IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 May 2023, accepted 2 June 2023, date of publication 8 June 2023, date of current version 20 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3283982

SURVEY

A Survey on Artificial Intelligence-Based Acoustic Source Identification

RUBA ZAHEER, IFTEKHAR AHMAD, (Member, IEEE), DARYOUSH HABIBI, (Senior Member, IEEE), KAZI YASIN ISLAM, (Member, IEEE), AND QUOC VIET PHUNG, (Member, IEEE)

School of Engineering, Edith Cowan University, Perth, WA 6027, Australia

Corresponding author: Ruba Zaheer (r.zaheer@ecu.edu.au)

This work was supported in part by the Department of Jobs, Tourism, Science and Innovation, Defence Science Center, Australia, under Grant G1006608.

ABSTRACT The concept of Acoustic Source Identification (ASI), which refers to the process of identifying noise sources has attracted increasing attention in recent years. The ASI technology can be used for surveillance, monitoring, and maintenance applications in a wide range of sectors, such as defence, manufacturing, healthcare, and agriculture. Acoustic signature analysis and pattern recognition remain the core technologies for noise source identification. Manual identification of acoustic signatures, however, has become increasingly challenging as dataset sizes grow. As a result, the use of Artificial Intelligence (AI) techniques for identifying noise sources has become increasingly relevant and useful. In this paper, we provide a comprehensive review of AI-based acoustic source identification techniques. We analyze the strengths and weaknesses of AI-based ASI processes and associated methods proposed by researchers in the literature. Additionally, we did a detailed survey of ASI applications in machinery, underwater applications, environment/event source recognition, healthcare, and other fields. We also highlight relevant research directions.

INDEX TERMS Acoustic source identification, feature extraction, machine learning, deep learning, sound classification.

I. INTRODUCTION

Acoustic data carry valuable insights for scientific and engineering research communities across different sectors that include human speech recognition [1], ocean exploration and localization [2], animal and birds localization [3] and underwater geographical imaging [4]. Acoustic data analysis is a complex process that encounters a number of challenges, including inaccurate data, inadequate measurements, noise/reverberation, and large amounts of data. For instance, multiple arrivals of an acoustic signal can result in poor source localization. Utterances and background noises in sound recordings make it difficult for machines to interpret an acoustic signal [5], [6].

In recent years, advances and developments in acoustic processing have been broadened through the application of AI principles. With the progress of AI, the capabilities of pattern recognition have tremendously increased in image processing, computer vision applications and speech processing. AI in acoustics has significantly contributed and progressed in the past few years. Advanced acoustic processing techniques can consolidate the strengths of AI methods to achieve better performance when it comes to recognition, identification and localization than conventional audio processing methods.

AI-based (ASI) can be tailored to meet the needs of a diverse range of applications. For example, AI-based ASI plays a vital role in the industrial sector through continuous condition monitoring [7]. It can be used to detect and identify faults in different components of machines, thereby improving their safety, efficiency and reliability.

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa M. Fouda.

The undersea domain is becoming more contested day by day; therefore, demanding constant surveillance and monitoring operations. The acoustic signature radiated by marine vessels has unique information that can be utilized to identify, detect and recognize a marine vessel. In addition, AI-based ASI can be employed in target detection and recognition that aids the navy in the crucial investigation of both deep and shallow underwater environments [8]. The surveillance of human activities in modern environments is predominately carried out in urbanized areas for the safety and security of the general public. Environmental sound recognition (ESR) systems are incorporating AI-based ASI to identify sound sources that exist in our everyday environment [9]. Additionally, ASI can also be tailored to meet the requirements of a wide range of healthcare applications, such as cardiac auscultation [10], fall detection [11] and hearing-impaired wearable devices [12]. This technology is not only limited to the above-mentioned applications, but is also useful for music genre classification [13], animal and bird species identification [14], [15], robotics [16], drone detection [17], and insect identification [18] as well.

Owing to the significance of ASI recently, a number of research works have been carried out and published in the literature in different fields of applications. We have summarized some of the latest review works in ASI, and they are outlined in Table 1.

The authors in [19] surveyed the environmental sound identification (ESI) and sound event recognition for surveillance applications in which various domain features have been compared that are suitable for sound events and scene identification systems. Moreover, AI model-based approaches have also been compared by using different available data sets for ESI and event detection. Lei et al. in [20] have presented a detailed review on intelligent fault diagnosis using the sound of machines in industrial settings and AI techniques. In addition, a detailed road map for future researchers is discussed to enhance the quality and outcome of AI models for intelligent fault diagnosis. AlShorman et al. [25] also published a study on fault diagnosis in components of motors using radiated acoustic patterns and AI methods. The authors in [21] surveyed different machine learning (ML) techniques used in the past few years and elucidated a deep learning (DL) framework for underwater target recognition. The information from underwater images is used to classify targets using DL. Similarly, Chen et al. [23] also reviewed the underwater target recognition application based on DL methods and discussed problems in feature extraction (FE) and captured underwater image quality.

In addition to this, the authors in [22] reviewed systematically and mostly used DL methods for the classification of heart sounds for cardiac auscultation. In this paper, two DL methods, convolutional neural network (CNN) and recurrent neural network (RNN) is emphasized over the course of the past five years. Nunes [24] published a detailed review on anomaly detection in the object based on its acoustic signature. In their review, ML techniques from 2010 to 2020 are studied and analyzed for anomalous detection. Most recently, Bansal and Garg [26] focused on ESI and classification using various traditional ML classifiers and deep neural networks(DNNs). They have also explicated various pre-processing and FE schemes for ESI. Lastly, the authors in [27] introduced and reviewed in their study the integration of internet of things (IoT) and ML approaches for smart environments in which acoustic sensing using IoT and ML algorithms has been outlined.

To the best of our knowledge, there is no other survey published so far that presents such a detailed overview of AI-based ASI. We have compiled a detailed overview of AI-based ASI along with its various applications to provide future researchers with a holistic understanding of this concept. Our contribution through this survey is organized as follows:

- We surveyed and compared recent reviews and surveys on AI-based ASI for various applications including, surveillance, healthcare, smart cities, underwater detection and machinery fault detection.
- We present a detailed overview of AI-based ASI process. We discuss data acquisition and traditional audio processing methodologies along with famous databases used in various fields by researchers. Moreover, we compiled and provide a detailed discussion on the significance and methodology of traditional audio processing techniques, FE methods, ML and DL algorithms that have been mostly used in the literature to aid readers in forming an effective model for a given problem.
- We provide a detailed overview of AI-based ASI in diagnosing faults in industrial machinery, underwater applications, event source detection (ESD), and ESI, healthcare, music-genre classification and wildlife monitoring applications. In our review, we provide a thorough comparison and analysis based on previous works' limitations and performance metrics.
- We discuss the possible future research directions in the light of this survey and some generalized methodological recommendations for future researchers to extend the pathways in this area and overcome problems in ASI-based ASI.

We organized the remaining paper as follows: In Section II, the methodology of data acquisition and audio data pre-processing has been discussed. We highlight some famous databases that have contributed to AI-based ASI. Next, we compile various popularly used FE techniques, AI algorithms and evaluation metrics. In Section III, we survey ASI in industrial fault detection, underwater applications, ESI, healthcare, music-genre classification, wildlife monitoring and forensics. Future research directions are discussed in Section IV followed by the conclusion of this paper in Section V.

TABLE 1. Brief summary of relevant review works.

Review work	Year	Application	Feature Extraction techniques	AI techniques	Overall Performance
Chandrakala *et al.* [19]	2019	Audio autonomous surveillance	Time Domain: (Zero crossing rate, Short time energy, wave features). Frequency domain: (Spectral roll-off, flatness, centroid, flux, pitch ratio and spectral dynamic features). Cepstrum: (MFCCs, MFCCs derivatives, LPCCs, LFCCs, GTCCs, PLP). Energy: (Signal energy and log energy), Image-based features.	ML and DL	Recognition accuracy: Highest accuracy rate of 73.7% is achieved when CNNs are trained on manually engineered spectrogram features.
Lei *et al.* [20]	2020	Machine fault diagnosis	Time domain: (Mean, Standard Deviation (SD), root amplitude, Root Mean Square (RMS), peak value, shape, skewness, kurtosis, clearance, impulse and crest indicators). Frequency domain: (mean, center, RMS, SD). Time-Frequency domain such as energy entropy.	ML and DL	Statistical analysis of performance is not presented.
Teng *et al.* [21]	2020	Underwater target recognition	Raw images	DL only	Highest recognition accuracy rate of 93.34% is achieved by compound CNNs in environmental interference conditions as compared to other DL approaches. It is the most stable approach for different underwater datasets.
Chen *et al.* [22]	2021	Cardiac auscultation	MFCCs, spectrograms, discrete wavelet transform (DWT) coefficients, basic time and frequency domain features.	DL only	CNN along with recurrent neural networks (RNNs) trained on improved MFCCs achieved approximately 98% accuracy on different datasets. Additionally, RNN (LSTM and BLSTM) approaches achieved approximately similar accuracy rates which are the highest in studies.
Chen *et al.* [23]	2021	Underwater target recognition	Raw images	DL only	In the review, Deep CNN achieved an average recognition rate of 90% to recognize underwater small targets only.
Nunes *et al.* [24]	2021	Anomalous sound detection	Mainly MFCCs, log-Mel energy and Mel-spectrogram are presented.	ML only	Among 33 ML algorithms, Auto Encoders (AEs) and CNNs outperform all other ML models and are most cited for their performance. Detailed statistical analysis is not presented in the review.
Alshorman *et al.* [25]	2021	Fault diagnosis in motors	Selection of amplitudes, spectral analysis and Shannon's entropy, envelop analysis, Empirical mode decomposition (EMD), acoustic spectral imaging and wavelet transforms.	ML and DL	DL-based approaches overcame the drawbacks of ML approaches. Statistical comparisons are not presented for different methods.
Bansal *et al.* [26]	2022	Surveillance, monitoring and security	MFCCs, Code Excited Linear Prediction (CELP), Temporal features (ZCR, autocorrelation, RMS, energy entropy, Short time energy and linear prediction coefficients (LPC), Image features such as log mel spectrogram, log Gammatone spectrogram and Spectral features.	ML and DL	Merits and demerits of different FE and AI methods are discussed in the study. CNN outperforms all MFCCs based ML methods with an accuracy score of 92.90%. Among DL approaches, CRNN has been proven effective for environment sound classification.
Peralta *et al.* [27]	2022	Smart cities	FE techniques are not discussed in the review.	ML only	The authors only presented the findings and trends of different studies.

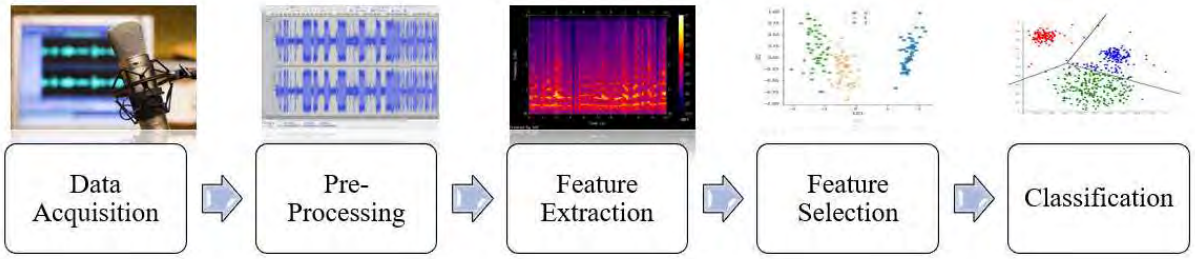

FIGURE 1. Acoustic source identification process.

II. ACOUSTIC SOURCE IDENTIFICATION OVERVIEW

AI-based acoustic source identification is a systematic process of recognizing an unknown source using the sound that it generates. Further, it includes five basic stages that comprise, data acquisition, data pre-processing, FE, feature selection, and identification or classification using AI algorithms. Generally, the model performs better when all of these steps are followed. These stages are illustrated in Figure 1. In this section, we have explained in detail all the steps of ASI and highlighted methods that have been previously used in literature frequently.

A. DATA ACQUISITION

In this subsection, we detail the methodology of data acquisition for the ASI process. Data acquisition can be defined as the process of collecting and gathering relevant information to drive the aims and objectives of an AI-based problem. Data collection is the fundamental step of the AI-ASI process. The methodology of data collection can impact the performance of an AI algorithm; therefore, it can alter the decision of a given problem.

There are two ways to generate data for further processing; synthetic data and real data. Supervised learning algorithms are limited by the scarcity of labeled data. There can be cases where a sufficient amount of data is unobtainable to characterize a particular problem, for example, underwater environment, hyper-diverse rain forests, etc. Therefore, an investigator can simulate a large amount of data to achieve an efficient recognition system. To generate realistic data, the investigator has to take into account natural noise and reverberation must be added to the generated sound.To limit the massive use of simulated data, data augmentation techniques [49] is a promising solution. Data augmentation generates additional training examples without more recordings, often leading to improved performance. Real acoustic data collection can be conducted via microphones and hydrophones. After collecting data from an experimental procedure, raw data needs to be initially labeled to avoid mixing. Real data can also be aggregated by using data augmentation techniques if the gathered samples are insufficient.

In AI-based ASI, public evaluation and benchmark datasets help the research community to investigate the performance of various proposed systems. We have categorized some of the popular data sets with respect to their applications. In addition, each data set's properties and web links are mentioned in Table 2.

B. DATA PRE-PROCESSING

Raw acoustic data is mostly not suitable for FE and needs to be preprocessed. Raw data in its original state is not suitable for the AI algorithm as it can compromise the performance of the model. There are various reasons that can affect the suitability of data for learning algorithms such as excess data, insufficient data and tampered data. Some of the data might be corrupted in case of loads of data. In contrast, insufficient data lacks the necessary attributes of the dataset. In both cases, the predictive ability of the model gets weakened resulting in poor accuracy. For example, the decision-tree algorithm splits the data set into training and testing sets and missing information may lead to an inaccurate decision [50]. Moreover, there are other important data pre-processing steps required to ensure the data is prepared for the next stage which are as follows:

1) DATA AUGMENTATION AND INTEGRATION

Data integration is defined as combining various heterogeneous data into unified data. This involves two techniques known as tight coupling and loose coupling [51]. Contrastingly, data augmentation is the technique of adding data by synthesizing new data from available data. Data augmentation can be carried out for time and frequency domain features. Recently, SpecAugment has become popular in audio processing as an effective data augmentation technique, especially for spectrograms [52].

2) DATA CLEANING

Data cleaning is done to enhance the quality of the signal. The process involves the identification of inaccurate or irrelevant data and the elimination/replacement of such unwanted data in a data set. Errors can occur during naming, missing entries, or human negligence while gathering. Denoising of data is also a part of the data cleaning process. Noisy components can be removed from audio samples by filtering [53]. Noise problems can also be countered by signal enhancement

TABLE 2. Datasets used for ASI in various applications.

Dataset	Application	Data quantity	Research Paper	Weblinks
MIMII-2019	Industrial fault detection	Source: Valves, pumps, fans, slide rails 26,092 recording samples of normal condition 6065 samples of anomalous condition	Purohit *et al.* [28]	[29]
DCASE Challenge (2013, 2016-2023)	Sound event detection	Latest DCASE challenge 2023: Task1: 2400 segments from 10 classes Task2: 70 segments from 7 classes Task3: 960segments from 13 classes Task4: 14412 segments from 10 classes Task 5: 1260 segments from 47 classes Task 6: 6972 segments from clotho dataset Task 7: 4850 segments from 7 classes	Mesaros *et al.* [30]	[31]
UrbanSound8K	Environment sound source identification	8732 recording samples from 10 classes	Salamon *et al.* [32]	[33]
ESC-10	Environment sound source identification	400 recordings from 10 classes	Piczak *et al.* [34]	[35]
ESC-50	Environment sound source identification	2000 recordings from 50 classes	piczak *et al.* [34]	[35]
PASCAL Challenge	Cardiac disease diagnosis	Challenge1: Crowd sourced from public via mobile application Challenge2: 656 audio files 4 classes	Bentley *et al.* [36]	[37]
2016 PhysioNet/ CinC Challenge	Cardiac disease diagnosis	Total 4430 recordings from 2 classes (Normal and abnormal)	liu *et al.* [38]	[39]
Daily Sounds	Activity monitoring/Surveillance	1049 Audio recordings from 18 sound classes	Sehili *et al.* [40]	N/A
RWCP Sounds	Sound scene detection	100 recording samples for 90 kinds of sounds	Nakamura *et al.* [41]	[42]
Insect Sounds	Agriculture development	95 audio recordings from 9 classes based on insect activities	Zhang *et al.* [43]	[44]
Dairy Cow Jaw movement sounds	Animal behaviour analysis	52 audio files from 3 classes of jaw movements	Vanrell *et al.* [45]	[46]
GTZAN Music-genre	Music-genre classification	100 audio recordings from 23 music and speech classes	Tzanetak *et al.* [47]	[48]

techniques [54]. Similarly, silence can be easily detected in audio samples and removed using amplitude-based silence detection algorithms [55].

3) DATA TRANSFORMATION

Data transformation is required when acoustic data attributes need to be scaled at the same level. Multiple features present in a data set, that might be mapped to different scales needs to be scaled at standard value. Therefore, normalization can be used to normalize all features to the same scale such as min-max normalization, z-score normalization and decimal scaling.

4) DATA LABELING

Pre-processing of the dataset also involves annotation/labeling of data after denoising and transformation. Usually, this is carried out by expert acousticians who are familiar with the targetted sounds and able to track sounds in an audio file. This involves the identification of targetted sound in an audio file and assigning it with a label also known as class. These labels are used to train an AI algorithm. Annotation can also be done by visually inspecting spectrograms of audio files.

C. FEATURE EXTRACTION AND SELECTION

AI models require discriminatory and distinct features to learn information about any particular sound. Therefore, FE is defined as the process of extracting meaningful information from raw data by removing most of the redundant data. The extent of training decides the performance of an AI algorithm. The effectiveness of these features results in accurate predictions from an algorithm. Therefore, FE and selection is the method of finding the features that possess most of the information of a particular data set. In this

subsection, we discuss the popularly used audio FE methods as illustrated in Figure 2.

1) TIME DOMAIN FEATURES

- Zero Crossing Rate (ZCR) is defined as the rate of change of an audio signal from positive to negative and negative to positive crossing zero level in the middle. In simple terms, it is the count of signals crossing zero level in one second period of time. The ZCR for k^{th} frame is represented mathematically as:

$$Z(k) = \frac{1}{2N} \sum_{n=1}^{M} |sgn[x_k(n)] - sgn[x_k(n-1)] \quad (1)$$

where M is the length of the frame and sgn(.) is the sign function that is

$$sgn[x_k(M)] = \begin{cases} 1 & x_k(n) \geq 0 \\ 0 & x_k(n) < 0 \end{cases} \quad (2)$$

ZCR estimates fundamental frequency and is proven efficient for voice-based systems [56]. ZCR conveys important information about the voiced and silent frames of a voice signal. Due to its ability to give discriminating frequency information, this feature can be designed as a classifier [57].

