doi.org/10.24425/aoa.2020.133155.
25. Lennartsson R.K., Dalberg E., Persson L., Petrović S. (2009), Passive acoustic detection and classification of divers in Harbor environments, [in:] *OCEANS 2009*, https://doi.org/10.23919/OCEANS.2009.5422407.

26. Li P., Wu J., Wang Y., Lan Q., Xiao W. (2022), STM: spectrogram transformer model for underwater acoustic target recognition, *Journal of Marine Science and Engineering*, **10**(10): 1428, https://doi.org/10.3390/JMSE10101428.
27. Lin J., Kilgour K., Roblek D., Sharifi M. (2020), Training keyword spotters with limited and synthesized speech data, [in:] *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings*, https://doi.org/10.1109/ICASSP40776.2020.9053193.
28. Mahmoud S., Saleh L., Chouaib I. (2025), Experimental results of diver detection in Harbor environments using single acoustic vector sensor, *Archives of Acoustics*, **50**(2): 173–185, https://doi.org/10.24425/aoa.2025.153663.
29. Phung S.L. *et al.* (2019), Mine-like object sensing in sonar imagery with a compact deep learning architecture for scarce data, [in:] *2019 Digital Image Computing: Techniques and Applications (DICTA 2019)*, https://doi.org/10.1109/DICTA47822.2019.8945982.
30. Radford C.A., Jeffs A.G., Tindle C.T., Cole R.G., Montgomery J.C., (2005), Bubbled waters: the noise generated by underwater breathing apparatus, *Marine and Freshwater Behaviour and Physiology*, **38**(4): 259–267, https://doi.org/10.1080/10236240500333908.
31. Seo D., Oh H.-S., Jung Y. (2021), Wav2KWS: transfer learning from speech representations for keyword spotting, *IEEE Access*, **9**: 80682–80691, https://doi.org/10.1109/ACCESS.2021.3078715.
32. Stolkin R. *et al.* (2006), Feature based passive acoustic detection of underwater threats, [in:] *Proceedings SPIE 6204, Photonics for Port and Harbor Security II*, https://doi.org/10.1117/12.663651.
33. Sun Y. *et al.* (2022), Multi-resonance flextensional hydrophone for open-circuit scuba diver detection, *AIP Advances*, **12**(11): 115310, https://doi.org/10.1063/5.0101999.
34. Sun Y. *et al.* (2024), Feature extraction methods for underwater acoustic target recognition of divers, *Sensors*, **24**(13): 4412, https://doi.org/10.3390/s24134412.
35. Sutin A., Salloum H., DeLorme M., Sedunov N., Sedunov A., Tsionskiy M. (2013), Stevens passive acoustic system for surface and underwater threat detection, [in:] *2013 IEEE International Conference on Technologies for Homeland Security (HST)*, https://doi.org/10.1109/THS.2013.6698999.
36. Tu Q., Yuan F., Yang W., Cheng E. (2020), An approach for diver passive detection based on the established model of breathing sound emission, *Journal of Marine Science and Engineering*, **8**(1): 44, https://doi.org/10.3390/JMSE8010044.
37. Vincent P., Larochelle H., Lajoie I., Bengio Y., Manzagol P.-A. (2010), Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, *Journal of Machine Learning Research*, **11**: 3371–3408.
38. VoiceBosch (2024), Sonar, Pixabay, https://pixabay.com/sound-effects/sonar-183436/ (access: 21.05.2025).
39. Wang W., Huang Y., Wang Y., Wang L. (2014), Generalized autoencoder: a neural network framework for dimensionality reduction, [in:] *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, https://doi.org/10.1109/CVPRW.2014.79.
40. Wang X., Meng J., Liu Y., Zhan G., Tian Z. (2022), Self-supervised acoustic representation learning via acoustic-embedding memory unit modified space autoencoder for underwater target recognition, *The Journal of the Acoustical Society of America*, **152**(5): 2905–2915, https://doi.org/10.1121/10.0015138.
41. Wang Y. *et al.* (2021), Underwater communication signal recognition using sequence convolutional network, *IEEE Access*, **9**: 46886–46899, https://doi.org/10.1109/ACCESS.2021.3067070.
42. Zhang X., Lu Z., Kang C. (2003), Underwater acoustic targets classification using support vector machine, [in:] *Proceedings of 2003 International Conference on Neural Networks and Signal Processing*, https://doi.org/10.1109/ICNNSP.2003.1280753.
43. Zhao J. *et al.* (2023), Underwater target perception algorithm based on pressure sequence generative adversarial network, *Ocean Engineering*, **286**(Part 1): 115547, https://doi.org/10.1016/J.OCEANENG.2023.115547.
44. Zhao W. *et al.* (2016), Passive acoustic detection of diver based on SVM, [in:] *2016 IEEE International Conference on Mechatronics and Automation*, https://doi.org/10.1109/ICMA.2016.7558635.