IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 September 2024, accepted 10 October 2024, date of publication 14 October 2024, date of current version 30 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3480788

SURVEY

Navigating the Depths: A Comprehensive Survey of Deep Learning for Passive Underwater Acoustic Target Recognition

NILS MÜLLER[1], JENS REERMANN[1], AND TOBIAS MEISEN[2]
[1]ATLAS ELEKTRONIK GmbH, 28309 Bremen, Germany
[2]Institute for Technologies and Management of Digital Transformation, Bergische Universität Wuppertal, 42119 Wuppertal, Germany

Corresponding author: Nils Müller (nils.mueller@atlas-elektronik.com)

Nils Müller and Jens Reermann report financial support was provided by ATLAS ELEKTRONIK GmbH.

ABSTRACT The field of deep learning is a rapidly developing research area with numerous applications across multiple domains. Sonar (SOund Navigation And Ranging) processing has traditionally been a field of statistical analysis. However, in the past ten to fifteen years, the rapid growth of deep learning has challenged classical approaches with modern deep learning-based methods. This survey provides a systematic overview of the Underwater Acoustic Target Recognition (UATR) domain within the area of deep learning. The objective is to highlight popular design choices and evaluate the commonalities and differences of the investigated techniques in relation to the selected architectures and pre-processing methods. Furthermore, this survey examines the state of UATR literature through the identification of prominent conferences and journals which points new researchers in directions where to allocate UATR related publications. Additionally, popular datasets and available benchmarks are identified and analysed for complexity coverage. This work targets researchers new to the field as well as experienced researchers that want to get a broader overview. Nonetheless, experienced sonar engineers with a strong background within classical analysis also benefit from this survey.

INDEX TERMS Deep learning, passive sonar classification, underwater acoustic target recognition.

I. INTRODUCTION

Underwater Acoustic Target Recognition (UATR) describes techniques of classifying, categorizing, and identifying unknown surface vessels or submarines through the analysis of their emitted acoustic signals [1], [2]. Acquiring knowledge about the source type of underwater noise has applications in maritime safety [3], [4], maritime traffic -monitoring [4], [5] and -management [6], navigation, surveillance [7], and in the recognition [8], [9] as well as protection of marine life [10], [11]. In many domains, remote sensing is mostly accomplished in the visual, or RADAR domain. However, most modalities are not able to penetrate the water column, effectively making them not applicable for underwater sensing. In contrast, acoustic waves are able to propagate long ranges underwater [12], [13]. Up to this date, classification of unknown maritime vessels are mostly performed manually by highly experienced sonar operators [14], [15], [16]. High performance sonar systems are able to detect, localize and differentiate subtle noise sources. Accompanied by the increasing number of maritime vessels through the ongoing process of globalization [17], [18], the number of targets sonar operators have to identify, manage and classify has been steadily increasing. The manual assessment has become infeasible from a pure time and cost perspective. Initial automated approaches utilized handcrafted features and classical statistical analysis. These approaches often result in insufficient robustness and generalization capabilities encountered at different recording conditions. Therefore, a great demand for reliable, automatic classification systems is present. Deep learning based approaches have been shown to increase the generalization

The associate editor coordinating the review of this manuscript and approving it for publication was Sawyer Duane Campbell.

capabilities. Also in other sonar domains like active sonar deep learning techniques are of interest [19], [20] which is not scope of this survey.

This paper aims to give a systematic up-to-date overview of the UATR field. In the past, there have already been various contributions to this subject that differ significantly from this work in terms of their objectives or level of detail. In [21], the authors gave a highly regarded review on shoreline surveillance methods with active and passive sonar systems, with deeper focus on the latter. In their work, the authors focused on various feature selection methods, as well as common architectures and learning strategies of various research works. Another, more recent overview that extends the study of [21] based on the commonalities and differences between utilized architectures and methods is provided by [14]. Despite the comprehensive analysis of various UATR methods considered here, we have identified difficulties with regard to the reproducibility and comparability of the research examined in [14]. Therefore, we have adapted the criteria of our survey to address both points, and to allow a direct comparison between the different contributions. The intention behind this objective is to provide the reader with an overview of the UATR field from the literature point of view. The data scarcity is a widely discussed difficulty in the deep learning based UATR [8], [11], [22], [23], [24], [25], [26], [27], [28], [29]. Therefore, a deep analysis on the available datasets and benchmarks with an extensive explanation regarding the complex underwater acoustic environment is provided in this work. In summary, this survey outlines three major contributions.

- **Examine the state of the UATR literature.**
  Show the historic development of the UATR field and describe how and where to allocate research works regarding UATR. Identify prominent conferences and journals of high quality papers and derive their key contributions.
- **Extend the technical analysis of UATR approaches from existing surveys.**
  The intention lies in identifying commonalities and differences between various approaches in terms of utilized architectures, feature representations, and learning strategies, while retaining reproducibility and comparability.
- **Provide in-depth analysis of the available datasets and benchmarks.**
  Investigate popular datasets and benchmarks with a spotlight on the complexity coverage. The in-depth analysis of the publicly available datasets highlights gaps that potential future datasets should cover in order to obtain an overall pool of well generalizable datasets that enables the UATR domain to advance similar to other domains like computer vision and speech recognition.

This survey is structured as follows. Section II gives an overview of the literature search process and the selection criteria applied. This section shows the historic development of the field and outlines the most qualitative journals and conferences for UATR considered in this survey. Section III gives a concise recap of the theoretical background knowledge for underwater acoustics, followed by a brief description of the selection of classical approaches and their corresponding strengths and limitations in section III-C. Subsequently, the benefits and promises of deep learning approaches are outlined, followed by a comparison of common approaches in section IV. The reproducibility of the examined publications is given in section V. Section VI covers the different available databases and gives insight to their complexity coverage. At last some concluding remarks are given in section VII followed by a summary of this work in section VIII.

II. METHODOLOGY

For the acquisition of publications, Elsevier's Scopus and Clarivate's Web Of Science databases were selected due to their multidisciplinary topics and extensive corpus for our literature search. The search term for both databases composed of the union of ''Underwater Acoustic Target Recognition'' and ''UATR''. Since the focus is put on deep learning approaches, the publication time span was set from 2012 to 2023, where 2012 marks the initial boom of deep learning with the introduction of the ImageNet [30] benchmark and AlexNet [31].

A. SEARCH RESULTS

The results obtained with the Scopus and Web of Science database were identical, returning the same 102 publications, respectively. Of these 102 publications, 89 had the topic of classification, 13 had different recognition topics other than UATR (like marine mammal identification and target detection/ localization), five dealt with augmentation methods, four mainly with generative approaches that focused on extending the data quantity, and one publication was one of the two aforementioned surveys [14]. The literature search was conducted in April 2024. Henceforth, these 102 publications will be referred to as the ''UATR literature corpus''.

B. SELECTION PROCESS

The research papers considered in this work have been carefully selected according to the described UATR topic. Accordingly, this mainly includes research on clustering, classification or identification of unknown maritime vessels, solely based on their emitted acoustic fingerprint. Additionally, criteria such as comprehensibility and reproducibility are also taken into account in the selection process. These include studies with comprehensive descriptions of the architectures used, a detailed description of the data pre-processing chain and a reproducible training procedure. Furthermore, the included research papers should make reasonable comparisons to similar approaches. This type of selection process is fairly subjective. However, this selection criteria highlights well-structured and comprehensible research works on which other, especially new research can easily build upon. This

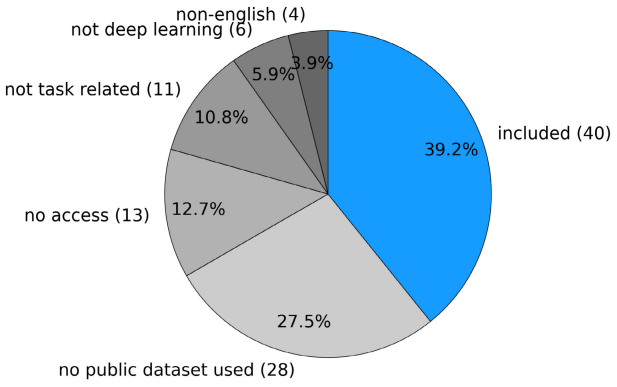

FIGURE 1. Amount of included and excluded papers from this survey.

selection process indicates a general structure of research works in the UATR domain, which new researchers can adapt and orientate to. The selection criteria are summarized as follows.

1) **UATR Task Related**.
   The work should focus on the detection of maritime vessels based solely on their noise. While other modalities could be taken into account as complementary, the classification based on their passive sonar signature is required.
2) **Deep Learning Approach**.
   While many works exist on classical statistical analysis, the objective of this work is to compare deep learning based approaches.
3) **Reproducibility and Comprehensibility**
   In order for a publication to be considered as comprehensible, the model architecture, method, and data pre-processing strategy has to be described. Additionally, for the sake of reproducibility, the research work must conduct their work on a publicly available dataset. To this work, we define the requirement of the usage of either the ShipsEar- [32], DeepShip- [33] or the Ocean Networks Canada dataset.

As stated above, this work is focused on the classification of unknown maritime vessels. An extensive review of deep learning based detection of unknown passive targets is given in [34]. A comprehensive overview of deep learning based detection and classification of sonar targets in image based sonar is given by the authors in [19].

We skimmed the publications on the basis of their title, abstract and experimental section, in order to gain an initial overview of studies relevant to the objective of this survey according to the criteria stated above and to narrow down the considered publications. After the skimming process, roughly two thirds of the publications of corpus A were discarded. Fig. 1 displays the distributions of included and discarded publications and gives insight into the reason certain publications were excluded.

Fig. 1 show that roughly a quarter (27.5%) of the resulting publications were discarded, as they did not include experiments on one of the three mentioned public datasets conflicting with the reproducibility criterion. We could not gain access to 13 of the 102 publications (12.7%) despite given access to most publishers. Nonetheless, despite appropriate effort these publications were only accessible with a costly subscription, or the original papers were untraceable. A total of eleven papers (10.8%) were not related to the UATR task of identifying maritime vessels. These publications focused on either the detection and localization of maritime vessels or on the identification of different noise sources, such as marine mammals. Six papers (5.9%) did not apply any deep learning based method to the UATR task, and four papers (3.9%) were not published in English. In total, 40 (39.2%) of the initial 102 publications remained which matched the selection criteria and therefore are considered as relevant to this survey. We will refer to these 40 publications as the ''relevant studies corpus''.

It is important to mention that a significant amount of promising publications of well known representatives of the passive sonar domain have been excluded from this survey on the basis of the aforementioned selection criteria. Additionally, some publications were simply not indexed by the SCOPUS and Web Of Science Database. It is advisable for new researches to extend their literature search outside the scope of this survey and directly look for publications of some prominent research institutes. A selection of well represented institutions in the underwater acoustic domain are given in section II-D.

C. DEVELOPMENT OF THE UATR FIELD

This section provides a short overview of the historic development of the UATR domain. We aim to contrast the development of UATR approaches based on classical statistical and deep learning based methods. A second literature search using the ''passive sonar classification'' (PSC) keywords was conducted with the Web of Science Database. We will refer to the result of this query as the ''PSC literature corpus'' in the remainder of this study. The query resulted in a total of 100 publications. After a similar skimming process as with the first literature search, we were able to divide the publications into 23 deep learning and 77 non-deep learning related papers. The 23 deep learning papers were already present in UATR literature corpus. The following Fig. 2 shows the development of the research field in the last 30 years. Note that six of the publications in UATR literature corpus were not deep learning related (see section II-A) resulting in the 96 papers shown in the following Fig. 2. We will refer to this subset of the UATR literature corpus as the ''deep learning literature corpus''. This development is demonstrated using the deep learning corpus (96 papers) and the non-deep learning publications obtained with the ''PSC'' keyword search (77 papers).

Fig. 2 clearly shows an increase in the number of publications from a handful of publications every few

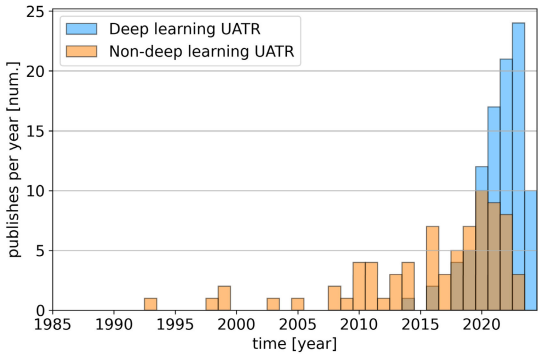

FIGURE 2. Development of the number of published articles under the keywords "Underwater Acoustic Target Recognition", "UATR" and "passive sonar classification" from Elseviers SCOPUS and Web of Science database up to the year 2023.

years up to ten publications per year under the "PSC" keyword. The number of publications under the "UATR" keyword underwent a similar development, but much later and more drastically. Currently, the total accumulated number of publications roughly match. The 92 publications with the "UATR" keywords consist mainly of research conducted on deep learning methods. While the research work found under the "PSC" keyword focuses on classical statistical approaches on handcrafted features. An interesting milestone in the development lies around the year 2020, where the number of "UATR" publications (12) firstly surpasses the number of "PSC" publications (10). In addition, the year 2020 also marks the turning point where the number of papers published under the secondary "PSC" keyword has been rapidly decreasing ever since. Fig. 2 highlights three important developments. Firstly, the number of publication in the UATR domain is rapidly increasing every year. Secondly, newly published research obviously shifted from classical statistical approaches to deep learning based methods. Lastly, the term "Underwater Acoustic Target Recognition" has become the associated keyword that addresses classification of unknown maritime vessels and has replaced "passive sonar Classification" which covers a more broad spectrum of applications.

D. BIBLIOGRAPHIC ANALYSIS

All analysis are only applied to the initial UATR search query. This section gives a brief overview of the most significant journals and conferences for UATR. Additionally, this bibliographic analysis should identify the placement of the UATR subject within the general research field. In order to identify the associated research field, we utilized the "Scimage Journal & Country Rank" website,[1] (last accessed: 30.04.2024, 11:23) and the associated research categories. For this survey, the identification of qualitative publishers is based on the total number of papers published under the "UATR" corpus and the number of papers included in this survey according the aforementioned selection criteria. This metric yields high quantity but low quality publishers and rewards the contrary. The significance σ of a publisher to this work is determined as:

$$\sigma = n_{relevant} + \frac{n_{relevant}}{n_{total}} \tag{1}$$

where σ describes the significance score, $n_{relevant}$ is the number of publications from the UATR corpus considered in this work and n_{total} is the total number of journal/ conference publications found under the UATR keyword. Fig. 3 displays the most significant journals and conferences to this work, based on the aforementioned score.

Fig. 3 shows the highest publishing journals and conferences in the UATR domain in terms of quantity. Only seven journals published more than a single UATR related paper. The "Journal of Marine Science and Engineering" is the largest UATR publisher with a total amount of eight publications, of which seven are included in this survey. The second-highest number of publications are from the "Sensors" journal, however scoring lower than the "Electronics" journal it includes irrelevant papers to this study. This penalization can also be seen in the last two displayed journals "Remote Sensing" and "Entropy" where the majority of the published papers were excluded from this survey. The same characteristic can be seen in the distribution of conferences proceedings. The major conference in the UATR field is the "OCEANS" Conference. Even though many papers were published as proceedings of various "Inter-Noise" conferences, the majority are considered as irrelevant to this study, resulting in a low score.

In addition to the metric σ, the mean number of citations for each journal is given by μ_{cites} to evaluate the visibility of publications within each journal and conference proceeding by comparing to the mean average citation count (MACC) of the corresponding journal. All bibliographic metrics are also displayed in Table 12 (see appendix). The "Journal of Marine Science and Engineering" and "Electronics" score the best in terms of relevant publications. However, the publications are only cited on average by 4.12 and, 1.5 respectively. In contrast, the two publication within the "IEEE Access" journal are cited 17 times on average. The distribution of the average citation count does not accompany the distribution of the number of relevant papers. The visibility of publications within the regarded journals is determined by subtracting the MACC from the mean number of citations μ_{cites}. Fig. 4 shows the paper visibility for each journal.

