Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19785-6

A parametric survey on polyphonic sound event detection and localization

Sallauddin Mohmmad[1,2] · **Suresh Kumar Sanampudi**[3]

Received: 2 March 2023 / Revised: 17 June 2024 / Accepted: 19 June 2024

Abstract
With the rapid growth in technology, everything will become automatic. Automatic detection is needed in health, security, smart homes, smart cities, and various environments. Sound is an essential function of automatic detection, which can be created in different environmental situations. To handle this, various sound detection devices are incorporated that run alone or with a combination of video. This auditory detection system is becoming more emerging in research today. This study is a survey on sound event detection and localization-related models. This survey research included details about different datasets, feature extractions, and classification methods that are implemented in sound-related research. Distinct and typical environmental situations comprise multiple sound sources that generate different sounds and noise. Our research compares distinct algorithms to detect sound events and localization, such as HMM, CNN, SVM, Random Forest, NMF, CRNN, VGG16, etc. Sound-related feature extraction methods, such as MFCC, Mel-Spectrogram, Mel band elements, ZCR, and wavelet features, and their importance in sound classification are also explained with a comparative analysis approach. The sections end with challenges with the existing approaches in different environments and feature extraction combinations.

Keywords Sound Event Detection · CNN · MFCC · SVM · HMM

1 Introduction

Sound Event Detection is one of the emerging research areas that can be implemented in multimedia, IoT, Robotics, Software modules, etc. Sound detection can help develop voice assistants, sound detection in various environments, border security systems, health systems,

✉ Sallauddin Mohmmad
sallauddin.md@gmail.com

Suresh Kumar Sanampudi
sureshsanampudi@jntuh.ac.in

[1] JNTU, Hyderabad, India

[2] School of Computer Science and Artificial Intelligence, SR University, Warangal, TS, India

[3] Department of Information Technology, JNTUH University College of Engineering, Nachupally, Kondagattu, Jagtial, TS, India

Published online: 02 August 2024

machinery defects, etc. Apart from that, the detection of sounds in various environments is a challenging task for researchers. The detection becomes complicated based on environmental conditions such as forest, rain, and under closed areas. Here, the sound can be generated in any direction from the environment and sometimes create reflections. A better approach is that the system should also need to find sound localization and detection. So, the research will be more complicated to find sound event detection along with the direction of arrival; in some applications like smart homes, robots, and the forest, localization should be an essential parameter for the system. We also need to process the polyphonic rather than the monophonic sound event in which the framework yields an arrangement of non-overlapping sound events. Polyphonic SED is fit for recognizing various sound events at a similar time. The quantity of sound events dynamic in an occurrence isn't known as prior, which presents an alternate degree of difficulty in detection. Polyphonic SED system requires multi-label categorization, which isn't broadly tested in sound data processing tasks. Acoustic sound detection provides a better result when machine learning techniques are implemented. The models should be developed with supervised and unsupervised learning-based algorithms. The current research strategies are mainly conducted with supervised learning-based algorithms. Here, we need different datasets concerning the model we need to evaluate. The learned data sets will be labeled or labeled weekly in the machine.

Several datasets are available for sound event detection and localization tasks, catering to various environments and challenges. These datasets include UrbanSound, AudioSet, DCASE challenge dataset with subsets like DCASE2013-2018, Freesound Dataset, TUT Sound Events 2017 and 2018, CHiME-Home dataset, MIMII Dataset for industrial audio anomaly detection [1–4], Detection of Sound Events in Urban Areas dataset, BUMD and etc. [5–7]. These datasets encompass a wide range of real-world sounds, annotated with labels for different sound events, enabling researchers and practitioners to develop and evaluate sound event detection and localization algorithms effectively.

In the classification process,, feature extraction has abecomes very important in bringing better classification results.

In sound event detection, various feature extraction techniques are employed to capture essential characteristics of audio signals. These techniques include Mel-Frequency Cepstral Coefficients (MFCCs) and Log Mel Spectrograms, which represent spectral features on a Mel frequency scale. Additionally, methods such as Gammatone Filterbank Features and Auditory Spectrograms are utilized to mimic human auditory perception. Deep Learning-based approaches, including Convolutional Neural Networks (CNNs), offer direct extraction of features from raw waveforms or pre-trained models like VGGish for log mel spectrogram embeddings [8–13]. Other techniques, such as Wavelet Transform, decompose signals into time–frequency representations, while rhythm-based features capture temporal patterns. Statistical properties, zero crossing rate, energy distribution, pitch-related information, and harmonic/timbral features further enrich the feature set. Combining these techniques provides a comprehensive representation of audio data essential for accurate sound event detection and localization tasks [14–17].

The existing research applied different algorithms such as the Hidden Markov model (HMM), non-negative matrix factorization (NMF), support vector machine (SVM), and random forest. Recent approaches use deep learning-based methods using deep neural networks (DNN), convolutional neural networks (ConvNet), recurrent neural networks (RNN), and convolutional recurrent neural networks (CRNN) [18–22]. All algorithms are implemented with key techniques like 1D,2D ConvNet, Multilayer CNN,GCCPHAT, BiGRU, LSTM and etc. The sound will be fragmented, and features extracted with MFCC, Mel Spectrogram, RMS, etc., then implemented various classification algorithms.

1.1 Motivation

In the domain of research on sound event detection, considerable survey studies have been conducted, with a primary focus on classifying sounds or acoustics across diverse environmental contexts. Despite the wealth of information provided in these studies, there are often certain gaps or missing components that researchers have acknowledged. However, recognizing the importance of addressing these limitations, researchers have made efforts to include relevant information based on the specific needs of their studies.In exploring the literature, we examined several survey papers that have high citations and a significant impact factor. Through this process, we have identified specific limitations in each of these papers. These limitations may range from methodological constraints to gaps in coverage or analysis. In doing so, we have tried to provide a comprehensive overview of the state of the art in sound event detection research, taking into account the challenges highlighted in previous studies.

Gabriel et al. [66] thoroughly categorized the techniques applied in the aforementioned scientific domains. It used standards from the literature to categorize sound source localization systems. Additionally, a comparison between traditional approaches predicated on the propagation model and approaches based on deep learning and machine learning techniques has been done. The most comprehensive knowledge possible about the potential applications of mathematical relationships, artificial intelligence, and physical phenomena in determining accurate source localization has been carefully considered. The paper also emphasizes the importance of these techniques in both military and civilian settings. However, the authors did not focus on the Datasets and feature extraction techniques.

Dang et al. [69] expressed the survey of sound event detection that involved the deep learning models and the challenge initiated by the DCASE 2016 to 2017. In this paper, the authors mainly focused on only various deep-learning models that are used for Sound classification, such as RNN,CNN, and CRNN. They did not explain the feature extractions, datasets and various results comparisons.

Nunes et al. [70], finding out if an object's sounds are typical or odd is part of their survey on detecting anomalous sounds. A Systematic Review (SR) examining research on anomalous sound detection employing machine learning (ML) methods presented in this paper. Between 2010 and 2020, 31 documents analyzed for this investigation. The most recent developments are covered, including evaluation techniques AUC and F1-score, ML models like Autoencoder (AE) and Convolutional Neural Network (CNN), and datasets like ToyADMOS, MIMII, and Mivia. The authors are not focused on the comparative study of various models.

Chandrakala et al. [72] surveyed sound event and scene representation and recommended appropriate machine-learning methods for audio surveillance projects. Different benchmark datasets are categorized based on the actual audio surveillance application scenarios. Several state-of-the-art methods are evaluated on two benchmark datasets intended for the sound event and audio scene detection tasks to obtain a quantitative understanding. Finally, future directions for improving environmental audio scene and sound event detection are delineated.

Teck Kai et al. [73] surveyed the sound event classifications in various directions. The authors perfectly explained the model implementations with a comparison of results. Sections are prepared as the parameters. However, in this survey, the authors did focus little on the feature extraction methods and datasets.Table 1 presents the various review research on sound event detection and localization and their limitations.

Table 1 Various review research on sound event detection and localization and their limitations

References	Environment	Year	Limitations
[66]	Source Localization	2023	Not described Datasets Not included the feature extractions Concepts Focused on only Neural Networks models
[67]	Audio Surveillance	2016	Not described about Datasets
[68]	homeSound	2017	Not included the comparative study of various models
[69]	Polyphonic Sound event detection	2017	Not described about Datasets Not included the feature extractions Concepts Focused on only Neural Networks models
[70]	Anomalous sound detection	2021	Not included the feature extraction techniques Less comparative study
[71]	Audio Surveillance	2020	Lack discussions on Machine learning models Lack of comparative study
[72]	Environmental Audio Scene	2019	Lack of comparative study on execution models
[73]	Polyphonic Sound Event Detection	2020	Not described about Datasets

1.2 Contributions

This paper endeavors to enhance the understanding of Sound Event Detection and Localization through a comprehensive review. Our primary focus lies in conducting a valuable comparative analysis of various research endeavors about sound event detection. We examine numerous recent models, highlighting their contributions and addressing significant challenges they present. Moreover, in a well-structured manner, we discuss key components of sound, including datasets, feature extraction techniques, machine learning models, and localization methodologies.

Below are the notable contributions of our research:

1. We conducted a thorough analysis of existing models in Sound Event Detection and Localization to enrich our review of the current state of research.
2. By identifying gaps in the literature, we have contributed to the progression of SED knowledge. Furthermore, we recommend future research directions and strategies to address these gaps effectively.
3. Our comparative study, through a meticulously prepared and extensive literature review, has provided valuable insights and enhanced the domain of Sound Event Detection.

The rest of my paper is presented as follows. Section 2 describes the different datasets that are used to detect the sound event. Section 3 discusses various feature extraction techniques used in various models. Section 4 illustrated the machine learning algorithms and Sect. 5 discussed the neural network models comparative study. Section 6 provides a review of the localization or direction of sound arrivals. Section 7 describes key parameters to evaluate sound event detection and localization. Section 8 discusses various environmental research scopes. Then followed by Sect. 9 with challenges. Finally, Sect. 10 illustrated the sound event detection related applications in the real world.

2 Datasets

The field of sound research has various kinds of specialized datasets, each prepared with different aspects of audio analysis and classification. Well-known datasets are prepared for environmental sounds, urban areas, parks, rooms, offices, motors, vehicles, traffic, the health sector, drones, animals, birds, etc. The UrbanSound dataset has ten classes of urban sounds, offering researchers a comprehensive collection ranging from street music to car horns and sirens [21, 26]. The AudioSet dataset by Google has millions of 10-s sound clips sourced from YouTube, covering over 600 labeled audio events, thus providing a rich resource for various research endeavors [28–30]. DCASE challenge datasets focus on real-world sound event detection and classification, offering recordings of various acoustic scenes and events. Freesound Datasets and FSD50K, drawn from the collaborative database Freesound, provide researchers with extensive collections of Creative Commons licensed sounds, facilitating tasks such as audio tagging and event detection [31, 32]. TUT Acoustic Scenes offers audio recordings from diverse acoustic environments supplemented with annotations for sound event detection and classification tasks. Speech Commands Dataset offers short audio clips of spoken words, key for keyword spotting and speech recognition research. ESC-10, smaller than ESC-50, streamlines experiments and educational purposes with its condensed 10-class environmental sound dataset [41, 42]. MIVIA Audio Events Dataset captures various events in indoor and outdoor settings, serving as a valuable resource for audio event detection and classification studies. CHiME Home and DESED datasets zoom into domestic environments, providing recordings of household activities and events for sound event detection and localization research [43–45].

