Multimedia Tools and Applications (2024) 83:8129–8143
https://doi.org/10.1007/s11042-023-15891-z

A review of deep learning techniques in audio event recognition (AER) applications

Arjun Prashanth[1] · **S. L. Jayalakshmi**[2] · **R. Vedhapriyavadhana**[1]

Received: 15 January 2022 / Revised: 17 May 2023 / Accepted: 22 May 2023 /
Published online: 14 June 2023

Abstract
In our day-to-day life, observation of human and social actions are highly important for public protection and security. Additionally, identifying suspicious activity is also essential in critical environments, such as industry, smart homes, nursing homes, and old age homes. In most of the audio-based applications, the Audio Event Recognition (AER) task plays a vital role to recognize audio events. Even though many approaches focus on the effective implementation of audio-based applications, still there exist major research problems such as overlapping events, the presence of background noise, and the lack of benchmark data sets. The main objective of this survey is to identify effective feature extraction methods, robust classifiers, and benchmark datasets. To achieve this, we have presented a detailed survey on features, deep learning classifiers, and data sets used in the AER applications. Also, we summarised the various methods involved in AER applications such as audio spoofing, audio surveillance, and audio fingerprinting. The future direction includes setting up a benchmark dataset, identifying the semantic features, and exploring the transfer learning-based classifiers.

Keywords Audio event recognition · Deep learning techniques · Features · Classifiers · Datasets · Convolutional neural network (CNN) · Mel frequency cepstral coefficients (MFCCs)

✉ S. L. Jayalakshmi
sathishjayalakshmi02@pondiuni.ac.in

Arjun Prashanth
arjun.prashanth2020@vitstudent.ac.in

R. Vedhapriyavadhana
vedhapriyavadhana.r@vit.ac.in

[1] School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India

[2] School of Engineering and Technology, Department of Computer Science, Pondicherry University (Main Campus), Puducherry, India

1 Introduction

In recent studies, audio signal processing and deep learning has been of great interest and a growing field in Machine learning [2]. Audio signal processing and deep learning are essential in many places such as voice assistants, music detectors, animal species recognition [17], healthcare [11], audio forensics [23], and in surveillance where recognition of events through video alone is inadequate [22].

This survey paper reviews completely about Audio Event Recognition (AER) tasks and their various applications. Even though so many developments are happening in human-being-to-machine communication, an audio signal examination is very imperative in joint audio-visual recognition. When human-to-machine interaction happens, anomalous sound event recognition provides new opportunities and it is useful in various applications such as investigation schemes, industrial error recognition, and safety observing both indoor and outdoor environments [21].

There are many advantages in audio event recognition, still there exist many research problems such as the presence of noise and overlapping of events thereby affecting the results [7, 8]. Furthermore, the sparsity of the data set is also a hurdle, that requires more attention if a particular audio event is happening only for a very short duration (less than 2 milliseconds). Hence, a robust machine learning algorithm is highly needed to handle the presence of noises and overlapping events in complex environments such as seminar halls, traffic areas, smart homes, and old age homes.

To overcome these challenges, many of the works have been carried out using deep learning-based approaches such as Convolutional Neural Network(CNN), Long short-term Memory (LSTM), Bi-directional Long Short Term Memory (Bi-LSTM) [13]. As per the literature survey, deep learning approaches are more insensitive to noise. To improve the recognition accuracy it uses a large number of examples with varying levels of noise [9].

Recently, for recognizing acoustic classes and their respective boundaries, audio segmentation and sound detection played a very important role. In the field of computer vision, the You Only Look Once (YOLO) algorithm is a familiar one, similarly, in the field of computer audition, the YOHO (You Only Hear Once) algorithm has been used to convert the acoustic boundaries into regression problem [38]. Thus, YOHO plays a vital role to recognise audio events in a complex environment. For the same, YOHO is added end-to-end for improving the speediness of AER applications.

The primary objective of this survey is to analyze the robust feature extraction methods and effective deep learning based classifiers to improve the performance of the AER system. For the same, this paper summarises the various feature extraction methods, deep learning classifiers, and AER applications. The structure of the paper is as follows: Section 2 presents the various features, classifiers, and datasets used for Audio Event Recognition (AER) task. Section 3 provides a review of the various applications of the AER task. Section 4 presents the conclusion with future directions.