Fig. 4 shows that the visibility of the papers within the regarded journals follows a similar trend as the mean number of citations, μ_{cites} which can be derived from Fig. 3. Therefore, the highest visibility of publications are currently achieved in the "IEEE Access", "Applied Science - MDPI" and "Geoscience and Remote Sensing Letters - IEEE" journals.

The 96 UATR publications of the deep learning literature corpus were published in over 60 different journals and conference proceedings, whereas the 40 publications of the relevant studies corpus were published in 29 different journals

[1] (https://www.scimagojr.com/journalrank.php

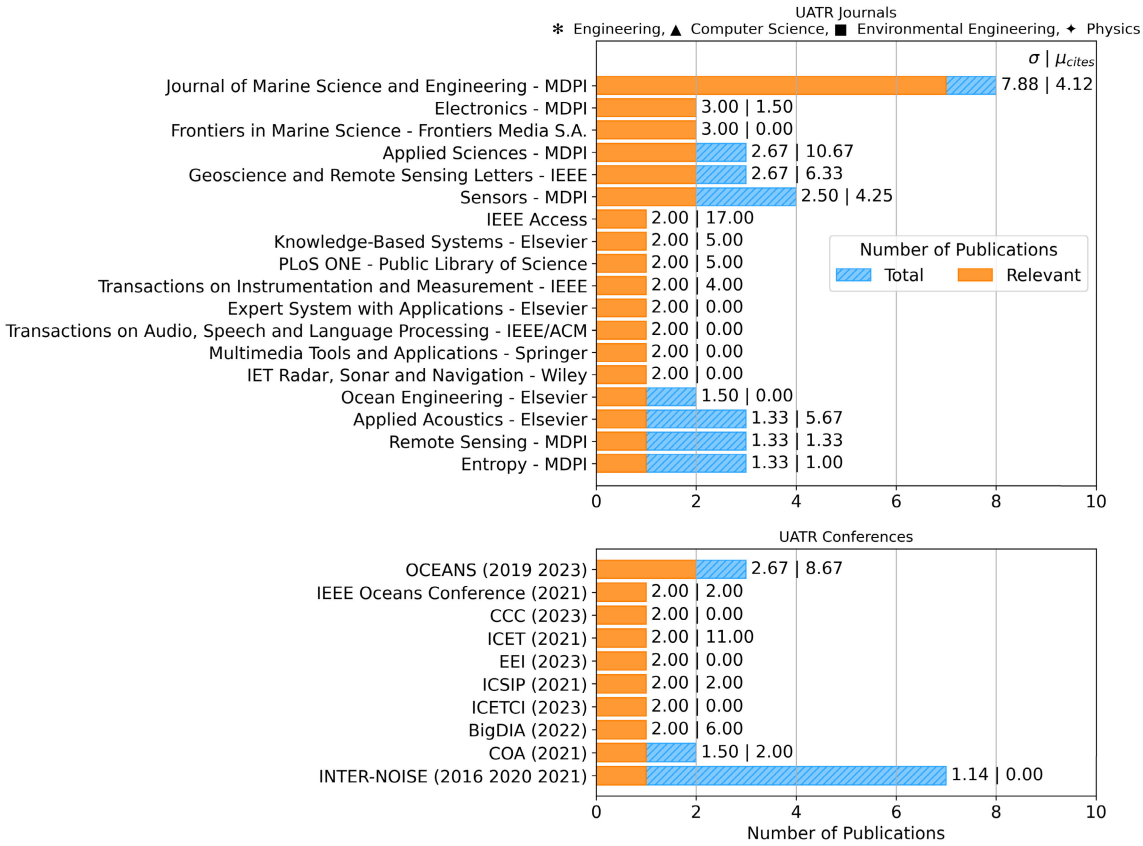

FIGURE 3. The most contributing journals and conferences to the UATR task, based on the total number of published work and the number of relevant work.

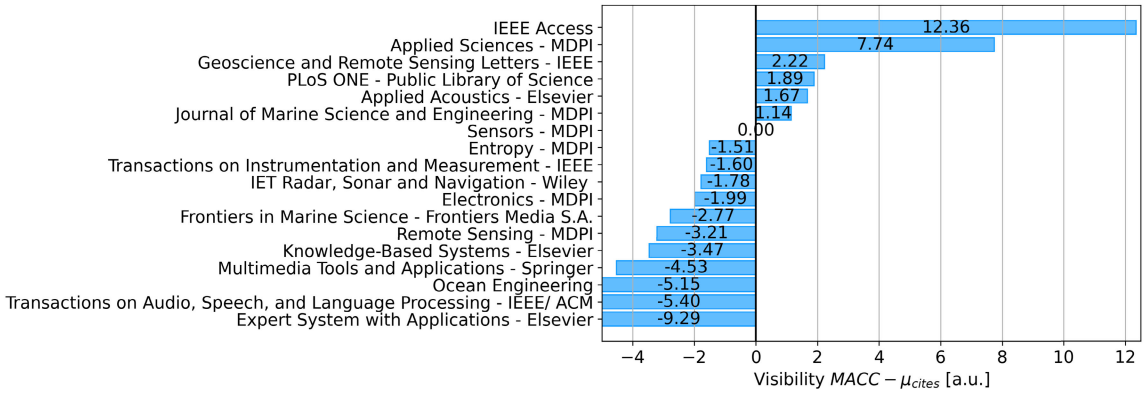

FIGURE 4. The visibility of publications within the included journals. The visibility is determined by subtracting the MACC from the mean number of paper citations for the corresponding journal.

and conference proceedings. This distribution accentuates the niche domain of UATR, indicating that a ''standard'' journal that focused solely on passive sonar processing is not present yet. Merely the ''Journal of Marine Science and Engineering - MDPI'' indicates a development towards this direction. Additionally, the associated research subject and category of the journal is indicated using the symbols: triangle for engineering, asterisk for computer science, square for environmental engineering and star for physics. The analysis using the ''Scimage Journal & Country Rank'' website showed that the UATR domain mainly is assigned to following research domains: engineering (15), computer science (8), environmental science/ engineering (5), and physics (5).

To elucidate the differences between the two keyword searches applied and to give a more extensive overview of

the research field, Table 1 presents a selection of the most frequently represented journals, authors and affiliations in the PSC- and UATR literature corpora, as well as the articles that were included in this analysis.

While a majority of journals appear in both PSC- and UATR literature corpora, the appearance of authors and the corresponding affiliations are more exclusive to the corpora. Despite the extensive range of research works reviewed in this study, several promising studies from the PSC literature corpus were deliberately excluded from this survey due to non-compliance with the inclusion criteria outlined in Section II-B. These included works on target detection and localization, sonar imaging, any recognition of non ship radiated noise sources, not deep learning related works and works that did not benchmark on the publicly available datasets. Therefore, after the cleansing, no publications from the journals like ''Sensors'', ''Journal of the Acoustical Society of America'', ''IEEE Journal of Oceanic Engineering'', ''Defence Science Journal'', ''Mechanical Systems and Signal Processing'' and ''IEEE Signal Processing Letters'' remained.

Nevertheless, we strongly encourage readers to consider relevant research beyond the scope of this survey. We aim to underscore significant and influential studies that were not included in this review. The following articles are derived from the literature reviews provided by [14], [21] who provide are more general machine learning oriented view on UATR and from highly cited articles that did not fulfil all selection criteria stated in section II-B. For instance, Hu et al. [23] pioneered the integration of depthwise separable convolutions within a 1D CNN architecture, facilitating the direct extraction of discriminative features from raw waveform signals. Similarly, Doan et al. [7] proposed a CNN architecture applied directly to raw waveform signals, demonstrating the effective use of skip connections to reuse features extracted by earlier layers. The authors in [35] addressed the challenge of data scarcity by leveraging unlabelled data with a deep belief network. Dynamic effects on the received signal, though often overlooked, are critical in real-world applications. Xue et al. [36] addressed these challenges by employing channel attention modules to manage Doppler shifts in a custom-acquired dataset. Additionally, the authors in [37] captured temporal correlations from time-frequency representations using a second-order pooling strategy within a CNN architecture applied to custom-recorded signals. Most studies on underwater acoustic target recognition (UATR) have focused on single-label target recognition, whereas in many practical scenarios, multiple targets are encountered simultaneously. Beckler et al. [38] addressed this limitation by employing a Bayesian formulation within a standard CNN to demonstrate multi-label recognition.

Furthermore, several prominent institutions in the fields of oceanography and sonar signal processing are notably absent from this literature review, as no publications utilizing deep learning for the recognition of ship-radiated noise were identified within either the initial or extended literature corpora under the ''Passive Sonar Classification'' keyword. The selection of the provided research institutions are widely represented from an applicative point of view in the sonar domain. We would therefore highlight the importance to direct the reader's attention to these institutions, which may contribute relevant research in future studies. A selection of these institutions includes:

- ''Center for Acoustics Research and Education'', University of New Hampshire[2]
- ''Institute of Sound and Vibration Research'', University of Southampton[3]
- ''United States Naval Research Laboratory''[4]
- ''Woods Whole Oceanographic Institution'' (WHOI)[5]
- ''Defence Science and Technology Laboratory'' (DSTL), UK Ministry of Defence[6]
- ''Applied Research Laboratories'' (ARLL), The University of Texas at Austin[7]
- ''Scripps Institution of Oceanography'', University of California San Diego[8]
- ''The Applied Research Laboratory'', Penn State University[9]
- ''Atlantic Research Center'' (ARC) Defence Research and Development Canada Research Center[10]

III. THEORETICAL BACKGROUND

The main intention of UATR is to differentiate between targets by analysis of the emitted acoustic noise. Unfortunately, the received acoustic signal is not only affected by the target vessel design and operating conditions, but also by various environmental influences. This section provides a brief overview on the physics of underwater acoustics and highlights the major difficulty in UATR that arises from this complicated domain.

In their studies, the authors in [39] and [40] demonstrated that different ship types have distinct spectral characteristics. Ship radiated noise follows a complex generation mechanism, where the three main contributors of noise have mechanical-, propeller-, and hydrodynamic origins [10], [35], [41]. Mechanical noise describes the noise generated by the propulsion system, machinery, gearbox, and other mechanical actuators on or within the vessel. Propeller noise describes the modulation of broadband cavitation noise caused by the propeller rotation. Hydrodynamic noise is generated through boundary interactions between the vessel's

[2](https://eos.unh.edu/center-acoustics-research-education

[3](https://www.southampton.ac.uk/about/faculties-schools-departments/school-of-engineering/institute-of-sound-and-vibration-research

[4](https://www.nrl.navy.mil/

[5](https://www.whoi.edu/

[6](https://www.gov.uk/government/organisations/defence-science-and-technology-laboratory

[7](https://wwwext.arlut.utexas.edu/labs.shtml

[8](https://scripps.ucsd.edu/

[9](https://www.arl.psu.edu/

[10](https://www.canada.ca/en/defence-research-development/services/capabilities.html

TABLE 1. Comparison of the article appearance quantity of widely represented journals, authors and affiliations under the "Passive Sonar Classification", "Underwater Acoustic Target Recognition" keywords and the research works included in this survey.

Category	Rank	Name	Publications in Corpus		
			PSC	UATR	included
Journal	1	Journal of Marine Science and Engineering	3	9	7
	2	Sensors	8	7	
	3	Applied Acoustics	2	3	1
	4	Journal of the Acoustical Society of America	7	1	
	5	IEEE Journal of Oceanic Engineering	5		
	6	Defence Science Journal	4		
	7	Sensors (MDPI)		4	2
	8	Entropy	4	3	1
	9	IEEE Geoscience and Remote Letters	2	3	2
	10	Applied Science – MDPI	2	3	2
	12	Mechanical Systems and Signal Processing	3		
	13	IET Radar Sonar and Navigation	3	1	1
	14	Remote Sensing		3	1
	15	Frontiers in Marine Science		2	2
	16	IEEE Signal Processing Letters	2		
	17	IEEE Sensors Journal	2		
	18	Journal of the Acoustical Society of Korea	2		
	19	Measurement		2	
	10	Ocean Engineering	2	2	1
	20	Lecture Notes in Computer Science	2		
Author	1	De Seixas JM	9		
	2	Yang HH		9	2
	3	De Moura NN	6		
	4	Li Jh		5	1
	5	Luo XW		5	5
	6	Kamal S	5		
	7	Zhao DX		4	2
	8	Supriya MH	4		
	9	Zhang W	3	1	1
	10	Liu F		4	2
	11	Liu DL		4	
	12	Shen TS		4	2
	13	Chandran CS	4		
	14	Xu J		4	1
	15	Xie Y		4	1
	16	Chen L		4	3
	17	Li DH		3	2
	18	Wang Y		2	3
	19	Varut RM		3	
	20	Liu JH		3	
Affiliation	1	Northwestern Polytechnical University	11	17	7
	2	Universidade Federal Do Rio de Janeiro	12		
	3	Chinese Academy of Sciences	4	10	2
	4	Harbin Engineering University	7	8	4
	5	University of Chinese Academy of Sciences CAS	8	7	1
	6	Cochin University Science Technology		7	1
	7	Institute Of Acoustics CAS	7		
	8	Southeast University China		6	
	9	Applied Science – MDPI		5	5
	10	United States Navy	4		
	11	Chinese ACAD MIL SCI		4	2
	12	United States Department Of Defence	4		
	13	Tiangong University		4	
	14	Zhejiang University		3	2
	15	Naval Physical Oceanographic Laboratory NPOL	3		
	16	Indian Institute Of Technology System IIT System	3		
	17	Portland State University	3	2	
	18	University Of Electronic Science technology Of China		3	1
	19	Defence Research Development Organisation DRDO	3		
	20	University of Sheffield	3		

hull and the ocean medium. The composition of these noise sources commit to unique acoustic fingerprints of maritime vessels. The effects on a received underwater waveform, with respect to amplitude, frequency and phase, are displayed in Fig. 5.

A sound wave propagating in water can either be subject to dynamic effects that alter the features of the emitted wave or to additive effects, like noise and other sources. The origin of and degree of these effects are highly dependent on the corresponding environment, geographic region, season

as well as target- and receiving platform design. The high level influences segment into further intermediate influences. In Fig. 5, these low level features are separated into a controllable (marked green) and non-controllable set (marked red). When generating a dataset, one would technically have information regarding the controllable parameters, whereas the non-controllable parameters can only be influenced indirectly. An ideal dataset for the underwater acoustic domain should account for sufficient variability in all influences to ensure an appropriate complexity coverage. This is later covered in section VI. For more in-depth explanation on the cross-correlation of the aforementioned variables, we highly suggest reviewing some standard literature for underwater acoustics, such as [42], [43], and [44].

As any mechanical wave that transports energy through a medium, the most important quantities to describe underwater acoustic waves are amplitude, frequency, and phase [4]. The amplitude is an umbrella term for the received pressure to a certain area. Analogous to every physical wave, underwater acoustic waves are also subject to interference, attenuation, as well as frequency and phase shifting effects. Despite the basic similarity, many approaches for classical acoustics are not applicable to underwater acoustics, mainly due to the inhomogeneous nature of the oceanic environment [4]. Underwater acoustic signals experience various propagation and transmission losses [3] as well as additive effects like ambient noise.

Regular inhomogeneities like time varying underwater-, and surface channels, as well as antiwaveguides such as shadow zones are caused by different sound velocity profiles [42, p. 2-9], [29]. The sound velocity profile typically depends on the water temperature, salinity, and pressure [42, p. 1], and is consequently dependent on the geographic region and time.