- ADSR envelop detection stands for Attach, Delay, Sustain and Release. This FE method is mostly used in music-genre classification and is not applicable to real-time sounds due to the absence of decay envelop. Additionally, it does not work with environmental sounds since they lack sustain temporal envelop. Therefore, this kind of envelope is known as AR envelope which is mostly used in timbre analysis in musical instruments [58].
- Log attack time: As its name implies, this is the logarithmic (base 10) of the time interval between the start time until it has reached to its stable stage. If T_0 is the starting time of the signal and T_1 is the maximum time then the range can be found by the length of the signal as follows:

$$Range = log_{10} \frac{1}{samplingrate} \quad (3)$$

$$LAT = log_{10}(T_1 - T_0) \quad (4)$$

Among its applications are the detection of musical onsets [59] and the detection of environmental and event sounds [60].

- Energy-based: Energy-based FE is applied on non-stationary audio signals. This is due to the fact that the energy of non-stationary signals varies at different segments; therefore, cannot be defined by a single energy value. Windowing is usually used to segment non-stationary signals into quasi-stationary frames. Short-time energy is calculated as the average energy of those frames [61]. Short-time energy is a promising detector of energy contents of un-voiced and voice signals [62]. As compared to voiced frames, it is relatively low in unvoiced frames. A number of applications can be found in audio analysis, including environmental sound and event detection [63], music systems [59], and acoustic monitoring systems [9].
- Auto-correlation is the extent of similarity of a signal with its delayed version. This measure is represented by $+1$ and -1 values. The maximum relation is given by $+1$, the minimum relation is given by -1, and the absence of any relation is represented by 0. For example, the correlation at some value of lag is less than 1 but greater than 0 depending on the extent of similarity [64]. Therefore, the correlation at zero lag will always be 1 since the signal is repeated undelayed. Music analysts use auto-correlation as a method of analyzing beats, tempo, and pitch.

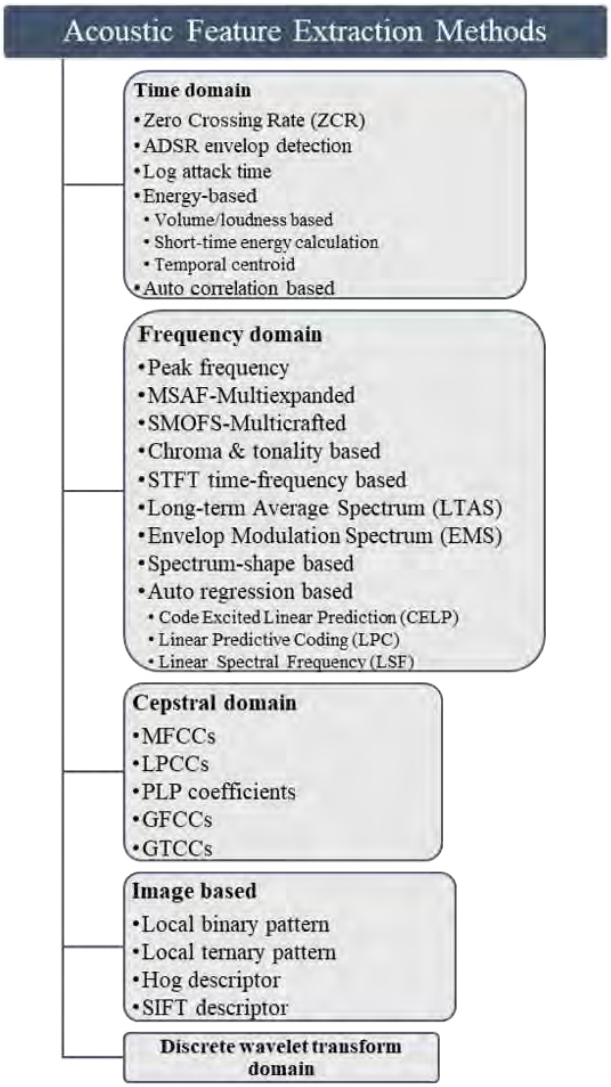

FIGURE 2. Acoustic feature extraction methods.

2) FREQUENCY DOMAIN FEATURES

- Peak frequency: Peak frequency conveys information about the most dominant frequency and the fundamental frequency of the signal. This is defined as the frequency of maximum power. In the case of music and speech classification, peak frequency information is used since vocal sounds have pure tones (sine wave). Peak frequency provides the best estimate of the pitch in this case.
- MSAF-Multiexpanded stands for the method of selection of amplitudes of frequency multi-expanded filter. These features are mostly used in fault diagnosis in electrical drilling motors [65] and commutator motors [66]. These acoustic features are handcrafted and generated by computing the difference between Fast Fourier Transform (FFT) spectra of different classes. The absolute value of differences forms a feature vector that is used to construct classes.
- SMOFS-Multicrafted is shortened method of frequency selection which is the same as MSAF-multi expanded and is being applied in the industrial sector. This is also used to classify faults in motors [65]. The only difference between SMOFS-multi crafted and MSAF-multi expanded FE method is the selection of frequency components after FFTs computation.
- Short-time Fourier transform (STFT) is the time-frequency transform of a signal represented as time-frequency distribution (TFD). In the time-frequency analysis of an audio signal, time is on one axis and frequency is on another axis. Changes in the amplitude of the signal over time can be observed along with the magnitude of frequency content in the signal. With the use of STFT, a time-frequency analysis can be performed on audio signals with abrupt discontinuities and patterns, which is a promising method for non-stationary signals. There are different types of TFD techniques depending on the requirement such as linear [67], quadratic [68], positive [69] and matching pursuit TFDs [70]. TFDs are used in audio processing in the detection of industrial gear faults [71], seismic data processing [72] and environmental sound source recognition [67].
- Chroma and tonality based: Chroma-based features represent an audio signal for example music audio in the form of 12 chroma segments mapped from the spectrum. Logarithmic STFT is used to compute these bin/segments. This representation of mapping is called chromagram. As the statistics from chroma energy distribution also have information about the audio, it is an important method for obtaining chroma-based features.

  Tonality-based features depend on the fundamental frequency of the harmonic audio signal. Tonality-based FE is only applicable to stationary periodic audio signals. The fundamental frequency is the lowest frequency of a periodic signal. For example, the pitch of music audio gives an estimate of the fundamental frequency. Tonal features find their applications in music onset detection [59], environmental sound source detection [73], and audio retrieval systems [74].
- Long-term Average Spectrum (LTAS): LTAS is the FFT generated unusual spectral information from an audio signal. Due to its ability to capture the spectrum of both glottal source and vocal tract, it is widely used in pathological speech [75]. LTAS acquires spectral information from every octave of a filtered speech signal. The spectral information comprises certain parameters which are combined to form a 99-dimensional feature vector. These parameters are Root mean square(RMS) values, normalized mean and standard deviation (SD) of segment RMS, segment SD normalized by full-band and band RMS, skewness, kurtosis, range of segment RMS and variation in RMS energy in ensuing segments.
- Envelop Modulation Spectrum (EMS): This FE method uses amplitude modulated audio signal. EMS is a representation of the energy distribution in amplitude variations across different frequencies. In the first step, a Butterworth filter of 8^{th} order is used to generate octave bins centered at certain frequencies from the audio signal. Following this, the Hilbert transform is used to extract the envelope of the original signal and the filtered octave bin. Then power spectrum is estimated by taking Discrete Fourier Transform (DFT) of the envelope. A 60-dimensional feature vector is then constructed containing 6 features derived from the power spectrum, including peak frequency, peak amplitude, spectrum energy (0-4Hz and 4-10Hz), and energy ratio. EMS features can be utilized to solve classification problems in pathological and control speech [76], [77].
- Spectrum-shape based: In spectrum-based features, a spectral centroid is commonly used to describe the position of a spectrum's center of mass. Normalized amplitude is computed by the distribution of frequencies and probabilities across the spectrum. The spectral centroid is a brightness parameter that describes the brightness of an acoustic signal. Additionally, this also conveys information about musical timbre [78] which is why it is employed in music-mood classification [79] scenarios.

  The spectral center is another type of spectrum-based feature that relies on median frequency of the signal spectrum. Due to its energy balancing attribute, this feature is used in rhythm tracking in the music field [80]. The spectral roll-off feature is defined as a frequency under which 95 percent of the energy remains. Audio surveillance systems [9], music-genre classification [47] and speech-music [61] classification use this feature for discrimination. There are other spectrum-based features that have different characteristics of the spectrum. Spectral spread, for example, categorizes sounds according

to their spectral bandwidth, while spectral skewness and spectral kurtosis indicate the symmetry and flatness of the spectrum, respectively.

- Auto regression-based features commonly include linear prediction coding coefficients (LPCCs) synthesized using linear prediction analysis of a signal. In this way, it eliminates the problem of redundancy by estimating new values based on the previous coefficients. The linear prediction model generates a compressed spectral envelope of a digital speech; therefore, it is commonly used in audio segmentation and retrieval applications. Additionally, there is another modified version of LPCCs which is known as Code Excited Linear Prediction (CELP) that reassembles the human vocal tract using a linear prediction model. In linear prediction models, excitation signals are fed into adaptive or fixed code-book entries. Afterwards, the model performs the search in the perceptually weighted domain and closed iterations. Due to its promising ability to code speech, this delivers better quality than low bit-rate algorithms. Therefore, they are used in ESI applications [81].

3) CEPSTRAL DOMAIN FEATURES

Cepstrum represents the cepstral domain that is generated by taking the inverse Fourier transform of log spectrum of a waveform. Cepstrum is categorized into three types depending on different audio applications. In speech processing, power cepstrum features are used, while real cepstrum features are used for pitch detection [82]. Analyzing cepstrum features is called cepstrum analysis or quefrency analysis. Cepstral features have a number of benefits such as source-filter separation, orthogonality and conciseness. These attributes make them suitable for training ML algorithms. In this subsection, we discuss various types of cepstrum features and their potential applications.

- Mel spectrogram: Mel spectrograms are widely used features for DL algorithms. They convey useful information about an acoustic signal such as loudness or intensity over time at different frequencies. They are based on the Mel scale which is the logarithmic transformation of a signal's frequency. The behavior of mel scale reassembles to human's perception of sound at different frequencies. The relationship between mel scale and frequency is shown mathematically as:

$$m = 2595.log(1 + \frac{f}{700}) \tag{5}$$

  If a signal is denoted by $x(n)$ and k_a is the index of mel scale filter, then Log Mel spectrogram is denoted by $S_a(n_a, k_a)$ which can be computed by Figure 3. Mel spectrograms have been used in a variety of applications such as speech-emotion recognition [83], healthcare [84], underwater target recognition [85], industrial fault diagnosis [86] and many others.
- Mel Frequency Cepstral Coefficients (MFCCs) are mostly used cepstrum features for audio processing due to their ability to resemble the human auditory system. An audio frame is pre-emphasized and hamming windowed. Subsequently, the time domain signal is converted into frequency (N-point) by using DFT. If $s(n)$ is an audio signal then the energy spectrum in the frequency domain can be represented by the below equation.

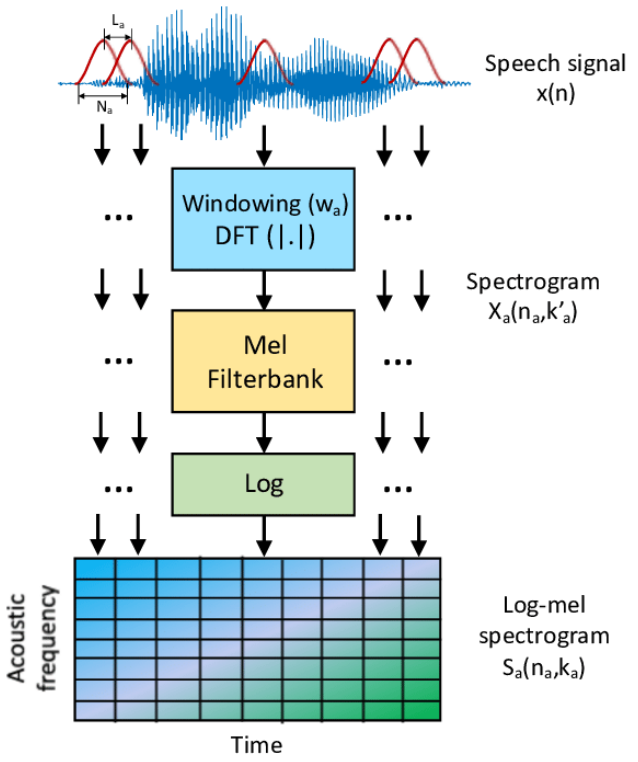

FIGURE 3. Computational process of log-Mel spectrogram [87].

$$|S(k)|^2 = |\sum_{n=1}^{N} s(n).e^{\frac{(-j2\pi nk)}{N}}|^2 \quad 1 \le k \le N \tag{6}$$

Then the filter banks are imposed on the frequency spectrum $S(k)$. Discrete Fourier Transform is taken again on filter bank energies and MFCCs are obtained that can be written as

$$c_m = \sqrt{\frac{2}{Q}\sum_{p=0}^{Q-1} log[e(p+1)].cos[m.(\frac{2p-1}{2}.\frac{\pi}{Q})]} \tag{7}$$

MFCCs are prominently used in speech and speaker recognition systems [88], [89], vowel detection [90], music-genre classification and audio similarity analysis [91].

- Linear Prediction Cepstral Coefficients (LPCCs) are generated when LPCs are transformed to a cepstral domain. LPCCs are less sensitive to numerical precision as compared to LPCs. It is very easy to convert LPCs to the cepstral domain. LPCCs carry lots of significance in

noise elimination [92], music-genre classification [93] and speech recognition systems [94].

- Perceptual linear prediction (PLP) cepstral coefficient is another form derived from Linear prediction coefficient. The PLP coefficients represent critical band spectral resolution, equal-loudness curve and intensity loudness power law [95]. To generate PLP coefficients, perceptual processing is performed; afterwards, auto-regressive modeling is done before converting those coefficients into cepstral coefficients. PLP coefficients are useful in animal sounds classification [96], emotion identification [97] and speech recognition systems [98].
- Greenwood function cepstral coefficients (GFCCs) use MEL features and are termed as a generalized form of MFCCs and deliver fine vocal representations of animals and birds. This is why, GFC features are primarily founded in terrestrial mammals. GFCCs are derived from the greenwood equation that closely maps the cochlear-frequency position for all terrestrial animals and birds species. Their primary applications include animal and bird sound identification and classification [99].
- Gammatone cepstral coefficients (GTCCs) is one of the most noise-robust features in automatic speech recognition systems. GTCCs are extracted in a similar way as MFCCs and are based on gammatone filter banks. These filter banks generate output which is the frequency-time domain representation of an acoustic signal. GTCCs can be derived to more features by taking first and second-order derivatives. They are employed in environmental sound recognition and automatic speech recognition systems (ASR) [99].

4) IMAGE BASED FEATURES

An image of an object contains patterns and points that help in the identification of that particular image. AI-based algorithms distinguish those objects with the help of those patterns. DL algorithms mostly use image-based features as inputs for recognition, identification and classification. In this subsection, we have discussed popularly used image-based features.

- Local Binary patterns (LBPs): In audio processing, local binary patterns are called visual descriptors and can be extracted from the spectrograms of audio signals. These patterns possess information on grayscale contrast and local spatial descriptors. The feature vectors extracted from the spectrogram are used by ML and DL algorithms for textural analysis. LBPs are powerful features used in computer vision applications. They have been proven useful in audio scene detection [100], psychological diseases analysis from speech [101] and emotion detection applications [102].
- Local Ternary patterns (LTPs): Local ternary pattern is an extended version of LBPs. Similar to LBPs, they are also extracted from spectrograms. The difference lies in the measurement scales of pixels. LBPs are scaled in binary pattern (0 and 1) only whereas, LTPs are scaled into three values that are -1,0 and 1. LTPs carry significance in audio scene detection and classification [103] and healthcare analysis [104].
- Histogram of gradients (HOG) descriptor: Histogram of gradients (HOG) is another feature descriptor that conveys information about the structure and shape of an object. These descriptor measures the magnitude and angle of the gradient and generates histograms. Similar to other image-based features discussed above, these features also extract information in the frequency-time domain. These features have been used in emotion detection [102], audio scene classification [105] and snore sound classification [106].
- Scale-invariant feature transform (SIFT) descriptor: SIFT is an image-based FE method used to generate local features for small and large-size objects. Their processing is efficient and close to real-time. Another benefit of SIFT features is its extensibility to a wide range of other types of features. SIFT features are used in computer vision applications, emotion detection [102] and audio/video concept classification [107].

5) DISCRETE WAVELET FEATURES

An audio signal can be converted into a time-frequency representation using a wavelet transform. This is merely a product of the audio signal with a wavelet. Wavelet transform works in two ways: continuous and discrete. Discrete Wavelet transform (DWT) is more efficient due to the frequency filter bank and can extract information from non-stationary signals such as audio signals. DWT delivers uniformity in time-frequency resolution. The coefficients generated by DWT are wavelet features and can also be extracted from wavelet packet decomposition. Discrete wavelet features are widely used in audio analysis [108], music classification [109], motor fault detection [110] and emotion recognition [111].

6) OTHER SPECIAL FEATURES

Researchers have combined a number of approaches to improve the extraction of discriminatory features and identification accuracy. The combination of DWT and MFCCs is usually done by concatenating both MFCCs and DWT features. A combination of DWT and MFCC method performs relatively better in noisy scenarios than either technique alone. For example, Authors in [112] and [113] used the hybrid method of MFCC and DWT in speaker recognition and speaker verification and achieved higher accuracy in different noisy cases. Similarly, Hidayat et al. [114] implemented the same combinational approach in the text-dependent speaker recognition system and achieved 96.67% overall recognition accuracy. Researchers have also combined MFCCs with GFCCs in various applications in order to achieve greater efficiency. In [115], authors have used the fusion of MFCCs, GFCCs and mel-spectrograms to classify heart conditions based on heart sounds. On the

PhysioNet2016 set, accuracy was achieved at 96%, which is higher than the accuracy achieved using MFCCs alone. Moreover, Al-Qaderi et al. [116] developed a two-stage speaker identification system using the fusion of MFCCs and GFCCs and various classification approaches and analyzed them under different environment noises. The fusion of FE techniques and classifier is evaluated vs SNR. The proposed fusion method demonstrated better recognition rates as compared to base classifiers and FE methods.

There are various vector-based FE methods that are widely used by researchers with ML and DL algorithms. Initially, the i-vector approach is employed in speaker recognition which is constructed by a feature extractor or frontend implemented using Gaussian mixture models (GMM) and universal background models (UBM) and backend is implemented with a probabilistic linear discriminant analysis (PLDA) classifier [117]. After the early success of i-vector-based systems, a number of hybrid methods combined i-vector and DL architectures [118], [119]. Subsequently, researchers have implemented speaker embedding systems based on DL, d-vectors [120], x-vectors [121], and t-vectors [122] after the success of i-vectors. In order to train deep d-vectors, frame-level speech information is used. The deep-vector architecture includes 300ms speech frames that contain 40 filterbanks. In addition, there are four dense layers (or fully connected layers) in the network, each containing 256 nodes. Contrastingly, the x-vector algorithm produces embedded speaker data based on variable-length speech input [121]. X-vector systems achieve a lower equal error rate (EER) than i-vector systems and d-vector systems. Inspired by the performance of FaceNet [123], various domains have implemented embeddings specific to facial images. The t-vector system also known as triplet network, is trained on a shared DNN triplet network and the triplet loss function is normally applied. T-vector systems do not perform better than x-vector systems, but these systems usually compete with one another [124].