2 Methodology

This section presents the methodology for sound signal feature extraction and design of deep learning–based classifiers. The general structure of the methodology includes the following steps: Data collection, preprocessing,, feature extraction, dataset preparation,

deep learning model design, training, model evaluation, hyper-parameter tuning, and testing. Figure 1 shows the methodology of the audio event recognition system with feature extraction, event modeling (training), and classification (testing) approaches. A standard dataset is considered as an input for the AER system. Then robust features are extracted from the given input dataset. After that, the event modeling will construct a reference model for each class separately using the classifier. During the classification, a particular test input is given and its corresponding label is predicted with the help of a classifier [8].

2.1 Audio features

In this section, the various feature extraction methods are reviewed. Feature extraction is a very important process for audio event recognition (AER). It extracts semantic information from the preprocessed audio signal data. It represents various discriminative natures of the audio signal, such as spectral content, temporal patterns, and statistical features. The choice of feature extraction techniques depends on the nature of the audio event applications [31]. Some of the commonly used feature extraction methods for audio signals include the following:

- Fourier Transform: Convert the signal from the time domain to the frequency domain using techniques like the Fast Fourier Transform (FFT). This can reveal the spectral content of the signal.
- Cepstral Features: Compute a representation that captures the spectral shape of the signal and its temporal variations, mimicking human auditory perception.

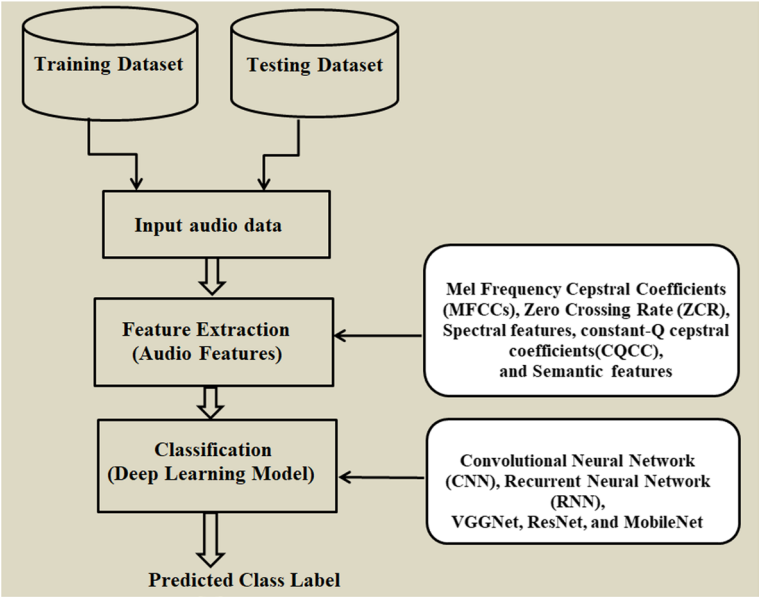

Fig. 1 Methodology of Audio Event Recognition(AER) System

- Spectrogram: Create a visual representation of the signal's frequency content over time by dividing the signal into short time segments and computing the FFT for each segment.
- Wavelet Transform: Analyze the signal using wavelets to capture localized spectral information at different scales.
- Statistical Features: Extract statistical properties such as mean, standard deviation, skewness, or kurtosis of the signal.

The ability of the cepstral features to distinguish between the effects of source and filter on a speech signal makes them generally advantageous [24]. The primary acoustic feature utilized for audio analysis tasks has been Mel Frequency Cepstral Coefficients (MFCCs). MFCC algorithm closely resembles the way the human ear functions and detects the same crucial bandwidths, which accounts for its popularity. MFCCs are most commonly used in speech recognition models such as identifying numbers spoken into telephones and voice recognition for security locks. To retain the phonetics of speech, MFCCs with frequency filters placed linearly and logarithmically at low and high frequencies respectively have been used [3].

The MFCC format represents audio samples by first performing a Fourier transform on them, then projecting the resulting spectral powers onto the mel-scale and calculating the log of those values at each of the mel frequencies. Then, using the list of mel log powers as a signal, the discrete cosine transform is applied, and the MFCCs are the amplitudes of the resulting spectrum. Log-mel spectrum, which is more frequently employed, is produced by skipping the final step [25].