Transmission loss refers to the reduction of received signal strength as a wave travels through and interacts with a medium. The received pressure is affected by many effects. Primarily, every mechanical wave suffers absorption by the propagation medium, where some energy is absorbed in terms of heat as the wave interacts with the molecules. This effect increases at larger propagation distances. Secondly, transmission losses such as scattering, geometric spreading and reflections decrease the amplitude. Scattering losses describe the redistribution of energy when acoustic signals scatter of particles, bubbles and other irregularities in the water column. As acoustic waves propagate omnidirectionally, geometrical losses occur where the energy is dispersed spherically. In the special case of an underwater acoustic channel, the energy is dispersed cylindrically. Reflection losses occur at boundaries of media with different propagation velocities, such as the seabed or the water surface. This results in the signal's attenuation as partial energy quantities are either transmitted or reflected [42]. Lastly, mechanical properties and operational states of the transducer can alter the emitted amplitude [11]. All the aforementioned effects contribute to the decrease of signal strength mutually. Generally, underwater acoustics is regarded as a low Signal-to-Noise-Ratio (SNR) scenario [7], [16], [27], [41], [45], [46]. The typical expected spreading losses under different circumstances in dependency on the target distance are displayed in Table 2. Real transmission losses are mainly comparable to spherical spreading.

TABLE 2. Underwater acoustic spreading loss dependent on target distance, r adapted from [47, p. 102].

Type	Intensity varies as	Transmission loss [dB]
No spreading	r^0	0
Cylindrical	r^{-1}	$10 \log r$
Spherical	r^{-2}	$20 \log r$
Hyperspherical	r^{-3}	$30 \log r$

Furthermore, the propagation of underwater acoustic signals is usually accompanied by multiple frequency alternating effects. First and foremost, relative dynamic movements between the signal transducer and receiver cause frequency shifts due to the Doppler effect [27]. Secondly, vessel operating conditions like engine revolutions per minute, gear choice etc., but also general ship conditions contribute differently to the emitted frequency spectrum [11], [29]. Dissipation and absorption of signal energy is typically frequency-dependent, where high frequencies are attenuated stronger compared to lower frequencies. Reflection, scattering, and refraction of an incident wavefront at boundaries cause frequency dependent generations of a secondary wavefront that superimpose with the original wave [43, p. 352]. Additive effects like ambient noise sources contribute additional frequency components to the spectrum of a received signal. Fig. 6 displays the frequency dependant attenuation modelled by the empirical derived Francois-Garrison formula for various seawater temperatures.

Lastly, the received phase of an underwater acoustic signal can be distorted or altered through various effects. Acoustic turbulence such as local inhomogeneities of the water column cause fluctuations in the velocity and direction of the sound wave. Boundary reflections at the seabed, water surface or submerged objects can cause phase shifts. Dynamic properties such as movement of the emitter and receiver also alter the phase component. At last, acoustic signals can take multiple propagation paths due to reflection, refraction, and scattering at the surface and seabed. This results in different paths and therefore distances the acoustic signal can propagate from the emitter to the receiver, consequently resulting in phase shifts. Multipath propagation especially occurs in shallow water regions and can add a significant temporal complexity to the signal [3], [7], [10], [49]. At last, the characteristics of the transducer and receiver also influence the received phase due to design, mounting and processing choices.

In addition to the preceding losses and modifications caused by the wave propagation in water, different additive contributions of ambient noise influence the SNR and spectrum of a received signal. At low frequencies 0.1 –

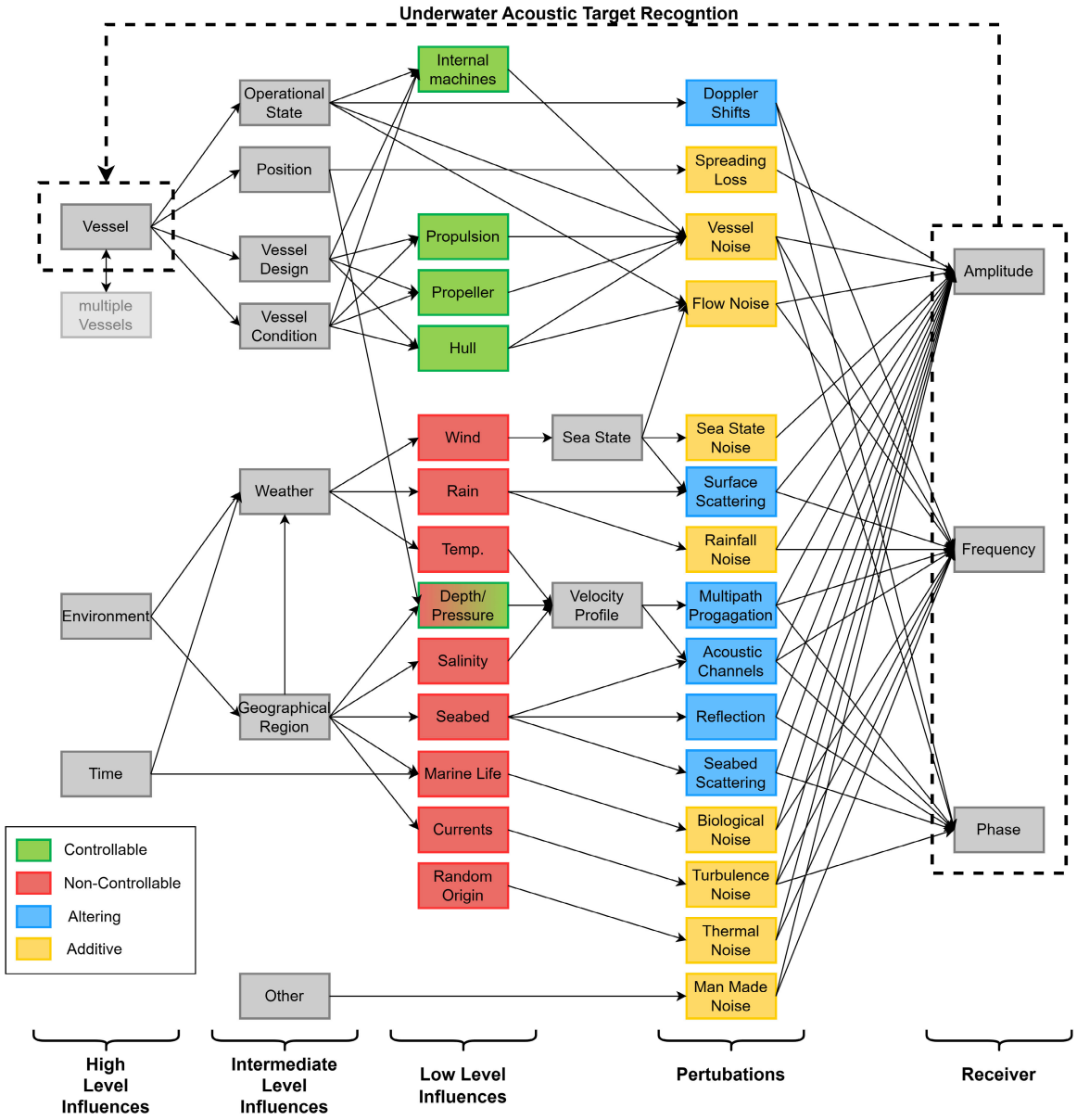

FIGURE 5. The controllable and non-controllable parameters of received acoustic signals, and their corresponding dependencies and effects on the amplitude, frequency and phase. The effects are divided into additive effects like noise and altering effects caused by transmission losses. Controllable parameters are considered controllable if a degree of choice exists in during the dataset generation.

10 Hz turbulence noise is present which includes alterations caused by ocean, atmosphere, and geophysical noise sources like underwater volcanic eruptions and earthquakes [42, p. 30, 31]. In the frequency band between, 10 − 300 Hz the main noise present is caused by remote ship traffic. The significant increase in commercial shipping, and the low attenuation of sound at these frequencies, generate a continuous background noise [4], [10], [13], [41], [50]. The main source for traffic/ vessel noise are the propulsion machinery, propeller interactions in water and the flow noise generated by the water movement at the ship's hull. This type of noise is the main of interest to UATR. The state of the sea itself is also a major noise contributor with many mechanisms. This mainly includes the breaking of waves on the sea surface and the accompanying cavitation noise of collapsing bubbles. This effect is highly associated with the wind conditions [42, p. 30, 31]. Other weather condition such as rainfall also generate broadband noise on the surface, typically in the range between 1 kHz and 5 kHz. Additionally, biological noise produced by the marine fauna can have a considerable impact on the noise level in certain regions and seasonal time spans [17]. Furthermore, the random movement of atoms and molecules due to thermal energy, also referred to as molecular agitation or thermal noise, generates

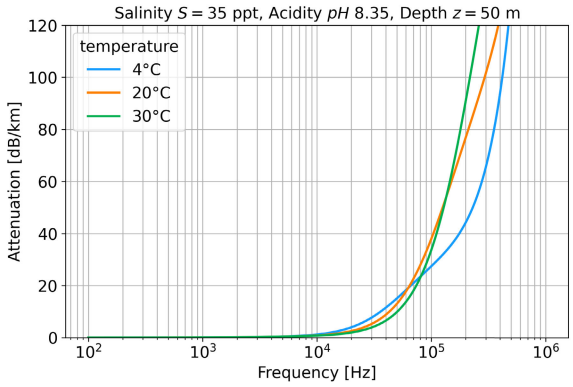

FIGURE 6. Attenuation according empirical francois-garrison formula for various sea water temperatures [48].

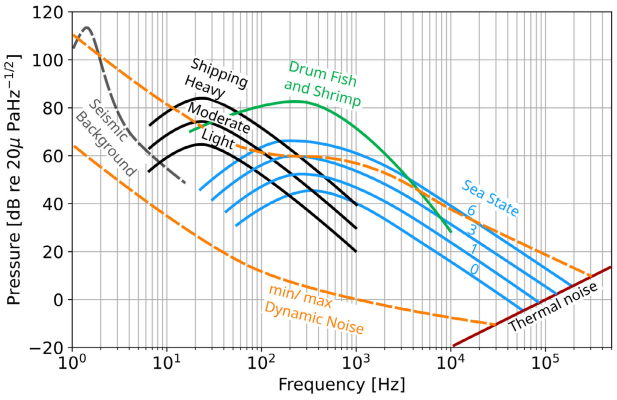

FIGURE 7. Contributions to the isotropic noise level adapted from [42, p. 32] and [47, p. 212].

a noise spectrum in the range between 100 kHz and 1 MHz. Lastly, uncontrollable man-made noise contributions, such as sea floor mining/ drilling, active sonar systems etc. can be present throughout a very broad frequency spectrum. All the aforementioned noise sources contribute differently to the isotropic noise level. The frequency dependant noise level for various selected ambient noise contributions is displayed in Fig. 7.

A. CHARACTERISTICS OF SHIP RADIATED NOISE

Despite the many influences to a propagating underwater acoustic wave, some key properties of ship emitted noise are evident. This section provides a short overview of these features and also provides general assumptions that can be applied for the signal processing.

Ship radiated noise consists of a wideband continuous spectrum with discrete narrowband frequency components [14], [41], [51]. The propeller cavitation and the flow noise of the vessel generate the broadband component. The target's internal machinery and propulsion systems are the main contributors to distinct, narrowband spectral frequency lines [52], [53]. While target information is predominantly distributed in lower frequencies [10], [39], [50], [54], especially below 5 kHz [51], high frequency components also include target information [53]. Most approaches focus on either determining the low frequency line spectrum or the high frequency modulation caused by the target's propeller [5].

While varying vessel operating conditions result in time dependant irregularities [55], [56], [57], ship radiated noise is mostly considered static in short time [14], [22], [41], [41], [58], [59], where the authors in [13] narrow down the stability between $10 - 30$ ms. The authors in [41] even state that ship radiated noise can be considered static up to 100 ms, making the underwater acoustic signals more stable compared to human speech. Nonetheless, short time variations, like engine start-up, de- and acceleration of the target vessels etc. can still give insight to useful vessel specific information as demonstrated by [55] and [56].

As stated above, the field of UATR is considered a low SNR scenario, where many of the influences described in this section decrease the effective amplitude of a propagating acoustic wave. Despite the distinct features that can be extracted to categorize unknown vessels, ship radiated noise is still subject to severe intra-class variance, where targets of the same category can vary extensively in speed, tonnage, power system and propeller characteristics [3].

B. FEATURES REPRESENTATIONS IN UNDERWATER ACOUSTICS

Traditionally, UATR systems are composed of independent feature extraction and a classification processes [23]. Due to the complexity of the oceanic environment, the extraction of meaningful characteristics is the major key in the vessel classification process [24], [41], [56]. This section is intended to cover the various feature representations commonly used in UATR and highlight the corresponding limitations and strengths. The various feature representation in the UATR domain can roughly be assigned to the following categories; time domain (waveform), frequency (spectrum), time-frequency (spectrogram) and auditory features [15], [24], [60].

Digitized acoustic signals are constructed as one-dimensional signals that describe amplitude oscillations over time. The raw waveform itself contains the highest possible information content of the received signal. Various waveform based features, such as the zero-crossing-rate and the waveform structure itself, are applied in acoustic processing [61]. However, features based solely on the plain temporal waveform are difficult to describe and comprehend. As the classification of unknown maritime vessels predominantly rely on the manual assessment of sonar operators, more sophisticated comprehensible features, such as spectral based characteristics, are extracted from the original signal. The main tool for the extraction of spectral components is the Fourier Transform (FT). The major shortcoming of spectral features is the inability to represent temporal context. Consequently, two-dimensional

time-frequency representations are often applied [62], [63]. The most basic time-frequency representation is the Short-Time-Fourier-Transform (STFT). The STFT is a temporally segmented Fourier Transform (FT) over short consecutive time frames. The output of such time-frequency representations are referred to as spectrograms. In audio processing, especially in speech recognition, the harmonic behaviour of the spectral components are exploited through the extraction of the cepstrum. The cepstrum can capture periodic structures within the spectrogram, and is obtained through the inverse Fourier transform of the logarithmic spectrogram.

As the most energy is concentrated in lower frequencies, research such as [39] and [64] focuses on the extraction of the characteristic narrowband line spectrum that is generated by the ship's internal machinery [14]. A common approach is the extraction of the LOw Frequency Analysis Recording (LOFAR)-spectrogram [41], [57]. In contrast to the narrowband detection, the Detection of Envelope Modulation On Noise (DEMON) focuses on the extraction of the modulation frequencies of the broadband noise caused by the propeller cavitation. The DEMON analysis is therefore utilized for the extraction of distinct propeller characteristics, such as number of propellers and number of blades per propeller [27].

Especially in speech recognition, auditory inspired features achieved satisfactory results [65]. Experienced human operators are able to distinguish different vessel types by timbre, fluctuations in frequency and beat [27]. Inspired by the capabilities of sonar operators, auditory inspired approaches have also made inroads into the UATR domain. As humans do not perceive sound linearly, the logarithmic Mel-Scale is often applied to mimic the human auditory perception [66] in the form of a Log Mel Spectrogram (LMS). The Mel-Scale emphasizes lower- over higher frequencies. Most often, a lower dimensional representation that captures the periodic cepstral properties of the Mel spectrogram are extracted with Mel Frequency Cepstral Coefficients (MFCC) [67], [68]. Besides the MFCC, a similar feature representation does not incorporate the Mel scale, but rather the Gammatone frequency scale to formulate Gammatone Spectrogram (GST). Analogous to the Mel approach, the Gammatone Frequency Cepstral Coefficients (GFCC) can be extracted from the GST [69]. Both MFCC and GFCC utilize a set of pre-defined filters of various bandwidths and amplitude to emphasize different spectral regions of an input signal. In contrast to the MFCC, the Gammatone filter bank applies smooth filters, which allow for more overlap and therefore better correlation between the independent filters [51]. The constant Q-Transform (CQT) is another time-frequency representation closely related to the LMS. Similarly, the CQT offers geometrically spaced frequency filters, where the resolution varies logarithmically. It has its origins in music processing, therefore it aims to have equal resolutions across all octaves [70].