7) PERFORMANCE OF ACOUSTIC FEATURE EXTRACTION TECHNIQUES

Traditional ML algorithms use almost all aforementioned features from time, frequency, and cepstral domains to solve various problems. Features need to be handpicked based on the performance of each model. DL algorithms work on unstructured audio representations. Sound features, such as spectrograms and MFCCs, are capable to extract patterns on their own. Furthermore, they are supported by a vast amount of data and computing power [125]. Therefore, widely used feature representations by DL algorithms that can be directly fed into neural network architectures are spectrograms, mel-spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs).

In order to solve a specific problem, researchers have used a variety of features. The best ones were selected by their performance and evaluation metrics. Researchers in [126] have compared the performance of MFCC, PLP, and LPC techniques for speaker recognition. Among these, PLP has performed better at low SNR values thus it is proven as a robust technique in the presence of noise. In [127], it has been validated that MFCCs are good at computing the distance between sounds. Moreover, the performance of a FE technique also depends on the nature of the sound. For example, speech and music sound share similarities in phones and notes, and harmonic structures in spectra. Unlike speech and music, environmental sounds have obscure periodicities and an indefinite dictionary of sounds which is why they are complex. Therefore, FE techniques that mimic human perception of sounds such as LPCCS [98], MFCCs [91], and GTCCs [99] are mainly used in speaker and music systems. Researchers have proven in one study [128] that the right combination of time and frequency-based techniques can outperform well-known techniques. In their paper, Robert et al. compared the performance of the combined approach of line spectral frequencies (LSF), ZCR and spectral ux (SFX) with MFCCs in a standard audio recognition system. F-scores indicated that the proposed approach achieved 97.5% and MFCC 78.9%. The Chromagram has demonstrated great potential in one study [129] of Thai classical music instruments and has proven a good representation of the complex internal structure of Thai music. One investigation [130] compared different acoustic FE techniques based on robustness to noise and spectro-temporal representation. According to the results, spectrogram, gammatone filterbank and Zweig impedance function-based linear transmission line generated good outcomes against noise at -5dB whereas, wavelet feature scored worst at $+2$dB. In terms of spectro-temporal representation, Zweig impedance function-based linear transmission line and wavelet feature performed better than the Mel spectrogram. Gammatone filterbank and spectrogram performed satisfactorily during the test phase.

D. AI METHODS

This subsection discusses some of the traditional and widely used ML and DL algorithms. After appropriate features are extracted successfully, these AI algorithms are trained on the informative features to learn about the acoustic signature generated by a particular sound source. Figure 4 represents the traditional machine learning and deep learning algorithms that are used in ASI in various scenarios.

1) MACHINE LEARNING ALGORITHMS

To date, few researchers have compared ML algorithms in research work [131], [132]. Some of the prominent ML algorithms are discussed as follows:

- K-Nearest Neighbour (KNN): K-nearest neighbour is an instance-based, non-parametric, and supervised ML algorithm. This is used to solve regression and classification-based problems mostly. An audio sample is assigned with a class label when most of the nearest neighbours belong to that class. This is known as majority voting which is the core concept of the KNN

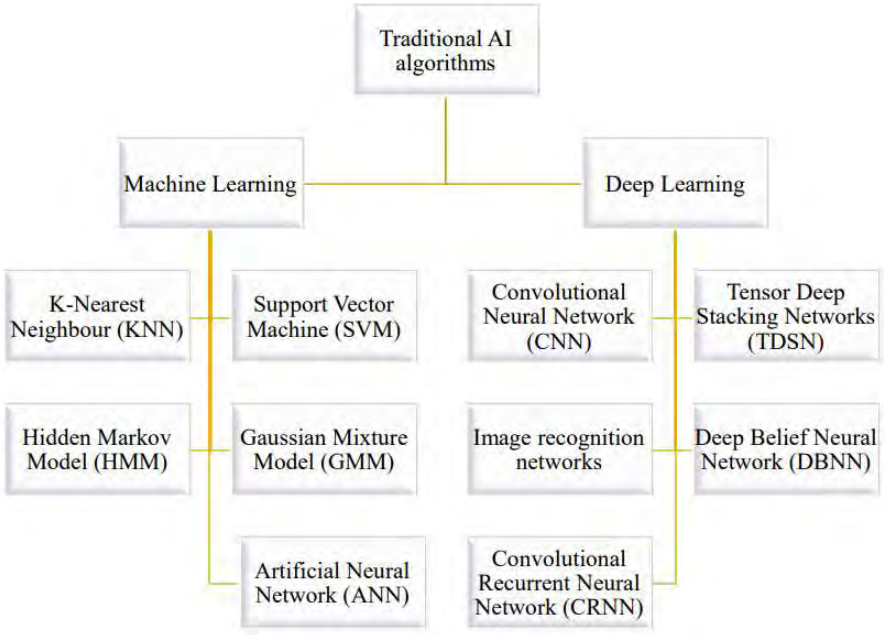

FIGURE 4. Traditional machine learning and deep learning methods.

algorithm. KNNs are used in identifying patterns in texts. [133], ESI [131] and finance studies [134].

- Support Vector Machines (SVM): SVM is another supervised learning classifier used for classification and regression analysis. SVM works with the use of various kernels based on the number of classes. There are different SVM kernels used for various problems such as linear, polynomial, Radial Basis Function (RBF), and gaussian. SVM uses a set of hyper-planes or decision boundaries in N-dimensional space that classifies different classes. The number of features is a decisive factor that determines the dimensions hyperplane. SVM has been proven a promising classifier for generalization problems. This can work well with small datasets. Some of the applications of SVM include text categorization [135], image classification [136] and environmental source detection [54].
- Hidden Markov Model (HMM): Hidden Markov Model (HMM) is a statistical classifier with its ability to consume less computational power as compared to other classifiers. HMM works on the principle of Markov chains. The Markov chain stays hidden in the process of observing events in different states of the Markov chain. In HMM, variables can be continuous or discrete. HMM learns the path of trajectory from an existing dataset containing classified trajectories. Therefore, for classification and recognition purposes, a flying object (hidden model) is classified knowing only its trajectory. HMM is employed in various fields such as speech recognition [137], gesture recognition [138] and target classification [139].
- Gaussian Mixture Model (GMM) is another probabilistic unsupervised learning model i.e. it does not need the prior information of the data points labeled with classes. GMM can approximate complex class density functions with random precision. Further, it can also be employed as a supervised classifier. However, its performance has shown to be lesser than the KNN and SVM in various applications [140].
- Artificial Neural Networks (ANN) are subsets of ML that work like biological neurons signaling each other. ANNs are made up of a set of artificial neurons also called nodes. These nodes form layers comprising an input layer and one or more hidden layers and an output layer. Every node has a weight and threshold value to communicate between layers. That's how they can transfer data between layers and the network gets trained on it and gains accuracy over time. ANN is a supervised algorithm and it learns by using examples. For instance, a network can identify a dog in an image when they are trained with manually labeled dog images using outputs obtained from other dog images. The learning rate in ANN varies over time. ANNs are massively used in almost every other field such as in healthcare [141], stocks and finance [142], 3D reconstruction [143] and environmental sound source identification [131].

2) DEEP LEARNING ALGORITHMS

This subsection outlines commonly used DL approaches which are as follows:

- Convolutional Neural Network (CNN): CNN is a type of DL also known as a feed-forward neural network. CNNs

are primarily used to detect, identify and classify objects based on given visual image data [144]. CNN comprises four types of layers, convolution layer, pooling layer, fully connected layer, and non-linear layer. In CNNs, nodes are capable to do weight sharing thus possessing an important property called shift-invariance. This is why CNNs have convolutional layers along with linear filter banks on input layers. They have applications in video recognition, recommender systems, object detection, image classification and natural language processing.

- Tensor Deep Stacking Network (TDSN) is a DL algorithm that learns from parallel hidden layers in each unit. TDSN is an extension of a deep stacking network that has sequential layers only in its modules. There is no change in the stacking operation of TDSN compared to the Deep stacking network(DSN). TDSN and DSN have the same computational complexity and scalability. In addition to this, TDSN offers training in hidden representations to encode speaker and environment information to include their factors [145]. Khamparia et al. performed sound classification using T-DSN and obtained an accuracy of 56.00% [146].
- Image recognition network: Image-based recognition networks are very deep CNNs specifically designed for image features. some of the image recognition networks include AlexNet, GoogLeNet, LeNET, and VGG16. AlexNet used a gradient descent optimization function with all the layers using a uniform learning rate of 0.001. AlexNet has eight layers whereas GoogleNet is deeper as it has 22 layers. GoogleNet is a promising deep CNN that can avoid the problem of overfitting due to many deep layers with the use of multiple-size filters at the same level of operation [147]. These image recognition networks are used on the ImageNet dataset for various applications.
- Deep Belief Neural Network (DBNN): DBNN is a traditional deep neural network (DNN) that faces problems like slow learning and works on big databases only. DBNN has multiple connected layers. When a network is trained unsupervised, it can construct its input layer depending on probabilities. Other layers can detect features thus they can further be trained under supervision for accurate predictions. DBNN has performed better than HMM and neural networks for event source detection [148].
- Convolutional Recurrent Neural Network (CRNN): CRNN combines convolutional neural network (CNN) and RNN and has presented better results in the audio processing domain.

E. EVALUATION METRICS

The evaluation metric is a useful criterion to study the quality of an AI algorithm. Evaluation of an AI algorithm is essential for any project. Many different performance metrics can be used to test a model. It is very important to include multiple evaluation metrics in a study to deliver a detailed performance report of an AI model. In this subsection, we have discussed some of the prominent evaluation metrics relevant to AI-based AI from the literature. The confusion matrix is the mostly used and prominent evaluation method for classification-based problems. A confusion matrix is a 2-dimensional *NxN* matrix that is used to summarize the classification results. In this matrix, one dimension represents a predicted class and the other dimension represents the corrected or true class in a given problem. In a binary classification problem, there are four important terms that are involved to describe each entry of the confusion matrix. True positive (TP) is the correct prediction of the positive class, True negative (TN) is the correct prediction of the negative class; false positive (FP) is the wrong prediction of the positive class, and false negative (FN) is the wrong prediction of the negative class.

The confusion matrix forms the basis for the other types of metrics. The classification accuracy rate is a usual metric based on calculating rates from subsets of these values. In simpler terms, the accuracy of the matrix can be evaluated by taking an average of the values lying across the main diagonal.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FN + FP)} \tag{8}$$

The precision score from the confusion matrix can be calculated by the ratio of true positives and total positives predicted. A low precision score of less than 0.5 depicts the outcome of a high number of false positives due to imbalanced class or untuned model hyperparameters.

$$Precision = \frac{TP}{(TP + FP)} \tag{9}$$

Recall or sensitivity is defined as the number of correct positive outcomes divided by the total number of positive instances identified by the classifier. The area under the precision-recall curve delivers an average precision score (APS). The mean of the average precision score is termed mAP and can be computed by taking the mean of APS over all classes. The F-1 score is another evaluation metric that measures the test's accuracy. This provides a harmonic mean of precision and recall scores for a classification task.

$$Recall = \frac{TP}{(TP + FN)} \tag{10}$$

$$F1 = \frac{2}{(\frac{1}{precision} + \frac{1}{recall})} \tag{11}$$

Area under the curve (AUC-ROC) is one of the frequently used metrics to analyze the performance of a classifier. AUC gives outcomes in binary classification problems. According to the definition of an AUC, it is the probability that a randomly chosen positive instance will rank higher than a randomly chosen negative instance by the classifier. There are three terms that explain the characteristics of AUC; True

	True class: Yes	**True class: No**
Predicted: Yes	True Positive (TP)	False Positive (FP)
Predicted: No	False Negative (FN)	True Negative (TN)

FIGURE 5. Confusion matrix.

positive rate (TPR), True negative rate (TNR) and False positive rate (FPR) are defined mathematically as:

$$TPR = \frac{TP}{(FN + TP)} \tag{12}$$

$$TNR = \frac{TN}{(TN + FP)} \tag{13}$$

$$FPR = \frac{FP}{(TN + FP)} \tag{14}$$

These AUC metrics are displayed on the receiver operating characteristic (ROC) graph. The AUC-ROC graph is drawn as FPR on the x-axis and TPR on the y-axis. The values of FPR and TPR range from 0 to 1. The greater the value of the AUC, the higher the performance of the model.

Mathews Correlation coefficient (MCC) [149] is a more reliable metric than the aforementioned traditional metrics. MCC is the measure of the difference between true and predicted classes that is analogous to x^2 statistics on a confusion matrix.

$$MCC = \frac{(TN.TP - FN.FP)}{\sqrt{(TP + FP).(TP + FN).(TN + FP).(TN + FN)}} \tag{15}$$

MCC achieves a high score when prediction in all categories (TP, FP, TN, and FP) is true with respect to the size of positives and negatives. MCC offers numerous advantages over the F1 score and accuracy in binary classification problems [150].

III. APPLICATIONS OF AI-BASED ASI

This section discusses the significance of AI-based ASI and its applicability to a variety of applications. As we mentioned before, this section provides a detailed discussion of AI-based ASI in fault diagnosis, underwater detection, ESR, healthcare, music-genre classification and wildlife monitoring. For all these applications, we compared previous works based on evaluation metrics, limitations and advantages. In addition, we also provide statistical analysis of proposed models from the literature review.

A. MACHINERY FAULT DETECTION

This subsection provides a detailed discussion of ASI methods that have been applied in industrial settings for intelligent fault diagnosis in machines and their components. In addition to the discussion of ASI methods for fault diagnosis, we present state-of-the-art artificial intelligence models for the identification and detection of faults. We have also summarized relevant studies in Table 3 to give a clear overview of their works. With the progress and development in production processes, science, and technology, machines, and equipment is getting advanced and automated. Modern machines are complex and their components are linked to each other. A slight fault can raise a chain of issues in a machine if not diagnosed timely. For instance, the crash of a US space shuttle occurred due to a slight problem in its component. Therefore, research is needed in the fields of urgent condition monitoring and intelligent fault diagnosis so that prompt maintenance activities can be done to ensure the smooth functioning of equipment. This shall increase the reliability and safety of the industrial environment and reduction of costs as well. Fault diagnosis has been conducted using vibration, thermal, current, and sound signals from the machinery in different components of machines such as motors, gearboxes, bearings, transformers, etc.

In 2013, Pandya et al. [151] discussed fault diagnosis in rolling element bearings in one of the earliest ASI-related papers. In their work, acoustic signals from bearings are captured and time-frequency features are derived using intrinsic mode functions. Then, supervised machine learning classifiers such as KNN and weighted KNN are used for the classification of faults and KNN has been chosen as the best classifier with 92.77% accuracy. Later on, Yoon and He [152] and Yao et al. [163] investigated the fault diagnosis in planetary gearbox using acoustic emissions and supervised

TABLE 3. Summary of ASI in industrial machinery fault detection.

Research publication	Machinery Type	Dataset	Feature extraction method	AI Technique	Performance evaluation
Pandya *et al.* [151]	Rolling element bearings	Self-collected	Statistical feature extraction from intrinsic mode function (IMD)	KNN, weighted KNN and APF-KNN	APF-KNN achieved highest Classification accuracy rate (CAR) of 96.66%.
Yoon *et al.* [152]	Planetary gearbox	Self-collected	Empirical mode decomposition	KNN, BP and LAMSTAR	LAMSTAR indicated less sensitivity with lowest diagnostic error rate of 0.5%.
Li *et al.* [153]	Gearbox	Self-collected	Wavelet packet transform	Deep random forest fusion (DRFF)	DRFF outperformed other data fusion methods with the highest CAR of 97.68%.
Waqar *et al.* [154]	Gear	Self-collected	Fast Fourier transform	Multilayer Perceptron Artificial neural network (MLP-ANN)	MLP-NN achieved 99.88% training and 94.24% detection rate.
Adam *et al.* [66]	Three phase induction motor	Self-collected	(SMOFS-32-MULTIEXPANDED-2-GROUPS and SMOFS-32-MULTIEXPANDED-1-GROUP	Nearest neighbour, BP-NN, modified words coding classifier	Recognition efficiency of proposed method resulted in the range of (88.19-100)%.
Yao *et al.* [155]	Gear	Self-collected	Time and frequency domain analysis	Convolutional neural networks (CNN)	Their model achieved an accuracy of 100% in the first two evaluation cases, 98.5% and 96.5% in other two cases respectively.
Islam *et al.* [156]	Bearings	Self-collected	Statistical features from time and frequency domain	K means clustering	Average sensitivity per class (ASPC) is 95.21% for dataset-1 (300RPM) and ASPC is 95.06% for dataset-2(350RPM)
Duong *et al.* [157]	Bearings	Self-collected	Continuous wavelet transform with envelope analysis	Convolutional neural networks (CNN)	Their proposed DCNN achieved 98.79% CAR higher than state-of-the-art.
Pham *et al.* [158]	Bearings	Self-collected	MEL spectrogram	Convolutional neural networks (CNN) with optimizers	Confusion matrix demonstrated average diagnostic accuracy (98.21,96.08,95.36,93.33)% for SNR (Noiseless,10,5,0) respectively
Potovcnik *et al.* [159]	Heating system valves	Self-collected	Statistical features from LDA and wrapper based feature selection, MFFCs	Discriminant Analysis, DT, Neural networks, SVM, Gaussian processes	The proposed method offered balanced classification rates (BCR) for valve cases i.e. 0.92, 0.96,0.95 and 1.00.
Yaman *et al.* [160]	Induction motor (Rotor and bearings	Self-collected	Discrete wavelet transform (DWT) Local binary Pattern (LBP)	SVM and K-nearest neighbourhood	Evaluated their FE methods with SVM and KNN and achieved higher than 99% diagnostic accuracy for a number of cases.
Brusa *et al.* [161]	Bearings	Self-collected	MEL spectrogram	YAMNet CNNs	Analyzed model with four datasets (A,B,C,D) and resulted in test accuracy 100%, 100%, 83.0% and 91.5% respectively with overfitting.
Cai *et al.* [162]	Substation transformer	Self-collected	Mel-frequency cepstral coefficients (MFCC)	Gaussian mixture models (GMMs)	Detecting operating point (DOP) at different SNRs is observed in spectrogram from 0dB to -100dB with 0 false alarms.
Yao *et al.* [163]	Planetary gearbox	Self-collected	Energy, Time and Envelope spectrum kurtosis (TESK)	BPNN, SVM, ELM and Random forests	Best fault classification accuracy 96.32% is achieved by RF among all other classifiers.
Santos *et al.* [164]	Induction mo-tor(bearings)	Self-collected	Time and frequency signature	MLP-NN & SVM	For unbalanced cases, MLP and SVM obtained 97% and 80% diagnostic accuracy equally for healthy and faulty motors.