When dealing with a small dataset in machine learning, the risk of overfitting becomes more evident. Overfitting occurs when a model learns the training data too well, capturing noise and irrelevant patterns that do not generalize to new data. Researchers have applied several strategies to handle this issue. Wang et al. [48] opted for fewer parameters in their approach to reduce the risk of overfitting. Hu et al. [49] implemented the cross-validation techniques, like k-fold cross-validation, to overcome the overfitting and estimate the model's performance by evaluating it on multiple validation sets.

Augmenting the dataset through techniques like data augmentation increases its effective size and helps the model generalize better. Additionally, feature selection can reduce model complexity by selecting relevant and eliminating redundant features. Monitoring the model's performance on a validation set during training and preventing early when performance declines can prevent overfitting. Furthermore, ensuring data using pre-trained models through transfer learning can leverage existing knowledge to improve model performance on small datasets.

Bubashait et al. [57] compared the various model accuracies on the Urbansound8K dataset. The features are extracted from urban sounds using Mel scale cepstral analysis (MEL) spectrum images. Sound processing is facilitated through an open-source library known as Librosa. The performance of CNN and LSTM models against a baseline ANN model in classifying. Evaluation of model performance is conducted using the UrbanSound8k dataset. The CNN model exhibits a lower performance, achieving an accuracy rate of 87.15% and an f1 score of 85.63%, compared to the DNN baseline and the LSTM model. Conversely, the LSTM model outperforms the CNN model, demonstrating superior accuracy on test data with a rate of 90.15% and an f1 score of 90.15%.

Fonseca et al. [63], FSD50K is an open dataset with over 51,000 audio clips, totaling more than 100 h of audio content, manually annotated across 200 classes from the

AudioSet Ontology. They provided a comprehensive account of the FSD50K creation process, tailored specifically to the unique characteristics of Freesound data. This included insights into encountered challenges and the solutions implemented. Furthermore, they conducted sound event classification experiments, presenting baseline systems and key insights into factors to consider when partitioning Freesound audio data for SER purposes.

Piczak et al. [64], the paper presented ESC-50 dataset, a fresh annotated collection of 2000 short clips spanning 50 categories of common sound events. It also offers a comprehensive collection of 250,000 unlabeled audio excerpts from recordings available through the Freesound project. Furthermore, it evaluates human accuracy in classifying environmental sounds, contrasting it with the performance of selected baseline classifiers utilizing features derived from mel-frequency cepstral coefficients and zero-crossing rate. Table 2 presents the various dataset and description.

2.1 Challenges

Diverse Characteristics Datasets are created based on specific environments with unique characteristics, such as environmental sounds, human sounds, Urban sounds, and vehicle sounds)—the process of preparing a generalized dataset across different types of sounds.

Feature Extraction Different sound types may require different preprocessing techniques and feature extraction methods based on the environment and sound type.

Domain Discrepancies Synthetic datasets reduce the complexity in processing pre-defined models but perform poorly in real-time environments.

Class Imbalance Datasets are prepared with unequal samples per class based on requirements and resource availability. This leads to biased models that perform poorly. For that, researchers need to perform augmentation and preprocessing additionally.

Overfitting With the limited data samples, the models cannot train properly.

3 Feature extraction

Sound feature extraction methods are techniques for extracting relevant information or characteristics from audio signals. These features are then used for various purposes, such as speech recognition, music analysis, sound classification, and more.

Feature extraction is vital for sound processing as it reduces high-dimensional sound data into representative features, reducing computational complexity while preserving essential information. These features facilitate efficient analysis, enabling speech recognition, music classification, and sound event detection [25, 26]. Moreover, they enhance noise robustness by focusing on discriminating aspects less affected by noise, promoting interpretability by revealing underlying sound characteristics and ensuring adaptability across diverse scenarios [27].

Time-domain features, derived directly from the amplitude values of the sound waveform, provide insights into the signal's temporal characteristics. Examples include zero-crossing rate (ZCR), energy, root mean square (RMS) amplitude, and temporal statistics such as mean and variance.On the other hand, frequency-domain features represent

Table 2 Different dataset with number of classes

Reference	Dataset	Number of classes	Number of samples	Durations of each sample	Sound Types
[54]	ARCA23K	70	23,727	Each audio clip is from 0.3 to 0.6 s	1.Music 2.Sounds of things 3.Natural sounds 4.Human sounds Animal 5.Source-ambiguous sounds
[55]	AudioSet	632	2982	0.3 to 30 s for each audio clip	1.Sound of things 2.Human sounds 3.Crackle 4.Animal 5.Wild animal 6.Music 7.Natural sounds 8.Home sounds
[55]	Usm -sed	26	24000	5 s for each audio clip	1.Miscellaneous sounds 2.Animals 3.human made sounds 4.vechile sounds 5.Construction site sounds 6.climate sounds
[57]	UrbanSound8k dataset	10	8732	<=4 s for each audio clip	1.vechile sounds 2.Music 3.Animals sounds 4.Children playing sounds 5.Environmental sounds
[58]	TUT Rare Sound Events 2017	3	1500	Around 3.5 min	Environmental sounds Nature sounds
[59]	Sound Events for Surveillance Applications	3	585	33 s for each audio clip	Casual sounds Environmental sounds

Table 2 (continued)

Reference	Dataset	Number of classes	Number of samples	Durations of each sample	Sound Types
[60]	GISE-51	51	16357	5 s for each audio clip	1.bird sounds 2.Nature sounds 3.Sound of things
[61]	FSDnoisy18k	20	18532	30 s for each audio clip	1.sounds of things 2.Environmental sounds 3.music
[62]	FSDKaggle2019	375	29266	0.3 to 30 s for each audio clip	1.sounds of things 2.Environmental sounds 3.music 4.animal sounds 5.insects sounds 6.human sounds
[63]	FSD50K	200	51,197	0.3 to 30 s for each audio clip	1.human sounds 2.crash cymbal 3.interior/domestic sounds 4.home sounds
[64]	ESC-50 dataset	50	2000	30 s for each audio clip	1.Animals 2.nature sounds 3.human sounds 4.interior/domestic sounds 5.exterior/urban sounds
[65]	CHiME Home	8	4378	4 s for each audio clip	Various sounds of home environment

the frequency content of the sound signal [28–30]. Techniques like STFT are employed for their extraction. Examples of frequency-domain features include spectral centroid, spectral bandwidth, spectral roll-off, and spectral flux. Cepstral features, such as MFCC, are derived from the spectral envelope of the sound signal. Pitch and harmonic features describe the pitch and harmonic structure of the sound signal. They include fundamental frequency (pitch), harmonic-to-noise ratio (HNR), and cepstral peak prominence (CPP), providing insights into the tonal properties of the audio.Temporal features include rhythm features like beat and tempo, as well as onset detection features that identify the starting points of sound events.Spectral features include spectral flatness, spectral contrast, and spectral entropy, which are useful for tasks like sound classification and acoustic scene analysis [31–33]. Wavelet and time–frequency features are derived from time–frequency representations of the sound signal obtained using techniques like wavelet transform or spectrograms [34, 35]. These features simultaneously capture time and frequency information, offering a detailed representation of the signal. Deep learning-based features represent a recent advancement. In this approach, features are learned directly from raw sound data using neural networks. Features extracted from CNN trained on spectrograms or RNN for sequence modeling have shown promising results in various audio processing tasks [56, 75]. Figure 1 illustrate the different kinds of feature type and relevant techniques.

Wang et al. [48] discussed an approach to detecting and locating sound events in real-world environments. For the audio-only part, they used ResNet-Conformer architecture as the primary acoustic model. For the audio-visual task, they utilized object and human body detection algorithms in videos to identify potential sound events, combining these findings with acoustic features to enhance detection. The model mainly used log-mel spectra features extracted from multichannel audio. They augmented the data using the ACS strategy and obtained about 192 h of data on the dev-test set of the STARSS23 dataset.

Jinbo et al. [49] report explained how they tackled Task 3 of the DCASE 2023 Challenge, which deals with identifying and locating sounds in real environments. They assessed the suggested approach using STARSS23's dev-test set. Using the data above generating strategy, they produce a significant amount of data, comprising 50000 5-s clips (dataset C) from computationally generated SRIRs and 2700 1-min clips (datasets A and B) from TAU-SRIR DB, where PANNs clean the sound event examples of B and C. The model implemented the CNN model to classify the sound by extracting the MFCC and Mel Spectrogram features. The model gained an accuracy of 82.2%.

Cheimariotis et al. [50] created a system designed to identify sound occurrences in-home sound classification. The model dealt with Task 4a of the "DCASE 2023, which is to identify 10 typical events that take place in homes within 10-s audio samples. The main components of the methodology were the application of data augmentation techniques to the mel-spectrograms that represented the audio clips, the use of BiGRU for sequence modeling, the fusion of these features with BEATs embeddings, and feature extraction through the use of a frequency-dynamic convolutional network enhanced with an attention module at each convolutional layer. The model has achieved the 0.798 accuracy.

Changmin et al. [86] suggested a model with a frequency dynamic CRNN structure. They first adjusted the sigmoid function by a temperature parameter to get a soft confidence value. Secondly, they employed a weak SED, which sets the timestamp to the duration of the audio clip and only makes weak predictions. Third, the PSDS scenario 2 benefited from adding the FSD50K dataset to the poorly labeled dataset. Next, the expanded dataset extracts features from the log-mel spectrogram. They used 128 mel-frequency bands, 256 sample hop lengths, and 2048 sample frame lengths to extract features. FDY-CRNN was used to implement the student and instructor models.

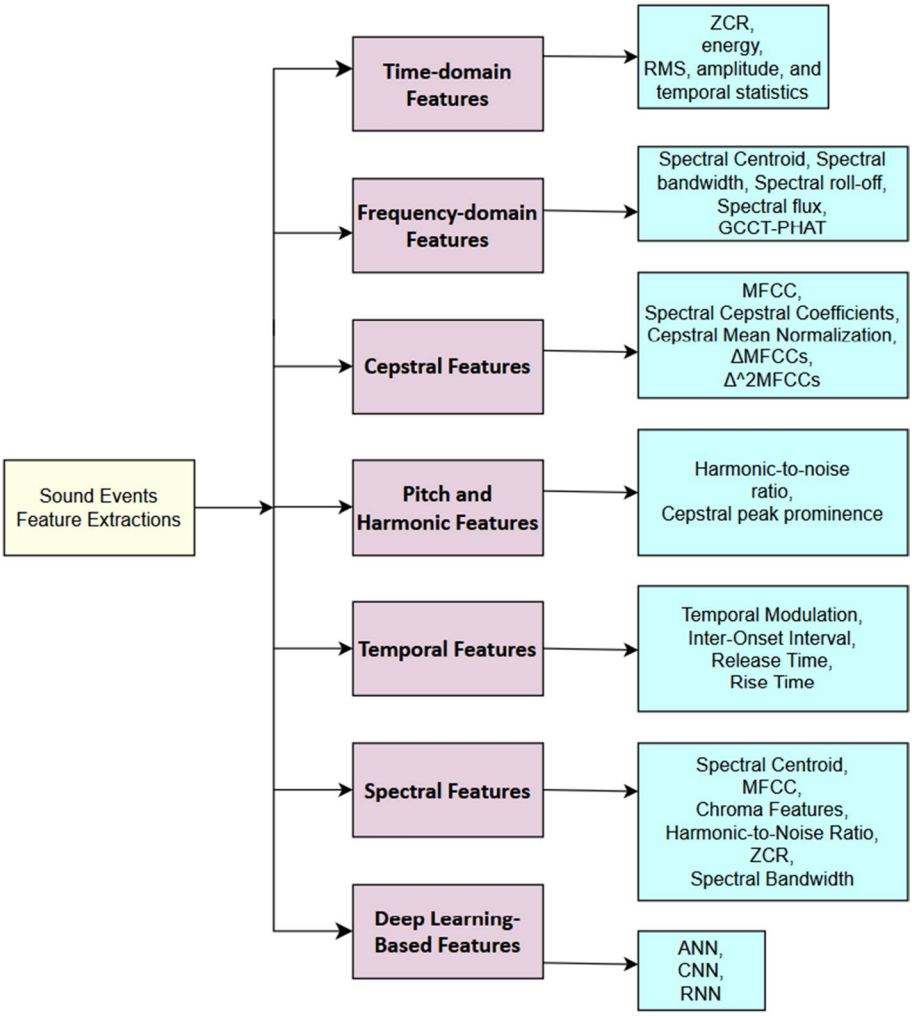

Fig.1 Different kinds of feature type and relevant techniques

The best PSDS scenario 1 of 0.473 and PSDS scenario 2 of 0.695 on the domestic environment SED real validation dataset.