All spectrum or cepstrum based approaches are able to capture energy, frequency and time information. The foundations of these feature representations are build on the Fourier transform, which describes a complex wave structure as a sum of superimposed sinusoidal waves with infinite length. This periodic property of the sinusoidal wave leads to an insufficient time localization capability. Moreover, this characteristic is underlined by the fixed window length of the STFT, where low- and high frequencies are resolved equally in time. The Wavelet Transform (WT) allows for better localization in both time and frequency, allowing for a more precise analysis of non-stationary signals [71]. Thus avoiding the conflict in balancing low- and high frequencies resolution by choosing an adequate window length. The WT adapts to the signal by applying shorter window lengths for high frequencies, and lower window lengths for lower frequencies, leading to a more effective representation of a wide range of signal dynamics. This is especially advantageous for analysing signals with transient components and varying frequencies. However, the Wavelet Transform still requires a priori information to select a proper mother wavelet. More recently, the Hilbert-Huang Transform (HHT) has been used to analyse underwater acoustic signals [72]. The HHT is an adaptive approach that extracts subtle oscillatory-like features from a signal. It focuses only on a certain set of frequencies, which have a high energy response. The main strengths of the HHT lies in the analysis of non-linear and non-stationary signals. It uses the Empirical Mode Decomposition (EMD) to decompose a signal into intrinsic mode functions (IMFs) [73]. The frequency-time analysis is performed by applying a Hilbert Transform on the mode functions [74]. The HHT is preferred over the WT when instantaneous frequencies are of great interest, as the WT assumes a certain level of frequency stability in the time span of the wavelet.

The choice of the different aforementioned time-frequency representations strongly depends on the assumption whether ship radiated noise is considered stationary or non-stationary. Research works, such as [14], [22], [41], [41], [58], and [59] consider ship radiated noise as stationary. In contrast, the researchers in [55], [56], and [75] consider ship radiated noise as non-stationary. There is still quite a discrepancy among the researchers within the relevant studies corpus, however the number of publications that consider ship radiated noise as stationary outweighs the contrary by two thirds. Nonetheless, this is not a simple assumption to make, as it is strongly dependent on the specific use-case. If constantly moving and distant vessels are investigated, the noise is rather stationary, whereas strong de- and acceleration movements and transients like engine startups etc. rather imply the presence of a time varying signal.

C. CLASSICAL CLASSIFICATION APPROACHES

The development of automated sonar processing approaches were predominantly orientated towards the human operator. Consequently, the focus has been historically put on comprehensible and interpretable, mostly time-frequency based features obtained through the STFT, LOFAR and DEMON analysis. The preceding feature representations are usually combined with simple statistical-, more recently with

machine learning and even more recently with deep learning to classify or categorize a received acoustic signal.

This section will briefly cover a selection of statistical classifiers in UATR to map unknown underwater acoustic signals to known labels. Statistical models have been shown to work sufficiently in data-scarce applications while being less prone to over-fitting [14]. They are fairly simple and offer comprehensible parameters and working principles, and are often utilized as supportive decision–systems to human operators.

Table 3 summarizes a selection of several statistical classification approaches build on top of the aforementioned different time, frequency, time-frequency and auditory features. The approaches were selected from the accumulated references of deep learning literature corpus. It is important to note that the provided scores and metrics in this section are difficult to compare directly, due to the lack of comparable, publicly available datasets and benchmarks at the time of most publications. All approaches are evaluated on custom and enclosed datasets (except for [32]). The algorithms of many of the selected approaches are available as of the shelf methods in most statistic toolboxes across various programming language. Nonetheless, no access to the used source code is available for any of the displayed methods.

According to the selected approaches, the Support Vector Machine (SVM) emerges as an often utilized classifier [61], [72], [76], [77], [78]. The SVM is a versatile supervised model, mostly used for regression and classification tasks. The main objective of the SVM lies in finding an optimal decision boundary (or hyperplane) that separates input data points with a maximal margin. This margin is based on the Euclidean distance between the data point and the decision boundary [79]. In [61] the authors constructed a nine-dimensional feature representation which included statistical properties extracted from the raw wave-structure. These included zero-crossing wavelength, peek-to-peek amplitude, zero crossing wavelength difference and wave train areas. Through the application of the SVM on the nine-dimensional feature, they were able to achieve a recognition rate of 89.5% on a two class classification of a custom real world dataset. In [76], the authors fit a non-linear polynomial to the continuous power spectrum of a received acoustic signal. They demonstrated that the power spectrum reflects certain ship specific properties. Additionally, they state, that no a priori information of the frequency distribution is required to achieve a recognition accuracy of up to 94.0%. Moura and Seixas [62] target scenarios where negative samples (targets not present in the dataset) are difficult to obtain. They apply a single-class SVM for the detection of novel ships types. They use the LOFAR analysis followed by a principal component analysis (PCA) to reduce the dimensionality. The PCA transforms correlated variables into linear uncorrelated variables, referred to as principal components. The principal components represent the variables of highest variance and are subsequently used as input to the SVM classifier. For the evaluation of the novel detection, they introduce the SP-index metric which is composed of the number of known events classified as novelty and the number of novelty events classified as known classes. With their approach, they were able to achieve an SP-index of 77.9%. The authors in [77] use the modified GFCC feature as input to the SVM, illustrating the noise robustness of over the MFCC and plain GFCC features. The modified GFCC, they refer to as MGFCC is extracted from a Gammatone Cochleargram in combination with a discrete cosine transform for dimensionality reduction. They are able to achieve a classification accuracy of 98.6% on a four class, custom dataset. The authors in [72] also utilize time-frequency representations in combination with SVM for the classification of unknown vessels. In contrast, however, they make use of the Hilbert-Huang transformation on the cepstral coefficients extracted using a Gammatone filter bank, achieving a classification accuracy of 96.7%. Additionally, the authors [72] were able to demonstrate the stability of the HHT features under various SNR scenarios.

The Gaussian Mixture Model (GMM) represents another often utilized classification model in UATR. The GMM is a probabilistic model that assumes that input samples are generated from a mixture of normal distributions. The key objective is the determination of the distribution parameters such as mean and variance to model the underlying mixed Gaussian distribution. Therefore, the GMM is often applied in supervised and unsupervised clustering and classifications scenarios. GMMs are highly flexible and can model a wide range of complex data distributions. Additionally, due to the probabilistic nature, GMMs can model the uncertainty, providing confidence levels. In [32] the authors set up baseline results utilizing GMMs for the publicly available ShipsEar Dataset they introduce. With the GMM and Cepstrum Coefficients (CC) as feature input, they are able to detect vessel presence with 100% accuracy. Additionally, their method is able to differentiate between four vessel types with an accuracy of 75.4%.

In [63] the authors model the human decision process of sonar operators based on ship specific characteristics extracted from the line spectrum of a DEMON analysis combined with a Fuzzy logic classifier. Fuzzy logic deals with probabilistic logic with approximate reasoning. In contrast to classical logic where the variables have fixed binary values, fuzzy logic incorporates non-fixed values. This has applications in decision-making systems where a human like reasoning under uncertainty is required. The implemented logic can model nonlinear relations of arbitrary complexity while remaining comprehensible. Nonetheless, fuzzy logic systems are somewhat subjective and reliant on expert knowledge in the implementation stage.

The statistical methods and the feature extraction methods presented demonstrate that the classification and categorization of unknown vessels based solely on the emitted noise is possible. The classical approaches require only a handful of data samples to train and are less subject to over-fitting while still being interpretable due to their pure

TABLE 3. Overview of UATR approaches using classic statistical and machine learning based classifiers for various feature representations.

Feature		Classifier	Reference	Contributions
Time	Wave Structure	SVM	Meng et al. (2014) [61]	Obtain loudness and timbre like characteristic from pure wave structure
Frequency	Spectrum	SVM	Liu et al. (2014) [76]	Fit a nonlinear polynomial to the spectrum. Demonstrate that power spectrum includes target information. No priori information of the feature distribution is necessary.
Time-Frequency	LOFAR	SVM	Moura et al. (2015) [62]	Apply a SVM on LOFARgrams for (single class) detection of novel ships.
	DEMON	Fuzzy Logic	Kummert (1993) [63]	Extraction of ship properties like propeller shaft rage, number of blades based on line spectrum of the DEMON analysis
	HHT	SVM	Zeng et al. (2014) [72]	Utilize HHT on top of GFCC features for time-frequency analysis. Indicate feature extraction robustness for low SNR conditions.
Auditory	MFCC	GMM	Santos- Dominguez et al. (2016) [32]	Introduce ShipsEar dataset and respective baseline recognition accuracy of 75.4%. Demonstrate that ship types can be identified by various hull sizes.
	GFCC	SVM	Lian et al. (2017) [77]	Utilize modified Gammatone Filter bank. Demonstrate noise robustness of GFCC features over MFCC

mathematical and statistical nature [14]. However, it has been shown that a fairly sophisticated and highly engineered feature extraction process is necessary to extract meaningful patterns from within the data. The feature extraction is of significant importance and requires high level of domain expertise to craft [7], [23], [26], [35], [45], [46], [80] and extensive manual labour [45]. The handcrafted features in combination with the shallow classifiers are not able to extract deep features from within the data, due to the limited model capacity [81] and strongly compressed information within the features [41]. In reality, the properties of underwater acoustics are unable to fulfil the assumptions made in the feature crafting process, resulting in insufficient generalization capabilities [41], [52], [82], [83].

IV. DEEP LEARNING APPROACHES

With the accumulation of accessible datasets, deep learning approaches have been increasing in popularity within the UATR field. Deep Learning based methods are purely data driven models that can learn hierarchical features [3] with little to no a priori information [56] allowing the development of end-to-end recognition models. The automatic feature extraction process directly tackles the issue of manual feature crafting confronted in the application of statistical methods [24], [55]. Irfan et al. [33] compared various statistical- to deep learning based approaches and demonstrated that the purely data driven deep learning approaches outperform the former. This section covers the deep learning based UATR approaches for various benchmarks taken from the relevant studies corpus. The objective is to identify common architectures, feature representations and learning strategies. The general data processing pipeline common for all investigated approaches is displayed in Fig. 8.

In general, the investigated publications follow a unified data processing procedure. At first, the recordings of various lengths are separated into smaller, equally sized batches. In the second step, pre-processing methods are applied to the fixed size windows. These typically include resampling to a unified sample rate, standardization and applying pre-emphasis filters. If required, the waveform samples are transformed into a desired representation form. The third and fourth step describe the extraction of meaningful features and classification of the processed input. Typically, these steps are aggregated in a single model with corresponding components.

All the included approaches evaluate their methods on common benchmarks. The benchmarks are derived from ShipsEar [32] and DeepShip [33] datasets. In addition to the recognition of the twelve different noise classes provided by the ShipsEar dataset, it has become common to categorize the vessel noise in subcategories of five or nine categories. Concerning readability and simplicity of comparison, the various approaches are separated into individual tables 4, 5, 6, 7, respective to the utilized dataset. Note, the displayed approaches are sorted in chronological order and certain approaches that utilize both datasets are listed in both tables accordingly.

A. MODEL ARCHITECTURES

Deep networks have been shown to more effective in capturing deep hierarchical feature representations as shallow networks with sonar data [28], [50], [57]. Despite the

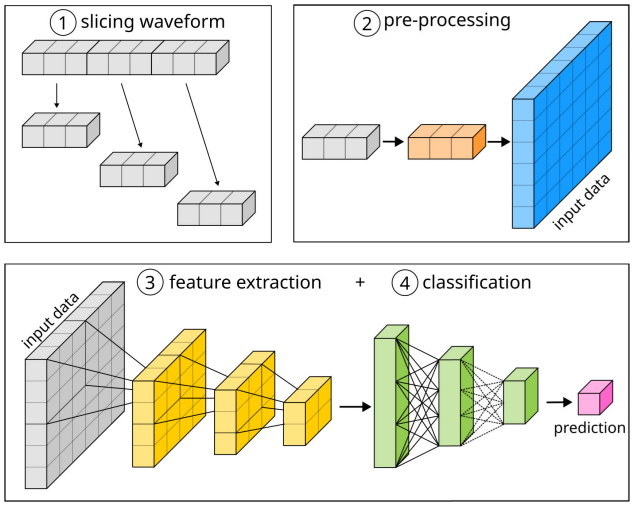

FIGURE 8. The standard processing pipeline in UATR. (1) In the first step, the collected samples in the dataset are sliced into equally sized batches. (2) pre-processing measures, such as resampling, standardization and pre-emphasis are usually applied in the second step. Additionally the sliced waveform samples are transformed into the desired representation form. Step (3) and (4) describe the feature extraction and classification incorporated in within the selected model. Some approaches specifically apply these two steps individually. Most of the research is concentrated on the pre-processing and feature -crafting and extraction.

performance increase, issues like the vanishing gradient problem complicate the training process of deep neural networks [7], [28], [55]. At first, popular network architecture choices in the UATR field and the motivating properties that lead to these decisions are highlighted. Fig. 9 summarizes the distributions of the model architectures listed in the tables 4 5, 6, 7 and 8.

Regarding the considered deep learning based UATR publications, it is evident that convolutional neural networks (CNN) are the most popular architecture, with a share of two thirds. This characteristic is carried along all investigated benchmarks. 26 of a total 40 publications utilize CNNs for the encoding layers. The CNN architectures have their origin in the computer vision domain and have been very successful ever since [49], [58]. On the basis of this success, a common practice has evolved in the audio and sonar domain, where CNNs are applied by adapting the input modalities to the image like inputs typically expected by CNNs. In total, 24 of the 26 CNN approaches transform the original waveform into a two-dimensional time-frequency representation similar to a gray scale image. Fig. 9 shows that all studies of the relevant studies corpus were published between 2019 and April 2024. Additionally, an increase of the CNN usage in the UATR field can be clearly derived from Fig. 9. In the years 2022 and 2023 transformer architectures have first made ground in the UATR domain. The usage of the Restricted Boltzmann Machine (RBM) and Multi Layer Perceptron (MLP) based approaches are focused around single research groups respectively, resulting in relatively low number of publications.

The introduction of residual connections and the corresponding Residual Neural Networks (ResNet) by [102] have become a widely utilized backbone within the UATR field. 12 publications of the 26 using CNNs make use of this backbone. The authors in [3], [5], [15], [22], [26], [50], [91], [97], and [99] make use of ResNet18. The authors in [6] make use of a ResNet50 architecture, proving that deeper networks are able to extract more semantic meaningful features compared to their shallow counterparts. Yang et al. [6] demonstrate that a sophisticated feature choice can elevate a smaller scale ResNet10 to achieve comparative results obtained with a larger ResNet18 architecture.

As previously mentioned in section III vessel noise is composed of narrowband and broadband frequency components. Capturing broadband frequency relations contrasts the local feature extraction property of CNNs [7], [49], [85]. The usage of large convolution kernels and very deep networks can overcome this issue, however, at the cost of expensive computation and difficult training [23], [52]. The authors in [85] utilize differently sized kernels in parallel to extract features with different resolutions to increase robustness towards capturing complex broadband targets. Li et al. [60] introduce a basic block module based on separated patches of the input spectrogram, where an MLP captures the globally distributed features across these patches. In [82] the authors make use of multiple differently sized dilated kernels to increase the receptive field, without significantly increasing the computational effort. Besides the model driven solutions, the authors in [6] are able to increase their receptive field from a data point of view. In their work, three different LOFAR spectrograms of various frequency resolutions are fused as input to the model in a multichannel array.

Tables 4,5,6,7,8 indicate that only a minority of approaches employ simple architectures. The majority of methods incorporate mechanisms designed to extract more globally extended or temporally related features. Attention mechanisms have been successfully introduced into CNNs to extract globally distributed features within the frequency domain. The authors in [10], [26], [56], [58], [90], and [97] introduce channel-wise attention modules that are able to highlight dominant frequency bands, containing the most useful information. The authors in [26] demonstrate the effective ocean noise suppression using channel attention. Wang et al. [13] utilize a self-attention module on top of a shallow multiscale network. They prove that the attention module is able to alleviate the limits of shallow networks by increasing the recognition accuracy by 2.3%.