TABLE 3. *(Continued.)* **Summary of ASI in industrial machinery fault detection.**

Research publication	Machinery Type	Dataset	Feature extraction method	AI Technique	Performance evaluation
Thoidis *et al.* [165]	Random mechanical machinery units	MIMII dataset and Self-collected	One class SVM+DOC feature learning	Deep temporal CNN	The proposed method is analyzed with AUC metric which showed improvement up to 7% and 5.3% under various SNR conditions for the pump and slider classes.
Orhan *et al.* [166]	UAV motors	Self-collected	Mel-frequency cepstral coefficients (MFCC)	Support Vector Machines (SVM)	Their method achieved 100%, 100%, 99.06% 90.53% CAR in a helicopter, duocopter, tricopter and quadcopter models respectively.
Liu *et al.* [167]	Rotor bearings	Self-collected	Fast Fourier transform (1/3 Octave method), Fourier synchrosqueezed transform	KNN, SVM decision trees	Average recognition efficiency (ARE) of three classifiers are analyzed in which SVM scored highest ARE 82.2% than KNN (76.86%) and DT (78.4%).
Fu *et al.* [168]	Power transformer	Self-collected	MEL spectrogram and filtered MFCCs	DNN (Dual-layer & dual channel)	Light FD indicated diagnostic precision 94.64% and recall score 96.33%.

learning algorithms. To accomplish this, Yoon and team set up the power gearbox (PGB) test rig experiment, created faults in gears artificially, and collected acoustic samples using acoustic sensors. In their study, KNN, back propagation (BP), and LAMSTAR learning algorithms are used to compare their performances based on their error rates. Subsequently, Yao utilized four classification models back propagation neural networks (BPNN), Extreme learning, random forests (RF), and SVM for fault classification and compared the results.

Right after Yoon's work, Waqar and Demetgul [154] did another experiment with worm gear in motors. Both vibration and sound signatures from the motors at different speeds are acquired. The collected signatures were preprocessed and classified using Multilayer Perceptron Artificial neural network (MLP-ANN). The trained algorithm achieved successful prediction of 2 different speeds and 4 different oil levels. Adam et al. [66] used real acoustic data of four states of faulty three-phase induction motors, extracted two types of features, and used nearest neighbour, BPNN, and a modified classifier based on words coding for recognition.

The authors in [156] proposed a fault diagnosis model that can identify new fault modes through K-means unsupervised clustering and store the real-time data for future fault diagnosis. An experiment has been conducted to validate their study in which they recorded acoustic signals from bearings at different speeds to create a database. Further, KNN classifier has been used to estimate the prediction performance.

In recent literature of the last three years, Potovcnik et al. [159] performed the classification based on the condition of the system valve using acoustic features and various ML algorithms. To accomplish this, an experimental setup of the valve assembly with a microphone is established in a semi-anechoic chamber. The proposed methodology also involves the comparative analysis of feature selection using different classification models. Later, Yaman [160] and Santos et al. [164] investigated the faults in the bearings of three-phase induction motors. Audio data is collected by setting up experiments and classifying the data using supervised learning algorithms such as SVM, KNN, and MLP respectively. Later in 2022, Orhan et al. [166] continued and proposed a lightweight method for the detection of faults in unmanned aerial vehicle (UAV) motors. Audio datasets are collected from healthy and faulty motors of various (UAV) sources. SVM classifier is used for fault diagnosis in UAV motors. Moreover, Cai et al. [162] and Fu et al. [168] did their research on anomaly detection in transformers using acoustics. Fu and team developed a method namely lightFD to perform SVM classification on edge devices with limited computing power. Most recently, Liu et al. [167] also studied rotor-bearing fault analysis and use sound data acquired from faulty rotating machinery and classified by SVM, KNN, and decision trees.

B. UNDERWATER APPLICATIONS

ASI in the underwater medium has a wide range of applications. For example, ASI capability plays a vital role in the military to identify friendly/adversarial objects (e.g., submarines, torpedoes) in water. ASI is also beneficial for scientists studying marine ecology, geology, oceanography, and seismology. Moreover, the ability to localize objects and analyze transients in ocean acoustics can be utilized by the mining industry for offshore oil and gas discovery and plant maintenance.

Traditionally, ASI has been performed successfully in the underwater medium using matched-field processing (MFP) [183]. However, one of the severe limitations of MFP is its sensitivity to the mismatch between model-generated datasets and real-world conditions [169], [183]. In other words, MFP may not be flexible enough to adapt its parameters to changing channel conditions (e.g., sound speed profiles, bathymetry, and chemical composition of water) which is an inherent and peculiar nature of the underwater

TABLE 4. Comparison of AI-based ASI in underwater applications.

Category	Technique	Year	Advantages	Limitations
FFNN	FFNN, SVM, and RF with supervised ML [169].	2017	Three machine learning algorithms (FFNN, SVM, and RF) are used to solve the range estimation problem both as a classification problem and as a regression problem.	The amount of training data available affects the performance of classifiers.
	FFNN and support vector machine (SVM) [170].	2017	MFP fails at about 4 km without accurate environmental information, while classifiers do well up to the 10 km range with limited environmental information.	The effect of large environmental variations which could include fluctuations in the sound speed profile was not investigated.
	Beamforming-based DoA estimation with FFNN [171].	2022	Accurate with little precise environmental information.	DoA estimation is sensitive to target distances resulting in a degraded performance with long distances.
MLP	MLP-NN with Grey Wolf Optimization (GWO) [172].	2016	Offers better convergence speed, lower possibility of trapping in a local minimum and improved classification accuracy compared to PSO, GSA and hybrid PSOGSA using three different sets of data.	Does not deal with complex MLP NNs.
	MLP-NN with biogeography-based optimization (BBO) [173].	2016	Performs better compared to other algorithms in terms of avoiding local minima, classification accuracy, and convergence speed.	BBO performance is sensitive to the migration model, especially for high-dimensional problems.
	MLP-NN with DA [174].	2019	Precise classification, greater convergence speed and better local optimum avoidance as compared to BBO, GWO, ALO, ACO, GSA and MVO algorithms.	Classification accuracy is not that impressive against that of BBO for the Gorman and Sejnowski dataset.
	MLP-NN with improved Whale Optimization Algorithm (WOA) [175].	2019	Precise classification, greater convergence speed and better local optimum avoidance as compared to PSO, GSA, ACO, GWO, and WOA.	-
	MLP-NN and adaptive MFCC with SSO [176].	2019	In terms of classification accuracy, local minima entrapment, and convergence speed, the proposed classifier outperforms GWO, BBO, interior search algorithm (ISA), and group method of data handling (GMDH).	-
	MLP-NN and Local Wavelet Acoustic Pattern (LWAP) [177].	2021	Proposed method outperforms other benchmark classifiers on classification accuracy, convergence speed, and entrapment in local minima.	-
DNN	DNN with feature extraction CNN–FFNN [178].	2018	Suitable for locating sources in shallow water environments that are complex and varied.	May not be applicable for deep water environments
	Auditory perception-inspired Deep CNN (ADCNN) [179].	2019	New method based on human auditory neural systems.	Classification accuracy is not that great compared to the methods used for benchmarking.
	DNN in shallow water environment with high frequency signals [180].	2019	There is no need for a forward acoustic propagating model.	Results may not be generalizable to the real world since the experiment was conducted in a small aquatic laboratory tank.
	DTL. [181].	2019	Transfer learning can be used to train DNNs where real-world deep-sea trial data are not readily available.	-
	SVM and CNN [182].	2021	Uses five dimensions to train the model.	Since NL and PSD are the most significant characterizers of underwater noise, could simplify to only two dimensions

medium. In order to combat this challenge, many works in the literature have resorted to data-driven ML techniques which can learn from and adjust themselves to rapidly varying channel conditions, yielding better results such as improved accuracy for ASI.

In this work, we have conducted a comprehensive literature review on ML techniques for ASI in the underwater medium. Table 4 compares the AI-based ASI in underwater applications based on their advantages and limitations. Additionally, based on the type of AI technique, we have organized the works into the following main categories:

- Feed-Forward Neural Networks (FFNN)
- Multi-Layer Perceptrons (MLP)
- Deep Neural Networks (DNN)

1) FEED-FORWARD NEURAL NETWORKS (FFNN)

The authors in [169] have formulated the ASI problem as a ML problem where the ML model learns directly from observed data. In this work, the authors utilize a vertical linear array for building a normalized covariance matrix which is used as a training dataset. Three ML techniques - FFNN, SVM and ensemble learning-based random forest (RF) are evaluated against the traditional

MFP technique. Results indicate that ML algorithms yield better results when ASI is posed as a classification problem rather than a regression problem. Moreover, FNN yields better predictive performance at multi-frequency inputs with SNRs above 0 dB despite a small number of training samples.

In their follow-up work in [170], the same authors show that ML-based classifiers deliver better results in estimating ship range for up to 10 km when MFP fails at approximately 4 km range when environmental information is limited.

More recently, the work in [171] has proposed a two-stage process of underwater target detection where the first stage involves beamforming-based direction of arrival (DoA) estimation and the second stage involves taking the DoA information and feeding it into an FFNN which develops a detection model. The proposed method can yield a detection accuracy of as high as 97.32% at particular locations of the ocean.

2) MULTI-LAYER PERCEPTRONS (MLP)

Traditional recursive algorithms such as gradient descent used in neural networks (NN) encounter multiple issues such as low accuracy, slow convergence, and local minima entrapment. This has led researchers to use heuristics/meta-heuristics-based algorithms for training NNs. The work in [172] has used grey wolf optimization (GWO) for training NN for target classification. Results show that compared to the Particle Swarm Optimization (PSO) algorithm, Gravitational Search Algorithm (GSA), and the hybrid algorithm (i.e. PSOGSA), the multi-layer perceptron (MLP NN) using GWO yields better results across all three datasets used (i.e., Iris, Lenses, and Sonar-1988) in terms of higher accuracy (>95% for Sonar), lower probability of local minima entrapment and higher convergence rate.

The authors in [173] have used another meta-heuristic-based algorithm (i.e., biogeography-based optimization (BBO)) for classification with NNs for the same three datasets as [172] (except Sonar-2015 has been used in [173]). Non-linear migration models offer two-pronged advantages: one to search agents for better exploration of the solution space resulting in local optima avoidance; and two to accelerate the search agents towards global optimum enhancing convergence rate without sacrificing accuracy of classification.

More recently, the authors in [174] have used another meta-heuristic Dragonfly Algorithm (DA) on active, passive, and Gorman and Sejnowski sonar datasets and compared its performance against BBO, GWO, Ant Lion Optimization (ALO), ACO, GSA and Multi-verse Optimization (MVO) algorithms where DA outperforms the rest in terms of accuracy and convergence speed.

The work in [175] presents a method for classifying targets in passive sonar using MLP trained by a salp swarm algorithm (SSA). The authors have also used MFCC to improve the dataset's dimensions. The proposed method utilizes SSA to optimize the weights and biases of the MLP, which are then used to classify sonar signals. SSA allows for faster and more efficient training of the MLP even with the presence of noise. The limitation of the proposed method is that it is based on passive sonar, which may be limited in its ability to detect targets that are quiet or have low echo strength. Additionally, the method relies on the SSA, which may be sensitive to initial conditions and may not be able to find the global optimal solutions.

More recently, the Whale Optimization Algorithm (WOA) and an improved WOA with Local Wavelet Acoustic Pattern (LWAP) have been utilized in [176] and [177], respectively, for training MLP NNs. These works indicate that meta-heuristics-based MLP NN training for sonar target classification is a promising technique.

3) DEEP NEURAL NETWORKS (DNN)

Besides MLP NNs, researchers have also used DNNs for ASI. The work in [178] uses DNN for ranging and depth determination of acoustic source in shallow water (100 m) using DNNs. They propose two methods: a) a two-stage FE and model building process b) a direct process of training a convolutional neural network (CNN)–FNN (CNN-FNN) architecture by using raw acoustic data. Both methods provide better performance compared to MFPs under mismatched environments.

Inspired by the human auditory system of sound perception, the authors in [179] propose a deep CNN for underwater target recognition. The architecture maps various frequency components into a bank of multi-scale deep filter sub-networks. Then it mimics the neuro-plasticity mechanism of the human brain to train those multi-scale deep filter sub-networks using raw time-domain ship noise. The proposed method, when trained with raw time domain waveforms, achieves better classification accuracy as compared with standard CNN/DNN methods with other types of training input such as MFCCs.

The work in [180] has proposed a DNN-based source localization technique for very shallow water environments (a 1.1 × 1.4 m laboratory tank with depth 0.1 m) with high-frequency components, while the work in [181] have devised a deep transfer learning technique which can be adapted by models such as that proposed in [180] to translate its capabilities for real-world deep-sea environments. The transfer learning approach in [181] can open new possibilities for effectively training DNNs since real-world deep-sea trial data are difficult to obtain.

More recently, the work in [182] has proposed a method where learning features are extracted from five different dimensions, i.e., noise spectrum level (NL), time–frequency spectrum (Spec), power spectral density (PSD), Mel-frequency cepstral coefficient, and Mel filter bank energy (FBANK). Then the authors compared the performance of SVM and CNN on noise-added data with various SNR levels. They have found that underwater noise can be best

characterized by NL and PSD features. Additionally, CNN outperforms SVM in noise classification.

C. EVENT AND ENVIRONMENTAL SOUND SOURCE DETECTION

In this subsection, we provide a detailed discussion of the significance of acoustic source identification (ASI) in environmental sound source recognition and event detection. First, we discuss the applications and benefits of environmental sound source identification. Then, we discuss in detail most of the FE and AI methods available in the literature. Later, we present a summary of the event and environmental source identification works and highlighted their limitations in Table 5.

ASI can be used to recognize events and scenes. Context recognition is becoming popular and becoming an important research area. Acoustic scene identification is defined as the recognition and classification of acoustic scenes such as schools, offices, hospitals, and trains/buses based on the generated sounds [200]. The aim is to create applications that can improve urban environments if the activities in the surroundings are detected and identified. Environment sound source identification is a complex activity as compared to speech and music because environment sounds are very dynamic in nature and sometimes it is difficult to identify targeted sounds due to background noise. Some of the sounds can have a low signal-to-noise ratio that can make it difficult for some sources to be recognized properly [201].

ESR is a promising method for audio surveillance applications [202]. In addition, ESR can be used in robotics to improve their navigating abilities and interactions with the environments [203], [204]. ESI along with video analysis contributes its benefits majorly in surveillance applications of homes and cities [205], [206]. Home surveillance is very important, especially for elderly people who are living alone or other smart home applications [207], [208]. ESI has been used to recognize animal species [209], [210], [211], bird species [212], [213] by their distinct acoustic signatures for bioacoustic applications and wildlife monitoring. Recently in a study [214], hive health has been monitored by analyzing hive sounds using AI algorithms.

In our study, we have reviewed the literature extensively to compile the renowned work which has been done in the field of ESI and sound event detection (SED). The authors in [19] did the survey and compiled the works which have been investigated in environment audio scene and sound event recognition for surveillance purposes. Recently, a detailed review [26] has been published in the area of ESI. This research work highlighted the environmental and events sound datasets, FE methods, and different ML and DL algorithms used in recent studies. In the past five years of literature, the prominent works accomplished by authors in urban sound event detection and environment sound source identification are as follows.

Authors in [184] have developed an efficient urban sound classification mechanism using ML. Local and global FE techniques are employed to process the most discriminant information-carrying features for the ML algorithm. A mixture of expert model techniques is also introduced to assemble information from local and global features. Zhu et al. [187] performed multi-scale FE on audio data. In multi-scale convolution, the signal waveform is convolved with filters at different scales and performed feature fusion. CNNs are trained after the pooling of useful features. By employing these setups, improvements in sound recognition are achieved that yield better results than previous methods. The authors in [188] improved sound recognition in hearing aids by using ensemble techniques. Moreover, automation in devices is introduced with respect to sensing and recognizing sounds and their sources. Similarly, Ahmed et al. published his work [192] to establish an automatic environmental sound recognition system using deep learning. Image-based features are used to train the DL algorithm. Moreover, the performance of various FE methods is compared and achieved different recognition accuracy rates using publicly available data sets.

Lately, in 2021, many researchers actively worked in this area and published their works. The authors in [194] recorded ambient sounds in an indoor environment for the purpose of recognizing an activity based on the produced sound. Spectral information has been extracted from the data collected by using smart IoT sensors. CNN-DL model has been used by the research team along with fuzzy logic to accomplish a coherent recognition of activities. Nanni et al. [195] presented their idea of combining ensembles of classifiers exploiting six data augmentation schemes for the training of CNNs. Further, those ensembles are tested on open-sourced environmental sound datasets. The performance of ensembles are compared extensively with the ones mentioned in the literature resulting in a high-performing ensemble with high accuracy. Zinemanas and team [196] proposed a novel interpretable architecture employing a DNN for ESI. Audio domain knowledge is used to improve the distinction in classes. The key idea was to incorporate frequency-dependent similarity by assigning different weights to each frequency bin in the latent space. Due to the system's interpretability, it can be evaluated and debugged easily. Moreover, the authors in [197] proposed an intelligent forest monitoring system that applies signal processing techniques such as dynamic time warping and ML algorithms trained by MFCCs and spectral feature spaces.

D. HEALTHCARE APPLICATIONS

This sub-section outlines the usefulness of AI-based ASI in the healthcare field and research that has been done in the past few years. AI-based ASI has been widely used in the healthcare field in fall detection, health, and fitness wearable devices, equipment for hearing-impaired patients, and cardiac auscultation.

TABLE 5. Summary of event or environmental source identification using artificial intelligence techniques.

Research Work	Dataset	Feature Extraction Method	AI Technique	Performance evaluation	Limitations
Ye *et al.* [184]	UrbanSound8k	Mel spectrogram, spectrogram entropy	Class conditional dictionary learning	Proposed framework achieved 77.36% mAP score for classification which is higher than other DCNNs.	Ineffective for audio data with low SNR. Not robust to noise.
Boddapati *et al.* [185]	ESC-50, ESC-10 and Urban-Sound8k	Images of audios in CRP, MFCCs, spectrogram form	CNN AlexNet and GoogLeNet	Results indicated CAR on ESC-50, ESC-10 and UrbanSound8K at 73%, 91%, and 93% respectively.	Limited to image-based features only.
Salamon *et al.* [186]	UrbanSound8k	Log mel spectrogram with data augmentation	CNN	Proposed SB-CNN with augmentation achieved mean accuracy 0.79 higher than SKM and PiczakCNN.	Data augmentation is not class conditional.
Zhu *et al.* [187]	ESC-10 and ESC-50	multi-scale time-domain convolution	CNN(WaveMsNet)	Proposed CNN obtained accuracy of 93.75% and 79.10% on ESC-10 and ESC-50 respectively.	Works only with raw waveform as input.
Shah *et al.* [188]	UrbanSound8k	MFCCs, LPCs and non-spectral	Ensemble methods with base learners	Results indicated mean recognition rate of AdaboostM1-RF is high i.e. 88.25% and less standard deviation 0.31.	Computationally inefficient Cannot work on low processors.
Khamparia *et al.* [146]	ESC-10 and ESC-50	Spectrogram images	CNN and TDSNN	Compared to existing methods, CNN (ESC-10) & ESC-50 and TDSN (ESC-10) achieved 77%, 49% and 56% CAR respectively.	System is limited to compressed images only. It doesn't work well with HD images.
Sang *et al.* [189]	UrbanSound8k	Raw waveform	CRNN	Architecture outperformed the deep networks by 7.38% accuracy with fewer parameters.	Works only with the raw waveform as input.
Chi *et al.* [190]	UrbanSound8k and ESC-50	LMS and LGS	CNN	Their method performed better than Piczak's method on ESC-50 and UrbanSound8k by 18.9% and 7.6% and achieved 83.8% and 80.3% accuracy respectively.	Proposed algorithm cannot perform well with low depths.
Mendoza *et al.* [191]	UrbanSound8k	Constant Q transform(CQT)	Parallel and sequential CNN	Results demonstrated that the parallel CNN has gained 83.79% accuracy which is highest compared to sequential CNN, end-to-end CNN and deep CNN.	Not energy efficient due to continuous use of sensors.
Ahmed *et al.* [192]	ESC-50, ESC-10 and Urban-Sound8k	Log Mel spectrogram	CNN	Proposed optimized Stacked CNN achieved highest accuracy of 92.9% with UrbanSound8K dataset than ESC-10 (91.7%) and ESC-50(65.8%).	Computational cost is high. Ineffective for audio samples with varying frequency and SNR.
Demir *et al.* [193]	DCASE-2017 ASC and Ur-banSound8K	Spectrogram	Deep CNN	Their proposed model achieved an accuracy rate of 96.23% on DCASE2017-ASC dataset and 86.70% with UrbanSound8K and	Computational cost is high.
Polo *et al.* [194]	Self-collected	Log Mel spectrogram MFCCs	CNN	They compared the performance CNN+MFCCs and CNN-LM models with audio and ad-hoc datasets and demonstrated that CNN-MFFC (ad-hoc dataset) obtained highest accuracy 99%, precision 99%, recall 99% and F-1 score 99%.	Not robust to noise. Complex processing is required for large datasets.