The authors in [1] proposed a framework using a featured ensemble of Zero crossing rate, spectral contrast, spectral centroid, spectral roll-off, spectral bandwidth, and MFCC features to recognize the anomalous event in an audio forensics application. To perform feature reduction, the Principal Component Analysis (PCA) is used. It selected MFCC as the best feature to efficiently recognize anomalous events in forensic investigation.

The proposed work in [36] and [5]used a Constant Q Cepstral Coefficient(CQCC) for extracting features from sound events by involving Constant Q-Transform (CQT) instead of taking short-time-Fourier-transform. Authors in [31] recently developed a common subspace learning (CSL) based method to extract the semantic information of multiple events from a complicated environment. The subspace is calculated using the fusion of content information and the temporal sequence of various events to arrive at the optimum solution. This proposed work emphasized the importance of semantic feature extraction when compared with the state-of-the-art bag of audio words (BoAW) approach.

In another work [25], the feature extraction with raw waveforms are infrequently utilized by keeping away the hand-designed features. Instead, the modeling ability of deep learning models and learning representations are employed for the AER task.

Audio event detection is highly needed for health, surveillance, and many other security applications. Hand-crafted representations can be used for resolving any AED tasks. To make the hand-crafted results better way, non-hand-crafted representations such as spectrogram, mel spectrogram, log mel spectrogram, and mel frequency cepstral coefficients are studied efficiently. Furthermore, usually used window and hop sizes do not deliver the optimal presentations for the hand-crafted demonstrations [18].

From the literature, it can be observed that, deep learning based features bring the high computational complexity and data necessities. Instead, we can use the compactness of the semantic feature to recognize the audio events with some background noises in an optimal way. From this section we can conclude that, MFCC features are more robust and best feature for AER task. Table 1 gives the features and methodologies used in various AER tasks.

Table 1 Features and methodologies used in AER task

S. No	Ref. No	Features	Classifiers	Methodologies
1	[1]	Zero crossing rate features, Spectral contrast features, Spectral centroid features, Spectral rolloff features,Spectral band-width features, and MFCC	Support Vector Machine (SVM), K-Nearest Neighbor Algorithm (KNN), Extreme Gradient Boosting (XGB), Multi-layer Perceptron (MLP), Random Forest (RF), and Logistic Regression (LR)	A rigorous and systematic method to construct and evaluate the dataset, which is for developing and testing new algorithms in audio forensics tasks. The suggested method extracts mel-frequency cepstral coefficients (MFCCs) features from the newly produced dataset and uses principal component analysis to decrease the number of features(PCA).
2	[34]	Mel spectrogram	Convolutional neural networks (CNNs) and recurrent neural networks (RNNs)	Some preprocessing techniques, like spectrogram generation and data augmentation, are used to standardise and enhance the data. The effectiveness of various deep learning models is then assessed.
3	[31]	Common subspace learning (CSL) based approach for semantic features	Support vector machine (SVM)	A method for acoustic event recognition based on common subspace learning (CSL) and semantic feature extraction is used to extract the more discriminative feature.
4	[6]	Mel spectrogram (MEL) and mel frequency cepstral coefficients (MFCC)	Extreme Gradient boosting (XGBoost) and Convolutional Neural Network (CNN)	A system for employing forest acoustic sounds to monitor the forest environment and to explore the ML based XGBoost algorithm and DL based CNN algorithm for recognizing the forest acoustics.

2.2 Deep learning classifiers

Previously, the Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Support Vector Machine (SVM), and Artificial Neural Network(ANN) are the most widely used models for the audio event recognition (AER) task [8].

It can be observed from the literature that, in automatic speech recognition, music information retrieval, and sound event analysis, deep learning models are being replaced with support vector machines for sequential data classifications, and Gaussian Mixture Models (GMMs)-Hidden Markov Models (HMMs) for sequence transduction. Commonly used architectures for sound classification include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or combinations of both. Even though, in audio enhancements, de-noising, and source separation, deep learning is being employed over Non-negative Matrix Factorization(NMF) and Wiener methods [25, 35]. Deep modeling concepts such as Convolutional Neural Networks (CNN) [6, 40], Recurrent Neural Networks (RNN) [19, 34], and variations of RNN structures such as Long Short Term Memory (LSTM) [4] networks or the Gated Recurrent Unit(GRU) [39] deployed in various audio event recognition applications.