Another approach to overcome the small receptive field of the CNN is addressed by the authors in [11], [50], [55], [87], and [101] by modelling the received time sequence using recurrent neural networks (RNN). In contrast to the aforementioned works, these approaches focus on capturing features distributed over the time domain rather than across the frequency domain. In particular, the authors make use of the Long-Short Term Memory Neural Networks (LSTM), which reduces the influence of gradient decay of during the

TABLE 4. Deep learning based UATR approaches evaluated using the ShpsEar dataset with five categories. The benchmark results are measured according to the recognition provided accuracy score.

Reference	Architecture	Feature	Additional Dataset	Results [Accuracy]
Hong et al., (2021) [84]	CNN	Log Mel Spectrogram + MFCC + (Chroma+Contrast+Tonnetz)	-	94.3%
Liu et al., (2021) [3]	CNN (MCNN-DAN)	Log Mel Spectrogram + MFCC + (Chroma+Contrast+Tonnetz)	-	94.3%
Yi et al., (2021) [85]	CNN (MsR-CNN)	MFCC	-	99.25%
Luo et al., (2021) [59]	CNN (MWSA)	LOFAR	-	96.32%
Luo et al., (2021) [28]	RBM + BP	Spectrum + DEMON	-	92.6%
Liu et al., (2021) [25]	CNN	MFCC + Log Mel Spectrogram, + CQT + GTS Spectrogram	DCASE [86] ImageNet	82.0%
Zhang et al., (2022) [87]	CNN + LSTM	STFT + Log Mel Spectrogram + 1st deriv + 2nd deriv	-	96.67%
Li et al., (2022) [88]	Transformer	Filter bank	ImageNet AudioSet [89]	97.7%
Li et al., (2022) [26]	CNN	Log Mel Spectrogram + MFCC + (Chroma+Contrast+Tonnetz)	DeepShip	98.0%
Li et al., (2022) [58]	CNN (FEM-ATNN)	Waveform	-	95.3%
Han et al., (2022) [55]	CNN + LSTM	Log Mel Spectrogram	-	92.14%
Feng et al., (2022) [49]	Transformer (UATR-Transf.)	Log Mel Spectrogram	DeepShip	96.9%
Qi et al., (2023) [50]	CNN + LSTM	MFCC + WT	-	97.98%
Li et al., (2023) [60]	CNN + MLP	Log Mel Spectrogram	-	92.6%
Yang et al., (2023) [11]	CNN + LSTM	Waveform	-	62.21% – 82.78% (for all types)
Sun et al., (2023) [9]	MLP	Frequency Spectrum	DeepShip	98.79%
Yang et al., (2023) [6]	CNN (LW-SEResNet10)	MFCC + 1st deriv + 2nd deriv	-	97.7%
Wang et al., (2023) [10]	CNN (AMNet)	STFT	custom	0.99 weighted avg F1
Wu et al., (2023) [15]	CNN	Filterbank + 1st deriv + 2nd deriv	ImageNet	95.0%
Chen et al., (2024) [90]	CNN (ARescat)	MFCC + 1st deriv + 2nd deriv	-	95.8%
Feng et al., (2024) [91]	CNN	MFCC + GFCC + STFT	-	98.34%
Feng et al., (2024) [92]	Transformer (UATR-Transf.)	Log Mel Spectrogram	DeepShip	67.89% (on 10% of the dataset)
Liu et al., (2024) [93]	CNN (RACNN)	MFCC	-	99.44%
Wang et al., (2024) [94]	Transformer (Swin)	Log Mel Spectrogram	custom	88.64%

training process of common recurrent neural networks [80]. The authors in [101] underline the effective extraction of temporally correlated features from time-frequency representations by introducing recurrence. This improved time sequence modelling capability is additionally supported by the research results of [50] and [87]. Yang et al. [11] demonstrate that these temporally related features can also be extracted from the time domain signal.

Another more recent approach to overcome the small receptive size of CNNs circumvent the usage of convolution operators. The first transformer architectures in the UATR domain were introduced by [49] with the

TABLE 5. Deep learning based UATR approaches evaluated using the ShipsEar dataset with nine categories. The benchmark results are measured according to the recognition provided accuracy score.

Reference	Architecture	Feature	Additional Dataset	Results [Accuracy]
Xie et al., (2023) [5]	CNN	Log Mel Spectrogram + CQT	DeepShip custom	85.34%

TABLE 6. Deep learning based UATR approaches evaluated using the ShipsEar dataset with twelve categories. The benchmark results are measured according to the recognition provided accuracy score.

Reference	Architecture	Feature	Additional Dataset	Results [Accuracy]
Luo et al., (2020) [41]	RBM + BP	Power Spectrum	Noisex-92 [95]	93.17%
Dong et al., (2022) [16]	CNN (BDAE)	EMD	-	75.28%
Wang et al., (2023) [13]	CNN	Waveform + Log Mel Spectrogram + 1st deriv + 2nd deriv	-	96.8%
Wang et al., (2023) [4]	Transformer (ADDTr)	Log Mel Spectrogram + 1st deriv + 2nd deriv	-	96.82%
Wang et al., (2024) [96]	CNN + Transformer (DWStr)	Log Mel Spectrogram	DeepShip	96.5%
Xie et al., (2024) [97]	CNN (CMoE)	STFT	DeepShip custom	86.21%
Cui et al., (2024) [98]	CNN	Log Mel Spectrogram	DeepShip	76.91% (15-shot)

TABLE 7. Deep learning based UATR approaches evaluated using the DeepShip dataset with five categories. The benchmark results are measured according to the recognition provided accuracy score.

Reference	Architecture	Feature	Additional Dataset	Results [Accuracy]
Li et al., (2022) [26]	CNN	Log Mel Spectrogram + MFCC + (Chroma + Contrast + Tonnetz)	ShipsEar (5)	99.0%
Feng et al., (2022) [49]	Transformer (UATR-Transf.)	Log Mel Spectrogram	ShipsEar (5)	95.3%
Sun et al., (2023) [9]	MLP	Frequency Spectrum	ShipsEar (5)	97.74%
Yao et al., (2023) [99]	CNN (DCGAN)	MFCC	-	96.37
Tian et al., (2023) [100]	CNN (MSRDNN)	Waveform STFT	ONC	79.5%
Xie et al., (2023) [5]	CNN	Log Mel Spectrogram + CQT	ShipsEar (9) custom	80.02%
Huang et al., (2023) [53]	CNN	Log Mel Spectrogram	-	0.51% – 0.73% f1 (for all classes)
Feng et al., (2024) [92]	Transformer (UATR-Transf.)	Log Mel Spectrogram	ShipsEar	66.03% (on 10% of the dataset)
Wang et al., (2024) [96]	CNN + Transformer (DWStr)	Log Mel Spectrogram	ShipsEar	92.75%
Xie et al., (2024) [97]	CNN (CMoE)	CQT	ShipsEar custom	78.59%

"UATR-Transformer", by [58] with the "STM: Spectrogram Transformer" in 2022 and more recently by [4], [92], [94], and [96] in 2023 and 2024, as depicted in Fig. 9. These purely attention based approaches demonstrate that comparative results can be obtained without the use of any convolutional operators. In contrast to the aforementioned

TABLE 8. Deep learning based UATR approaches evaluated using the Ocean Networks Canada dataset. The benchmark results are measured according to the recognition provided accuracy score.

Reference	Architecture	Feature	Additional Dataset	Results [Accuracy]
Yang et al., (2019) [101]	CNN + LSTM	Log Mel Spectrogram	-	90.0%
Tian et al., (2023) [100]	CNN (MSRDNN)	Waveform STFT	ONC	95.2%

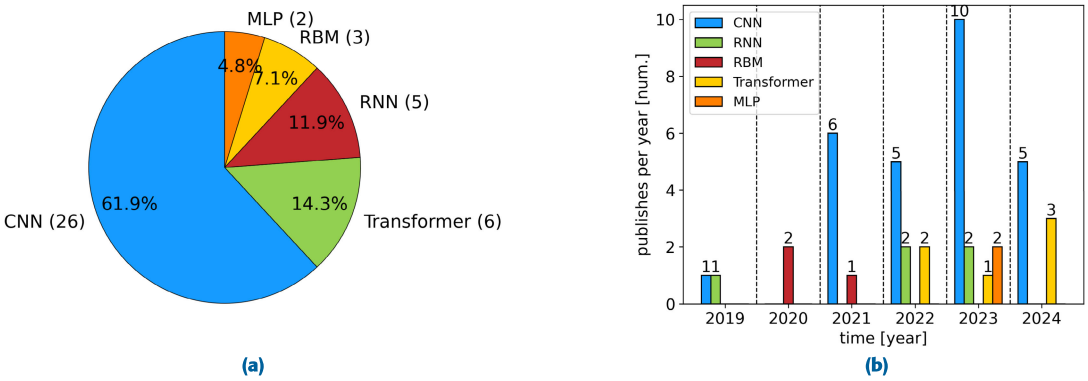

FIGURE 9. Distribution of the model architectures applied in UATR (a) and the historic development of the deep learning architectures (b) from the relevant studies corpus.

channel emphasizing attention, the Transformer makes use of self-attention, which effectively correlates all the input tokens against each other. All the three transformer based approaches are build on top of the Vision Transformer (ViT) introduced by [103] in 2020. Feng and Zhu [49] introduce a novel tokenization scheme that includes information from neighbouring tokens. The transformer based approaches in [49] and [58] are built on the self-attention mechanism, which has a quadratic computation complexity $O\left(N^2D\right)$, where N marks the sequence length and D the dimension of the feature vector. The authors in [13] introduce an additive attention mechanism with linear complexity $O\left(ND\right)$ to reduce the computational effort required. Additionally, [58] demonstrate pre-training of the transformer can be achieved with cross domain pre-training on the ImageNet and AudioSet [89] dataset, despite the large domain shift. Therefore, addressing the issue of large data requirement, notable to the transformer architecture [24], [58]. This is also demonstrated for non transformer based approaches in [15].

The majority (75%) of the UATR approaches integrate supervision into their training process. The research group around Xinwei Luo and Yulin Feng published three papers build on top of the Restricted Boltzmann Machine: [1], [41], [59] in the years 2020 and 2021. In their research, the authors were able to prove that an automated feature encoding can be achieved by layer-wise training without the need of supervision. The target recognition was accomplished by sending the encoded feature vector through a back-propagation network. The authors in [101] also applied non supervised pre-training approaches, demonstrating that the combination of CNN and LSTM is able to extract meaningful features without the need of annotated data. Recently in 2023 and 2024 the first contrastive learning approaches have gained a foothold in the UATR domain [9], [94], [98]. Sun and Luo [9] investigated a modified supervised version of SimCLR, a popular contrastive learning approach from the vision domain that is focused on self-supervised learning [104]. In their work, the authors demonstrated the superior intra-class accumulation and inter-class separation capability of the contrastive approach. They compared their results to an implemented a version of the DBM proposed in [59] and gained a recognition accuracy of 98.79% over 77.32% with the DBM on the ShipsEar dataset. The authors in [94] effectively integrate a sophisticated patch-level masking technique in combination with a Swin Transformer architecture, preventing the model to perform basic interpolation between adjacent patches. Nonetheless, the authors in [5] and [97] demonstrated, that the effectiveness of contrastive learning arises with sparse datasets. Otherwise, they are still outperformed by classic supervised learning methods if sufficient annotated data is given.

B. UATR FEATURE REPRESENTATIONS

This subsection aims to highlight various data representation methods utilized in the UATR domain which are derived from tables provided in section IV, with reference to section III-B. The results are categorized into time-, frequency-, time-frequency- and other representations and are displayed in Fig. 10.

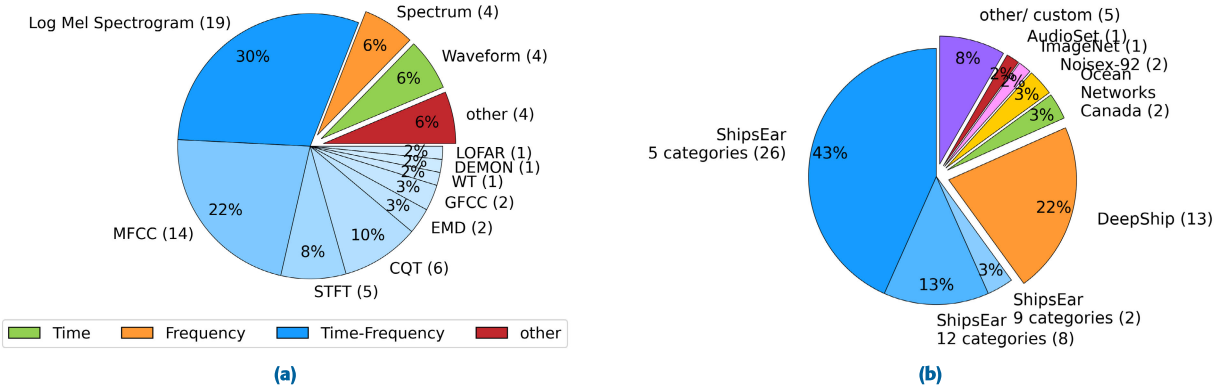

FIGURE 10. Distribution of the data representation forms (a) and utilized datasets (b) utilized in deep learning based UATR models.

The biggest share among the utilized data representations is taken by time-frequency representations. In total, over three out of four approaches make use of this form of data representation. Single time- and frequency- representations alongside the other photometric representations such as chroma, contrast, tonnetz and zero crossing rate are only applied in one out of four approaches with equal shares of four publications each. Regarding the spectral representation, most spectrogram-based approaches transform the received signal onto a logarithmic frequency scale like the Log-Mel Spectrogram or MFCCs, emphasizing lower frequency components. The choice of logarithmically scaled frequencies representations are not subject to independent benchmarks, but are rather present among all publicly available benchmark results. This choice aligns with the low-frequency energy concentration characteristic of ship-radiated noise stated in section III. The two most utilized representations are the Log Mel Spectrograms 30% and the derived MFCCs, 2% utilized over half of the investigated publications. In total 61% of the approaches are derived from a logarithmic frequency representation such as Log Mel Spectrogram, MFCC, CQT and GFCC. The large share of time-frequency based features coincides with the large share of CNN based approaches (see Fig. 9). However, the question remains whether the choice of the data representation form is driven by the model choice. 24 of the 26 CNN approaches choose a time-frequency representation as the input feature to their models.

Approximately half of the reviewed approaches utilize multiple features concurrently, demonstrating superior recognition performance when combining more than one feature representation. Approaches like [5] and [25] utilize frequency representations with different resolutions and emphasis characteristics, whereas the authors in [4], [6], [13], [15], [87], and [90] integrate the first and second derivatives of the frequency over time representations to capture temporally related properties of the signal. Additionally, the authors in [3], [26], and [84] utilize features inspired from computer vision based approaches like Chroma, Contrast and Tonnetz proving the effectiveness of regarding the time-frequency representations as an image. These observations highlight that single features are often insufficient to capture all relevant signal characteristics. Notably, the study by Tian et al. [100] illustrates that using highly diverse input features, such as raw waveform and STFT, can significantly enhance recognition accuracy due to their strong complementarity.

Fig. 10 displays the distribution of the utilized datasets for pre-training, fine-tuning and validation of the investigated UATR approaches. Three datasets are actual passive sonar dataset, namely ShipsEar, DeepShip and Ocean Networks Canada and make up 84%. The ShipsEar dataset is the most commonly utilized dataset. with a share of 59%. However, various approaches make use of different vessel type groupings within the same dataset. 26 publications 43% grouped the original eleven vessel types into four groups and two publications 3% into nine groups, plus with an extra background noise class. In the original paper [32] the authors also utilize the four category grouping. The objective of eight publications 13% was to classify the original eleven vessel types. The second most utilized dataset is the DeepShip dataset, with 13 publications 12%. Only two publications 3% made use of the Ocean Networks Canada dataset. This small share is caused by the fact that the Ocean Networks Canada is an unlabelled dataset. Therefore, this dataset was only of interest of the two approaches that investigated non supervised learning approaches [100], [101].