TABLE 5. *(Continued.)* **Summary of event or environmental source identification using artificial intelligence techniques.**

Research Work	Dataset	Feature Extraction Method	AI Technique	Performance evaluation	Limitations
Nanni *et al.* [195]	ESC-50	DGT, MEL, GA, CO	CNN (AlexNet, GoogleNet, VGGNet, ResNet and InceptionV3	They compared best-performing ensembles and obtained competitive performance to state-of-the-art. They achieved 97%, 90.51% and 88.65% recognition accuracy on the bird dataset, cat dataset and ESC- 50 respectively.	Proposed system is based on DL approaches only but not on texture features. Requires large computational power.
Zinemanas *et al.* [196]	UrbanSound8K Medley-solos-DB google speech commands	LMS in latent space	APNet	Their study compared and analyzed the performance of APNet and refined APNet models. The accuracy of APNet is the lowest i.e. 65.8%. Among refined models, APNet (R.channels) achieved the highest accuracy 69.1% with least parameters 1.5M.	There is a trade-off between the interpretability and accuracy of the proposed system. This system can't perform SED and audio-tagging.
Segarceanu *et al.* [197]	Self-collected	MFCCs Fourier power spectral coefficients	GMM, LSTM,DTW, FFNN	Among all classifiers, FFNNs outperformed with an average identification rate of over 78% with about 10% higher performance than other competitive classifiers.	Adopted empirical approach only to choose the length of the analysis window. Proposed NNs cannot work with several classes of data jointly.
Mushtaq *et al.* [198]	ESC-50, ESC-10 and Urban-Sound8k	Mel spectrogram	CNN with transfer learning	Their proposed ResNet-152 and DesNet-161 systems performed better with data augmentation (NAA). The achieved results indicated ResNet-152 accuracy of 99.04% (ESC-10) and 99.49%. DesNet-161 achieved 97.57% on ESC-10 dataset.	Computational costs are high as it has so many layers.
Zhang *et al.* [199]	ESC-50 and ESC-10	Mel spectrogram	Dilated CRNN	The proposed D-CNN based ESC outperformed state-of-the-art ESC results with an absolute error rate less than 10% compared to previous methods. Their method gained accuracy of 81.9% (UrbanSoun8K), and 68.1% (ESC-50) and 87.1% (CICESE).	The proposed model is not robust to noise.

Elderly people have the tendency to fall down in their later ages and sometimes they do not have access to any external help. The fall can be serious and can lead to severe injuries that may take longer to heal in old age. In this case, early first aid is very crucial to reduce the risk of death [215], [216]. Therefore, these incidents can considerably be avoided using modern AI techniques along with traditional methods. In past, various methods for fall detection have been proposed such as by using cameras [217], sensors [218], [219] and radars [220], [221]. Health support wearable devices for hearing impaired people provides a promising solution for them to interact with their surroundings [222], [223]. Nowadays, heart auscultation has been used massively in conjunction with AI techniques for the diagnosis of cardiovascular diseases [224] and condition monitoring of arteries and valves [225].

The authors in [103] proposed a fall detection framework that is based on signal processing methods such as silent zone suppression and acoustic ternary pattern FE. SVM has been used to classify and detect fall events. Their proposed method works well in a multi-class environment. Yauganouglu et al. [226] developed a real-time detection system for hearing-impaired people using a wearable device in which sound events' information is conveyed to the user through vibrations. A combination of pre-processing methods is used for FE. Correct perception and recognition of sound have been made delivered using KNNs classifier and audio fingerprinting. Later, Ramadhan and team [227] published their work in which acoustic event recognition is investigated as part of a smart home system for elderly people. In their work, spectrograms are extracted from practically collected

audio data to train a DL model i.e. CNN. During their investigation, an accuracy rate of 97.5% in silence and 85% in normal scenarios are achieved respectively.

Recently, Jain et al. [12] developed an interactive tool namely ProtoSound for the hearing impaired or people with hard-of-hearing problems. The system has the ability to personalize a sound recognition model from user recordings. User recordings undergo the same set of steps of FE and classification performed by the chosen model. CNN architecture is being used in the ProtoSound system for the prediction of sounds. Their proposed ProtoSound achieved an average accuracy of 88.9%. Authors in [228] also presented a review of sound recognizer tools for hearing-impaired individuals. In their review, a user-driven automated sound recognition system is studied using ML techniques. The potential use of personalizable sound recognition systems is also highlighted for prospective research.

Authors in [229] used curve fitting and a KNN algorithm. The normal and abnormal heart sounds have been classified with 92% accuracy. Nouman et al. [230] developed a framework for automatic heart sound detection by using neural networks. An optimal combination of 1D-CNN and 2D-CNN is employed which exhibited an accuracy of 89.22%. Furthermore, Zhang and team [231] proposed a novel method based on temporal quasi-periodic features and LSTM algorithm. The popular 2016 PhysioNet dataset is used and an accuracy of 94.66% is achieved.

In addition, the authors in [232] performed their research in determining the heart condition based on its sounds and developed an AI-enabled tool for automatic quality assessment. Two datasets (2016 PhysioNet/CinC Challenge and self-collected) are used to compute necessary de-noised features and trained MLP classifier to perform binary classification of heart sounds. Then in the same year Zhiming et al. [233] proposed a heart sound recognition method to identify congenital heart disease in patients. Two classifiers, SVM and BP are used to train MFCC features extracted from heart sounds obtained from the 2016 Heart sound Challenge dataset. Their work improved the accuracy of detection of the congenital disease up to 93.52%. Kui et al. [234] proposed a promising approach in the classification of heart sounds using the duration-dependent hidden Markov model (DHMM) in the segmentation of heart sounds. Additionally, dynamic frame length is used to extract MFCCs from heart audio. Then, the extracted features are classified using DL i.e. (CNN). A majority voting optimization algorithm is used to optimize the classification results. They achieved 93.89% and 86.25& accuracy for binary class and multi-class respectively.

Another research performed by Bilal and team [235] classified heart sounds using 1D-CNN. He proposed a classification model employing Local Binary Pattern (LBP) and Local Ternary Pattern (LTP) features. Using PASCAL and PhysioNet 2016 datasets, he scored 91.66% and 91.78% classification accuracy respectively. Recently, the authors in [236] published their work investigating heart sound classification aimed at the diagnosis of disease due to heart failure. Two heart sound data sets (PhysioNet and PASCAL) are used and then preprocessed to generate MFCC features. Principal Component analysis and linear discriminant analysis has been used for feature selection and dimensionality reduction. Eventually, SVM, gradient boosting algorithm(GBA), and random forests classifiers are trained on those features to perform the classification task.

E. OTHER APPLICATIONS

This subsection discusses some of the recent literature reviews of the research work in various other applications such as music-genre classification, wildlife monitoring and forensic applications.

1) MUSIC-GENRE CLASSIFICATION

Moreover, inspired by the advancements in natural language processing (NLP), Zhuang et al. [237] designed a transformer classifier for music-genre classification. They used the famous GTZAN dataset and the transformer model is fed with mel-spectrograms as features and achieved an accuracy of 76.0%. Later, Mounika et al. [238] applied CNN and CRNN to classify music into various genres. The proposed classification is performed on GTZAN dataset and generated mel-spectrograms as distinguishing features. Their proposed model indicated a classification accuracy of 73.2% with train accuracy being 12% lower than validation accuracy due to the overfitting problem in the model. The authors in [239] developed a transformer model-based music recognizer in which they used MFCCs to recognize the genre of audio. The performance of their proposed model is analyzed on GTZAN original dataset and data augmented set which resulted in a better accuracy rate of 75.1%.

Furthermore, Shah and team [240] classified music into different genres using various time and frequency domain features. They extracted spectral centroid, onset strength, ZCR, tempo, spectral contrast, spectral bandwidth, roll-off contrast, and flatness to train SVM, random forests and gradient-boosting ML algorithms. In addition, Spectrograms are extracted to train DCNNs. Their classification performance is compared which proves CNN outperforms ML algorithms. Lately, Cheng et al. [13] performed their research to understand the music-genre classification problem using visual mel spectrogram with YOLOv4 neural network which is based on CNN. Their model is evaluated on various metrics such as precision, recall, F1-score, mAP, and confusion matrix. The average mAP results indicated 97.93% accuracy on the test set and 91.49% on the training set. They achieved better accuracies on GTZAN dataset however, the used graphical spectrum feature increases hardware cost. Most recently, the authors in [241] introduced a hybrid approach of CNN, multimodal and transfer learning based model. In this approach, GTZAN and Ballroom dataset has been used for analysis and benchmarking. Wavelet features are

extracted and mel-spectrograms as visual representations in CNN. Results have demonstrated that their proposed hybrid model scored 81% on GTZAN and 81% on ballroom datasets. Added to this, the computational performance of their model is also analysed in their study on a laptop and a supercomputer with a supercomputer having much lower computational time.

2) WILDLIFE MONITORING AND BIO-ACOUSTICS

Furthermore, in wildlife monitoring and bio-acoustics, several studies have been performed to study animal behavior and classification problems based on their acoustic features. The behavior of domestic cats was studied by Pandeya et al. [242] by developing an automated classification system over cats' generated sounds. The cat sounds dataset has been increased with the help of a data augmentation technique and extracted mel spectrograms as features for classification. Transfer learning of CNN and convolutional deep belief network (CDBN) has been carried out due to the close relation of cat sound and music. This resulted in an overall good performance in classification accuracy and receiver operating characteristic (ROC) metrics. Later on, the same authors in [243] investigated cow sounds as SED technique and developed an autonomous monitoring system. Using a data-driven approach, mel spectrogram is selected as a potential feature for video object description models (VODMs) and this approach is compared with conventional CNN. The proposed approach achieved better quantitative and qualitative scores. In addition to this, Li et al. [244] proposed an automatic sound recognition system in dairy cows that classifies the ingestive behavior in them. A publicly available jaw movements dataset of two forage species is used. Time, frequency domain, and MFCCS features are formed and a statistical model is developed. Then, three DL models CNN(Conv1D), (Conv2D), and LSTM are trained and optimized to classify the ingestive behaviors. The resultant performance under different forage species and heights came out to be 0.93, and the difference between the best and poorest obtained was 0.4-0.5.

The authors in [43] did their research on insect sound recognition. An ARS center dataset is used which comprises sound files of various activities of insects such as moving, feeding, and calling. CNNs are trained with feature maps created with MFCCs and obtained 92.56% recognition rate. Sun et al. [245] developed a reliable rainforest monitoring system using data augmentation and CNN-based transfer learning due to the scarcity of datasets. This system enables the detection and classification of various animal species (birds, amphibians, invertebrates, mammals). Their model achieved an average accuracy of $\geq$ 90% with mel-spectrogram features. However, their model included limited sonotypes of rainforest only. Moreover, Echinski and team [246] also investigated birds species using the sounds of birds and established a recognizer model. The spectrograms of bird sounds are fed into Resnet34 CNN for training. Performance metrics indicated a macro average F1 score of 0.74. However, their system could not recognize new entries in the test dataset which needs to be addressed. Recently, Jiang et al. [247] solved the classification problem as SED of ape calls using LSTM neural network. Three types of input features are used i.e. raw waveform, spectrogram and wave2vec 2.0 for the training of NNs. In their study, the results demonstrated that wave2vec 2.0 outperformed the raw waveform of than spectrogram in the classifier.

3) SPEAKER IDENTIFICATION (SID) SYSTEM FOR FORENSICS

AI-based ASI is not limited to the above-mentioned applications. Over time, it has gained much significant attention in speaker identification and verification system for forensics and surveillance applications. Authors in [248] proposed a novel model to detect disguised voices for forensic identification systems. GMM supervector obtained from Gaussian distribution of the speaker's voice and extracted MFCCs are used as features to train SVM classifier. The proposed model achieved good identification rates and lower error rate than 7%. Later, authors in [249] presented another method based on the evaluation of speech quality data. Three experiments are performed on SRE dataset to assess the impact of quality data on forensic speaker recognition (FSR). GMM universal background model (UBM) is trained on MFCCs and delta MFCCs-based vectors. The results indicated their proposed model obtained an Equal Error Rate (EER) of 0.6% as compared to state-of-the-art performance on TIMIT dataset. Rozario et al. [250] implemented a speaker recognition system using ANNs. The performance of Relative Spectral Amplitude (RASTA) PLP, MFCCs and Power Normalized Cepstral Coefficient (PNCC) features are compared on TIMIT database. The results demonstrated that MFCCs outperformed PNCC and RASTA-PLP in speaker identification with the highest accuracy score of 90.66% on the full speech segment.

Subsequently, authors in [251] proposed a DL-based speaker identification mechanism using improved shuffled MFCC (SHMFCC). The data augmentation approach is used in conjunction with the extraction of shuffled MFCC features. Three different datasets; LibriSpeech, TSP and VoxCeleb1 are used to conduct experiments for the study. The tuned Feed forward DNN was trained and tested under various noisy conditions. The proposed method demonstrated high accuracy in all noisy scenarios. Later, Bakir et al. [252] presented a forensic voice application. Data sets comprising recordings from 1000 people have been gathered and MFCCs features are extracted. The identification rates of CNN and DBN trained on these features are compared. CNN performed better than DBN on all MFCC vector lengths. Recently, Authors in [253] performed speech enhancement firstly by employing spectral and log Minimum Mean Square Error (MMSE) techniques. Then, the task of speaker identification on the Australian Forensic Voice comparison database is carried out by training GMM on MFCC features. The average scores of log MMSE are observed much higher than of

spectral subtraction. The trained model achieved a 63.5% accuracy score for speech signals enhanced by using the log-MMSE technique. Babu et al. [254] presented a short review of the forensic speaker identification (FSID) system. The authors highlighted the physical properties of speech signals. Several FE techniques and AI algorithms for FSID are also discussed in the paper.

- *Computational complexity of SID systems:* Speaker recognition is an important technology when it comes to forensics, access control systems and the financial sector. AI-based approaches introduce a new direction to this technology in terms of recognition accuracy, computational complexity and identification rates. Over the past decades, there has been plenty of research being done to solve problems in speaker recognition and speaker verification systems. Inspired by the remarkable performance of DL algorithms in SID systems, researchers have applied DL algorithms in speaker recognition [255], [256] and delivered high accuracy. However, advanced DL algorithms are computationally intensive; therefore, restricting its implementation on hardware platforms.

  ML algorithms implemented in SID systems such as GMM and SVM are less computationally complex as compared to DL algorithms. The execution time of GMM trained on MFCCs is 0.8ms at 48MHz frequency for a speech set of length 20 [257]. Moreover, the execution time of SVM and MFCCs is 4.6ms at 50Hz [258]. Another group of researchers [259], developed a SID framework based on MFCCs and SVM and concluded an execution time of 9.10ms per frame.

  DL algorithms comprise several hidden layers which add additional time complexity to a system. In speaker recognition, deep CNNs, RNNs and ResNet models are mostly employed by the researchers. Cai et al. in [260] presented a DNN-based framework with i-vector approach for the Speaker verification system. The inefficient use of DNN layers resulted in the high computational complexity of this method. Authors in [261] modified ID-ResNet20 by changing its convolutional kernels from 3×3 to 1×3 which reduced the computational complexity of the system by two-thrids approximately. The peak computation-to-communication ratios of layers resulted in 3.75 Gb/s for a speech length of 3s. In addition, the authors further modified the ResNet20 by adding a pooling layer after the convolutional layer. In comparison to the original ResNet20, the modified model achieved 51% reduced parameters and 64% computational complexity. In [262], authors developed a less complex attacking toolkit namely PhoneyTalker for DNN-based speaker recognition systems. The results from proposed framework demonstrated a low average time cost (ATC) of 0.03s and 15% ASR improvement than state-of-the-art methods. In addition, authors in [263] proposed a lightweight Few-shot speaker identification (FSSI) based on recurrent convolutional block (RCB) on the backbone of Bidirectional LSTM. A softmax layer is introduced in the proposed model and evaluated on three datasets (VoxCeleb1, VoxCeleb2, and LibriSpeech). The performance metrics model size (MS) and the number of multiplication and addition operations (MACs) indicated improvements as compared to state-of-the-art methods. The method achieved the highest accuracy scores 92.89%, 92.74% and 98.51% (V2-set, V1-set, L-set) on feature subset size 4 with low values of MS (54.14k) and MACs (103.16M).

In addition to above-mentioned applications, there are a few more applications that we discussed here. Authors in [264] classified sonar targets of different shapes and sizes in the air using MLP neural networks. In their method, they generated feature vectors after extraction of raw echos' spectrograms and other spectral features using STFT. After training MLP-NN, the performance is compared with narrowband and wideband excitation signals. Jin et al. [265] developed an object recognition framework for robots in an open environment based on their acoustic signatures collected by using the dynamic contact method. K-nearest neighbour ML algorithm is trained with MFCC features. Their framework proved that robots can detect objects by their acoustic waveforms and gives the best results with 180° joint rotation and 180° horizontal rotation. Moreover, He et al. [266] investigated drone sound identification in a noisy environment. Feature vectors of drone sounds are created by employing harmonic line association (HLA) and wavelet packet transform (WPT) FE methods. SVM along with optimized parameters by genetic algorithm (GA) is used to identify drones. They achieved 100% identification probability during trials.

IV. FUTURE RESEARCH DIRECTIONS

With numerous challenges and limitations faced by ASI, there exists immense potential for future research in various areas. This section discusses some recommendations based on our perspective for future consideration. We believe these directions would be interesting to investigate which would improve the performance of ASI and enhance a better understanding of this concept. Some of these recommendations appeal to general methodological problems and some are specific to ASI in the light of the earlier analysis in this survey:

- Real-time database expansion: Sometimes, required datasets are not available to solve a particular ASI problem for example in the underwater domain. Due to scarcity of real databases, it is very challenging to address undersea problems. DL approaches cannot be applied in this case because of insufficient data. In these cases, the underwater research community can expand underwater databases by collecting new reliable real-time audio datasets.
- Poor generalization ability of DL: Another problem that demands real-world data sets is the poor generalization

ability of the DL model. DL models trained on simulated data perform poorly when tested on real-world datasets. This is called train-test data mismatch. This problem also occurs when acoustic data acquisition trials are unrealistically performed and room geometries are not considered. Therefore, in this line domain adaption [267] and transfer learning [268] techniques must be investigated which ensures improving the performance of the network for one problem (real data) but actually trained for another problem (simulated data).