Soo-Jong et al. [87] addressed the challenge of weakly labeled datasets using a novel time–frequency (T-F) segmentation framework. They utilized a CNN for segmentation and global weighted rank pooling for classification, and features are extracted by log mel spectrogram. Validation on DCASE 2018 data showed significant performance improvements over baseline scores, with F1 scores of 0.534, 0.398, and 0.167 achieved in audio tagging, frame-wise SED, and event-wise SED, respectively. Additionally, our method achieves an F1 score of 0.218 in T-F segmentation, a task previously unattainable. Table 3 presents the different feature extraction techniques and outcome with respect to ML based algorithms.

Table 3 Different feature extraction techniques and outcomes

Ref	Dataset	Augmentation	Features	Model	Result	Platform
[47]	Sound Event Detection for Driver Safety From Kaggle	No	Deep Features using CRNN	YOLOv5	Accuracy:81%	driving scenarios
[48]	Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset	Yes	log-mel spectra features	ResNet	Accuracy:92%	DCASE2023 Challenge
[49]	TAU-SRIR DB, FSD50K	No	MFCC and Mel Spectrogram	CNN	Accuracy:82.2%	REAL SPATIAL SOUND SCENES
[50]	DCASE2023	Yes	Mel-Spectrogram	BiGRU-CNN	F1 Score:61.3%	Domestic Environment
[51]	Synthetic	Yes	Log Mel Spectrogram	ResNet	Accuracy:82.9%	Urban Sounds
[52]	DCASE 2019 Task	Yes	Log Mel Spectrogram	RNN	F1 Score:57.2%	Home Sounds
[53]	MAESTRO Real data set	No	log-mel frequency	CRNN	F1 Score: 69.26%	cafe, city center, grocery store, metro station, and residential area
[84]	DCASE 2018 task1 and task 2	No	Only one Audio feature	U- Net	F1 Score: 64.4%	Natural sounds, Field recordings, Birds vocalizations
[85]	Urbansound8k,Custom Database	No	Mel Band Energies, Mel spectrogram	CNN	F1 Score:55.7%	home sounds, Vehicle sounds,Environmental sounds
[86]	FSD50K dataset	Yes	Log mel -spectogram	CNN	F1 Score:69.1%	human sounds, crash cymbal interior/domestic sounds, home sounds
[87]	DCASE Challenge 2020	No	log-Mel spectrograms	CNN	F1 score:58.2%	healthcare monitoring, bird call detection, urban sound analysis, multimedia content
[88]	DCASE 2016 Task3 dataset and DCASE 2017 Task3 dataset	No	Log-Mel spectrograms, MFCCs	CRNN	F1 score:63.4%	Environmental sounds, Vehicle sounds
[89]	DCASE 2018 Task 1 acoustic scene dataset, Freesound dataset	No	Log-Mel spectrograms	CNN	F1 score 53.4%	Vehicle sounds, Music,Animals sounds,Children playing sounds

Table 3 (continued)

Ref	Dataset	Augmentation	Features	Model	Result	Platform
[90]	acoustic events database JDAE TUKE	No	MFCC	HMM	Accuracy:94%	gunshot and breaking glass
[91]	ITC-Irst dataset	Yes	MFCC, LPCC	SVM	Accuracy:84%	Glass break, baby cry and gunshot
[92]	TUT-SED 2009 and TUT-SED 2016 dataset	No	log mel-band energies	CNN RNN	F1 score:68.0%	interior/domestic sounds, exterior/urban sounds
[93]	DESED dataset for DCASE 2020 Task 4	No	MFCC	CRNN	F1 Score:62.5%	Environmental sounds,human sounds
[94]	DCASE 2021 dataset	Yes	log-mel spectrograms	BiLSTM	F1 Score: 62.1%	Environmental sounds

3.1 Challenges

Feature extraction is essential in many machine learning and signal processing applications, especially audio processing. It involved transforming raw data into a set of measurable elements that a model can use as input for prediction.

Feature Selection The selection of features significantly impacts model performance. Different features may capture different values from the sound signal. A model's feature selection can significantly affect its accuracy and performance.

Computational Complexity Extracting deep features or complex spectrograms requires significant computational resources.

Real-Time Applications In the real-time scenarios such as live audio streaming or real-time speech recognition, audio data must be processed.

4 Machine learning sound event detection

Several types of research have been conducted on Sound Event detection with multi-channel and polyphonic sounds. They have been used to implement machine learning models such as HMM,NMF, SVM,Linear Regression, etc. SVM is utilized for its robustness in classification tasks, while KNN offers simplicity and effectiveness by classifying based on neighboring data points. Random Forest, an ensemble method, combines multiple decision trees for accurate classification. Decision trees are favored for their interpretability and ability to handle non-linear relationships in data. Linear regression, though primarily for continuous target variables, can be adapted for classification, although it's more commonly used in related tasks like sound source localization.

T. Heittola et al. [2] have discussed that they have implemented the 15 different types sound of 30-s length acoustic sounds as data sets for their research. They divided the data into two parts, the development and evaluation sets. The development set is again divided into training and tested sets to be used for cross-validation during system development in the implementation process of MFCC calculated as 40 ms frames and 40 mel bands, and 50% overlaps each. The classification of GMM has been used on the data set. The GMM classification is measured based on the accuracy and correctly classified segments. The authors considered the error rate and F-Score in fixed time intervals. The sound events in one second are compared with output and ground truth values. The author evaluated the scenario based on precision, recall, and F-score. The F1 score fro wind blowing is 14.2 in the residential area and water tap running F1 score is 41.2 in Home environment. *Kawaguchi* et al. [7] have implemented a model to classify the sound by using non-negative matrix factorization (NMF) and this model also compared with Semi-supervised NMF(SSNMF). This models mainly relate their results non-negative matrix under approximation(NMU).

Selver Ezgi et al. [8] have defined a model to detect the multimedia event. For that, they initially extracted the MFCC feature from sound-related data samples. They have used SVM to perform the classification. The occurrences in an actual class are represented in the rows, although the examples in a predicted class are represented in each column of the

matrix. The confusion matrix shows that significant gains may be achieved by executing the appropriate parameter optimizations. The system's overall recognition rate is generally good for different classes.

Parathai et al. [10] proposed the solution to classify events from a single noisy mixture. It consists of two main steps: separating the acoustic event and classifying the acoustic event. Complex Matrix Factor (CMF) is expanded through cooperation with optimal adaptation. L1 scattered offered adaptive CMF to decompose a noisy single-channel mixture, where the method encodes the spectrum plot and predicts the phase of the incoming signal in the time–frequency response. A Vector Machine Strategy (SVM) was applied on a one-to-one (OvsO) basis with an average supervisor to classify the unmixed audio into the category of the matched audio event. By moving the unmixed signals, the MSVM method divides the independent signals into blocks, after which the three features of each block are coded. OvsO uses the SVM approach to learn cepstral coefficients of frequency inclination, short-time energy, and short-time zero interference rate from several classes of audio events.

Huy Dat et al. [11] have introduced a model with SVM classification that used a distribution of subdomain temporal envelope (STE) and kernel technologies for subdomain probability distance (SPD). The generalized gamma modeling, well designed to characterize the sound and probability distance core provides the closed-shape solution for calculating the convergence distance, greatly reducing the computations price. Experiments were carried out using a database of 10 various categories of sound events. The proposed classification style outperformed traditional SVM classifiers with cepstral frequency inclination coefficients significantly, according to the findings (MFCCs).

Xianjun et al. [14] introduced a strategy outlined in their study involving the utilization of a pre-trained CNN to extract bottleneck features coupled with random forest classifiers for event detection. The study comprehensively details these techniques along with their practical applications. Additionally, the authors propose a method to incorporate context into the classification process by modeling the temporal evolution of event classes using an HMM. Through rigorous evaluation of two publicly available datasets, TUT Acoustic Scenes 2017 and TUT Sound Events 2017, the authors demonstrate the effectiveness of their methodology and achieve an accuracy of 91%. *Yuanjun Zhao* et al. [15] introduced a novel approach to sound event identification leveraging multiple optimized kernels. Their method demonstrates improved categorization performance through the integration of diverse kernels. The technique involved training several SVM utilizing various kernel functions and aggregating their outcomes for decision-making, as elaborated in their research. Additionally, the authors advocated for a grid search approach to fine-tune kernel parameters effectively. Through extensive evaluation on publicly available datasets—TUT Acoustic Scenes 2017 and TUT Sound Events 2017—the authors showcase the effectiveness of their methodology, achieving state-of-the-art results in terms of accuracy and F1-score. Table 4 illustrate the Different Machine Learning Models to Find Polyphonic Sound Event Detection.