Pre-training has become a viable approach in UATR [4], [5], [15], [25], [26], [41], [49], [88] to overcome the data scarcity issue, implying that the given datasets are not sufficient to capture all feature variations of ship radiated noise. The number of publications pre-training on cross domain Datasets, such as ImageNet, AudioSet, DCASE and Noisex-92 [4], [15], [25], [41], [88] outweigh the number of research works pre-training on actual sonar data [26], [49], [101] five to three. In [15] the authors were able to increase the recognition accuracy by 2.3% by pre-training on the ImageNet dataset indicating that similar features are present in time-frequency representations of ship radiated

noise and general images. Especially the transformer based approaches [4], [88] rely heavily on sufficient pre-training with a large data corpus like ImageNet and AudioSet due to the high training data requirements of the Transformer architecture [24], [88]. The authors of [88] were able to increase the recognition accuracy from 85.7% up to 97.7% by pre-training on ImageNet and AudioSet.

The generalization capability is an often discussed topic in the UATR domain [9], [24], [25], [28], [100], [105]. Despite being an aware issue, not many approaches undergo deep investigations on the generalization capabilities of their proposed methods. The general approach is to train and evaluated only a single dataset (23 publications). This is especially dangerous as a common approach to enhance the available sample quantity is to subsample the recorded files into fixed size with overlapping frames. A naive assignment of the subsampled frames into a train, test- and validation dataset is successive to data leakage [5]. This information leakage diminishes the validity of potential generalization statements. The results given by [10], [14], and [85] have to be taken under a grain of salt regarding this aspect of data leakage. This supposition could be counteracted with a detailed description of the data preparation process, which is not given or fully comprehensible. Only nine of the investigated publications cross-evaluated their models on different available underwater acoustic datasets [5], [9], [49], [100], [105], whereas the authors in [56] evaluated their results on a custom, not accessible dataset. An impact to the results on the evaluation dataset compared to the single-dataset approaches in not visible in the tables 4, 5, 6, 7, and 8.

Few-Shot learning methods, where the objective is to train on as few target samples as possible, have also been the subject of recent publications. The authors in [92] and [98] investigate the recognition performance on fine-tuning on only a handful of samples. The authors in [98] consider fine-tuning sets consisting of only 3, 5, 10, or 15 examples per class in the downstream task. The authors in [92] fine-tune on 10% of the original dataset. These approaches have shown promising results on the transfer ability of well pre-trained networks.

C. SUMMARY OF DEEP LEARNING BASED APPROACHES

To gain a better overview of the different methodical approaches given in UATR research, table 9 gives a structural summary of the aforementioned publications.

The literature review of this section reveals several key findings and corresponding suggested research articles in the domain of deep learning-based recognition of underwater acoustic targets:

1) **Global Features**. Ship-radiated noise is characterized by globally extended features. Traditional Convolutional Neural Network (CNN)-based methods, which are adept at extracting local features, have been successfully extended to capture broader contextual information through various techniques. These include the use of attention mechanisms, multi-spectral features, and recurrent neural network structures that incorporate extended temporal contexts. Transformer-based approaches, with their inherent self-attention mechanisms, naturally excel at detecting global features across the input sequence. At the feature extraction level, incorporating first- and second-order time derivatives of input spectrograms effectively captures short-term temporal effects, such as reverberation. The presence of globally extended features is additionally supported by the broadband characteristics of ship-radiated noise discussed in Section III. The study by Feng and Zhu [49] illustrates the use of Transformers to identify global features in this context. For research that incorporates a global perspective in their approach, Chen et al. [90] present a CNN-based method augmented with an attention mechanism, while Yang et al. [6] provide a comprehensive Transformer-based approach. Additionally, Zhang et al. [87] demonstrate that architectures incorporating recurrent structures can enhance recognition performance by leveraging stronger temporal sequence modelling capabilities.
2) **Auditory based Features** Ship-radiated noise predominantly occupies the low-frequency spectrum. This characteristic is evidenced by the widespread use of logarithmic frequency scaling in various approaches, which is consistent with the theoretical framework outlined in our manuscript. Although ship-radiated noise has distinct properties, its concentration in the lower frequency range is analogous to the patterns observed in other auditory-based deep learning applications, such as speech processing [106], [107]. Consequently, we recommend leveraging the rapidly evolving methodologies from speech processing when designing recognition systems for ship-radiated noise. The studies by [6], [55], and [96] highlight the superior effectiveness of logarithmically spaced spectrograms over linearly spaced ones for capturing relevant features in such contexts.
3) **Feature Fusion** The extensive use of feature fusion techniques highlights the complexity of acoustic data and suggests that single features alone are inadequate to fully characterize the acoustic signals. Integrating multiple features at different scales, or combining features with distinct properties — such as fusing raw, time-domain signals with their corresponding frequency representations — has consistently been shown to outperform approaches relying on single features. For instance, Tian et al. [100] demonstrated the effectiveness of combining frequency-based features with raw waveform signals, validating the advantages of using diverse input features in conjunction. Another common fusion technique involves incorporating temporal derivatives of the time-frequency representations. This approach has been shown to improve recognition accuracy by up to 5% when using Transformer

TABLE 9. Systematic overview of classical-, and deep learning based approaches according to the applied method and utilized dataset.

Approach	Method	Description	Utilized dataset		
			ShipsEar	DeepShip	other
Classical Approaches	SVM	Waveform			[61]
		Time-Frequency			[62], [72], [76], [77]
	GMM				[32]
	Fuzzy Logic				[63]
Deep Learning	MLP		[9]	[9]	
	CNN	shallow net	[13], [85]		
		ResNet	[3], [5], [6], [15], [26], [50], [92]	[5], [26], [97], [99]	[5], [97]
		Attention	[10], [26], [56], [58], [90], [97]	[10], [26], [97]	[10], [97]
	RNN	LSTM	[11], [50], [55], [87]		[101]
	Transformer	UATR, STM, SWIN	[4], [49], [58], [92], [94]	[49], [92], [96]	[94]
	Multi-Features	temp. derivatives	[4], [6], [13], [15], [90]		
		Spectral	[5], [13], [25], [85]	[100]	[100]
		Image Features	[3], [26], [84]	[26]	
	Pre-Training	Sonar Data	[26], [49]	[26], [49]	[101]
		Non Sonar Data	[4], [15], [25], [41], [88]		
	Unsupervised Learning	DBM	[41], [59]		
	Contrastive Learning	Contrastive Coding, SimCLR	[5], [9]	[5], [9]	[5]
	Few-Shot	CNN, Transformer	[92], [98]	[92]	

architectures [4], and it is also effective in CNNs [6] and recurrent neural networks [87]. Furthermore, Wang et al. [13] demonstrated a potential increase in recognition accuracy by up to 12.5% through the fusion of raw waveforms with log Mel spectrograms and their corresponding derivatives.

4) **Cross-Domain Analogies**. The integration of image-based features such as chroma, contrast, and tonnetz [3], [84], as well as the successful pre-training on image datasets, suggests the existence of common low-level properties between general image representations and the time-frequency representations of ship-radiated noise. Pre-training on large-scale datasets, such as ImageNet, has proven highly beneficial for Transformer-based approaches [88], and has also been applied to CNN architectures [15]. While the effectiveness of pre-training on image data has been substantiated in various studies, Li et al. [88] further demonstrated that pre-training on acoustic data from AudioSet yields superior performance compared to image-based pre-training for acoustic recognition tasks.
5) **Data Scarcity** Pre-training, along with semi-supervised and unsupervised learning approaches, presents promising solutions to the challenge of limited data availability in underwater acoustic target recognition. These methods enable models to utilize existing unlabelled data, extract generalizable features, and reduce reliance on extensive labelled datasets. Unsupervised learning, in particular, has emerged as a productive research direction. Studies by Luo et al. [41], [59] and Feng and Luo [1] have demonstrated the successful extraction of meaningful features from unannotated data. More recently, contrastive learning methods have gained prominence within the unsupervised learning domain. Sun and Luo [9] achieved an impressive recognition accuracy of 98.79%, highlighting the effectiveness of their approach in intra-class aggregation and inter-class separation. In addition, Feng et al. [92] showcased the few-shot recognition capabilities of their Transformer-based approach, achieving a recognition accuracy of up to 67.89% using only 10% of the ShipsEar dataset. Similarly, Cui et al. [98] demonstrated a 15-shot scenario on the DeepShip dataset, achieving a recognition accuracy of up to 79.91%.
6) **Generalisation** Stability is a critical factor in the evaluation of recognition systems, particularly given the challenges associated with generalisation to unseen data across all deep learning domains. The complex nature of underwater acoustics and the numerous factors that can influence the received signal underscore the importance of this evaluation criterion. However, we identified that numerous articles focus on achieving high recognition results on single datasets. While these approaches give good results, they rarely address one of the most discussed difficulties of UATR. We therefore recommend that readers prioritize studies that evaluate their approaches on multiple sonar datasets, such as those given by [5], [9], [49], [100], and [105], despite not always achieving the highest benchmark results.

V. REPRODUCIBILITY AND TRANSPARENCY

While all the investigated approaches provide sufficient description of the applied architectures and data pre-processing steps, not a single source code is made available in the regarded publications. Despite the sufficient descriptions, the reproducibility of the approaches are severely limited as the manual implementation and rebuilding of the proposed models are theoretically doable, yet remain immensely time

and labour-intensive and error-prone. Transparency and open source characteristic is a major contributor to the prosperous development of AI in numerous fields [108]. However, the UATR field lacks this transparency characteristic. This unfortunate property can be explained as sonar processing remains a niche field where many defence industries take part. Licensing issues could hinder the source code provision as it would result in a conflict of interest. In addition, most custom datasets remain confidential, as sonar- and especially UATR-applications are mainly of interest to the defence industry.

VI. COMPLEXITY COVERAGE OF THE SHIPSEAR AND DEEPSHIP DATASET

The data distribution of ship radiated noise is very large. In order for models to learn robust and discriminate features, the UATR field requires sufficient amounts of training data that cover the complex data distribution covered in section III. Only two publicly available and annotated datasets exist to this purpose.

The ShipsEar dataset was published in 2016 by [32] and includes over 90 recordings of about 40 different vessels of eleven categories. In addition to the eleven vessel types, fishing boat, trawler, mussel boat, pilot ship, tugboat, dredger, ro-ro, ocean liner, passenger ferry, sailboat, motorboat, an additional class of background noise is also made available. The recordings took place at three different recording sites on the Atlantic coast in the northwest of Spain. In particular, at entry routes to the port of Vigo and other ports, as well as in the middle of the Ria de Vigo. The dataset includes of many docking and undocking manoeuvres, as well as some background and biological noise contributors. The time span of the recordings was between autumn 2012 and summer 2013.

The DeepShip dataset was published in 2021 by [33]. It includes recordings of 265 different vessels of the four different categories: cargo, passenger ship, tanker, tugboat. The complete data scope covers about 47 hours of recordings. The measurements took place in the time span between May 2016 and October 2018 at three different locations around the Georgia delta node. Due to the long timespan, the recordings include various background and biological noise sources, as well as various weather and tidal conditions. The labels were acquired using the automatic identification system (AIS) on the basis of the location and timestamp.

Table 10 compares the ShipsEar, and DeepShip dataset according to the intermediate- and low level features derived in section III. The properties were derived from the original papers. As the Ocean Networks Canada dataset is not as common (see section IV) and as it lacks of detailed information about the acquired signals, it is intentionally excluded from the following comparison.

Table 10 shows that many of the intermediate- and low level features derived in section III are covered by the two dataset to some extent. As the ShipsEar dataset was acquired in a close distance to a port, many docking and undocking manoeuvres, as well as continuously moving vessels are present in the dataset. The operational states covered by the DeepShip dataset are mostly vessels passing by. However, the DeepShip dataset includes targets in a radius of 2 km, whereas the ShipsEar only covers targets in a relatively close distance of less than 300 m. In terms of variety in vessel numbers, the DeepShip datasets clearly outweigh the ShipsEar dataset, with 265 over roughly 40 recorded vessels. However, the ShipsEar dataset offers more variability in the different vessel types, where eleven different vessel types are recorded over only four different in the DeepShip dataset. Both datasets cover various weather and temperature conditions. While the ShipsEar dataset offers recordings of wind and rain noise without the presence of vessel noise, the DeepShip recordings always include the vessel noises on top. As the DeepShip dataset was acquired over two and a half years, more variety in weather and temperature conditions are expected than in the ShipsEar dataset, which was only acquired over half a year. Regardless, it is important to mention that this is not clearly stated in the dataset descriptions. Both datasets were acquired over three different on-site locations. Nonetheless, the recordings sites are geographically close, where similar salinity, marine life, current, and seabed conditions are expected. Both datasets mention other noise contributors, including some anthropogenic noises like a suction dredge in the ShipsEar dataset. However, the presence of these noise sources are only roughly mentioned as they are considered ''to be expected'' and not described in detail. Concerning recording properties, the DeepShip dataset clearly offers a larger quantity of recordings of up to 47 h over approximately 3 h provided by the ShipsEar dataset.

As stated above, both datasets cover many of the intermediate- and low level influences of underwater acoustics. Nevertheless, the variability of these conditions are only given in only a handful of occasions. Additionally, these alternating conditions are mostly present as a single occasions and not contemporaneous with other influences. Understandably, this is an almost impossible task to accomplish. The lack of sufficient variability however impedes the development of models that learn good generalizable features [25], [28], [100]. This is also reflected in the literature, where only six approaches cross-validate their results on other acoustic datasets [5], [9], [49], [56], [100], [105]. Table 11 displays the mean, minimum, maximum, standard deviation accuracies, and the top three publication references that achieve the highest results on the corresponding benchmarks.

VII. RESEARCH GAPS AND FUTURE DIRECTIONS

Deriving recognition systems purely from data using deep learning has given promising results in the UATR field. Nonetheless, some key issues could be addressed to encourage further development, especially when comparing to more popular research domains such as computer vision and natural language processing. A significant amount of publications was not considered in this work, as they were not comparable in terms of dataset usage (see Fig. 1). As stated

TABLE 10. Comparison of the two most widely utilized passive sonar datasets in the UATR field regarding the low- and intermediate- level features derived in section III.

Influence		ShipsEar		DeepShip	
High Level	**Intermediate Level**	**Score**	**Comments**	**Score**	**Comments**
Operational State	-	3	ships maneuvering in port, de- and accelerating, pass-by	1	mostly pass-by vessels
Position/ Distance	-	max. 350 m	targeted hydrophones at vessels for high SNR	max. 2000 m	only recorded in 2 km radius
Vessel design, condition and, type	**Propulsion System**	40	recorded 11 different vessel types (assigned to 4 categories)	265	recorded 4 different vessel types
	Propeller				
	Hull				
Weather	**Wind**	yes	single recording without vessels	yes	recordings with vessel noise
	Rain	yes		yes	
	Thunder	-	-	yes	
	other	-	-	-	
	Temperature	(2)	not directly mentioned, only 2 seasons are covered	(4)	not directly mentioned, all 4 seasons are covered
Region	**Salinity**	-	three different geographical close locations	-	three different geographical close locations
	Depth	max. 45 m	-	min. 175 m	inferred from hydrophone depth
	Seabed	1	some rocks	2	sand and silt
	Marine Life	yes	-	yes	-
	Current	yes	single recording without vessels	yes	-
Time	**Seasons**	2	-	10	-
Other	**Noises**	(yes)	suction dredge waves crashing against port wall	(yes)	only mentioned
Recording	**Duration**	approx. 3 h	-	approx. 47 h	-
	Background	yes	-	yes	-

in [14] and [21] the largest leap forward of the UATR domain will be achieved by creating more versatile, available and curated datasets. Additionally, more transparency in the UATR field is desirable, as this would undoubtedly increase the reproducibility of the proposed methods, on which future researchers can easily build upon. Adding to this, it is essential to give insight on why the code publication might not be possible, i.e. in terms of licensing and conflict of interest.