- Improvements in audio processing (AP): Multiple acoustic data acquisition trials conducted in various situations can introduce background noise and polyphonic sounds. Robust techniques need to be developed to identify and eliminate such anomalous sounds, especially in the case of multiple sound sources. New robust FE hybrid approaches with better discrimination for real-time ASI applications need to be explored. Our survey revolves around ASI only and we haven't considered the cases of Acoustic source separation (ASS), diarization and sound source enhancements that are all connected to ASI. In this survey paper, we have shown, how AI-based data-driven approach to ASI can replace conventional AP techniques. We believe that a combination of AP techniques and powerful DL models in particular deep generative models such as (GANs) [269], variational autoencoders (VAEs) [270] and dynamical VAEs [271], can model the temporal and/or spectral characteristics of sounds. Therefore along with AP these DL approaches can improve the performance of aforementioned problems and may be implemented by future researchers.
- Multi-task learning approach (MTL): Multi-task training is a general method used to improve the performance DNNs on a given problem by training the model to simultaneously handle other several tasks [272]. As per our knowledge and our survey, no one has used the MTL approach to tackle tasks jointly. In an ASI-based problem, this approach is implemented in the following way: First part of the model (e.g., FE module for several blocks) is common for different tasks, afterwards the model divides into different modules each one performing a different specialized task. The common module ensures the discovery of efficient signal representation which is used for other tasks. This approach offers data efficiency and shared representations and reduces overfitting problem as well.
- We noticed in our survey, many deep networks are presented which are computationally inefficient and thus require high computing power. Therefore future research can consider developing powerful DNNs using computing power for big datasets.
- In this paper, we have carried out a detailed survey but in future, meta-analysis, simulations and results can be added as well that will foster pathways to better research and development.

V. CONCLUSION

ASI is facing numerous challenges in accuracy, automation and robustness. AI methods have evolved and serve as a promising solution to these problems. In the past decade, considerable research has been carried out in various domains to identify and recognize sound sources from their acoustic signatures but these research works have not been surveyed and compiled to give a comprehensive review. Our work serves as a detailed guide for future researchers. In our work, we have attempted to study and review the past few research works towards acoustic source identification using AI methods and organized them in terms of different applications.

In this paper, we have presented an in-depth survey of ASI in the industry for fault detection, underwater for target recognition, surveillance, medical for disease diagnosis and fall detection and some others. Initially, we highlighted potentially available databases for future research to start with. Then, we highlighted a few basic audio processing steps. Afterwards, an overview of FE techniques in time, frequency and cepstral domains was presented to aid researchers to choose the best technique as per the given problem, dataset and AI algorithm. We have also discussed briefly some of the traditional ML and DL algorithms that have been mostly used in literature. Added to this, we have given a comprehensive survey of the ASI works along with its significant contributions in various fields in subsequent sections. Lastly, we have discussed some future research directions for the readers after explaining the thorough idea of the concept and its significance.

REFERENCES

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[2] K. L. Gemba, S. Nannuru, and P. Gerstoft, "Robust ocean acoustic localization with sparse Bayesian learning," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 49–60, Mar. 2019.

[3] W. W. Au and M. O. Lammers, *Listening in the Ocean*. Springer, 2016.

[4] P. Gerstoft and D. F. Gingras, "Parameter estimation using multifrequency range-dependent acoustic data in shallow water," *J. Acoust. Soc. Amer.*, vol. 99, no. 5, pp. 2839–2850, May 1996.

[5] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. Hoboken, NJ, USA: Wiley, 2018.

[6] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 48, pp. E7856–E7865, Nov. 2016.

[7] R. Liu, B. Yang, E. Zio, and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: A review," *Mech. Syst. Signal Process.*, vol. 108, pp. 33–47, Aug. 2018.

[8] X. Cao, X. Zhang, Y. Yu, and L. Niu, "Deep learning-based recognition of underwater target," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Oct. 2016, pp. 89–93.

[9] A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze, "Using one-class SVMs and wavelets for audio surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 4, pp. 763–775, Dec. 2008.

[10] W. Zhang, J. Han, and S. Deng, "Heart sound classification based on scaled spectrogram and partial least squares regression," *Biomed. Signal Process. Control*, vol. 32, pp. 20–28, Feb. 2017.
[11] L. Ren and Y. Peng, "Research of fall detection and fall prevention technologies: A systematic review," *IEEE Access*, vol. 7, pp. 77702–77722, 2019.
[12] D. Jain, K. Huynh Anh Nguyen, S. M. Goodman, R. Grossman-Kahn, H. Ngo, A. Kusupati, R. Du, A. Olwal, L. Findlater, and J. E. Froehlich, "ProtoSound: A personalized and scalable sound recognition system for deaf and hard-of-hearing users," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2022, pp. 1–16.
[13] Y.-H. Cheng and C.-N. Kuo, "Machine learning for music genre classification using visual Mel spectrum," *Mathematics*, vol. 10, no. 23, p. 4427, Nov. 2022.
[14] J. N. Oswald, C. Erbe, W. L. Gannon, S. Madhusudhana, and J. A. Thomas, "Detection and classification methods for animal sounds," *Exploring Animal Behav. Through Sound*, vol. 1, pp. 269–317, Jan. 2022.
[15] P. Anusha and K. ManiSai, "Bird species classification using deep learning," in *Proc. Int. Conf. Intell. Controller Comput. Smart Power (ICICCSP)*, Jul. 2022, pp. 1–5.
[16] Y. Sudo, K. Itoyama, K. Nishida, and K. Nakadai, "Sound event aware environmental sound segmentation with mask U-Net," *Adv. Robot.*, vol. 34, no. 20, pp. 1280–1290, Oct. 2020.
[17] J. Kim, D. Lee, Y. Kim, H. Shin, Y. Heo, Y. Wang, and E. T. Matson, "Deep learning based malicious drone detection using acoustic and image data," Tech. Rep., 2022. [Online]. Available: https://www.easychair.org/
[18] J. J. Noda, C. M. Travieso-González, D. Sánchez-Rodríguez, and J. B. Alonso-Hernández, "Acoustic classification of singing insects based on MFCC/LFCC fusion," *Appl. Sci.*, vol. 9, no. 19, p. 4097, Oct. 2019.
[19] S. Chandrakala and S. L. Jayalakshmi, "Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies," *ACM Comput. Surveys*, vol. 52, no. 3, pp. 1–34, May 2020.
[20] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, "Applications of machine learning to machine fault diagnosis: A review and roadmap," *Mech. Syst. Signal Process.*, vol. 138, Apr. 2020, Art. no. 106587.
[21] B. Teng and H. Zhao, "Underwater target recognition methods based on the framework of deep learning: A survey," *Int. J. Adv. Robotic Syst.*, vol. 17, no. 6, pp. 1–12, 2020, doi: 10.1177/1729881420976307.
[22] W. Chen, Q. Sun, X. Chen, G. Xie, H. Wu, and C. Xu, "Deep learning methods for heart sounds classification: A systematic review," *Entropy*, vol. 23, no. 6, p. 667, May 2021.
[23] Y. Chen, H. Niu, H. Chen, and X. Liu, "A review of underwater target recognition based on deep learning," *J. Phys., Conf.*, vol. 1881, no. 4, Apr. 2021, Art. no. 042031.
[24] E. C. Nunes, "Anomalous sound detection with machine learning: A systematic review," 2021, *arXiv:2102.07820*.
[25] O. AlShorman, F. Alkahatni, M. Masadeh, M. Irfan, A. Glowacz, F. Althobiani, J. Kozik, and W. Glowacz, "Sounds and acoustic emission-based early fault diagnosis of induction motor: A review study," *Adv. Mech. Eng.*, vol. 13, no. 2, pp. 1–19, 2021, doi: 10.1177/1687814021996915.
[26] A. Bansal and N. K. Garg, "Environmental sound classification: A descriptive review of the literature," *Intell. Syst. with Appl.*, vol. 16, Nov. 2022, Art. no. 200115.
[27] J. J. P. Abadía and K. Smarsly, "An introduction and systematic review on machine learning for smart environments/cities: An IoT approach," *Machine Learning for Smart Environments/Cities*. 2022, pp. 1–23.
[28] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," 2019, *arXiv:1909.09347*.
[29] *Mimii Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection*. Accessed: Mar. 27, 2023. [Online]. Available: https://zenodo.org/record/3384388#.zcejyhzbyuk
[30] 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.
[31] *detection and Classification of Acoustic Scenes and Events*. Accessed: Mar. 27, 2023. [Online]. Available: https://dcase.community/
[32] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1041–1044.
[33] *Urbansound8k Dataset*. Accessed: Mar. 27, 2023. [Online]. Available: https://urbansounddataset.weebly.com/urbansound8k.html
[34] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1015–1018.
[35] *ESC: Dataset for Environmental Sound Classification*. Accessed: Mar. 27, 2023. [Online]. Available: https://dataverse.harvard.edu/dataset.xhtml?persistentid=doi:10.7910/dvn/ydeput
[36] P. Bentley, G. Nordehn, M. Coimbra, S. Mannor, and R. Getz. (2011). *Classifying Heart Sounds Challenge*. [Online]. Available: http://www.peterjbentley.com/heartchallenge
[37] B. Peter. *Classifying Heart Sounds Challenge*. Accessed: Mar. 27, 2023. [Online]. Available: http://www.peterjbentley.com/heartchallenge/#downloads
[38] C. Liu et al., "An open access database for the evaluation of heart sound algorithms," *Physiological Meas.*, vol. 37, no. 12, pp. 2181–2213, Dec. 2016.
[39] *Classification of Heart Sound Recordings: The Physionet/Computing in Cardiology Challenge 2016*. Accessed: Mar. 27, 2023. [Online]. Available: https://physionet.org/content/challenge-2016/1.0.0/
[40] M. A. Sehili, D. Istrate, B. Dorizzi, and J. Boudy, "Daily sound recognition using a combination of GMM and SVM for home automation," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Jun. 2012, pp. 1673–1677.
[41] S. Nakamura, K. Hiyane, F. Asano, Y. Kaneda, T. Yamada, T. Nishiura, T. Kobayashi, S. Ise, and H. Saruwatari, "Design and collection of acoustic sound data for hands-free speech recognition and sound scene understanding," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jan. 2002, pp. 161–164.
[42] *RWCP Sound Scene Database in Real Acoustical Environments (RWCP-SSD)*. Accessed: Mar. 28, 2023. [Online]. Available: http://research.nii.ac.jp/src/en/rwcp-ssd.html
[43] M. Zhang, L. Yan, G. Luo, G. Li, W. Liu, and L. Zhang, "A novel insect sound recognition algorithm based on MFCC and CNN," in *Proc. 6th Int. Conf. Commun., Image Signal Process. (CCISP)*, Nov. 2021, pp. 289–294.
[44] *Bug Bytes Sound Library: Stored Product Insect Pest Sounds*. Accessed: Mar. 28, 2023. [Online]. Available: https://data.nal.usda.gov/search/type/dataset
[45] S. R. Vanrell, J. O. Chelotti, L. A. Bugnon, H. L. Rufiner, D. H. Milone, E. A. Laca, and J. R. Galli, "Audio recordings dataset of grazing jaw movements in dairy cattle," *Data Brief*, vol. 30, Jun. 2020, Art. no. 105623.
[46] V. Sebastian. *Dataset-Jaw-Movements*. Accessed: Mar. 28, 2023. [Online]. Available: https://github.com/sinc-lab/dataset-jaw-movements
[47] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.
[48] T. Carl. *GTZAN Genre Collection*. Accessed: Mar. 28, 2023. [Online]. Available: https://www.kaggle.com/datasets/carlthome/gtzan-genre-collection
[49] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proc.*, vol. 3, no. 1, pp. 91–99, Jun. 2022.
[50] J. Davis, M. Piovoso, K. Hoo, and B. Bakshi, "Process data analysis and interpretation," in *Advances in Chemical Engineering*, vol. 25. Amsterdam, The Netherlands: Elsevier, 1999, pp. 1–103.
[51] P. Ziegler and K. R. Dittrich, "Data integration—Problems, approaches, and perspectives," in *Conceptual Modelling in Information Systems Engineering*. Cham, Switzerland: Springer, 2007, pp. 39–58.
[52] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," 2019, *arXiv:1904.08779*.
[53] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 559–563.
[54] J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin, "Robust environmental sound recognition for home automation," *IEEE Trans. Autom. Sci. Eng.*, vol. 5, no. 1, pp. 25–31, Jan. 2008.
[55] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Automatic recognition of urban soundscenes," in *New Directions in Intelligent Interactive Multimedia*. Cham, Switzerland: Springer, 2008, pp. 147–153.
[56] B. Kedem, "Spectral analysis and discrimination by zero-crossings," *Proc. IEEE*, vol. 74, no. 11, pp. 1477–1493, Nov. 1986.

[57] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 1996, pp. 993–996.

[58] J. José Burred, A. Robel, and T. Sikora, "Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 663–674, Mar. 2010.

[59] D. Smith, E. Cheng, and I. Burnett, "Musical onset detection using MPEG-7 audio descriptors," in *Proc. 20th Int. Congr. Acoust. (ICA)*, vol. 2327, Sydney, NSW, Australia, 2010, p. 1014.

[60] X. Valero and F. Alías, "Applicability of MPEG-7 low level descriptors to environmental sound source recognition," in *Proc. 1st Euroregio Conf.*, 2010, pp;. 1–11.

[61] A. I. Al-Shoshan, "Speech and music classification and separation: A review," *J. King Saud Univ. Eng. Sci.*, vol. 19, no. 1, pp. 95–132, 2006.

[62] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Z. Kégl, "Aggregate features and ADABOOST for music classification," *Mach. Learn.*, vol. 65, no. 2, pp. 473–484, 2006.

[63] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2002, pp. II-1941–II-1944.

[64] Y. Ando, "Autocorrelation-based features for speech representation," in *Proc. Meetings Acoust.*, 2013, Art. no. 060033.

[65] A. Glowacz, "Fault detection of electric impact drills and coffee grinders using acoustic signals," *Sensors*, vol. 19, no. 2, p. 269, Jan. 2019.

[66] A. Glowacz, "Acoustic based fault diagnosis of three-phase induction motor," *Appl. Acoust.*, vol. 137, pp. 82–89, Aug. 2018.

[67] K. Umapathy, B. Ghoraani, and S. Krishnan, "Audio signal processing using time-frequency approaches: Coding, classification, fingerprinting, and watermarking," *EURASIP J. Adv. Signal Process.*, vol. 2010, no. 1, pp. 1–28, Dec. 2010.

[68] D. Preis and V. C. Georgopoulos, "Wigner distribution representation and analysis of audio signals: An illustrated tutorial review," *J. Audio Eng. Soc.*, vol. 47, no. 12, pp. 1043–1053, 1999.

[69] L. Cohen and T. Posch, "Positive time-frequency distribution functions," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 1, pp. 31–38, Feb. 1985.

[70] B. Ghoraani and S. Krishnan, "Time–frequency matrix feature extraction and classification of environmental audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2197–2209, Sep. 2011.

[71] N. Baydar and A. Ball, "A comparative study of acoustic and vibration signals in detection of gear failures using Wigner–Ville distribution," *Mech. Syst. Signal Process.*, vol. 15, no. 6, pp. 1091–1107, Nov. 2001.

[72] P. Boles and B. Boashash, *Application of the Cross-Wigner–Ville Distribution to Seismic Data Processing*. Harlow, U.K.: Longman Cheshire, 1992.

[73] G. Muhammad and K. Alghathbar, "Environment recognition from audio using MPEG-7 features," in *Proc. 4th Int. Conf. Embedded Multimedia Comput.*, Dec. 2009, pp. 1–6.

[74] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE MultimediaMag.*, vol. 3, no. 3, pp. 27–36, Mar. 1996.

[75] E. Mendoza, N. Valencia, J. Muñoz, and H. Trujillo, "Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS)," *J. Voice*, vol. 10, no. 1, pp. 59–66, Jan. 1996.

[76] V. Berisha, S. Sandoval, R. Utianski, J. Liss, and A. Spanias, "Selecting disorder-specific features for speech pathology fingerprinting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7562–7566.

[77] J. M. Liss, S. LeGendre, and A. J. Lotto, "Discriminating dysarthria type from envelope modulation spectra," *J. Speech, Lang., Hearing Res.*, vol. 53, no. 5, pp. 1246–1255, Oct. 2010.

[78] G. Agostini, M. Longari, and E. Pollastri, "Musical instrument timbres classification with spectral features," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 1, pp. 1–10, Dec. 2003.

[79] F. Wang, X. Wang, B. Shao, T. Li, and M. Ogihara, "Tag integrated multi-label music style classification with hypergraph," in *Proc. ISMIR*, 2009, pp. 363–368.

[80] W. A. Sethares, R. D. Morris, and J. C. Sethares, "Beat tracking of musical performances using low-level audio features," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 275–285, Mar. 2005.

[81] E. Tsau, S. Kim, and C.-C. J. Kuo, "Environmental sound recognition with CELP-based features," in *Proc. Int. Symp. Signals, Circuits Syst. (ISSCS)*, Jun. 2011, pp. 1–4.

[82] A. M. Noll and M. R. Schroeder, "Short-time 'cepstrum' pitch detection," *J. Acoust. Soc. Amer.*, vol. 36, no. 5, p. 1030, 1964.

[83] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-Mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.

[84] Q. Zhou, J. Shan, W. Ding, C. Wang, S. Yuan, F. Sun, H. Li, and B. Fang, "Cough recognition based on Mel-spectrogram and convolutional neural network," *Frontiers Robot. AI*, vol. 8, May 2021, Art. no. 580080.

[85] F. Liu, T. Shen, Z. Luo, D. Zhao, and S. Guo, "Underwater target recognition using convolutional recurrent neural networks with 3-D Mel-spectrogram and data augmentation," *Appl. Acoust.*, vol. 178, Mar. 2021, Art. no. 107989.

[86] G. Hong and D. Suh, "Mel spectrogram-based advanced deep temporal clustering model with unsupervised data for fault diagnosis," *Exp. Syst. Appl.*, vol. 217, May 2023, Art. no. 119551.

[87] A. Gallardo-Antolín and J. M. Montero, "On combining acoustic and modulation spectrograms in an attention LSTM-based system for speech intelligibility level classification," *Neurocomputing*, vol. 456, pp. 49–60, Oct. 2021.

[88] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[89] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Commun.*, vol. 54, no. 4, pp. 543–565, May 2012.

[90] Y. Korkmaz, A. Boyacı, and T. Tuncer, "Turkish vowel classification based on acoustical and decompositional features optimized by genetic algorithm," *Appl. Acoust.*, vol. 154, pp. 28–35, Nov. 2019.

[91] M. Müller, *Information Retrieval for Music and Motion*, vol. 2. Springer, 2007.

[92] J. Chen, Y. Huang, Q. Li, and K. K. Paliwal, "Recognition of noisy speech using dynamic spectral subband centroids," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 258–261, Feb. 2004.

[93] N. C. Maddage, C. Xu, and Y. Wang, "A SVM C based classification approach to musical audio," in *Proc. Int. Soc. Music Inf. Retr. Conf.*, 2003, pp. 1–2.

[94] A. Bernard and A. Alwan, "Source and channel coding for remote speech recognition over error-prone channels," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, pp. 2613–2616.