5 Neural networks models

Neural network architectures such as CNN and RNN play a vital role in accurately classifying large datasets. Emerging CNN variants like VGG16, VGG19, ResNet, AlexNet, MobileNet, DenseNet, and EfficientNet, among others, have been used in classification

Table 4 Different machine learning models to find polyphonic sound event detection

Reference	Scope	Algorithm	Feature Extraction	Sampling Rate	signal-to-noise ratio (SNR)	Achieved
Küçükbay et al. [8]	Audio based event detection	SVM	MFCC	44.1 kHz		recognition performance of the baseline system (48%) is increased by 7%
Parathai et al. [10]	Noisy Sound-Event Mixture Classification	SVM and NMF	MFCC,ZCR	44.1 kHz	20 dB	The average performance improvement of the proposed adaptive CMF method against the uniform constant sparsity was 1.32 dB SDR
Tran et al. [11]	Sound Event Detection	SVM	MFCC	16 kHz	SNR of 12–18 dB	Achived p-value 0.20
Phan, Huy et al. [13]	Acoustic Event Detection and Classification	Random Forests	ZCR	16 kHz		Improvements of approximately 9% and 10.8% over the best baseline system HMM
Xia et al. [14]	acoustic event detection (AED)	SVM and Random Forest	log-spectral parameters, zero-crossing rate, spectral bandwidth	44.1 kHz		accuracy of 91.56%
Komatsu et al. [23]	Acoustic Event Detection	non-negative matrix factorization (NMF)	Croma features and spectral bandwidth	16 kHz	10,15,and 20 dB	improvement in F-measure by up to 60%
Komatsu et al. [26]	Acoustic Event Detection	semi-supervised non-negative matrix factorization (NMF)	STFT and MFCC	44.1 kHz	-6,0, 6 dB	F-measure by the proposed method with semi-supervised NMF is improved by as much as 11.1%

Table 4 (continued)

Reference	Scope	Algorithm	Feature Extraction	Sampling Rate	signal-to-noise ratio (SNR)	Achieved
Grondin et al. [28]	Bird sound spectrogram decomposition	non-negative matrix factorization (NMF)	MFCC	22.05 kHz)	20 dB	relative error reduction with respect to the conventional MFCC system of approximately 20.14% is achieved

tasks. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models are employed to preserve previous layer outputs in current layers, often extended with Bi-directional LSTM and Bi-directional GRU for enhanced performance [4, 36, 37]. These architectures extract complex features from audio data, capture temporal dependencies, and sense hierarchical patterns within sound events. Large labeled datasets and techniques such as transfer learning and SED models achieve good accuracy and generalization across diverse applications, from environmental monitoring to smart home systems. Despite noise robustness and scalability challenges, neural network models hold promise in elevating audio analysis systems and driving real-world implementations of SED forward. The CNN-related models are adequate for capturing local spectrogram features, representing sound as a function of time and frequency. RNNs, LSTM, and GRU models are adept at handling temporal dynamics in sequential data, making them suitable for modeling long-range dependencies in audio sequences. On the other hand, CRNNs merge the strengths of CNNs and RNNs, enabling them to capture local and temporal features simultaneously [38–40]. The choice of model depends on factors such as data characteristics, computational resources, and task specifics. Each architecture has its advantages and limitations, and selecting the most appropriate model requires careful consideration of these factors to ensure optimal performance in sound event detection tasks.

Annamaria et al. [1] have investigated "Sound Event Detection in the DCASE 2017 Challenge," which presented an analysis of the DCASE 2017 challenge's sound event detection task, which aimed to advance state-of-the-art sound event detection. The authors describe the dataset and the evaluation metric used in the challenge in detail, highlighting the difficulties associated with sound event detection, such as varying acoustic conditions and class imbalances. They also talk about how deep neural networks (DNNs), CNN and RNN are used in high-performing systems, as well as data augmentation and ensemble learning. Overall, the paper provides a comprehensive analysis of the DCASE 2017 challenge's sound event detection task, providing insights into the task's challenges and the state-of-the-art techniques used to address these challenges.

Hyungui Lim et al. [3] introduced the 1D ConvNet to detect rare sound events. The authors used the 1D convolutional neural network, RNN, and LSTM algorithms with long-amplitude mel-spectrogram as input acoustic features. Frame-wise log-amplitude mel-spectrogram fed into our proposed model, and the model returns the output for every incoming sequence. They implemented the spectral-side 1D ConvNet that enables frame-level investigation. Their research used the two layers of RNN with each 128 LSTM unit. They have applied the unidirectional backward RNN-LSTM procedure to produce more accuracy in the system. The performance on the test set of the development dataset yields an error rate of 0.07 and an F-score of 96.26 on the event-based metric.

Adavanne et al. [5], this paper proposed utilizing CBRNN to detect bird calls, treating the Bird Audio Detection (BAD) challenge as a SED task. CRNN architecture combines the modeling capabilities of CNN, RNN, and fully connected (FC) layers. In this, CRNNs expanded to handle multiple feature classes, with CNN feature maps processed using a bidirectional RNN, forming the convolutional bidirectional RNN. The model with CBRNN achieves an AUC measure of 95.5% on five cross-validations of the development data and 88.1% on unseen evaluation data.

Qiuqiang et al. [6] proposed a model integrated with CNN-Transformer, which is similar to CRNN.In their approach, they implemented threshold optimization like mean average precision(mAP) for SED. The authors implemented the improved architecture of LSTM called BiGRU with CNN. Model automatic threshold optimization system achieves state-of-the-art results, with an audio tagging F1 score of 0.646, surpassing the score of 0.629

obtained without threshold optimization, and a sound event detection F1 score of 0.584, outperforming the score of 0.564 without threshold optimization.

Kyoungjin et al. [24] have implemented a model for SED using CNN. For that, they have chosen the multi-channel environment, and they have drawn the STFT coefficient from them. This model also extended with weighted prediction error (WPE). MVDR beamforming is carried out with the source and noise masks estimated by the DNN. Likewise, the experiment goes on with multiple test cases. In this paper, evaluation of the metrics has been done on binary analysis on the entered test data of true positives (TP), the number of false positives (FP), and the number of false negatives (FN) is aggregated. Here the authors evaluated the metrics for Precision (P), Recall (R), and F-Score. The entire proposed system executed end is explained in three parts. Here 1st, the first part is dereverberation, 2nd, the part is MVDR beam forming. The final stage will be CRNN-based SED. By the final stage, the proposed system detects the presence and absence of sound events.

Xu et al. [27] introduced a gated convolutional neural network and a temporal attention-based localization technique for audio classification. The model employed a CRNN with learnable gated linear units (GLUs) applied to the log Mel spectrogram. Additionally, they introduced a temporal attention mechanism across frames to predict the event locations within a chunk derived from the weakly labeled data. The model excelled in both sub-tasks of the DCASE 2017 challenge, achieving an F-value of 55.6% and an Equal Error of 0.73, respectively.

Adavanne et al. [29] aimed to focus on Sound Event Localization and Detection (SELD) for the DCASE 2019 challenge. A baseline method utilizing a convolutional recurrent neural network establishes benchmark performance on this reverberant dataset. The results consider different numbers of overlapping sound events and varied reverberant environments. Overall, SELDnet demonstrated slightly superior performance on the FOA dataset compared to the MIC dataset. Furthermore, SELDnet exhibits enhanced performance in scenarios devoid of polyphony across datasets. Notably, the SELDnet model trained with data from all five environments displays the best performance, particularly excelling in the initial environment with an F1-Score of 85.0 and an error rate of 0.25.

Jingyang et al. [30] presented a comprehensive approach to tackle the SELD task, consisting of data augmentation, network prediction, and post-processing stages. Our approach employed the CRNN architecture for model prediction. Given the scarcity of data in the challenge setting, we advocate for data augmentation to enhance the system's performance. Evaluation of the DCASE 2019 Challenge Task 3 Development Dataset reveals our system achieves approximately a 59% reduction in Sound Event Detection (SED) error rate and a 13% reduction in directions-of-arrival (DOA) error compared to the baseline system, specifically on the Ambisonic dataset.

Turab Iqbal et al. [31] focused on two-stage polyphonic sound event detection and localization, employing log mel features for event detection and intensity vector along with Generalized Cross Correlation (GCC) for localization. These features are fed into a microphone array system. Log mel features were primarily utilized for event detection, while intensity vector and GCC features employed for precise localization. Additionally, an intensity vector in log mel space and GCC with phase transform (GCC-PHAT) features was utilized for DOA estimation. The methodology involved constructing 2DCNN layers, referred to as feature layers, comprising four groups of 2D CNN layers with subsequent 2×2 average pooling. Each group of CNN layers comprised two 2D Convs with a receptive field of 3×3, a stride of 1×1, and a padding size of 1×1. The two-stage approach yielded promising results with an error rate of 0.13, an F1-Score of 0.930, and a DOA error of 6.61 degrees.

Ying Tong et al. [32] proposed a model operated by taking consecutive spectrogram time frames as input and generating two outputs simultaneously. Firstly, it conducts Sound Event Detection (SED) through multi-label classification on each time frame, effectively capturing temporal activity for all sound event classes. Secondly, it performs localization by estimating the 3-D Cartesian coordinates of the direction-of-arrival (DOA). Compared to various baselines, including SED and DOA estimation methods, the proposed approach showcases robustness across diverse structures, adaptability to unseen DOA values, resilience to reverberation, and effectiveness in low SNR scenarios. Within this architecture, local shift-invariant features within the spectrogram are learned through multiple layers of 2D Convolutional Neural Networks (CNNs). Each CNN layer utilizes Rectified Linear Unit (ReLU) activation on dimensions of $3 \times 3 \times 2C$ receptive fields along the time–frequency-channel axis. This model achieves an accuracy of 87%.

Thi Ngoc Tho et al. [34], a novel approach was proposed to estimate Sound Event Localization and detection by employing a CRNN-based Sequence Matching Network (SMN). The authors accounted for overlapping sounds and their onset and offset parameters, aligning them with the active segments of the output and incorporating a DOA estimator alongside sound classes. Implementation involved utilizing BiGRU coupled with fully connected layers. In the second phase, a CRNN-based SMN was trained to align the output sequences of the event detector and DOA estimator. The estimated DOAs were then associated with relevant sound classes. This modular and hierarchical approach significantly enhanced the performance of the SELD task across all evaluation metrics. The proposed ensemble achieved a localization error of 9.3°, a localization recall of 90%, and secured the second position in the team category of the DCASE2020 sound event localization and detection challenges. Table 5 presents the Different Neural Network Models to Find Polyphonic Sound Event Detection. Table 6 illustrates the different Neural network models with various parameters.

6 Approaches for sound event localization

Many overlapping sound waves in different frequency bands make up noise. Sound event detection and localization are two tasks that work together to identify the actions of sounds like horns and dogs barking heavy engines when they're active, calculating their separate geographical position courses, and connecting textual labels with sound events. Generally, the arrival angle of sound direction is challenging to detect. Several researchers have been involved in finding the localization of sound events. Sound event localization, crucial in various applications like surveillance, robotics, and augmented reality, uses several approaches to accurately determine sound sources' spatial coordinates. One common method involves microphone arrays, where the Time Difference of Arrival (TDOA) or Direction of Arrival (DOA) of sound signals across multiple microphones is analyzed. GCC-PHAT is one widely used signal-processing algorithm for sound event localization, especially with microphone arrays. It works by finding the time delay between signals received by different microphones due to variations in sound source arrival times, known as Time Delay of Arrival (TDOA) [45–47]. GCC-PHAT calculates cross-correlation between microphone signals to identify the time delay that maximizes correlation. Before computing cross-correlation, signals undergo a preprocessing step called Phase Transform (PHAT), which normalizes signals based on their phase to reduce the influence of signal magnitude. This normalization improves TDOA estimation accuracy and sound source

Table 5 Different neural network models to find polyphonic sound event detection

Reference	Scope	Algorithm	Key Operation	Sampling Rate	signal-to-noise ratio (SNR)	Achieved
Hyungui et al. [3]	Rare Sound Event Detection	CRNN	MFCC,Mel Spectrogram	44.1 kHz	-6, 0, 6 dB	The proposed system has achieved the 1st place in the challenge with an error rate of 0.13 and an F-Score of 93.1
Xia et al. [15]	Multiple Optimized Kernels in CNN model	CNN	log mel-band energies,MFCC	44.1 kHz	-	The model achieved the F1 Score 61.7% on five multiple kernals
Stoller et al. [16]	Audio Source Separation	Wave-U-Net	Mel-Spectrogram	22.05 kHz	20 dB	signal-to-distortion (SDR) achieved 50% of the time
Cao et al. [31]	Sound Event Localization and Detection	CNN	log mel-band energies,MFCC	32 kHz	20 dB	Achieved Two Stage DOA error 6.61° only
Ying Tong et al. [33]	Sound Event Localization and Detection	CNN	Mel Spectrogram and GCC-PHAT	48 kHz	-	Achieved DOA error 3.7° only