Furthermore, the developed architectures and data representations shown in section IV demonstrate that a lot of focus is either put on increasing the information content of the utilized data representations or in the creation of highly engineered and sophisticated feature extractors. Many approaches use time-frequency to cope with the CNNs image-like input, and therefore craft a suitable input feature through multifeature fusion. This hints that the time-frequency representations strongly compresses the original signal, where a lot of information is lost. Creating time-frequency representations is always accompanied by a balance between time- and frequency resolution. This is not evident from pure waveform approaches. As it can be seen in Fig. 5 a lot of information is contained in the phase, which is neglected when using the magnitude of the time-frequency representations. Additionally, the crafting of these features also requires domain knowledge for the parameter selection, such as window size, windowing function, and number of frequency bins. Therefore, more focus should be put on extracting meaningful features from the data representations that offer the highest information density.

TABLE 11. The minimal, maximal, mean and standard deviation of the achieved accuracies in the corresponding benchmarks. Note, the results from [10], [14], and [85] were not considered in this analysis due to the incomparable results caused by data leakage, as well as the publications that did not provide an accuracy score and the approaches that provided few-shot classification results [92], [98].

Dataset	mean accuracy [%]	min. accuracy [%]	max. accuracy [%]	standard deviation [%]	top three publications [reference]
ShipsEar 5 categories	94.39	72.49	99.44	5.96	[13], [56], [93]
ShipsEar 9 categories	85.34	85.34	85.34	0.00	[5], -, -
ShipsEar 12 categories	87.48	75.28	96.82	8.78	[4], [16], [58]
DeepShip	89.90	78.59	99.0	8.34	[8], [9], [99]

TABLE 12. Bibliographic metrics of the included publications and Journals. Numbers derived from https://www.scimagojr.com/journalrank.php, (last accessed: 30.04.2024, 11:23).

Journal	mean citations μ_{cites}	Relevance score σ	mean average citation count (impact factor)
Journal of Marine Science and Engineering - MDPI	4.12	7.88	2.98
Electronics - MDPI	1.50	3.00	3.49
Frontiers in Marine Science - Frontiers Media S.A.	0.00	3.00	2.77
Geoscience and Remote Sensing Letters - IEEE	6.33	2.67	4.11
Applied Sciences - MDPI	10.67	2.67	2.92
Sensors - MDPI	4.25	2.50	4.25
IEEE Access	17.00	2.00	4.64
Multimedia Tools and Applications - Springer	0.00	2.00	4.53
Transactions on Instrumentation and Measurement - IEEE	4.00	2.00	5.60
Knowledge-Based Systems - Elsevier	5.00	2.00	8.47
IET Radar, Sonar and Navigation - Wiley	0.00	2.00	1.78
Expert System with Applications - Elsevier	0.00	2.00	9.29
Transactions on Audio, Speech, and Language Processing - IEEE/ ACM	0.00	2.00	5.40
PLoS ONE - Public Library of Science	5.00	2.00	3.11
Ocean Engineering	0.00	1.50	5.15
Entropy - MDPI	1.00	1.33	2.51
Remote Sensing - MDPI	1.33	1.33	4.55
Applied Acoustics - Elsevier	5.67	1.33	4.00

Additionally, models utilizing multiscale kernels and attention mechanism have been shown to perform well on extracting and highlighting more meaningful components in the data. Especially the transformer approaches address these issues quite differently than the CNN based approaches achieving promising results.

At last, as the amount of available dataset is still the largest burden of the UATR domain, other learning strategies like self-supervised approaches that do not require annotated labels should be investigated more deeply. These approaches have shown remarkable success in other domains [104], [109], [110] and offer an alternative approach to address the data scarcity issue. Recently self-supervised approaches have also set foot in the UATR domain [9], [94], [98] giving promising recognition results and a possibility to circumvent the data scarcity issue.

VIII. CONCLUSION

This survey provided an overview of comparable deep learning based approaches in the UATR field. In total, 40 scientific publications were assessed in detail according to three key questions. The first questions aimed to analyse the available literature field regarding the identification of maritime vessels based on the emitted acoustic signature. The results showed that the UATR field can be assigned to the general field of engineering, computer science, environmental engineering, and physics. These four categories can support the literature search regarding UATR related topics. The reader is pointed towards significant Journals and Conferences that have a high qualitative standard regarding published research works. Nonetheless, the analysis demonstrated that the publishing field is distributed among numerous journals of various domains, demonstrating the lack of a major UATR related publisher.

The second contribution of this survey is the identification of commonalities between UATR related papers regarding model architecture, applied datasets and feature representations. The results of this study demonstrated that the majority of research is orientated towards the computer vision domain, applying CNN based architectures in a supervised learning

procedure. More recently, contrastive learning approaches and non-convolutional models like the Transformer architecture have achieved first promising results in the UATR domain following a similar domain development as the audio and vision domain. Nonetheless, a significant variance in the applied architectures and corresponding modifications is present. An identical behaviour can be seen in the data representation usage. Generally, a lot of effort is put in the selection of the utilized representation form. This highlights the importance and dependency on robust and meaningful feature representations in the UATR domain. While most approaches utilize time-frequency representations, the corresponding results range from the worst to the best scoring approaches, complicating the identification of the most promising representation. Still, the results obtained using deep CNN backbones like ResNets and features such as the MFCCs and fusion with the first and second time derivatives tend to stand out by a handful of publications. However, a best-performing architecture and data representation can not be clearly derived as the number of publications are too small, and the variance is too high to perform any meaningful statistical analysis. This is also underlined by the lack of transparency and reproducibility.

At last, in depth analysis on the two major public datasets is undertaken. While all studies were evaluated on at least one of these real world measurements, many publications lack an in-depth analysis of the generalization capabilities of the proposed methods, especially considering the complex data distribution of underwater acoustics. Evaluation on multiple benchmarks is not common, which is partially related to the fact that only two major benchmarks exists in the UATR field. This highlights the necessity for more variability in available datasets and benchmarks. For the present benchmarks, the best highest accuracy scores are achieved on the 5 category ShipsEar dataset, followed by the DeepShip, ShipsEar 12 categories and ShipsEar 9 categories benchmarks with a mean recognition accuracy of 94.29%, 93.15%, 89.93%, and 85.34% respectively.

In conclusion, we demonstrated that the UATR field is a rapidly growing research field. The study conducted in this work highlights the great potential of purely data driven approaches for the recognition task of maritime vessels. Compared to the baseline of 75.4% ACC provided in [21] utilizing a statistical model, most deep learning approaches are able to elevate the recognition accuracy by at least 20%.

In order for the UATR field to thrive similar to other fields like computer vision and natural language processing, some key requirements, like minimizing data scarcity and increasing transparency in the proposed approaches, need to be assessed.

APPENDIX

BIBLIOGRAPHIC METRICS OF THE INCLUDED JOURNALS

Table 12 displays the bibliographic metrics utilized in section II-D and especially in Fig. 3. The mean average citation count (impact-factor) are derived from the "Scimage Journal & Country Rank" website https://www.scimagojr.com/journalrank.php, (last accessed: 26.02.2024, 14:55).

REFERENCES

[1] Y. Feng and X. Luo, "Underwater acoustic feature extraction based on restricted Boltzmann machine," *Proc. Int. Congr. Noise Control Eng.*, 2020, pp. 4001–4012.

[2] F. Wang, Y. Zhang, and Y. Wang, "Reliable LPS features of multiple airborne radars against PRS," *IET Radar, Sonar Navigat.*, vol. 13, no. 10, pp. 1747–1754, Oct. 2019.

[3] F. Liu, H. Ding, D. Li, T. Wang, Z. Luo, and L. Chen, "Few-shot learning with data enhancement and transfer learning for underwater target recognition," in *Proc. OES China Ocean Acoust. (COA)*, Jul. 2021, pp. 992–994.

[4] B. Wang, W. Zhang, Y. Zhu, C. Wu, and S. Zhang, "An underwater acoustic target recognition method based on amnet," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[5] Y. Xie, J. Ren, and J. Xu, "Guiding the underwater acoustic target recognition with interpretable contrastive learning," in *Proc. OCEANS*, Jun. 2023, pp. 1–6.

[6] S. Yang, L. Xue, X. Hong, and X. Zeng, "A lightweight network model based on an attention mechanism for ship-radiated noise classification," *J. Mar. Sci. Eng.*, vol. 11, no. 2, p. 432, Feb. 2023.

[7] V.-S. Doan, T. Huynh-The, and D.-S. Kim, "Underwater acoustic target classification based on dense convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[8] J. Li, B. Wang, X. Cui, S. Li, and J. Liu, "Underwater acoustic target recognition based on attention residual network," *Entropy*, vol. 24, no. 11, p. 1657, Nov. 2022.

[9] B. Sun and X. Luo, "Underwater acoustic target recognition based on automatic feature and contrastive coding," *IET Radar, Sonar Navigat.*, vol. 17, no. 8, pp. 1277–1285, Aug. 2023.

[10] Y. Wang, H. Zhang, and W. Huang, "Fast ship radiated noise recognition using three-dimensional mel-spectrograms with an additive attention based transformer," *Frontiers Mar. Sci.*, vol. 10, Nov. 2023, Art. no. 1280708.

[11] H. Yang, X. Huang, and Y. Liu, "Infogan-enhanced underwater acoustic target recognition method based on deep learning," in *Proc. Int. Conf. Autonomous Unmanned Syst.*, in Lecture Notes in Electrical Engineering, vol. 1010, 2023, pp. 2705–2714.

[12] P. Qi, J. Sun, Y. Long, and L. Zhang, "Underwater acoustic target recognition with fusion feature," in *Proc. Int. Conf. Artif. Intell. Comput. Sci.* Cham, Switzerland: Springer, 2021, doi: 10.1007/978-3-030-92185-9_50.

[13] H. Wang, C. Xu, and D. Li, "Underwater acoustic target recognition combining multi-scale features and attention mechanism," in *Proc. IEEE 3rd Int. Conf. Electron. Technol., Commun. Inf. (ICETCI)*, May 2023, pp. 246–253.

[14] X. Luo, L. Chen, H. Zhou, and H. Cao, "A survey of underwater acoustic target recognition methods based on machine learning," *J. Mar. Sci. Eng.*, vol. 11, no. 2, p. 384, Feb. 2023.

[15] J. Wu, P. Li, Y. Wang, Q. Lan, W. Xiao, and Z. Wang, "VFR: The underwater acoustic target recognition using cross-domain pre-training with FBank fusion features," *J. Mar. Sci. Eng.*, vol. 11, no. 2, p. 263, Jan. 2023.

[16] Y. Dong, X. Shen, and H. Wang, "Bidirectional denoising autoencoders-based robust representation learning for underwater acoustic target signal denoising," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–8, 2022.

[17] J. Hildebrand, "Anthropogenic and natural sources of ambient noise in the ocean," *Mar. Ecology Prog. Ser.*, vol. 395, pp. 5–20, Dec. 2009.

[18] United Nations Conference on Trade and Development, *The Trade and Development Report*, UNCTAD, Geneva, Switzerland, 2022. Accessed: Feb. 12, 2024. [Online]. Available: https://unctad.org/tdr2022

[19] Y. Steiniger, D. Kraus, and T. Meisen, "Survey on deep learning based computer vision for sonar imagery," *Eng. Appl. Artif. Intell.*, vol. 114, Sep. 2022, Art. no. 105157.

[20] M. Buß, S. Benen, D. Kraus, and A. Kummert, "False alarm reduction for active sonars using deep learning architectures," in *Proc. Undersea Defence Technol. (UDT)*, 2019, pp. 1–5.

[21] L. C. F. Domingos, P. E. Santos, P. S. M. Skelton, R. S. A. Brinkworth, and K. Sammut, "A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance," *Sensors*, vol. 22, no. 6, p. 2181, Mar. 2022.