Recently, Ozkan Inik [16] adapted the Particle Swarm Optimization (PSO) algorithm for CNN parameter optimization. Federico Colangelo et al. [9] proposed an algorithm for audio event recognition in noisy environments based on the use of the deep recurrent neural network (RNN). This algorithm initially processes the audio signal by applying Short Time Fourier Transform followed by Log-Mel Filtering. Finally, Two Long Short-Term Memory(LSTM) cells are fed with the filtered signal, one cell detects the presence of the relevant event and the other classifies it.

Stowell et al. [34] implemented a system for automatic acoustic recognition of birds through deep learning technique. The authors used two classifiers to attain a good performance of 85%. The first was a baseline classifier (code-named *smacpy*) using a proven method of converting audio samples to MFCC representation and then the distributions of the MFCCs are modeled via GMMs. It is a very simple, efficient, and adaptable approach. However, it has been replaced with more advanced techniques for accuracy in general-purpose sound recognition systems in [32].

The second classifier (code-named skfl) was a more modern and capable classifier that was introduced to recognize different bird species in [33]. Two layers of unsupervised feature learning were applied to the Mel Spectrogram to automatically turn the learned data into features. A group learning technique based on decision trees called random forest was utilized for classification. These concepts were picked because they are well known for performing well in challenging classification circumstances like multi-modal classes, unbalanced data sets, and outliers.

Another classifier used popularly in the literature is Residual Neural Networks (ResNets). Due to the issue of vanishing gradient with CNNs and RNNs, ResNets are chosen to mitigate this drawback. ResNets utilize skip connections in layers that connect two non-adjacent layers. This acts as a shortcut allowing training updates to back-propagate faster towards the lower layers during training [5, 15, 28].

Turab et al. [37] investigated the use of multi-feature ensemblers by taking the three cutting-edge audio features such as the Mel Spectrogram, Mel Frequency Cepstral Coefficients, and Zero Crossing Rate. The proposed system is trained by using cutting-edge DL models, including Convolutional Neural Network, EfficientNet, and MobileNet, along with conventional classifiers such as SVM, and Multi Perceptron.

Recently, the concept of knowledge distillation is attracted many researchers for transferring the knowledge of pre-trained complex models to a small network. Gao et al. [12] proposed a knowledge distillation method for audio event classification by considering the features such as constant-Q transform (CQT), log-scaled Mel-spectrogram (log Mel), and MFCCs. In which the integration of multiple representations with mutual complement information is converted into a single representation. Then it was given to VGGNet and ResNet for classification.

This section summarizes various deep learning classifiers and their importance in AER applications. Especially, the knowledge distillation method gives a fruitful direction for adapting representation learning to improve the performance of the AER task. Table 2 gives a summary of the deep learning classifiers used in various AER tasks.

This review suggests that the deep features and deep-learning-based classifiers proposed in the literature outperform previous machine-learning-based methods. However, identification of new deep learning architectures is crucial in enhancing the performance of real-time applications in the presence of background noises. Next, the development of real-time audio surveillance systems must explore the recognition of abnormal sound events in presence of multiple audio source conditions. Lastly, it is difficult to create a unified framework that can handle multiple categories of audio inputs from different locations and applications.

2.3 Data-sets

There are a large number of data sets available to demonstrate the different aspects of audio event recognition tasks. Some of the popular datasets are listed in Table 3.

3 Applications of audio event recognition task

In this section, Table 4 discusses the summary of the three emerging AER applications such as Audio Surveillance, Audio Fingerprinting, and Audio Spoofing with a detailed analysis.

Table 2 Summary of Deep learning classifiers used in AER task

S. No	Ref. No	Deep learning classifiers
1	[9]	Deep Recurrent Neural Network (RNN)
2	[34]	convolutional and/or recurrent neural nets (CNNs, RNNs, or CRNNs)
3	[5]	Convolutional neural networks (CNNs) and recurrent neural networks (RNNs)
4	[37]	Convolutional Neural Network, EfficientNet, and MobileNet
5	[12]	VGGNet and ResNet
6	[16]	Convolutional Neural Network (CNN)