Table 6 Different neural network models with various parameters

Reference	Feature Extraction	Algorithm	Frame size	Error Rate	F-Score
Jeongsoo Park et al. [3]	MFCC	1D Convolutional Neural Network,	46 ms	0.13	93.1
Adavanne et al. [5]	Deep Features	BiGRU-CNN	40 ms	0.84	43.3
Keisuke et al. [20]	GCC-PHAT,MFCC	CNN-BiGRU	40 ms	0.756	42.17
Hongning Zhu et al. [21]	MFCC,Mel-Spectrogram	CRNN	40 ms	0.609	59.0
Kyoungjin et al.[24]	MFCC,Mel Band Energies	DNN	40 ms	0.13	92.8
Kong et al. [27]	Mel Spectrogram	CNN		0.42	72.5
Grondin et al. [28]	MFCC,GCC-PHAT	BiGRU-CRNN	43 ms	0:14	92:2
Archontis Politis et al. [29]	STFT,MFCC	2D CNN	40 ms	0.25	85.0
Jingyang et al. [30]	Mel-spectrogram and STFT	CRNN	40 ms	Mel-0.18 STFT-0.27	Mel-89.0 STFT-84.1
Turab Iqbal et al. [31]	GCC-PHAT	BiGRU-CNN	40 ms	-	93.0
Adavanne et al. [32]	Mel Spectrogram and GCC-PHAT	CRNN	100 ms	0.13	89.8
Ying Tong et al. [33]	MFCC	CNN	40 ms	9.5	94.1
Nguyen et al. [34]	Wave-let transmission and Mel Spectrgram	CRNN	12.5 ms	0.359	71.2
Bongjun et al. [36]	Mel band Energies	CNN	25 ms	0.523	63.8
Xianjun et al. [37]	MFCC	CRNN,SVM	40 ms	0.59	48.32

localization, making GCC-PHAT effective in mitigating reverberation and noise for more precise localization results.

Grondin et al. [28] The two CRNNs to perform sound event detection with Time Differences Of Arrival (TDOA) and localization with DOA on the proposed system were based. In this paper, the system has four microphone arrays, thus combining results with six pairs of microphones to provide the 3-D Direction of Arrival (DOA) and the final classification. The proposed sound event detection and localization were submitted to the DCASE 2019 challenge. This research also performs CRNN architecture which uses both the spectrogram and the GCCPHAT features to perform the SED and estimate TDOA.

Archontis Politis et al. [29] have presented sound event localization and detection based on the DCASE 2019 challenge. Here the entire research done with a multi-room reverberant dataset is provided for the task. In this approach, the DNN has been implemented for classification and regression. The average SNR sound event was sampled at 30 dB, and in this research, authors considered the temperature also as a parameter when finding sound detection and localization. The model implemented CRNN with bi-directional GRU to identify the direction of arrival separate from sound event detections.

Zhang et al. [30], the main goal of this paper is to detect the polyphonic sound event and localization. The authors explained the concept: data augmentation, network prediction, and post-processing stage. In the last stage of post-processing, they proposed an idea like prior knowledge-based regulation(PKR). By using PKR concept, they brought the average value of localization prediction. They proved that their process reduces the mean square error. They have implemented the CRNN to find the localization and sound detection. The training set of SED jointed with STFT, and DOA jointed with Mel-spectrogram.

Adavanne et al. [32] have explained the key concept of direction of arrival to find the localization of sound by using the phase and magnitude spectrum of sound waves from multiple directions. In this research, the localization was identified by defining of 3-D Cartesian coordinates of DOA. In this method, the phase and magnitude of the sound signal are evaluated separately to achieve a better result on localization. The entire process on the baseline of CRNN.

Ying Tong et al. [33] have implemented a new SELD method based on multi directional of arrival beam forming and multitasking learning. Multiple-DOA beam forming is used to achieve signal separation and provides a varied sound field description. For SED and sound source localization (SSL), we plan two networks and utilize a multitasking tutorial for SED, where the task associated with SSL acts as regulation. Instead of estimating the signal from DOA For each source, they suggested doing several DOA for the formation of the beam, which directs the beams evenly towards different DOA, such as sources that distribute spatially and noise signal can be separated. DOA output signals are used to extract features for both SSL and SED. Based on CPS and SPP, the steering vector is calculated for each DOA and used to design beam converters for many DOAs. The three-task learning system is used, which uses both regression and Criterion SSL based on classification for organizing the network SED.

Nguyen et al. [34] Sound event detection and localization have to be done in two separate ways, one for detection and the other for localization. Here the detection depends on time–frequency patterns to distinguish different sound classes. Localization and direction of the sound estimation use magnitude or phase differences between microphones. Here they implemented the trained CNN to frequency patterns with the magnitude and phase of the signal to execute the model. The system also extended with a new concept: sequence matching network (SMN). Initially, the model detects the sound by using CRNN to detect the sound events and a single-source histogram method to estimate the DOAs. The next

level model was implemented with a trained CRNN-based sequence matching network to match the two output sequences of the event detector and DOA estimator.

Trowitzsch et al. [35] have proposed a system that uses a robotic binaural system to detect sound events and localization. Presents an approach that robustly binds localization with detecting sound events in a robotic binaural system. We use recreations of a complete set-up of test scenes with different co-happening sound sources and propose execution measures for deliberate examination of the effect of scene intricacy on this isolated identification of sound sorts. Investigating the impact of spatial scene plan, we show how a robot could work with a superior through an ideal head pivot. Besides, we explore the exhibition of isolated identification given conceivable restriction mistakes just as a blunder in assessing the number of dynamic sources.

Xianjun et al. [37], in this study, blanket representations of SELD are generated using traditional microphone array signal processing, they implemented a new SELD method based on multidirectional of arrival (DOA) beam forming and multitasking learning. Multiple-DOA beam forming is used to achieve signal separation and provides a varied sound field description. For sound event localization and sound source localization (SSL), they planned two networks and utilized a multitasking tutorial for SED, where the task associated with SSL acts as regulation. They evaluated the model that instead of estimating the signal from DOA For each source, we suggest doing several DOA for the formation of the beam, which directs the beams evenly towards different DOA, such as sources that distribute spatially and noise signals can be separated. DOA output signals are used to extract features for both SSL and SED. Based on CPS and SPP, the steering vector is calculated for each DOA and used to design beam converters for many DOAs. The three-task learning system is used, which uses both regression and Criterion SSL based on classification for organizing the network SED. Experimental results using the DCASE2019 SELD task database show that the suggested technique obtains the most current results. Table 7 describes the sound arrival direction, localization-related models, and key approaches.

7 Acoustic parametric analysis

Acoustic monitoring has become a widely used process for assessing the status and diversity of sound-producing. Different acoustic metrics are utilized to find the accuracy in sound detection, such as Acoustic Complexity Index(ACI), Acoustic Diversity Index(ADI), Acoustic Evenness Index(AEI) and etc. Extensive analysis needs to identify and detect the audio signal of the various types. This process consumes more time. Studies conducted in various environments and geography regions release errors in Correlation among audio diversity and biodiversity indicators, indicating a need for studies to evaluate acoustic monitoring.

Moreno-Gómez et al. [38] have investigated the concept of acoustic indices in the rainforest and biodiversity hotspots. Seven audio indicators are evaluated to assess the reliability as surrogate models for variations in the bird and the tadpole animals. They have used three automated voice recordings they are SM1, SM4,SM3, where every device is put into just one sampling station. As the first approach to assess the relationship between birds and the indicators of tadpole richness and vocal diversity has conducted correlations among the variables for every station, we used the bootstrap technique with the 1000 iterations. For every iteration, received feedback randomly with replacing and ran the Correlation analysis. The top-ranked model, M1, encompassing ACI, H, Hf, Ht, and BI, was identified as

Table 7 Models for sound event localizations

Reference	Features Extraction	Sources	Algorithm	Array	SELD	DOA	DOA error
Grondin et al. [28]	GCCPHAT features	One	BiGRU-CRNN	microphone arrays	Yes	Yes	7.4°
Archontis Politis et al. [29]	STFT, Mel Band Energy	Multiple	2D CNN	Microphone arrays	Yes	Yes	23.1° to 30°
Wenhao Ding et al. [30]	GCCPHAT, Mel Band Energy	Multiple	CRNN	Microphone arrays	Yes	Yes	24.8°
Qiuqiang Kong et al. [31]	GCC-PHAT	-	BiGRU-CNN	Microphone arrays	Yes	Yes	6.61°
Adavanne et al. [32]	Phase and Magnitude Spectrum	Multiple	CRNN	Generic	Yes	Yes	12.1°
Ying Tong et al. [33]	MFCC, GCC-PHAT	Multiple	CNN	Microphone arrays	Yes	Yes	3.7°
Nguyen et al. [34]	GCC-PHAT	Multiple	CRNN	Microphone arrays	Yes	Yes	12.1°
Trowitzsch et al. [35]	GCCPHAT, Mel Spectrogram	Multiple	CNN	Microphone arrays	Yes	No	-

the most suitable. This model featured fixed effects for intercept, bird richness, and anuran richness, random intercepts for station, hour, and month, and random slopes for birds and anurans by station. With AICc weights exceeding 0.95 and Delta-AICc values surpassing 7 compared to the second-ranked model, M1 demonstrated substantial support, suggesting that the factors within it effectively elucidate the variance observed across the five acoustic indices. Notably, ACI, BI, and Ht exhibited the highest effect sizes for species richness, with Ht particularly influencing bird richness significantly.Their AICc weights were below 0.5, and Delta-AICc values were less than 1 about the subsequent models, indicating that other model possibilities shouldn't be disregarded. Figure 2 present the result scenario of proposed model of [38].