[95] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[96] P. J. Clemins and M. T. Johnson, "Generalized perceptual linear prediction features for animal vocalization analysis," *J. Acoust. Soc. Amer.*, vol. 120, no. 1, pp. 527–534, Jul. 2006.

[97] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker, "Multiple classifier systems for the classification of audio-visual emotional states," in *Proc. 4th Int. Conf. Affect. Comput. Intell. Interact.* Memphis, TN, USA: Springer, Oct. 2011, pp. 359–368.

[98] N. Dave, "Feature extraction methods LPC, PLP and MFCC in speech recognition," *Int. J. Advance Res. Eng. Technol.*, vol. 1, no. 6, pp. 1–4, 2013.

[99] 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.

[100] S. Abidin, R. Togneri, and F. Sohel, "Spectrotemporal analysis using local binary pattern variants for acoustic scene classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 2112–2121, Nov. 2018.

[101] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *J. Biomed. Informat.*, vol. 83, pp. 103–111, Jul. 2018.

[102] B. Sun, L. Li, T. Zuo, Y. Chen, G. Zhou, and X. Wu, "Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild," in *Proc. 16th Int. Conf. Multimodal Interact.*, Nov. 2014, pp. 481–486.

[103] S. M. Adnan, A. Irtaza, S. Aziz, M. O. Ullah, A. Javed, and M. T. Mahmood, "Fall detection through acoustic local ternary patterns," *Appl. Acoust.*, vol. 140, pp. 296–300, Nov. 2018.

[104] M. S. Hossain, "Patient state recognition system for healthcare using speech and facial expressions," *J. Med. Syst.*, vol. 40, no. 12, pp. 1–8, Dec. 2016.
[105] W. Yang and S. Krishnan, "Combining temporal features by local binary pattern for acoustic scene classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 6, pp. 1315–1321, Jun. 2017.
[106] F. Demir, A. Sengur, N. Cummins, S. Amiriparian, and B. Schuller, "Low level texture features for snore sound discrimination," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 413–416.
[107] W. Jiang, C. Cotton, S.-F. Chang, D. Ellis, and A. Loui, "Short-term audio-visual atoms for generic video concept classification," in *Proc. 17th ACM Int. Conf. Multimedia*, Oct. 2009, pp. 5–14.
[108] G. Tzanetakis, G. Essl, and P. Cook, "Audio analysis using the discrete wavelet transform," in *Proc. Conf. Acoust. Music Theory Appl.*, vol. 66. Princeton, NJ, USA: Citeseer, 2001, pp. 1–6.
[109] Y. Liu, Q. Xiang, Y. Wang, and L. Cai, "Cultural style based music classification of audio signals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 57–60.
[110] Z. Duan, T. Wu, S. Guo, T. Shao, R. Malekian, and Z. Li, "Development and trend of condition monitoring and fault diagnosis of multi-sensors information fusion for rolling bearings: A review," *Int. J. Adv. Manuf. Technol.*, vol. 96, nos. 1–4, pp. 803–819, Apr. 2018.
[111] S. Noor, E. A. Dhrubo, A. T. Minhaz, C. Shahnaz, and S. A. Fattah, "Audio visual emotion recognition using cross correlation and wavelet packet domain features," in *Proc. IEEE Int. WIE Conf. Electr. Comput. Eng. (WIECON-ECE)*, Dec. 2017, pp. 233–236.
[112] V. Sabitha and P. Janardhanan, "Speaker verification system using MFCC and DWT," *IOSR J. Electron. Commun. Eng.*, pp. 24–29, 2013.
[113] F. Amelia and D. Gunawan, "DWT-MFCC method for speaker recognition system with noise," in *Proc. 7th Int. Conf. Smart Comput. Commun. (ICSCC)*, Jun. 2019, pp. 1–5.
[114] S. Hidayat, M. Tajuddin, S. A. A. Yusuf, J. Qudsi, and N. N. Jaya, "Wavelet detail coefficient as a novel wavelet-MFCC features in text-dependent speaker recognition system," *IIUM Eng. J.*, vol. 23, no. 1, pp. 68–81, Jan. 2022.
[115] S. Arora, S. Jain, and I. Chana, "A fusion framework based on cepstral domain features from phonocardiogram to predict heart health status," *J. Mech. Med. Biol.*, vol. 21, no. 4, May 2021, Art. no. 2150034.
[116] M. Al-Qaderi, E. Lahamer, and A. Rad, "A two-level speaker identification system via fusion of heterogeneous classifiers and complementary feature cooperation," *Sensors*, vol. 21, no. 15, p. 5097, Jul. 2021.
[117] P. Kenny, "Bayesian speaker verification with, heavy tailed priors," *Proc. Odyssey*, 2010. [Online]. Available: https://cir.nii.ac.jp/crid/
[118] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 1695–1699.
[119] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4814–4818.
[120] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4052–4056.
[121] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333.
[122] C. Zhang, F. Bahmaninezhad, S. Ranjan, H. Dubey, W. Xia, and J. H. L. Hansen, "UTD-CRSS systems for 2018 NIST speaker recognition evaluation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5776–5780.
[123] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[124] F. Bahmaninezhad, C. Zhang, and J. H. L. Hansen, "An investigation of domain adaptation in speaker embedding space for speaker recognition," *Speech Commun.*, vol. 129, pp. 7–16, May 2021.
[125] D. Ketan. *Audio Deep Learning Made Simple (Part 1): State-of-The-Art Techniques*. Accessed: Mar. 31, 2023. [Online]. Available: https://towardsdatascience.com/audio-deep-learning-made-simple-part-1-state-of-the-art-techniques
[126] J. Joshy and K. Sambyo, "A comparison and contrast of the various feature extraction techniques in speaker recognition," *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 9, no. 11, pp. 99–108, Nov. 2016.
[127] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognit. Artif. Intell.*, vol. 116, no. 1, pp. 374–388, 1976.
[128] R. Gubka and M. Kuba, "A comparison of audio features for elementary sound based audio classification," in *Proc. Int. Conf. Digit. Technol.*, May 2013, pp. 14–17.
[129] P. Boonmatham, S. Pongpinigpinyo, and T. Soonklang, "A comparison of audio features of Thai classical music instrument," in *Proc. 7th Int. Conf. Comput. Converg. Technol. (ICCCT)*, Dec. 2012, pp. 213–218.
[130] J. D. Krijnders and P. W. J. van Hengel, "A comparison of spectro-temporal representations of audio signals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 8212–8216.
[131] V. Bountourakis, L. Vrysis, and G. Papanikolaou, "Machine learning algorithms for environmental sound recognition: Towards soundscape semantics," in *Proc. Audio Mostly Interact. Sound*, Oct. 2015, pp. 1–7.
[132] N. Jekic and A. Pester, "Environmental sound recognition with classical machine learning algorithms," in *Proc. Smart Ind. Smart Educ., 15th Int. Conf. Remote Eng. Virtual Instrum.* Cham, Switzerland: Springer, 2019, pp. 14–21.
[133] Y. Wang, R. Wang, D. Li, D. Adu-Gyamfi, K. Tian, and Y. Zhu, "Improved handwritten digit recognition using quantum K-nearest neighbor algorithm," *Int. J. Theor. Phys.*, vol. 58, no. 7, pp. 2331–2340, Jul. 2019.
[134] M. A. Mukid, T. Widiharih, A. Rusgiyono, and A. Prahutama, "Credit scoring analysis using weighted k nearest neighbor," *J. Phys., Conf.*, vol. 1025, May 2018, Art. no. 012114.
[135] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. 10th Eur. Conf. Mach. Learn.* Chemnitz, Germany: Springer, Apr. 1998, pp. 137–142.
[136] L. Barghout, "Spatial-taxon information granules as used in iterative fuzzy-decision-making for image segmentation," in *Granular Computing and Decision-Making: Interactive and Iterative Approaches*. Springer, 2015, pp. 285–318.
[137] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
[138] T. Starner and A. Pentland, "Real-time American sign language recognition from video using hidden Markov models," in *Proc. Int. Symp. Comput. Vis. (ISCV)*, 1995, pp. 265–270.
[139] G. Kouemou and F. Opitz, "Hidden Markov models in radar target classification," in *Proc. IET Int. Conf. Radar Syst.*, 2007, pp. 1–5.
[140] H. Wan, H. Wang, B. Scotney, and J. Liu, "A novel Gaussian mixture model for classification," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 3298–3303.
[141] N. Ganesan, K. Venkatesh, M. Rama, and A. M. Palani, "Application of neural networks in diagnosing cancer disease using demographic data," *Int. J. Comput. Appl.*, vol. 1, no. 26, pp. 76–85, 2010.
[142] J. French, "The time traveller's CAPM," *Investment Analysts J.*, vol. 46, no. 2, pp. 81–96, 2017.
[143] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3D-R2N2: A unified approach for single and multi-view 3D object reconstruction," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 628–644.
[144] M. V. Valueva, N. N. Nagornov, P. A. Lyakhov, G. V. Valuev, and N. I. Chervyakov, "Application of the residue number system to reduce hardware costs of the convolutional neural network implementation," *Math. Comput. Simul.*, vol. 177, pp. 232–243, Nov. 2020.
[145] B. Hutchinson, L. Deng, and D. Yu, "Tensor deep stacking networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1944–1957, Aug. 2013.
[146] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, "Sound classification using convolutional neural network and tensor deep stacking network," *IEEE Access*, vol. 7, pp. 7717–7727, 2019.
[147] P. Kalaiarasi and P. E. Rani, "A comparative analysis of AlexNet and GoogLeNet with a simple DCNN for face recognition," in *Advances in Smart System Technologies: Select Proceedings of ICFSST*. Springer, 2021, pp. 655–668.

[148] O. Gencoglu, T. Virtanen, and H. Huttunen, "Recognition of acoustic events using deep neural networks," in *Proc. 22nd Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2014, pp. 506–510.

[149] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica Biophysica Acta (BBA)-Protein Struct.*, vol. 405, no. 2, pp. 442–451, 1975.

[150] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, Dec. 2020.

[151] D. H. Pandya, S. H. Upadhyay, and S. P. Harsha, "Fault diagnosis of rolling element bearing with intrinsic mode function of acoustic emission data using APF-KNN," *Exp. Syst. Appl.*, vol. 40, no. 10, pp. 4137–4145, Aug. 2013.

[152] J. Yoon and D. He, "Planetary gearbox fault diagnostic method using acoustic emission sensors," *IET Sci., Meas. Technol.*, vol. 9, no. 8, pp. 936–944, Nov. 2015.

[153] C. Li, R.-V. Sanchez, G. Zurita, M. Cerrada, D. Cabrera, and R. E. Vásquez, "Gearbox fault diagnosis based on deep random forest fusion of acoustic and vibratory signals," *Mech. Syst. Signal Process.*, vols. 76–77, pp. 283–293, Aug. 2016.

[154] T. Waqar and M. Demetgul, "Thermal analysis MLP neural network based fault diagnosis on worm gears," *Measurement*, vol. 86, pp. 56–66, May 2016.

[155] Y. Yao, H. Wang, S. Li, Z. Liu, G. Gui, Y. Dan, and J. Hu, "End-to-end convolutional neural network model for gear fault diagnosis based on sound signals," *Appl. Sci.*, vol. 8, no. 9, p. 1584, Sep. 2018.

[156] M. R. Islam, Y.-H. Kim, J.-Y. Kim, and J.-M. Kim, "Detecting and learning unknown fault states by automatically finding the optimal number of clusters for online bearing fault diagnosis," *Appl. Sci.*, vol. 9, no. 11, p. 2326, Jun. 2019.

[157] B. P. Duong, J. Y. Kim, I. Jeong, K. Im, C. H. Kim, and J. M. Kim, "A deep-learning-based bearing fault diagnosis using defect signature wavelet image visualization," *Appl. Sci.*, vol. 10, no. 24, p. 8800, Dec. 2020.

[158] M. T. Pham, J.-M. Kim, and C. H. Kim, "Intelligent fault diagnosis method using acoustic emission signals for bearings under complex working conditions," *Appl. Sci.*, vol. 10, no. 20, p. 7068, Oct. 2020.

[159] P. Poto nik, B. Olmos, L. Vodopivec, E. Susi , and E. Govekar, "Condition classification of heating systems valves based on acoustic features and machine learning," *Appl. Acoust.*, vol. 174, Mar. 2021, Art. no. 107736.

[160] O. Yaman, "An automated faults classification method based on binary pattern and neighborhood component analysis using induction motor," *Measurement*, vol. 168, Jan. 2021, Art. no. 108323.

[161] E. Brusa, C. Delprete, and L. G. Di Maggio, "Deep transfer learning for machine diagnosis: From sound and music recognition to bearing fault detection," *Appl. Sci.*, vol. 11, no. 24, p. 11663, Dec. 2021.

[162] R. Cai, Q. Wang, Y. Hou, and H. Liu, "Event monitoring of transformer discharge sounds based on voiceprint," *J. Phys., Conf.*, vol. 2078, no. 1, Nov. 2021, Art. no. 012066.

[163] J. Yao, C. Liu, K. Song, C. Feng, and D. Jiang, "Fault diagnosis of planetary gearbox based on acoustic signals," *Appl. Acoust.*, vol. 181, Oct. 2021, Art. no. 108151.

[164] H. Santos, P. Scalassara, W. Endo, A. Goedtel, J. Guedes, and M. Gentil, "Non-invasive sound-based classifier of bearing faults in electric induction motors," *IET Sci., Meas. Technol.*, vol. 15, no. 5, pp. 434–445, Jul. 2021.

[165] I. Thoidis, M. Giouvanakis, and G. Papanikolaou, "Semi-supervised machine condition monitoring by learning deep discriminative audio features," *Electronics*, vol. 10, no. 20, p. 2471, Oct. 2021.

[166] O. Yaman, F. Yol, and A. Altinors, "A fault detection method based on embedded feature extraction and SVM classification for UAV motors," *Microprocessors Microsyst.*, vol. 94, Oct. 2022, Art. no. 104683.

[167] Y. Liu, Y. Cheng, Z. Zhang, and J. Wu, "Acoustic fault diagnosis of rotor bearing system," *Shock Vibrat.*, vol. 2022, pp. 1–9, Apr. 2022.

[168] X. Fu, K. Yang, M. Liu, T. Xing, and C. Wu, "LightFD: Real-time fault diagnosis with edge intelligence for power transformers," *Sensors*, vol. 22, no. 14, p. 5296, Jul. 2022.

[169] H. Niu, E. Reeves, and P. Gerstoft, "Source localization in an ocean waveguide using supervised machine learning," *J. Acoust. Soc. Amer.*, vol. 142, no. 3, pp. 1176–1188, Sep. 2017.

[170] H. Niu, E. Ozanich, and P. Gerstoft, "Ship localization in Santa Barbara channel using machine learning classifiers," *J. Acoust. Soc. Amer.*, vol. 142, no. 5, pp. EL455–EL460, Nov. 2017.

[171] J. Jiang, Z. Wu, M. Huang, and Z. Xiao, "Detection of underwater acoustic target using beamforming and neural network in shallow water," *Appl. Acoust.*, vol. 189, Feb. 2022, Art. no. 108626.

[172] M. R. Mosavi, M. Khishe, and A. Ghamgosar, "Classification of sonar data set using neural network trained by gray wolf optimization," *Neural Netw. World*, vol. 26, no. 4, pp. 393–415, 2016.

[173] M. Khishe, M. R. Mosavi, and M. Kaveh, "Improved migration models of biogeography-based optimization for sonar dataset classification by using neural network," *Appl. Acoust.*, vol. 118, pp. 15–29, Mar. 2017.

[174] 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.

[175] M. Khishe and M. R. Mosavi, "Improved whale trainer for sonar datasets classification using neural network," *Appl. Acoust.*, vol. 154, pp. 176–192, Nov. 2019.

[176] M. Khishe and H. Mohammadi, "Passive sonar target classification using multi-layer perceptron trained by salp swarm algorithm," *Ocean Eng.*, vol. 181, pp. 98–108, Jun. 2019.

[177] W. Qiao, M. Khishe, and S. Ravakhah, "Underwater targets classification using local wavelet acoustic pattern and multi-layer perceptron neural network optimized by modified whale optimization algorithm," *Ocean Eng.*, vol. 219, Jan. 2021, Art. no. 108415.

[178] Z. Huang, J. Xu, Z. Gong, H. Wang, and Y. Yan, "Source localization using deep neural networks in a shallow water environment," *J. Acoust. Soc. Amer.*, vol. 143, no. 5, pp. 2922–2932, May 2018.

[179] H. Yang, J. Li, S. Shen, and G. Xu, "A deep convolutional neural network inspired by auditory perception for underwater acoustic target recognition," *Sensors*, vol. 19, no. 5, p. 1104, Mar. 2019.

[180] J. Yangzhou, Z. Ma, and X. Huang, "A deep neural network approach to acoustic source localization in a shallow water tank experiment," *J. Acoust. Soc. Amer.*, vol. 146, no. 6, pp. 4802–4811, Dec. 2019.

[181] W. Wang, H. Ni, L. Su, T. Hu, Q. Ren, P. Gerstoft, and L. Ma, "Deep transfer learning for source ranging: Deep-sea experiment results," *J. Acoust. Soc. Amer.*, vol. 146, no. 4, pp. EL317–EL322, Oct. 2019.

[182] G. Song, X. Guo, W. Wang, Q. Ren, J. Li, and L. Ma, "A machine learning-based underwater noise classification method," *Appl. Acoust.*, vol. 184, Dec. 2021, Art. no. 108333.

[183] A. B. Baggeroer, W. A. Kuperman, and P. N. Mikhalevsky, "An overview of matched field methods in ocean acoustics," *IEEE J. Ocean. Eng.*, vol. 18, no. 4, pp. 401–424, 1993.

[184] J. Ye, T. Kobayashi, and M. Murakawa, "Urban sound event classification based on local and global features aggregation," *Appl. Acoust.*, vol. 117, pp. 246–256, Feb. 2017.

[185] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Proc. Comput. Sci.*, vol. 112, pp. 2048–2056, Jan. 2017.

[186] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, Mar. 2017.

[187] B. Zhu, C. Wang, F. Liu, J. Lei, Z. Huang, Y. Peng, and F. Li, "Learning environmental sounds with multi-scale convolutional neural network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[188] S. A. Shah, A. Malik, S. Aziz, and W. Ahmad, "Sound recognition aimed towards hearing impaired individuals in urban environment using ensemble methods," *J. Inf. Commun. Technol. Robotic Appl.*, pp. 30–39, 2018.

[189] J. Sang, S. Park, and J. Lee, "Convolutional recurrent neural networks for urban sound classification using raw waveforms," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2018, pp. 2444–2448.

[190] Z. Chi, Y. Li, and C. Chen, "Deep convolutional neural network combined with concatenated spectrogram for environmental sound classification," in *Proc. IEEE 7th Int. Conf. Comput. Sci. Netw. Technol. (ICCSNT)*, Oct. 2019, pp. 251–254.

[191] J. M. Mendoza, V. Tan, V. Fuentes, G. Perez, and N. M. Tiglao, "Audio event detection using wireless sensor networks based on deep learning," in *Proc. Int. Wireless Internet Conf.* Cham, Switzerland: Springer, 2019, pp. 105–115.

[192] M. R. Ahmed, T. Islam Robin, and A. Ali Shafin, "Automatic environmental sound recognition (AESR) using convolutional neural network," *Int. J. Modern Educ. Comput. Sci.*, vol. 12, no. 5, pp. 41–54, Oct. 2020.