Table 3 Datasets available for AER task

S.no	Ref no, Year	Name of the Dataset	Description
1	[34], 2019	Chernobyl Dataset	The different locations in Chernobyl dataset with 6620 different audio files. The Abandoned village and Shrub area had low radiations, while the Pine forest and Mixed forest had high radiations and medium radiations were observed in Deciduous forest and Meadow area.
2	[34], 2019	Warblr dataset	It is a crowd-sourced dataset from UK-wide project. Warblr is a software application available on mobile devices which provides automatic bird species classification. 10,000 Audio files were used from the years 2015-2016. The files had varying level of noise.
3	[34], 2019	PolandNFC dataset	It contains one author's recordings from monitoring autumn nocturnal bird migration. Due to large data size (> 3200 hr of recording) and the files being unlabelled, only a small subset of 22 audio clips lasting 30 mins each were subjectively chosen. Manual annotation was performed by visual inspection of the spectogram and listening to the clips.
4	[13], 2020	Sound events for Surveillance Applications (SESA) dataset	The dataset is collected from Freesound and it is divided into 480 training files and 105 testing files. The following 4 classes are used for recognizing the suspicious activity in the given complex environment: Casual, Gunshot, Explosion, and Siren.
5	[13], 2020	MIVIA road audio events Dataset	It has 400 events in total for road surveillance applications, such as tire skidding and automobile accidents.
6	[4], 2021	GTZAN dataset	This dataset describes the various classes used in audio fingerprinting. It consists of following classes: Music, Genres, and speech.
7	[10], 2017	DCASE-2021:Automated audio captioning dataset	This dataset contains the sound events related to crowdsourcing environment. For the same, it contains the following classes such as muffled sound,the sound of a big car, people talking in a small and empty room,and ringing sound of a clock.
8	[6], 2023	FSC22 dataset	This forest sound dataset contains main 6 classes (mechanical sounds, forest threats sounds, environmental sounds, human sounds, animal sounds and vehicle sounds) and 34 subclasses.

Table 4 Summary of Audio Event Recognition applications

S. no	Ref no, Year	Features used	Classifiers used	Applications	Dataset used	Pros	Cons
1	[4], 2021	Fourier Transform of the spectral features	LSTM	Audio Fingerprinting	GTZAN	– Improved Data retrival Accuracy	– Lack of universal dataset. – Cannot identify context and environment. – Original Audio cannot be recreated.
2	[5], 2019	MFCC, CQCC, Log Magnitude of STFT	ResNet	Audio Spoofing	ASVSpoof 2019 Audio Spoofing	– The Fusion model proposed achieves 0 Equal Error Rate	– When the dataset is diverse and unknown, system provides less performance for logical access.
3	[5], 2019	STFT, Log-Mel Filtering	LSTM	Audio Surveillance and event recognition	MIVIA Audio Events	– High Accuracy even in the presence of noise.	– Overlapping events lead to more confusion among similar classes.
4	[13], 2020	Gammatonegram image extraction	CNN	Audio Surveillance and event recognition	MIVIA Audio Events, Freesound SESA	– Fully connected layers with pyramid structure. – Provided better generalization for small training dataset. – Cut down false positive rates. Low sensitivity to noise	– The real-time experiment produced lesser performance when compared to the conventional experiment.
5	[32], 2015	MFCC, Melspectogram	GMM and random forest	Audio Surveillance and event recognition	Chernobyl, Warblr, Poland-NFC	– Good performance of 85% AUC(area under ROC curve)	– Poor generalisation – GMM model had worse results as compared to more advance classifiers

Table 4 (continued)

S. no	Ref no, Year	Features used	Classifiers used	Applications	Dataset used	Pros	Cons
6	[2], 2023	Melspectogram	Audio MLPMixer (AMM)	Audio Surveillance	ESC-10, Urbansound8k (US8K), and DCASE-2019 Task-1(A)	– Efficient knowledge transfer with improved accuracy	– Complexity in implementation – Sensitivity to hyperparameters – Increased computational requirements

3.1 Audio surveillance

In many cases, video surveillance alone is not sufficient to detect events of interest. Audio samples are needed to be taken into account to ensure accurate event recognition. Also, audio analysis is cheaper than video analysis since it has lower band-with requirements for data streaming and requires less computational resources for data storing, processing, and manipulating. Furthermore, microphones in contrast to cameras can be both unidirectional and omnidirectional and thus providing a spherical field of view. In addition, Audio waves can be reflected and therefore can be picked up even when obstacles are present, which is a drawback for video processing as obstacles block the field of view and deter the event recognition [14].