[22] F. Hong, C. Liu, L. Guo, F. Chen, and H. Feng, "Underwater acoustic target recognition with a residual network and the optimized feature extraction method," *Appl. Sci.*, vol. 11, no. 4, p. 1442, Feb. 2021.
[23] G. Hu, K. Wang, and L. Liu, "Underwater acoustic target recognition based on depthwise separable convolution neural networks," *Sensors*, vol. 21, no. 4, p. 1429, Feb. 2021.
[24] A. Jin and X. Zeng, "A novel deep learning method for underwater target recognition based on res-dense convolutional neural network with attention mechanism," *J. Mar. Sci. Eng.*, vol. 11, no. 1, p. 69, Jan. 2023.
[25] C. Liu, F. Hong, H. Feng, and M. Hu, "Underwater acoustic target recognition based on dual attention networks and multiresolution convolutional neural networks," in *Proc. OCEANS*, Sep. 2021, pp. 1–5.
[26] J. Li, G. Zhao, B. Li, X. Wang, and M. Huang, "A reduced dimension multiple signal classification–based direct location algorithm with dense arrays," *Int. J. Distrib. Sensor Netw.*, vol. 18, no. 5, May 2022, Art. no. 155013292210975.
[27] D. Liu, W. Shen, W. Cao, W. Hou, and B. Wang, "Design of Siamese network for underwater target recognition with small sample size," *Appl. Sci.*, vol. 12, no. 20, p. 10659, Oct. 2022.
[28] X. Luo, Y. Feng, and M. Zhang, "An underwater acoustic target recognition method based on combined feature with automatic coding and reconstruction," *IEEE Access*, vol. 9, pp. 63841–63854, 2021.
[29] M. Ghavidel, S. M. H. Azhdari, M. Khishe, and M. Kazemirad, "Sonar data classification by using few-shot learning and concept extraction," *Appl. Acoust.*, vol. 195, Jun. 2022, Art. no. 108856.
[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Piscataway, NJ, USA, Jun. 2009, pp. 248–255.
[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, Eds. San Diego, CA, USA: Neural Information Processing Systems Foundation, 2012, pp. 1097–1105.
[32] D. Santos-Domínguez, S. Torres-Guijarro, A. Cardenal-López, and A. Pena-Gimenez, "ShipsEar: An underwater vessel noise database," *Appl. Acoust.*, vol. 113, pp. 64–69, Dec. 2016.
[33] M. Irfan, Z. Jiangbin, S. Ali, M. Iqbal, Z. Masood, and U. Hamid, "DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification," *Exp. Syst. Appl.*, vol. 183, Nov. 2021, Art. no. 115270.
[34] D. Neupane and J. Seok, "A review on deep learning-based approaches for automatic sonar target recognition," *Electronics*, vol. 9, no. 11, p. 1972, Nov. 2020.
[35] H. Yang, S. Shen, X. Yao, M. Sheng, and C. Wang, "Competitive deep-belief networks for underwater acoustic target recognition," *Sensors*, vol. 18, no. 4, p. 952, Mar. 2018.
[36] L. Xue, X. Zeng, and A. Jin, "A novel deep-learning method with channel attention mechanism for underwater target recognition," *Sensors*, vol. 22, no. 15, p. 5492, Jul. 2022.
[37] X. Cao, R. Togneri, X. Zhang, and Y. Yu, "Convolutional neural network with second-order pooling for underwater target classification," *IEEE Sensors J.*, vol. 19, no. 8, pp. 3058–3066, Apr. 2019.
[38] B. Beckler, A. Pfau, M. Orescanin, S. Atchley, N. Villemez, J. E. Joseph, C. W. Miller, and T. Margolina, "Multilabel classification of heterogeneous underwater soundscapes with Bayesian deep learning," *IEEE J. Ocean. Eng.*, vol. 47, no. 4, pp. 1143–1154, Oct. 2022.
[39] M. F. McKenna, D. Ross, S. M. Wiggins, and J. A. Hildebrand, "Underwater radiated noise from modern commercial ships," *J. Acoust. Soc. Amer.*, vol. 131, no. 1, pp. 92–103, Jan. 2012.
[40] W. S. Filho, J. M. de Seixas, and N. N. de Moura, "Preprocessing passive sonar signals for neural classification," *IET Radar, Sonar Navigat.*, vol. 5, no. 6, pp. 605–612, 2011.
[41] X. Luo and Y. Feng, "An underwater acoustic target recognition method based on restricted Boltzmann machine," *Sensors*, vol. 20, no. 18, p. 5399, Sep. 2020.
[42] L. M. Brechovskich and J. P. Lysanov, *Fundamentals of Ocean Acoustics*, 3rd ed., New York, NY, USA: Springer, 2003.
[43] F. J. Fahy, *Foundations of Engineering Acoustics*. Amsterdam, The Netherlands: Elsevier, 2007.
[44] X. Lurton, *Introduction To Underwater Acoustics: Principles and Applications*. Berlin, Germany: Springer-Verlag, 2016.
[45] J. Jiang, T. Shi, M. Huang, and Z. Xiao, "Multi-scale spectral feature extraction for underwater acoustic target recognition," *Measurement*, vol. 166, Dec. 2020, Art. no. 108227.
[46] S. Shen, H. Yang, and M. Sheng, "Compression of a deep competitive network based on mutual information for underwater acoustic targets recognition," *Entropy*, vol. 20, no. 4, p. 243, Apr. 2018.
[47] R. J. Urick, *Principles of Underwater Sound*, 3rd ed., Los Altos, CA, USA: Peninsula, 2010.
[48] R. E. Francois and G. R. Garrison, "Sound absorption based on ocean measurements. Part II: Boric acid contribution and equation for total absorption," *J. Acoust. Soc. Amer.*, vol. 72, no. 6, pp. 1879–1890, Dec. 1982.
[49] S. Feng and X. Zhu, "A transformer-based deep learning network for underwater acoustic target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[50] P. Qi, G. Yin, and L. Zhang, "Underwater acoustic target recognition using RCRNN and wavelet-auditory feature," *Multimedia Tools Appl.*, vol. 83, no. 16, pp. 47295–47317, Oct. 2023.
[51] Z. Lian and T. Wu, "Feature extraction of underwater acoustic target signals using gammatone filterbank and subband instantaneous frequency," in *Proc. IEEE 6th Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Oct. 2022, pp. 944–949.
[52] Y. Honghui, L. Junhao, and S. Meiping, "Underwater acoustic target multi-attribute correlation perception method based on deep learning," *Appl. Acoust.*, vol. 190, Mar. 2022, Art. no. 108644.
[53] Y. Huang, X. Kong, and M. Xu, "Mel spectrum feature recognition based on adaptive center frequency," in *Proc. 5th Int. Conf. Electron. Eng. Informat. (EEI)*, Jul. 2023, pp. 633–636.
[54] Q. Huang and X. Zeng, "An underwater acoustic target recognition method combining wavelet decomposition and an improved convolutional neural network," *J. Harbin Eng. Univ.*, vol. 43, no. 2, pp. 159–165, 2022.
[55] X. C. Han, C. Ren, L. Wang, and Y. Bai, "Underwater acoustic target recognition method based on a joint neural network," *PLoS ONE*, vol. 17, no. 4, Apr. 2022, Art. no. e0266425.
[56] Y. Ma, M. Liu, Y. Zhang, B. Zhang, K. Xu, B. Zou, and Z. Huang, "Imbalanced underwater acoustic target recognition with trigonometric loss and attention mechanism convolutional network," *Remote Sens.*, vol. 14, no. 16, p. 4103, Aug. 2022.
[57] P. Wang and Y. Peng, "Research on underwater acoustic target recognition based on LOFAR spectrum and deep learning method," in *Proc. 5th Int. Conf. Autom., Control Robot. Eng. (CACRE)*, Sep. 2020, pp. 666–670.
[58] P. Li, J. Wu, Y. Wang, Q. Lan, and W. Xiao, "STM: Spectrogram transformer model for underwater acoustic target recognition," *J. Mar. Sci. Eng.*, vol. 10, no. 10, p. 1428, Oct. 2022.
[59] X. Luo, M. Zhang, T. Liu, M. Huang, and X. Xu, "An underwater acoustic target recognition method based on spectrograms with different resolutions," *J. Mar. Sci. Eng.*, vol. 9, no. 11, p. 1246, Nov. 2021.
[60] Z. Li, Y. Ji, B. Guo, and K. Yang, "FMD-based feature extraction of underwater acoustic targets," *Harbin Gongcheng Daxue Xuebao/J. Harbin Eng. Univ.*, vol. 44, no. 9, pp. 1542–1548, 2023.
[61] Q. Meng, S. Yang, and S. Piao, "The classification of underwater acoustic target signals based on wave structure and support vector machine," *J. Acoust. Soc. Amer.*, vol. 136, p. 2265, Oct. 2014.
[62] N. N. de Moura and J. M. de Seixas, "Novelty detection in passive SONAR systems using support vector machines," in *Proc. Latin Amer. Congr. Comput. Intell. (LA-CCI)*, M. M. B. R. Vellasco, Y. J. T. Valdivia, and H. S. Lopes, Eds., Piscataway, NJ, USA, Oct. 2015, pp. 1–6.
[63] A. Kummert, "Fuzzy technology implemented in sonar systems," *IEEE J. Ocean. Eng.*, vol. 18, no. 4, pp. 483–490, Oct. 1993.
[64] P. T. Arveson and D. J. Vendittis, "Radiated noise characteristics of a modern cargo ship," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 118–129, Jan. 2000.
[65] X. Wang and R. Müller, "Pinna-rim skin folds narrow the sonar beam in the lesser false vampire bat (*Megaderma spasma*)," *J. Acoust. Soc. Amer.*, vol. 126, no. 6, pp. 3311–3318, Dec. 2009.
[66] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, Jan. 1937.
[67] M. Khishe and A. Safari, "Classification of sonar targets using an MLP neural network trained by dragonfly algorithm," *Wireless Pers. Commun.*, vol. 108, no. 4, pp. 2241–2260, Oct. 2019.
[68] L. Zhang, D. Wu, X. Han, and Z. Zhu, "Feature extraction of underwater target signal using mel frequency cepstrum coefficients based on acoustic vector sensor," *J. Sensors*, vol. 2016, pp. 1–11, Jan. 2016.

[69] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1684–1689, Dec. 2012.

[70] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, Jan. 1991.

[71] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 961–1005, Sep. 1990.

[72] X. Zeng and S. Wang, "Underwater sound classification based on gammatone filter bank and Hilbert–Huang transform," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Aug. 2014, pp. 707–710.

[73] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London. Ser. A, Math., Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, Mar. 1998.

[74] N. E. Huang, N. O. Attoh-Okine, and N. E. Huang, *The Hilbert–Huang Transform in Engineering*. Boca Raton, FL, USA: Taylor & Francis, 2005.

[75] P. Wang and Y. Peng, "Research on feature extraction and recognition method of underwater acoustic target based on deep convolutional network," in *Proc. IEEE Int. Conf. Adv. Electr. Eng. Comput. Applications (AEECA)*, Aug. 2020, pp. 863–868.

[76] J. Liu, Y. He, Z. Liu, and Y. Xiong, "Underwater target recognition based on line spectrum and support vector machine," in *Proc. Int. Conf. Mechatronics, Control Electron. Eng.* Paris, France: Atlantis Press, 2014, pp. 79–84.

[77] Z. Lian, K. Xu, J. Wan, and G. Li, "Underwater acoustic target classification based on modified GFCC features," in *Proc. IEEE 2nd Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, B. Xu, Ed., Piscataway, NJ, USA, Mar. 2017, pp. 258–262.

[78] J. Liu, Z. Liu, and Y. Xiong, "Underwater target recognition based on WPT and SVM," in *Proc. Int. Conf. Comput. Commun. Technol. Agricult. Eng.*, 2010, pp. 275–278.

[79] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1023/A:1022627411411.

[80] H. Yang and S. Yi, "Underwater acoustic target feature fusion method based on multi-kernel sparsity preserve multi-set canonical correlation analysis," *Xibei Gongye Daxue Xuebao/J. Northwestern Polytechnical Univ.*, vol. 37, no. 1, pp. 87–92, Feb. 2019.

[81] H. Dong, X. Shen, K. He, and H. Wang, "Nonlinear filtering effects of intrawell matched stochastic resonance with barrier constrainted duffing system for ship radiated line signature extraction," *Chaos, Solitons Fractals*, vol. 141, Dec. 2020, Art. no. 110428.

[82] S. Tian, D. Bai, J. Zhou, Y. Fu, and D. Chen, "Few-shot learning for joint model in underwater acoustic target recognition," *Sci. Rep.*, vol. 13, no. 1, p. 17502, Oct. 2023.

[83] S. Yang and X. Zeng, "Combination of gated recurrent unit and network in network for underwater acoustic target recognition," in *Proc. Int. Congr. Expo. Noise Control Eng.*, 2021, pp. 486–492.

[84] F. Hong, C. Liu, L. Guo, F. Chen, and H. Feng, "Underwater acoustic target recognition with ResNet18 on ShipsEar dataset," in *Proc. IEEE 4th Int. Conf. Electron. Technol. (ICET)*, May 2021, pp. 1240–1244.

[85] Z. Yi, L. Pingzheng, X. Shuidong, Y. Qiong, M. Yanxin, and L. Mengqi, "Multiresolution convolutional neural network for underwater acoustic target recognition," in *Proc. IEEE 6th Int. Conf. Signal Image Process. (ICSIP)*, Oct. 2021, pp. 846–850.

[86] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the DCASE 2017 challenge," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 6, pp. 992–1006, Jun. 2019.

[87] W. Zhang, B. Lin, Y. Yan, A. Zhou, Y. Ye, and X. Zhu, "Multi-features fusion for underwater acoustic target recognition based on convolution recurrent neural networks," in *Proc. 8th Int. Conf. Big Data Inf. Analytics (BigDIA)*, Aug. 2022, pp. 342–346.

[88] D. Li, F. Liu, T. Shen, L. Chen, X. Yang, and D. Zhao, "Generalizable underwater acoustic target recognition using feature extraction module of neural network," *Appl. Sci.*, vol. 12, no. 21, p. 10804, Oct. 2022.

[89] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Piscataway, NJ, USA, Mar. 2017, pp. 776–780.

[90] Z. Chen, G. Xie, M. Chen, and H. Qiu, "Model for underwater acoustic target recognition with attention mechanism based on residual concatenate," *J. Mar. Sci. Eng.*, vol. 12, no. 1, p. 24, Dec. 2023.

[91] H. Feng, X. Chen, R. Wang, H. Wang, H. Yao, and F. Wu, "Underwater acoustic target recognition method based on WA-DS decision fusion," *Appl. Acoust.*, vol. 217, Feb. 2024, Art. no. 109851.

[92] S. Feng, X. Zhu, and S. Ma, "Masking hierarchical tokens for underwater acoustic target recognition with self-supervised learning," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 1365–1379, 2024.

[93] D. Liu, H. Yang, W. Hou, and B. Wang, "A novel underwater acoustic target recognition method based on MFCC and RACNN," *Sensors*, vol. 24, no. 1, p. 273, Jan. 2024.

[94] X. Wang, P. Wu, B. Li, G. Zhan, J. Liu, and Z. Liu, "A self-supervised dual-channel self-attention acoustic encoder for underwater acoustic target recognition," *Ocean Eng.*, vol. 299, May 2024, Art. no. 117305.

[95] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[96] Y. Wang, H. Zhang, W. Huang, M. Zhou, Y. Gao, Y. An, and H. Jiao, "DWSTr: A hybrid framework for ship-radiated noise recognition," *Frontiers Mar. Sci.*, vol. 11, Feb. 2024, Art. no. 1334057.

[97] Y. Xie, J. Ren, and J. Xu, "Unraveling complex data diversity in underwater acoustic target recognition through convolution-based mixture of experts," *Exp. Syst. Appl.*, vol. 249, Sep. 2024, Art. no. 123431.

[98] X. Cui, Z. He, Y. Xue, K. Tang, P. Zhu, and J. Han, "Cross-domain contrastive learning-based few-shot underwater acoustic target recognition," *J. Mar. Sci. Eng.*, vol. 12, no. 2, p. 264, Feb. 2024.

[99] Q. Yao, Y. Wang, and Y. Yang, "Underwater acoustic target recognition based on data augmentation and residual CNN," *Electronics*, vol. 12, no. 5, p. 1206, Mar. 2023.

[100] S.-Z. Tian, D.-B. Chen, Y. Fu, and J.-L. Zhou, "Joint learning model for underwater acoustic target recognition," *Knowl.-Based Syst.*, vol. 260, Jan. 2023, Art. no. 110119.

[101] H. Yang, G. Xu, S. Yi, and Y. Li, "A new cooperative deep learning method for underwater acoustic target recognition," in *Proc. OCEANS*, Jun. 2019, pp. 1–4.

[102] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[103] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[104] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*.

[105] D. Li, F. Liu, T. Shen, L. Chen, and D. Zhao, "A robust feature extraction method for underwater acoustic target recognition based on multi-task learning," *Electronics*, vol. 12, no. 7, p. 1708, Apr. 2023.

[106] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020, *arXiv:2005.08100*.

[107] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "QuartzNet: Deep automatic speech recognition with 1D time-channel separable convolutions," 2019, *arXiv:1910.10261*.

[108] M. Langenkamp and D. N. Yue, "How open source machine learning software shapes AI," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, V. Conitzer, J. Tasioulas, M. Scheutz, R. Calo, M. Mara, and A. Zimmermann, Eds., New York, NY, USA, 2022, pp. 385–395.

[109] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[110] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-supervised audio spectrogram transformer," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 10, pp. 10699–10709.

NILS MÜLLER received the B.Sc. degree in technical and applied physics and the M.Sc. degree in electronics engineering from Hochschule Bremen, Germany, in 2019 and 2021, respectively. He is currently pursuing the joint Industrial Ph.D. degree with ATLAS ELEKTRONIK GmbH and the University of Wuppertal, Germany. During his studies, he specialized in optical metrology, with a focus on deep learning and machine learning-based data evaluation. From 2020 to 2021, he was a Research Assistant with the I3M Institute, Hochschule Bremen. Since 2023, he has been a full-time Ph.D. Student with ATLAS ELEKTRONIK GmbH. His research interests include data-driven sonar signal processing and deep learning.

JENS REERMANN received the joint M.Sc. degree in microelectronic systems from the University of Applied Sciences Hamburg, Germany, and the University of Applied Sciences Westküste, Germany, in 2013, and the Dr.-Ing. degree from Kiel University, Germany, in 2017. During his studies, he participated in the practical orientated support program at Lufthansa Technik AG. In 2017, he was a member of the Digital Signal Processing and System Theory Group and the Collaborative Research Group (CRC 1261), Kiel University, with an emphasis on magnetoelectric sensors. From 2017 to 2018, he was a Software Engineer with Dräger Safety AG Co. KGaA. In 2018, he joined ATLAS ELEKTRONIK GmbH as a Systems Engineer. Since 2020, he has been leading the Team "Signal Exploitation and Localization." His research interests include digital signal processing, sonar systems, and artificial intelligence (AI).

TOBIAS MEISEN received the degree in computer science with a specialization in data mining and data exploration and management and the Ph.D. degree in engineering.

From October 2015 to August 2018, he was a Junior Professor with RWTH Aachen University. He contributed his research results here as part of the Cluster of Excellence "Integrative Production Technology for High-Wage Countries." He has been a Professor with the Institute for Technologies and Management of Digital Transformation (TMDT), University of Wuppertal, since September 2018. He has been the Chair of the Interdisciplinary Center for Data Analytics and Machine Learning (IZMD) and the Founding Ambassador of the School of Electrical, Information and Media Engineering, since October 2018. He is currently the Co-Founder of Hotsprings GmbH, which is a part of umlaut. In his daily work, he is dedicated to digital transformation, especially modern information management in a networked and digital world. He is the co-author and the author of more than 100 scientific publications and regularly serves as a reviewer for various conferences and journals. In recent years, he and his team have successfully supported numerous research and development projects with partners from research and industry. His research interests include conceptual design, development, and implementation of autonomous technical systems, with a focus on deep learning and machine learning. In his second research area, he is dedicated to the collection and integration of digital data, with a special focus on the evolutionary construction and the management of knowledge graphs.

Dr. Meisen was awarded the Young Researcher Award as part of the first funding phase of the Excellence Initiative, in March 2010.