Eldridge et al. [39] have investigated sounding out acoustic metrics with an independent device recorder, which enables the large-scale monitoring of the audio and the audio scanning. Nearly 26 vocal indices are calculated, and a comparison has been made to the observed differences in species diversity. The Five Audio Diversity Indicators (Voice Dynamic Index, Audio Diversity Index, Audio Equivalence Index, Audio Entropy, and the normal difference audio index) and three simple audio descriptors were evaluated. Highly signified correlations are 65% among the audio indicators, and the richness of bird species was observed in temperate habitats. Poor bonding has been observed in geo-tropical habitats that host multiple types of sounds other than birds. Multivariate classification analysis showed that each habitat has a distinct acoustic scene, and AIs trace the differences that are observed in the community composition that depend on the habitat. Rapid Audio Survey (RAS), was suggested as not invasive and an approach to the assessment of biodiversity, and it is interesting gaining to research people and policymakers. They analyzed the ACI value with 0.49, ADI value greater than 0.5

Felipe Carmo et al. [40] have investigated acoustic indices in the rainforest. They arranged the model at the point of 12 stations. It follows the bird monitoring protocol using the GPS from the range of 350 to 500 m.They implemented the ARU method, which is called Autonomous recorder units. It is one of the sampling methods. They performed the automatic sound monitoring using the 9 ARU'S, SM2.One of the

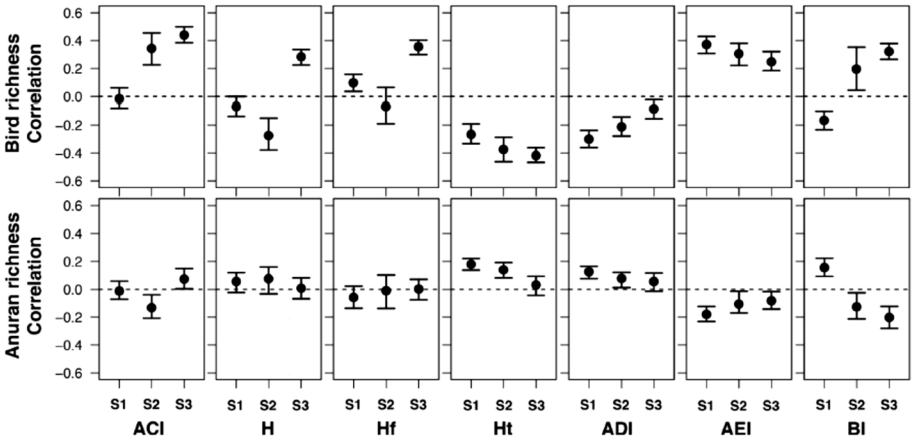

Fig. 2 Results from bootstrapped correlations computed by Spearman tests between bird and anuran richness and acoustic diversity indices in the three stations [38]

methods stopped recording, so it is not considered in the analysis part. To avoid the recordings of the noise of the humans, like when we stepped on the branches of the forest floor, they used the ARU method with the 12-point stations. Each recorder was used with 2 Omni directional microphones. They investigated their research with ARU method for the 18 days; the first nine were used with pc sampling. The data collected were used to differentiate between the acoustic indices during the researcher's presence and absence.

Fairbrass et al. [41] have investigated their research on acoustic indices measuring biodiversity in urban areas. They used acoustic recordings for 7 days to capture the weekly daily activities. In order to increase the variability in the recordings of the biotic sounds. The (acoustic indices) AI'S were tested using the threshold frequency.For consistency, they tested all the AIs using an upper threshold value of 12 kHz.They have acknowledged that frequencies are included above the threshold for the BI and NDSI. Acoustic diversity is identified by the various sound events associated with the same sound class identified in each recording. Most of the sites were influenced by both low and high-frequency sound activities. Anthropogenic is one of the sounds. In the dataset, it is composed of a wide variety of sound types, like traffic sounds, followed by human voices, crackles from the recorders, electrical buzzes, and the environment.

Machado et al. [42] have performed their survey on bird communication. They have assessed how two specific records (the acoustic variety list – ADI – and standardized distinction sounds cape file – NDSI) reflect bird species lavishness and organization in an ensured region close to Brasilia city. Their research has conjectured that ADI ought to mirror the qualities of birds in the cerrado and in the exhibition woods, i.e., with higher qualities in display timberland than in the cerrado. Based on natural surroundings structure, they have likewise guessed that NDSI ought to be lathey haver in less intricate territory, and lothey haver in regions near urbanized regions. They have evaluated 30 areas by introducing programmed recorders to create 15 min wave documents Manual investigation of the documents uncovered the presence of 107 bird species our outcomes shothey haver that ADI was altogether connected with species lavishness, being higher in exhibition woods than in the cerrado. Acoustic files for biological investigations and biodiversity checking are one of the programmed approaches for information examination. As per their assessment, the relationship of acoustic variety record and bio variety by applying a straight model looking at the mean ADI esteem and the bird species extravagance enrolled in every area.

Siddagangaiah et al. [46] have presented a noise-resilient approach for detecting biophonic sounds from fish choruses based on complexity-entropy (hereinafter referred to as C-H). The C-H approach was tested with data collected in Changhua and Miaoli (Taiwan) in the spring of 2016 and 2017. Miaoli was subjected to constant maritime traffic, which resulted in a 10 dB rise. They suggested that using the C-H technique could assist overcome the limits of acoustic indices in noisy maritime environments. They developed an approach for detecting fish choruses based on the C-H method and compared its detection performance to AIs such as ACI, ADI, and BI. The fish chorusing was shown to be favorably connected with C, but negatively linked with H, resulting in $|r| > 0.9$. The use of marine acoustic biological activity as a proxy for addressing trends in biodiversity levels and ecosystem functioning could be very useful. The C-H approach was developed and tested in marine habitats to fill in the gaps left by other indices originally designed and utilized for terrestrial settings. Noise from shipping operations or natural sources such as wind and tides had no effect on the C-H technique, which was found to be strongly linked with fish chorusing. When used in

conjunction with other current acoustic indices, the C-H technique could be a useful tool for managers and decision-makers to track changes in the makeup of animal communities. Table 8 presents the various parameters to evaluate the sound event detection in different accuracy levels.

8 Sound processing in different environments

The research on Sound event detection become more emerging in the present days and implemented in diverse domains for different purposes. SED finds application in environmental monitoring, which tracks environmental sounds like bird calls, animal noises, and weather patterns, aiding ecological studies and disaster management. In surveillance and security, SED identifies suspicious sounds such as glass breaking or gunshots, enhancing public safety measures. Integrating SED into smart home systems enables the recognition of specific events like smoke alarms or appliance malfunctions, enhancing home safety and convenience. The healthcare system also enhanced the technology with a sound classification that benefits in monitoring patient conditions by identifying medically suitable sounds such as heartbeats. SED detects equipment failures or anomalies in industrial environments as automotive safety systems utilize SED to detect sounds like horns or sirens, contributing to road safety. Speech recognition systems leverage SED to filter out background noise, improving accuracy. Entertainment and gaming applications utilize SED for audio experiences and interactive events. Additionally, echo monitoring systems serve various purposes, including sonar systems for underwater object detection, medical ultrasound imaging for visualizing internal structures, radar systems for tracking objects, and structural health monitoring for assessing structural integrity. These applications underscore the versatility and significance of SED and echo monitoring systems across multiple fields, promising further advancements in the future.

Imoto et al. [79] introduced a novel SED method based on multitask learning(MTL) of SED and ASC, employing soft labels for acoustic scenes to better represent the nuanced relationship between sound events and scenes. Experimental evaluations conducted on the TUT Sound Events 2016/2017 and TUT Acoustic Scenes 2016 datasets demonstrate that our proposed approach enhances SED performance by 3.80% in F-score compared to conventional MTL-based methods. Specifically, the proposed CNN-BiGRU model achieves an F1 score of 49.82% and an error rate of 0.691, outperforming the baseline model with an F1 score of 42.17% and an error rate of 0.756.

Mingying Zhu et al. [80] introduced a new method to classify bird sounds automatically. It starts by dividing the audio into sections using a sliding window, selecting the five sections with the highest energy. Then, it extracts important features using a technique called orthogonal matching pursuit. For a dataset with 14 bird species, we achieve a classification accuracy of 98.96% and an F1-score of 98.93% using a 2D-CNN-v2. The highest accuracy for another dataset with 18 species is 97.82%, and the F1-score is 97.47% 2D-CNN-v2 with Bark-scaled SFM as input. The dataset xenocanto encompasses 14 bird species prevalent in Queensland, Australia, sourced from the Xeno-Canto website and resampled at 11,025 Hz to accommodate the predominant frequencies below 5 kHz in these recordings.

Minhyuk et al. [81] proposed a model to classify human activities utilizing sound recognition, leveraging a residual neural network. The dataset encompassed ten classes of

Table 8 Various parameters to evaluate sound events

Reference	Scope	Key technique	ACI	ADI	AEI	acoustic index calculation	sampling rate	window length
Moreno-Gómez et al. [38]	Birds	linear mixed effects models (LMMs)	0.456 (0.020–0.960)	0.208 (0.052–0.815)	0.092 (0.006–0.787)	Seewave R, soundecology	44 kHz	512
Eldridge et al. [39]	acoustic indices	multivariate random forest classifier	0.49	> 0.5	> 0.5	Seewave R, soundecology		256–512
Felipe Carmo et al. [40]	forest monitoring	Autonomous recorder units (ARUs) sampling method	0.50	− 0.25 0.014	0.29 0.004	Seewave R, soundecology	48 kHz	512
Fairbrass et al. [41]	urban areas	GLMER Lambert-W	0.69	0.61		R	12–96 kHz	720
Machado et al. [42]	bird communities	normalized difference soundscape index − NDSI)		-0.581	− 0.999 and 0.703,	Seewave R	2–12 kHz	1024
Ross et al. [43]	landscape-scale sensor networks	normalized difference soundscape index	0.88	0.66	0.65	Seewave R soundecology	24 kHz	512
Gómez et al. [44]	disturbed habitats:	SVM multi-layer neural network	0.78	0.68	0.89	R	44.1 kHz	
Khanaposhtani et al. [45]	Bird Sounds	Light Detection and Ranging (LiDAR) Vertical Root Mean Square Error (VRMSE)	0.55	0.89		Seewave R soundecology	44.1 kHz	512
Siddagangaiah et al. [46]	Fish Choruses	noise resilient method based on complexity-entropy	0.67	0.81		Seewave R Sound ecology	44.1 kHz	512

daily indoor activities. After data collection, feature extraction was carried out using the Log Mel-filter bank energies technique. A robust residual neural network comprising 34 convolutional layers was then trained on this data. The findings revealed a remarkable accuracy rate of 87.6%. Precision scores varied between 76.8% and 92.6% across different activity classes, while Recall scores ranged from 75.8% to 98.6%. Additionally, the F1 score ranged from 78.6% to 93.7%.

Yuren et al. [82] proposed a model to classify the Borneo forest sounds using the CNN model, such as animal calls, wind sounds, and bird calls. They found that accuracy was better even with lots of data, but it got much better when they used data augmentation and transfer learning, even with very little data. This shows that CNNs can be useful for identifying animal sounds, even in small projects with many rare species. The modified version of the Keras VGG-19 model achieved 90.4% accuracy on balanced data and 93.2% on imbalanced data.

Messner et al. [83] proposed model to detect heart sound (S1) representing systole and heart sound (S2) marking diastole heart sound states using RNN. They used the PhysioNet/CinC Challenge 2016 dataset, comprising heart sound recordings and annotated states. They employed spectral and envelope features extracted from these recordings. The model achieved an average F1 score of approximately 96%. Table 9 illustrates the various environments that are included in the sound related research.

9 Current research challenges

Sound event detection poses several challenges in current research. Real-world environments are often filled with background noise from various sources. Developing robust sound event detection systems requires large amounts of annotated audio data for training and evaluation. This background noise can significantly degrade the performance of sound event detection systems by covering the target sound events. Robustness to domain shifts and adapting to new acoustic conditions are essential for practical deployment in diverse real-world scenarios. Building a robust SED model requires a large and diverse dataset covering various sound events in various acoustic environments. Collecting and annotating such datasets can be time-consuming and expensive. Feature extraction is a crucial step in SED systems. It is challenging to extract features that effectively represent sound events while suppressing irrelevant background noise and interference. Extracting high-level, semantically meaningful features that capture common characteristics across different sound event categories can improve the generalization ability of SED systems. Processing audio data and training complex machine-learning models for SED can be computationally challenging, especially when dealing with large datasets or deploying models on resource-constrained devices.