Shaer et al. [29] proposed a multi-stage machine learning (MML) approach for pipe leakage detection in an industry. The suggested MML pipeline reduces the data size initially through feature collection techniques and subsequently incorporates time correlations by removing time-based features. Support vector machines are used to verify the validity of the pipeline using the few extracted features. Renaud et al. [27] proposed a hybrid approach by combining convolutional neural networks (CNNs) with gradient boosting algorithms, such as XGBoost and LightGBM, to accurately classify and predict noise levels in smart cities. The authors highlighted the significance of monitoring noise pollution due to its more impact on health and quality of life in human beings.

Mnasri et al. [20] have stated in their paper that video monitoring alone is insufficient for detecting major accidents since any dangerous behavior on the road might be mistaken for an accident, resulting in a succession of false alarms being sounded. Instead, audio signal processing can recognize the sounds such as crashes, tires screeching, and harsh braking which can lead to more precise accident recognition. The supervision of video streams by a human operator is required by current solutions that use IP cameras placed in significant locations. Additionally, crash recognition based on computer vision and video surveillance may be unsuccessful due to inadequate lighting at night and in cloudy weather as well as the inability to cover all regions.

Antonio Greco et al. [13] designed and trained a CNN named AReN to automatically detect events such as screams, gunshots, and broken glasses. The Gammatonegram images are extracted from sound events with different signal-to-noise ratios to emulate the working of the human auditory system. Then, these images are fed in AReN(21-layer CNN). Thus the proposed system produced a better generalization for small training data sets. The authors employed the MIVIA Audio Events dataset to obtain a recognition accuracy of 99.42%. Also, Colangelo et al. [10] demonstrated a model to detect gunshots, screams, and glass breaking with an improved level of accuracy even in the presence of noise.

3.2 Audio fingerprinting

Audio Fingerprinting is the minimal signature of an audio stream (speech and non-speech sounds) that can be used to summarize and represent the audio stream compactly by extracting relevant features from the stream. This system is still in its infancy stage only. However, an effective audio fingerprinting framework must be constructed with the following properties -

1. Robustness - It must work even in the presence of noise and also be accurate to a certain degree.
2. Pairwise independence - No two different voice signals must have the same fingerprint, i.e two different voice signals must have different audio fingerprints
3. Quick Database Query - The database search to match against fingerprints must be quick and efficient.
4. Versatility - Regardless of audio extraction, the audio recognition system must be capable of audio recognition.
5. Reliability - The speech recognition system's approach must be competent and robust.
6. Fragility - The system must be able to detect alterations that happened in the original audio signals.

Altalbe and Ali [4] proposed a system involving Least Mean Square (LMS) filter to preprocess the audio signal. This output is improved by using the double-threshold segmentation method. After that, the preprocessed speech signal (frames and windows) is divided into Time Fourier Transform (TFT) to obtain a unique spectral article code. A Long Short Term Memory (LSTM) neural network model was used for classification. The accuracy of the system is measured as 98.56%. Lastly, a new direction for research is provided by expanding the use of various representation learning methods in audio-based monitoring applications for tasks like audio fingerprinting and acoustic source localization.

3.3 Audio spoofing

A speaker verification system can be tricked by recording, synthesizing, or altering an authentic original audio signal using external software. This method is known as audio spoofing [30].

Alzantot et al. [5] designed an audio spoofing recognition system using Deep Residual Neural Networks. The authors constructed three different models such as MFCC-ResNet, CQCC-ResNet, and Spec-ResNet using MFCCs, and Log Magnitude of STFT features. It was implemented using python's PyTorch module and was trained on a setup having TitanX GPU. The dataset used in [5] was divided into the following three types: training set (8 male, 12 female), development set (4 male, 6 female), and evaluation (21 male, 27 female). The spoofed audio was generated using 17 different speech synthesis and voice conversion toolkits.

When compared with the baseline models, the CQCC model was vulnerable to waveform filtering-based video conversion attacks. Next, the Spec-ResNet was found to be the best model closely followed by CQCC-ResNet.

When the speaker is close to the recording device and the user is using a high-quality playback device, the Spec-ResNet works better than CQCC-ResNet. However, the CQCC-ResNet performs better than the Spec-Resnet when the speaker is remote or close to the recording device and when the speaker is using replay devices that are of poorer quality. The two classifiers together result in an improvement of about 73%. The major finding is developing a unified framework to handle diverse environmental sounds is a significant obstacle.