[193] F. Demir, D. A. Abdullah, and A. Sengur, "A new deep CNN model for environmental sound classification," *IEEE Access*, vol. 8, pp. 66529–66537, 2020.

[194] A. Polo-Rodriguez, J. M. Vilchez Chiachio, C. Paggetti, and J. Medina-Quero, "Ambient sound recognition of daily events by means of convolutional neural networks and fuzzy temporal restrictions," *Appl. Sci.*, vol. 11, no. 15, p. 6978, Jul. 2021.

[195] L. Nanni, G. Maguolo, S. Brahnam, and M. Paci, "An ensemble of convolutional neural networks for audio classification," *Appl. Sci.*, vol. 11, no. 13, p. 5796, Jun. 2021.

[196] P. Zinemanas, M. Rocamora, M. Miron, F. Font, and X. Serra, "An interpretable deep learning model for automatic sound classification," *Electronics*, vol. 10, no. 7, p. 850, Apr. 2021.

[197] S. Segarceanu, G. Suciu, and I. Gavat, "Environmental acoustics modelling techniques for forest monitoring," *Adv. Sci., Technol. Eng. Syst. J.*, vol. 6, no. 3, pp. 15–26, May 2021.

[198] Z. Mushtaq, S.-F. Su, and Q.-V. Tran, "Spectral images based environmental sound classification using CNN with meaningful data augmentation," *Appl. Acoust.*, vol. 172, Jan. 2021, Art. no. 107581.

[199] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in *Proc. 22nd Int. Conf. Digit. Signal Process. (DSP)*, Aug. 2017, pp. 1–5.

[200] M. S. Hossain and G. Muhammad, "Environment classification for urban big data using deep learning," *IEEE Commun. Mag.*, vol. 56, no. 11, pp. 44–50, Nov. 2018.

[201] S. Chachada and C.-C.-J. Kuo, "Environmental sound recognition: A survey," *APSIPA Trans. Signal Inf. Process.*, vol. 3, no. 1, 2014.

[202] T. Virtanen and M. Helén, "Probabilistic model based similarity measures for audio query-by-example," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2007, pp. 82–85.

[203] S. Chu, S. Narayanan, C.-C. Kuo, and M. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2006, pp. 885–888.

[204] N. Yamakawa, T. Takahashi, T. Kitahara, T. Ogata, and H. G. Okuno, "Environmental sound recognition for robot audition using matching-pursuit," in *Proc. Int. Conf. Ind., Eng. Other Appl. Appl. Intell. Syst.* Cham, Switzerland: Springer, 2011, pp. 1–10.

[205] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 257–267, Feb. 2007.

[206] R. Sitte and L. Willets, "Non-speech environmental sound identification for surveillance using self-organizing-maps," in *Proc. 4th Conf. IASTED Int. Conf., Signal Process., Pattern Recognit., Appl.*, 2007, pp. 281–286.

[207] J. Chen, A. H. Kam, J. Zhang, N. Liu, and L. Shue, "Bathroom activity monitoring based on sound," in *Proc. Int. Conf. Pervasive Comput.* Cham, Switzerland: Springer, 2005, pp. 47–61.

[208] M. Vacher, F. Portet, A. Fleury, and N. Noury, "Challenges in the processing of audio channels for ambient assisted living," in *Proc. 12th IEEE Int. Conf. e-Health Netw., Appl. Services*, Jul. 2010, pp. 330–337.

[209] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," *Pattern Recognit. Lett.*, vol. 31, no. 12, pp. 1524–1534, Sep. 2010.

[210] S. Brodie, S. Allen-Ankins, M. Towsey, P. Roe, and L. Schwarzkopf, "Automated species identification of frog choruses in environmental recordings using acoustic indices," *Ecol. Indicators*, vol. 119, Dec. 2020, Art. no. 106852.

[211] C.-I. Kim, Y. Cho, S. Jung, J. Rew, and E. Hwang, "Animal sounds classification scheme based on multi-feature network with mixed datasets," *KSII Trans. Internet Inf. Syst. (TIIS)*, vol. 14, no. 8, pp. 3384–3398, 2020.

[212] F. Weninger and B. Schuller, "Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 337–340.

[213] T. Tuncer, E. Akbal, and S. Dogan, "Multileveled ternary pattern and iterative ReliefF based bird sound classification," *Appl. Acoust.*, vol. 176, May 2021, Art. no. 107866.

[214] B. S. Soares, J. S. Luz, V. F. De Macêdo, R. R. V. E. Silva, F. H. D. de Araújo, and D. M. V. Magalhães, "MFCC-based descriptor for bee queen presence detection," *Exp. Syst. Appl.*, vol. 201, Sep. 2022, Art. no. 117104.

[215] J. A. Stevens, P. S. Corso, E. A. Finkelstein, and T. R. Miller, "The costs of fatal and non-fatal falls among older adults," *Injury Prevention*, vol. 12, no. 5, pp. 290–295, 2006.

[216] R. J. Gurley, N. Lum, M. Sande, B. Lo, and M. H. Katz, "Persons found in their homes helpless or dead," *New England J. Med.*, vol. 334, no. 26, pp. 1710–1716, Jun. 1996.

[217] M. Grassi, A. Lombardi, G. Rescio, P. Malcovati, M. Malfatti, L. Gonzo, A. Leone, G. Diraco, C. Distante, P. Siciliano, V. Libal, J. Huang, and G. Potamianos, "A hardware-software framework for high-reliability people fall detection," in *Proc. IEEE Sensors*, Oct. 2008, pp. 1328–1331.

[218] C. Doukas and I. Maglogiannis, "Advanced patient or elder fall detection based on movement and sound data," in *Proc. 2nd Int. Conf. Pervasive Comput. Technol. Healthcare*, Jan. 2008, pp. 103–107.

[219] F. Bianchi, S. J. Redmond, M. R. Narayanan, S. Cerutti, and N. H. Lovell, "Barometric pressure and triaxial accelerometry-based falls event detection," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 18, no. 6, pp. 619–627, Dec. 2010.

[220] L. Liu, M. Popescu, M. Skubic, M. Rantz, T. Yardibi, and P. Cuddihy, "Automatic fall detection based on Doppler radar motion signature," in *Proc. 5th Int. Conf. Pervasive Comput. Technol. for Healthcare (PervasiveHealth) Workshops*, May 2011, pp. 222–225.

[221] M. Wu, X. Dai, Y. D. Zhang, B. Davidson, M. G. Amin, and J. Zhang, "Fall detection based on sequential modeling of radar signal time-frequency features," in *Proc. IEEE Int. Conf. Healthcare Informat.*, Sep. 2013, pp. 169–174.

[222] M. Chan, D. Estève, J.-Y. Fourniols, C. Escriba, and E. Campo, "Smart wearable systems: Current status and future challenges," *Artif. Intell. Med.*, vol. 56, no. 3, pp. 137–156, Nov. 2012.

[223] L. Findlater, B. Chinh, D. Jain, J. Froehlich, R. Kushalnagar, and A. C. Lin, "Deaf and Hard-of-hearing Individuals' preferences for wearable and mobile sound awareness technologies," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, pp. 1–13.

[224] A. K. Dwivedi, S. A. Imtiaz, and E. Rodriguez-Villegas, "Algorithms for automatic analysis and classification of heart sounds—A systematic review," *IEEE Access*, vol. 7, pp. 8316–8345, 2019.

[225] C. Liu and A. Murray, "Applications of complexity analysis in clinical heart failure," in *Complexity and Nonlinearity in Cardiovascular Signals*. Springer, 2017, pp. 301–325.

[226] M. Ya ano lu and C. Köse, "Real-time detection of important sounds with a wearable vibration based device for hearing-impaired people," *Electronics*, vol. 7, no. 4, p. 50, Apr. 2018.

[227] A. W. Ramadhan, A. Wijayanto, and H. Oktavianto, "Implementation of audio event recognition for the elderly home support using convolutional neural networks," in *Proc. Int. Electron. Symp. (IES)*, Sep. 2020, pp. 91–95.

[228] S. M. Goodman, P. Liu, D. Jain, E. J. McDonnell, J. E. Froehlich, and L. Findlater, "Toward user-driven sound recognizer personalization with people who are d/deaf or hard of hearing," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 2, pp. 1–23, Jun. 2021.

[229] M. Hamidi, H. Ghassemian, and M. Imani, "Classification of heart sound signal using curve fitting and fractal dimension," *Biomed. Signal Process. Control*, vol. 39, pp. 351–359, Jan. 2018.

[230] F. Noman, C. Ting, S. Salleh, and H. Ombao, "Short-segment heart sound classification using an ensemble of deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1318–1322.

[231] W. Zhang, J. Han, and S. Deng, "Abnormal heart sound detection using temporal quasi-periodic features and long short-term memory without segmentation," *Biomed. Signal Process. Control*, vol. 53, Aug. 2019, Art. no. 101560.

[232] V. Roquemen-Echeverri, P. G. Jacobs, S. Heitner, P. M. Schulman, B. Wilson, J. Mahecha, and C. Mosquera-Lopez, "An AI-powered tool for automatic heart sound quality assessment and segmentation," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2021, pp. 3065–3074.

[233] L. Zhiming and M. Sheng, "Heart sound recognition method of congenital heart disease based on improved cepstrum coefficient features," in *Proc. Int. Conf. Comput. Eng. Artif. Intell. (ICCEAI)*, Aug. 2021, pp. 319–324.

[234] H. Kui, J. Pan, R. Zong, H. Yang, and W. Wang, "Heart sound classification based on log Mel-frequency spectral coefficients features and convolutional neural networks," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102893.

[235] M. B. Er, "Heart sounds classification using convolutional neural network with 1D-local binary pattern and 1D-local ternary pattern features," *Appl. Acoust.*, vol. 180, Sep. 2021, Art. no. 108152.

[236] Y. Zeinali and S. T. A. Niaki, "Heart sound classification using signal processing and machine learning algorithms," *Mach. Learn. Appl.*, vol. 7, Mar. 2022, Art. no. 100206.

[237] Y. Zhuang, Y. Chen, and J. Zheng, "Music genre classification with transformer classifier," in *Proc. 4th Int. Conf. Digit. Signal Process.*, Jun. 2020, pp. 155–159.

[238] K. Mounika, S. Deyaradevi, K. Swetha, and V. Vanitha, "Music genre classification using deep learning," in *Proc. Int. Conf. Advancements Elect., Electron., Commun., Comput. Automat. (ICAECA)*, 2021, pp. 1–7.

[239] A. A. A. Harryanto, K. Gunawan, R. Nagano, and R. Sutoyo, "Music classification model development based on audio recognition using transformer model," in *Proc. 3rd Int. Conf. Artif. Intell. Data Sci. (AiDAS)*, Sep. 2022, pp. 258–263.

[240] M. Shah, N. Pujara, K. Mangaroliya, L. Gohil, T. Vyas, and S. Degadwala, "Music genre classification using deep learning," in *Proc. 6th Int. Conf. Comput. Methodologies Commun. (ICCMC)*, Mar. 2022, pp. 974–978.

[241] K. K. Jena, S. K. Bhoi, S. Mohapatra, and S. Bakshi, "A hybrid deep learning approach for classification of music genres using wavelet and spectrogram analysis," *Neural Comput. Appl.*, vol. 35, pp. 1–26, Jan. 2023.

[242] Y. R. Pandeya, D. Kim, and J. Lee, "Domestic cat sound classification using learned features from deep neural nets," *Appl. Sci.*, vol. 8, no. 10, p. 1949, Oct. 2018.

[243] Y. R. Pandeya, B. Bhattarai, and J. Lee, "Visual object detector for cow sound event detection," *IEEE Access*, vol. 8, pp. 162625–162633, 2020.

[244] G. Li, Y. Xiong, Q. Du, Z. Shi, and R. S. Gates, "Classifying ingestive behavior of dairy cows via automatic sound recognition," *Sensors*, vol. 21, no. 15, p. 5231, Aug. 2021.

[245] Y. Sun, T. Midori Maeda, C. Solís-Lemus, D. Pimentel-Alarcón, and Z. Bu ivalová, "Classification of animal sounds in a hyperdiverse rainforest using convolutional neural networks with data augmentation," *Ecol. Indicators*, vol. 145, Dec. 2022, Art. no. 109621.

[246] P. Eichinski, C. Alexander, P. Roe, S. Parsons, and S. Fuller, "A convolutional neural network bird species recognizer built from little data by iteratively training, detecting, and labeling," *Frontiers Ecol. Evol.*, vol. 10, p. 133, Mar. 2022.

[247] Z. Jiang, A. Soldati, I. Schamberg, A. R. Lameira, and S. Moran, "Automatic sound event detection and classification of great ape calls using neural networks," 2023, *arXiv:2301.02214*.

[248] Q. Chen, J. Li, and Y. Li, "Forensic identification for electronic disguised voice based on supervector and statistical analysis," in *Proc. Conf. Oriental Chapter Int. Committee Coordination Standardization Speech Databases Assessment Techn. (O-COCOSDA)*, Oct. 2016, pp. 147–150.

[249] G. Pop, S. Mihalache, and D. Burileanu, "Forensic speaker identification using speech quality data," in *Proc. Int. Conf. Commun. (COMM)*, Jun. 2018, pp. 509–512.

[250] M. S. Rozario, A. Thomas, and D. Mathew, "Performance comparison of multiple speech features for speaker recognition using artifical neural network," in *Proc. 9th Int. Conf. Adv. Comput. Commun. (ICACC)*, Nov. 2019, pp. 234–239.

[251] M. Barhoush, A. Hallawa, and A. Schmeink, "Robust automatic speaker identification system using shuffled MFCC features," in *Proc. IEEE Int. Conf. Mach. Learn. Appl. Netw. Technol. (ICMLANT)*, Dec. 2021, pp. 1–6.

[252] Ç. Bakir and M. Yüzkat, "A search on the importance of forensic voice studies in forensic and a example application," in *Proc. 30th Signal Process. Commun. Appl. Conf. (SIU)*, May 2022, pp. 1–4.

[253] S. Vivekananthan and P. Vijayalakshmi, "Forensic speech enhancement of voiceprints and speaker identification under mismatched conditions," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT)*, Jul. 2022, pp. 1–6.

[254] B. V. K. Babu, D. K. Bhargav, R. K. Sah, L. Regalla, and N. Singh, "Forensic speaker recognition system using machine learning," in *Proc. Int. Conf. Sustain. Comput. Data Commun. Syst. (ICSCDS)*, Mar. 2023, pp. 696–701.

[255] M. Hrúz and Z. Zajíc, "Convolutional neural network for speaker change detection in telephone speaker diarization system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4945–4949.

[256] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: An end-to-end neural speaker embedding system," 2017, *arXiv:1705.02304*.

[257] R. Ramos-Lara, M. López-García, E. Cantó-Navarro, and L. Puente-Rodriguez, "Real-time speaker verification system implemented on reconfigurable hardware," *J. Signal Process. Syst.*, vol. 71, no. 2, pp. 89–103, May 2013.

[258] E. Cantó-Navarro, M. López-García, R. Ramos-Lara, and R. Sánchez-Reíllo, "Flexible biometric online speaker-verification system implemented on FPGA using vector floating-point units," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 23, no. 11, pp. 2497–2507, Nov. 2015.

[259] X. Wei, C. Hao Yu, P. Zhang, Y. Chen, Y. Wang, H. Hu, Y. Liang, and J. Cong, "Automated systolic array architecture synthesis for high throughput CNN inference on FPGAs," in *Proc. 54th ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Jun. 2017, pp. 1–6.

[260] W. Cai, J. Chen, and M. Li, "Analysis of length normalization in end-to-end speaker verification system," 2018, *arXiv:1806.03209*.

[261] J. Xu, S. Li, J. Jiang, and Y. Dou, "A simplified speaker recognition system based on FPGA platform," *IEEE Access*, vol. 8, pp. 1507–1516, 2020.

[262] M. Chen, L. Lu, Z. Ba, and K. Ren, "PhoneyTalker: An out-of-the-box toolkit for adversarial example attack on speaker recognition," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, May 2022, pp. 1419–1428.

[263] Y. Li, H. Chen, W. Cao, Q. Huang, and Q. He, "Few-shot speaker identification using lightweight prototypical network with feature grouping and interaction," *IEEE Trans. Multimedia*, pp. 1–12, 2023.

[264] P. Kroh, R. Simon, and S. Rupitsch, "Classification of sonar targets in air: A neural network approach," *Sensors*, vol. 19, no. 5, p. 1176, Mar. 2019.

[265] S. Jin, H. Liu, B. Wang, and F. Sun, "Open-environment robotic acoustic perception for object recognition," *Frontiers Neurorobotics*, vol. 13, p. 96, Nov. 2019.

[266] Y. J. He, I. Ahmad, L. Shi, and K. Chang, "SVM-based drone sound recognition using the combination of HLA and WPT techniques in practical noisy environment," *KSII Trans. Internet Inf. Syst.*, vol. 13, no. 10, pp. 5078–5094, 2019.

[267] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," in *Advances in Data Science and Information Engineering*. Springer, 2021, pp. 877–894.

[268] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[269] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.

[270] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019.

[271] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, "Dynamical variational autoencoders: A comprehensive review," 2020, *arXiv:2008.12595*.

[272] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," 2020, *arXiv:2009.09796*.

RUBA ZAHEER received the Bachelor of Engineering degree in telecommunication from the National University of Sciences and Technology, Pakistan, in 2015. She is currently pursuing the master's degree in engineering science (research-based) with Edith Cowan University, Australia. She has been a Laboratory Engineer with COMSATS University, Pakistan. Her research interests include data analytics, applied AI, signal processing, and underwater communications.

IFTEKHAR AHMAD (Member, IEEE) received the Ph.D. degree in communication networks from Monash University, Australia, in 2007. He is currently an Associate Professor with the School of Engineering, Edith Cowan University, Australia. His research interests include 5G technologies, green communications, QoS in communication networks, software-defined radio, wireless sensor networks, and computational intelligence.

DARYOUSH HABIBI (Senior Member, IEEE) received the Bachelor of Engineering (Hons.) and Ph.D. degrees in electrical engineering from the University of Tasmania, in 1989 and 1994, respectively. His employment history includes Telstra Research Laboratories, Flinders University, Intelligent Pixels Inc., and Edith Cowan University, where he is currently a Professor, a Pro Vice-Chancellor, and the Executive Dean of the School of Engineering. He has more than 200 refereed publications in high-impact journals, conference proceedings, and book chapters. His research interests include engineering design for sustainable development, renewable and smart energy systems, environmental monitoring technologies, and reliability and quality of service in engineering systems and networks. He is a fellow of Engineers Australia and the Institute of Marine Engineering, Science and Technology.

KAZI YASIN ISLAM (Member, IEEE) received the B.E. degree in electrical and computer systems engineering from Monash University, Malaysia, in 2015, and the M.E. degree in electronics and communication engineering from Edith Cowan University, Australia, in 2019, where he is currently pursuing the Ph.D. degree. His research interests include underwater wireless communication, green communication, the IoT, and machine learning.

QUOC VIET PHUNG (Member, IEEE) received the Ph.D. degree in communication engineering from Edith Cowan University, Australia, in 2010. He is currently a Lecturer with the School of Engineering, Edith Cowan University. His research interests include smart sensors, 5G technologies, underwater communication, software-defined radio, data analytics, network security, and applied artificial intelligence.