Sound events in real-world environments are often accompanied by background noise, which can degrade the performance of SED models. Robust noise reduction and interference rejection techniques are necessary to improve the reliability of event detection. Table 10 presents the challenges of different models with various feature combinations.

Table 9 Various kinds of environments are included in the sound-related research

Reference	Feature	Model	Result	Dataset	Environment
[3]	Log-amplitude mel-spectrogram	RNN-LSTM	96.26%	DCASE 2017 Challenge Task 2 Dataset	Rare Sounds(glass breaking, baby crying and gunshot)
[4]	Log mel energies	CNN	94.2%	Synthetic	Machines working Status
[5]	Dominant frequency and log mel-band energy	BiLSTM	87.8%	freefield1010	Bird Audio Classifcation
[19]	Mel-spectrograms	NMF	F1 Score:56.3%	DCASE 2020 challenge dataset	Home sounds
[29]	GCC-PHAT	BiGRU-CRNN	92.3%	TAU Spatial Sound Events 2019—Microphone Array (MIC),	Room Environment Sounds
[41]	MFCC	Acoustic Indexes	87.5%	UrbanSound	Urban Areas
[54]	MFCC	VGG	94.2%	ARCA23K	real-world label noise
[56]	Mel-Spectrogram	SVM	84.2%	YT-NTL-U and Pub-S	Speech
[59]	Temporal and Spectral Features	SVM	90%	MIVIA dataset	Road Surveillance
[74]	Mel –Spectogram	CNN based multitask learning (MTL)	F1 score of 42.17%	TUT Acoustic Scenes 2016	Environmental sounds
[76]	log Mel-filter bank energies	CNN	Accuracy: 87.6%	CIFAR-10	Human activity
[77]	Deep Feature	VGG-19	Accuracy:90.4%	bioacoustic workbench Ecosounds	Forest
[78]	MFCC	CRNN	F1 Score:96%	PhysioNet/CinC Challenge dataset	Heart Sound Classification
[79]	Deep Features	CNN-BiGRU	F1 Score:49.82%	TUT Sound Events 2016/2017	residential area" and "city center
[81]	Log Mel-filter bank energies	CNN	93.7%	open-source video and audio platforms	Human activity
[82]	Mel-Spectrogram	VGG19	94.5%	Xeno-Canto	Rain Forest
[83]	spectral and envelope features	RNN	F1-Score:96%	PhysioNet/CinC Challenge 2016 dataset	Heart Sound Classification

10 Applications

Sound event detection using machine learning has led to the development of various related technologies with a wide range of applications.

a. **Acoustic Scene Analysis**: Technologies for analyzing the acoustic characteristics of environments, such as identifying the presence of specific sounds (e.g., sirens, alarms, speech) and categorizing acoustic scenes (e.g., indoor, outdoor, urban, rural).
b. **Keyword Spotting and Wake Word Detection**:These techniques for detecting specific keywords or wake words within audio streams. They are commonly used in voice-activated devices and virtual assistants to trigger actions or initiate interactions (e.g., Apple's Siri, Google Assistant, Samsung's Bixby,Amazon Alexa, etc.).
c. **Environmental Monitoring Systems**: Systems equipped with sensors and sound event detection algorithms for monitoring and analyzing sounds in natural habitats, urban areas, or industrial environments to track biodiversity, assess noise pollution, monitor traffic patterns, or detect anomalies.
d. **Healthcare Monitoring Devices**: Devices and applications capable of monitoring health-related sounds (e.g., coughing, snoring, breathing patterns) for telemedicine, sleep analysis, monitoring patients with respiratory conditions, or detecting signs of distress.
e. **Security and Surveillance Systems**: Technologies for detecting and classifying sounds related to security threats or abnormal events, such as glass breaking, footsteps, gunshots, or vehicle alarms, in surveillance camera footage or audio recordings for enhanced security monitoring.
f. **Smart Home Automation**: Integration of sound event detection capabilities into smart home systems to automate tasks based on detected sounds (e.g., turning on lights in response to doorbell rings and alerting homeowners to potential security breaches).
g. **Industrial Monitoring and Predictive Maintenance**: Solutions for monitoring machinery and equipment in industrial settings by analyzing sounds to detect anomalies, predict failures, schedule maintenance, and optimize performance to minimize downtime and improve operational efficiency.
h. **Assistive Technologies for People with Disabilities**: These technologies are designed to assist individuals with hearing or other disabilities by analyzing sounds and providing relevant feedback or alerts (e.g., sound-based navigation aids and assistance in identifying environmental sounds).
i. **Entertainment and Gaming**: Integrating sound event detection algorithms into gaming and entertainment systems to create immersive experiences, enhance virtual reality environments, or provide interactive gameplay based on detected sounds.
j. **Automotive Safety and Driver Assistance Systems**: Sound event detection capabilities are incorporated into vehicles to improve driver safety, detect potential hazards (e.g., sirens, horns, tire screeches), and enhance driver assistance features such as collision avoidance and emergency braking systems.

These technologies show the diverse range of applications enabled by sound event detection using machine learning, spanning across industries and domains to address various needs related to monitoring, safety, automation, and user experience enhancement.

Table 10 Challenges of different models with various feature combinations

Reference	Feature Extraction	Model	Challenges
[8]	MFCC	SVM	• Some sounds are heard more often than others, which can confuse SVMs • SVMs have settings that need to be just right, but it's hard to figure out what they should be • Sound patterns can be tricky, and SVMs might struggle to understand them without making things too complicated • Sounds change over time, but SVMs don't really get that • Using SVMs with lots of sound data can be slow and hard to manage • SVMs might get too focused on certain sounds and not be good at recognizing new ones
[10]	MFCCs, STE, and ZCR	NMF	• Using multiple features can make feature vectors large, leading to more computational complexity • MFCCs, STE, and ZCR features can be sensitive to background noise and environmental variations • NMF may struggle to capture the temporal dynamics of sound events when using static feature representations such as MFCCs, STE, and ZCR • NMF algorithms need a lot of computing power, especially when dealing with large datasets
[13]	ZCR	Random Forest	• Random forests can need help dealing with many different features in audio signals, especially when combining them with ZCR • Random Forests might learn too much from the training data, making them perform poorly on new data, especially if it is noisy or has many features like ZCR • Random Forests have several hyperparameters that need to be tuned for optimal performance • Random Forests are prone to overfitting
[15]	CNN	log mel-band energies,MFCC	• CNNs can be adapted for 1D convolution for temporal data, but capturing long-term temporal dependencies can be challenging • Handling variable-length inputs in CNNs requires additional preprocessing steps, such as padding or segmenting the input data • Log mel-band energies and MFCCs require careful consideration to preserve the temporal and spectral characteristics of the original audio
[18]	NMF-CNN	NMF	• Integrating NMF and CNN approaches requires ensuring the data representations are compatible and effectively combined • NMF decomposes the input audio into basic components, and CNN learns hierarchical representations directly from the data. So, carefully consider how the outputs of NMF are fed into the CNN • Maintaining this interpretability while integrating NMF with CNNs can be difficult, as CNNs typically operate as black-box models with less interpretable internal representations

Table 10 (continued)

Reference	Feature Extraction	Model	Challenges
[24]	CRNN	STFT,Deep Features	• Using STFT makes selecting appropriate parameters such as window size, overlap, and frequency resolution complex • CRNNs can be complex models, especially when combined with STFT for feature extraction
[30]	CRNN	Mel-spectrogram	• CRNN models can be complex, especially when dealing with large datasets and high-dimensional feature representations like Mel-spectrograms • Deploying CRNN models for real-time sound event classification applications requires low-latency inference, which can be challenging given the computational complexity of the model architecture
[90]	HMM	MFCC,Mel-filter bank coefficients, Mel-filter bankcoefficient	• HMMs struggle with understanding longer patterns in sound • They find it hard to tell similar sounds apart, leading to mistakes • If there are many sounds to detect, HMMs can be slow • They can't handle sounds of different lengths effectively • HMMs get confused by noisy data, making them less accurate • If there's not much training data, they might not perform well
[95]	GMM	MFCC	• GMM assumes sound data follows a certain pattern, but it might not always fit the real world's diverse sounds well • Sound data can be complex, making GMMs slow and prone to mistakes • If we don't have enough labeled sound examples, GMMs won't work well • GMMs don't understand how sounds change over time • GMMs struggle when some sounds are more common than others • GMMs are easily confused by noise in sound data

11 Conclusion

In this paper, we surveyed and analyzed different models of sound event detection. We also reviewed various algorithms and critical techniques to achieve better results. This paper also illustrates multiple parameters and metrics to evaluate the sound event and localization. Polyphonic sound detection has become one of the key reviews in this research. The significance of accurate definitions for evaluating sound metrics cannot be overstated. It comprises distinct algorithms and obtains acceptable results based on the benchmark databases being implemented using a uniform assessment process. The research performed in this paper is considered part of our effort toward getting the reference point and a better understanding of defining task-based metrics for implementing polyphonic sound event detection. In our future research, we plan to implement sound classification in forest areas to classify tree-cutting sounds and protect forest natural resources. This project may help government bodies to find illegal logging in the forest. In this case, forests consist of various sounds, such as bird calls, animal noises, vehicle sounds from nearby roads, wind sounds, and tree-cutting activities. Based on the environmental situation, our future model will classify and identify the use of total sound. Our research focuses mainly on preserving the classification of forests' natural resources based on sound classification.

Author's contributions All authors equally contributed and approved the final manuscript.

Funding Not applicable.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare that they have no conflicts.

References

1. Mesaros A, Diment A, Elizalde B, Heittola T, Vincent E, Raj B, Virtanen T (2019) Sound event detection in the DCASE 2017 challenge. IEEE/ACM Tran Audio Speech Lang Process 27(6):992–1006
2. Mesaros A, Heittola T, Virtanen T (2016) TUT database for acoustic scene classification and sound event detection. 24th European Signal Processing Conference (EUSIPCO), pp. 1128–1132, https://doi.org/10.1109/EUSIPCO.2016.7760424
3. Lim H, Park J, Han Y (2017) Rare sound event detection using 1D convolutional recurrent neural networks. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), pp. 80–84
4. Kawaguchi Y, Tanabe R, Endo T, Ichige K, Hamada K (2019) Anomaly detection based on an ensemble of dereverberation and anomalous sound extraction. In ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 865–869
5. Adavanne S, Drossos K, Çakir E, Virtanen T (2017) Stacked convolutional and recurrent neural networks for bird audio detection. In 2017 25th European signal processing conference (EUSIPCO), pp. 1729–1733. IEEE
6. Kong Q, Yong Xu, Wang W, Plumbley MD (2020) Sound event detection of weakly labelled data with CNN-transformer and automatic threshold optimization. IEEE/ACM Trans Audio Speech Lang Process 28:2450–2460
7. Kawaguchi Y, Endo T, Ichige K, Hamada K (2018) Non-negative novelty extraction: A new non-negativity constraint for NMF. 16th International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 256–260