The authors in [26] proposed a system to efficiently differentiate between synthetic and genuine speech and non-speech audio distribution. The authors used log-mel features to represent the speech and non-speech audio sounds. The main finding is that it is difficult to create a unified framework that can handle different environmental sounds.

4 Conclusion and future work

This survey paper provided a complete overview of current developments in the audio event recognition task. Also, it clearly explains the various kinds of features and deep learning methods available for AER-based applications. Subsequently, this paper gives the inference that the deep features and deep-learning-based algorithms proposed in the literature outperform previous machine-learning-based methods. However, when it comes to deep feature extraction and modeling, the necessity for a large amount of training data as well as parameter and hyperparameter tuning are major research problems that affect the performance of audio event recognition tasks. Additionally, the lack of a common framework to extract features from similar environments such as baby day-care center, and old age homes create a major research problem.

Many of the conventional methods in the literature perform well for benchmark datasets within a limited environment. Hence, very less exploration of real- world audio-based applications such as surveillance and forensics. The future work will mainly concentrate on setting up a benchmark datasets with more anomalous events from a complex environment. Then, to handle these kinds of critical applications, a common framework (YOHO) is required. For the same, we will employ transfer learning-based approaches to improve the performance of the AER system. Especially, the knowledge distillation method gives a fruitful direction for adapting representation learning to improve the performance of the AER task.

Funding No funds, grants, or other support was received.

Data availability Data will be made available on reasonable request.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article

References

1. Abbasi A, Javed ARR, Yasin A, Jalil Z, Kryvinska N, Tariq U (2022) A large-scale benchmark dataset for anomaly detection and rare event classification for audio forensics. IEEE Access 10:38885–38894
2. Achyut Mani Tripathi and Om Jee Pandey (2023) Divide and distill: new outlooks on knowledge distillation for environmental sound classification. IEEEACM Trans Audio, Speech, Language Process 31:1100–1113
3. Alim SA, Rashid NKA (2018) Some commonly used speech feature extraction algorithms. In: Lopez-Ruiz R (ed) From natural to artificial intelligence, chapter 1. IntechOpen, Rijeka
4. Altalbe A (2021) Audio fingerprint analysis for speech processing using deep learning method. Int J Speech Technol:1–7
5. Alzantot M, Wang Z, Srivastava MB (2019) Deep residual neural networks for audio spoofing detection. arXiv preprint arXiv:1907.00501
6. Bandara M, Jayasundara R, Ariyarathne I, Meedeniya D, Perera C (2023) Forest sound classification dataset: Fsc22. Sensors 23(4):2032
7. Bhatti UA, Yuan L, Zhaoyuan Y, Nawaz SA, Mehmood A, Bhatti MA, Nizamani MM, Xiao S et al (2021) Predictive data modeling using sp-knn for risk factor evaluation in urban demographical healthcare data. J Med Imaging Health Inform 11(1):7–14

8. Chandrakala S, Jayalakshmi SL (2019) Environmental audio scene and sound event recognition for autonomous surveillance: a survey and comparative studies. ACM Comput Surv (CSUR) 52(3):1–34
9. Colangelo F, Battisti F, Carli M, Neri A, Calabró F (2017) Enhancing audio surveillance with hierarchical recurrent neural networks. In *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE
10. Drossos K, Adavanne S, Virtanen T (2017) Automated audio captioning with recurrent neural networks. In *IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, new Paltz, New York, USA
11. Fang Y, Liu D, Jiang Z, Wang H et al (2023) Monitoring of sleep breathing states based on audio sensor utilizing mel-scale features in home healthcare. J Healthcare Eng 2023
12. Gao L, Kele X, Wang H, Peng Y (2022) Multi-representation knowledge distillation for audio classification. Multimed Tools Appl 81(4):5089–5112
13. Greco A, Petkov N, Saggese A, Vento M (2020) Aren: a deep learning approach for sound event recognition using a brain inspired representation. IEEE Trans Inform Forensics Sec 15:3610–3624
14. Greco A, Saggese A, Vento M, Vigilante V (2019) Sorenet: a novel deep network for audio surveillance applications. In 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), pages 546–551
15. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. ProceedIEEE Conf Comput Vision Pattern Recogn:770–778
16. Inik O (2023) Cnn hyper-parameter optimization for environmental sound classification. Appl Acoust 202:109168
17. Jiang Z, Soldati A, Schamberg I, Lameira AR, Moran S (2023) Automatic sound event detection and classification of great ape calls using neural networks. arXiv preprint arXiv:2301.02214
18. Küçükbay SE, Kalkan S et al (2022) Hand-crafted versus learned representations for audio event detection. Multimed Tools Appl:1–20
19. Lipton ZC, Berkowitz J, Elkan C (2015) A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019
20. Mnasri Z, Rovetta S, Masulli F (2020) Audio surveillance of roads using deep learning and autoencoder-based sample weight initialization. In *2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON)*, pages 99–103
21. Mnasri Z, Rovetta S, Masulli F (2022) Anomalous sound event detection: a survey of machine learning based methods and applications. Multimed Tools Appl 81(4):5537–5586
22. Mohaimenuzzaman M, Bergmeir C, West I, Meyer B (2023) Environmental sound classification on the edge: a pipeline for deep acoustic networks on extremely resource constrained devices. Pattern Recogn 133:109025
23. Mustafa A, Qamhan, Altaheri H, Meftah AH, Muhammad G, Alotaibi YA (2021) Digital audio forensics. Microphone and environment classification using deep learning. IEEE Access 9:62719–62733
24. Poorjam AH (2018) Why we take only 12-13 mfcc coefficients in feature extraction?, 05
25. Purwins H, Li B, Virtanen T, Schluter J, Chang S-Y, Sainath T (2019) Deep learning for audio signal processing. IEEE J Selected Topics Signal Process 13(2):206–219
26. Ray R, Karthik S, Mathur V, Prashant Kumar G Maragatham ST, Shankarappa RT (2021) Feature genuinization based residual squeeze-and-excitation for audio anti-spoofing in sound ai. In 2021 12th international conference on computing communication and networking technologies (ICCCNT), pages 1–5. IEEE
27. Renaud J, Karam R, Salomon M, Couturier R (2023) Deep learning and gradient boosting for urban environmental noise monitoring in smart cities. Expert Syst Appl:119568
28. Revay S, Teschke M (2019) Multiclass language identification using deep learning on spectral images of audio signals. CoRR, abs/1905.04348
29. Shaer I, Shami A , (2022) Sound event classification in an industrial environment: Pipe leakage detection use case. arXiv preprint arXiv:2205.02706
30. Shim H-J, Jung J-W, Heo H-S, Yoon S-H, Ha-Jin Y (2018) Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes. In 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI), pages 172–176
31. Shi Q, Deng S, Han J (2022) Common subspace learning based semantic feature extraction method for acoustic event recognition. Appl Acoust 190:108638
32. Stowell D, Giannoulis D, Benetos E, Lagrange M, Plumbley MD (2015) Detection and classification of acoustic scenes and events. IEEE Trans Multimedia 17(10):1733–1746
33. Stowell D, Plumbley MD (2014) Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. PeerJ 2:e488

34. Stowell D, Wood MD, Pamuła H, Stylianou Y, Glotin H (2019) Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. Methods Ecol Evol 10(3):368–380
35. Su C, Huang H-Y, Shi S, Guo Y, Wu H (2017) A parallel recurrent neural network for language modeling with pos tags. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 140–147
36. Todisco M, Delgado H, Evans N (2017) Constant q cepstral coefficients: a spoofing countermeasure for automatic speaker verification. Comput Speech Lang 45:516–535
37. Turab M, Kumar T, Bendechache M, Saber T (2022) Investigating multi-feature selection and ensembling for audio classification. arXiv preprint arXiv:2206.07511
38. Venkatesh S, Moffat D, Miranda ER (2022) You only hear once: a yolo-like algorithm for audio segmentation and sound event detection. Appl Sci 12(7):3293
39. Xu Y, Kong Q, Huang Q, Wang W, Plumbley MarkD (2017) Convolutional gated recurrent neural network incorporating spatial features for audio tagging. In 2017 international joint conference on neural networks (IJCNN), pages 3461–3466. IEEE
40. Zhao Y, Xia X, Togneri R (2019) Applications of deep learning to audio generation. IEEE Circ Syst Magaz 19(4):19–38