8. Küçükbay SE, Sert M (2015) Audio-based event detection in office live environments using optimized MFCC-SVM approach. In Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), pp. 475–480
9. Lei B, Mak M-W (2014) Sound-event partitioning and feature normalization for robust sound-event detection. In 2014 19th International Conference on Digital Signal Processing, pp. 389–394. IEEE
10 Parathai P, Tengtrairat N, Woo WL, Abdullah MAM, Rafiee G, Alshabrawy O (2020) Efficient Noisy sound-event mixture classification using adaptive-sparse complex-valued matrix factorization and OvsO SVM. Sensors 20(16):4368
11. Tran HD, Li H (2010) Sound event recognition with probabilistic distance SVMs. IEEE Trans Audio Speech Lang Process 19(6):1556–1568
12. Huang S-J, Liu C-C, Chen C-P (2023) Sound event detection system based on VGGSKCCT model architecture with knowledge distillation. Appl Artif Intell 37(1):2152948
13. Phan H, Maass M, Mazur R, Mertins A (2015) Early event detection in audio streams. In 2015 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE
14. Xia X, Togneri R, Sohel F, Huang D (2018) Random forest classification based acoustic event detection utilizing contextual-information and bottleneck features. Pattern Recogn 81:1–13
15. Xia X, Togneri R, Sohel F, Zhao Y, Huang DD (2020) Sound event detection using multiple optimized kernels. IEEE/ACM Trans Audio Speech Lang Process 28:1745–1754
16. Stoller D, Ewert S, Dixon S (2018) Wave-u-net: A multi-scale neural network for end-to-end audio source separation.arXiv preprint arXiv:1806.03185
17. Zhou Q, Feng Z, Benetos E (2019) Adaptive noise reduction for sound event detection using subband-weighted NMF. Sensors 19(14):3206
18. Chan TK, Chin CS, Li Y (2020) Non-negative matrix factorization-convolutional neural network (NMF-CNN) for sound event detection. arXiv preprint arXiv:2001.07874
19 Chan TK, Chin CS, Li Y (2021) Semi-supervised NMF-CNN for sound event detection. IEEE Access 9:130529–130542
20. Shin Y, Chun C (2023) Sound event localization and detection using imbalanced real and synthetic data via multi-generator. Sensors 23(7):3398
21. De La Torre Cruz J, Quesada FJC, Reyes NR, Galán SG, Orti JJC, Chica GP (2021) Monophonic and polyphonic wheezing classification based on constrained low-rank non-negative matrix factorization. Sensors 21(5):1661
22. Innami S, Kasai H (2012) NMF-based environmental sound source separation using time-variant gain features. Comput Math Appl 64(5):1333–1342
23. Komatsu T, Senda Y, Kondo R (2016) Acoustic event detection based on non-negative matrix factorization with mixtures of local dictionaries and activation aggregation. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2259–2263
24. Noh K, Chang J-H (2020) Joint optimization of deep neural network-based dereverberation and beam forming for sound event detection in multi-channel environments. Sensors 20(7):1883
25. Ferroni G, Turpault N, Azcarreta J, Tuveri F, Serizel R, Bilen Ç, Krstulović S (2021) Improving sound event detection metrics: insights from dcase 2020. In ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 631–635. IEEE
26. Komatsu T, Toizumi T, Kondo R, Senda Y (2016) Acoustic event detection method using semi-supervised non-negative matrix factorization with a mixture of local dictionaries. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), pp. 45–49
27. Xu Y, Kong Q, Wang W, Plumbley MD (2018) Large-scale weakly supervised audio classification using gated convolutional neural network. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 121–125. IEEE
28. Grondin F, Glass J, Sobieraj I, Plumbley MD (2019) Sound event localization and detection using CRNN on pairs of microphones. arXiv preprint arXiv:1910.10049
29. Adavanne S, Politis A, Virtanen T (2019) A multi-room reverberant dataset for sound event localization and detection. arXiv preprint arXiv:1905.08546
30. Zhang J, Ding W, He L (2019) Data augmentation and prior knowledge-based regularization for sound event localization and detection. DCASE 2019 Detection and Classification of Acoustic Scenes and Events 2019 Challenge
31. Cao Y, Iqbal T, Kong Q, Galindo M, Wang W, Plumbley M (2019) Two-stage sound event localization and detection using intensity vector and generalized cross-correlation. DCASE2019 Challenge, Tech Rep
32. Adavanne S, Politis A, Nikunen J, Virtanen T (2018) Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. IEEE J Sel Topics Signal Process 13(1):34–48
33. Xue W, Tong Y, Zhang C, Ding G, He X, Zhou B (2020) Sound event localization and detection based on multiple DOA beam forming and multi-task learning. Proc Interspeech 2020:5091–5095

34. Nguyen TNT, Jones DL, Gan W (2020) Ensemble of sequence matching networks for dynamic sound event localization detection and tracking. In Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)
35. Trowitzsch I, Schymura C, Kolossa D, Obermayer K (2019) Joining sound event detection and localization through spatial segregation. IEEE/ACM Trans Audio Speech Lang Process 28:487–502
36. Kim B, Pardo B (2019)Sound event detection using point-labeled data. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 1–5
37. Xia X, Togneri R, Sohel F, Huang D (2018) Auxiliary classifier generative adversarial network with soft labels in imbalanced acoustic event detection. IEEE Trans Multimedia 21(6):1359–1371
38. Moreno-Gómez FN, Bartheld J, Silva-Escobar AA, Briones R, Márquez R, Penna M (2019) Evaluating acoustic indices in the Valdivian rainforest, a biodiversity hotspot in South America. Ecol Ind 103:1–8
39. Eldridge A, Guyot P, Moscoso P, Johnston A, Eyre-Walker Y, Peck M (2018) Sounding out ecoacoustic metrics: avian species richness is predicted by acoustic indices in temperate but not tropical habitats. Ecol Ind 95:939–952
40. Jorge FC, Machado CG, da Cunha Nogueira SS, Nogueira-Filho SLG (2018) The effectiveness of acoustic indices for forest monitoring in Atlantic rainforest fragments. Ecol Indic 91:71–76
41. Fairbrass AJ, Rennert P, Williams C, Titheridge H, Jones KE (2017) Biases of acoustic indices measuring biodiversity in urban areas. Ecol Ind 83:169–177
42. Machado RB, Aguiar L, Jones G (2017) Do acoustic indices reflect the characteristics of bird communities in the savannas of Central Brazil? Landsc Urban Plan 162:36–43
43 Ross SRP-J, Friedman NR, Dudley KL, Yoshimura M, Yoshida T, Economo EP (2018) Listening to ecosystems: data-rich acoustic monitoring through landscape-scale sensor networks. Ecol Res 33(1):135–147
44. Gómez WE, Isaza CV, Daza JM (2018) Identifying disturbed habitats: a new method from acoustic indices. Ecol Inform 45:16–25
45. Khanaposhtani MG, Gasc A, Francomano D, Villanueva-Rivera LJ, Jung J, Mossman MJ, Pijanowski BC (2019) Effects of highways on bird distribution and soundscape diversity around Aldo Leopold's shack in Baraboo, Wisconsin, USA. Landsc Urban Plan 192:103666
46. Siddagangaiah S, Chen C-F, Wei-Chun Hu, Pieretti N (2019) A complexity-entropy based approach for the detection of fish choruses. Entropy 21(10):977
47. Castorena C, Cobos M, Lopez-Ballester J, Ferri FJ (2023) A safety-oriented framework for sound event detection in driving scenarios. Appl Acoust 215:109719
48. Wang Q, Chai L, Wu H, Nian Z, Niu S, Zheng S, Wang Y et al (2022) The NERC-SLIP system for sound event localization and detection of DCASE2022 challenge. DCASE2022 Chall Tech Rep
49. Hu J, Cao Y, Wu M, Yang F, Wang W, Plumbley MD, Yang J (2023) A data generation method for sound event localization and detection in real spatial sound scenes. Tech Rep DCASE2023 Chall
50. Cheimariotis G-A, Mitianoudis N (2023) Sound event detection in domestic environment using frequency-dynamic convolution and local attention. Information 14(10):534
51. Diez I, Saratxaga I, Salegi U, Navas E, Hernaez I (2023) NoisenseDB: an urban sound event database to develop neural classification systems for noise-monitoring applications. Appl Sci 13(16):9358
52. Yuan S, Yang L, Guo Y (2023) Sound event detection with perturbed residual recurrent neural network. Electronics 12(18):3836
53. Zhang H, Zuo L, Chen J, Cai X, Wu M (2023) Sound event detection based on soft label. Detect Classif Acoust Scenes Events (DCASE) Chall
54. Iqbal T, Cao Y, Bailey A, Plumbley MD, Wang W (2021) ARCA23K: an audio dataset for investigating open-set label noise. arXiv preprint arXiv:2109.09227
55. Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Lawrence W, Moore RC, Plakal M, Ritter M (2017) Audio set: An ontology and human-labeled dataset for audio events. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776–780. IEEE
56. Zhang Y, Han W, Qin J, Wang Y, Bapna A, Chen Z, Chen N et al (2023) Google usm: scaling automatic speech recognition beyond 100 languages. arXiv preprint arXiv:2303.01037
57. Bubashait M, Hewahi N (2021) Urban sound classification using DNN, CNN & LSTM a comparative approach. In 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), pp. 46–50. IEEE
58. Shuyang Z, Heittola T, Virtanen T (2020) Active learning for sound event detection. IEEE/ACM Trans Audio Speech Lang Process 28:2895–2905
59. Almaadeed N, Asim M, Al-Maadeed S, Bouridane A, Beghdadi A (2018) Automatic detection and classification of audio events for road surveillance applications. Sensors 18(6):1858
60. Yadav S, Foster ME (2021) GISE-51: a scalable isolated sound events dataset. arXiv preprint arXiv:2103.12306

61. Fonseca E, Plakal M, Ellis DPW, Font F, Favory X, Serra X (2019) Learning sound event classifiers from web audio with noisy labels. In ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 21–25. IEEE
62. Fonseca E, Plakal M, Font F, Ellis DPW, Serra X (2019) Audio tagging with noisy labels and minimal supervision. arXiv preprint arXiv:1906.02975
63. Fonseca E, Favory X, Pons J, Font F, Serra X (2021) Fsd50k: an open dataset of human-labeled sound events. IEEE/ACM Trans Audio Speech Lang Process 30:829–852
64. Piczak KJ (2015) ESC: Dataset for environmental sound classification. In Proceedings of the 23rd ACM international conference on Multimedia, pp. 1015–1018
65. Foster P, Sigtia S, Krstulovic S, Barker J, Plumbley MD (2015) Chime-home: A dataset for sound source recognition in a domestic environment. In 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 1–5. IEEE
66. Jekateryńczuk G, Piotrowski Z (2023) A survey of sound source localization and detection methods and their applications. Sensors 24(1):68
67. Crocco M, Cristani M, Trucco A, Murino V (2016) Audio surveillance: a systematic review. ACM Comput Surv (CSUR) 48(4):1–46
68 Alsina-Pagès RM, Navarro J, Alías F, Hervás M (2017) homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring. Sensors 17(4):854
69. Dang A, Vu TH, Wang J-C (2017) A survey of deep learning for polyphonic sound event detection. In 2017 International Conference on Orange Technologies (ICOT), pp. 75–78. IEEE
70. Nunes EC (2021) Anomalous sound detection with machine learning: a systematic review. arXiv preprint arXiv:2102.07820
71. Shreyas N, Venkatraman M, Malini S, Chandrakala S (2020) Trends of sound event recognition in audio surveillance: a recent review and study. The Cognitive Approach in Cloud Computing and Internet of Things Technologies for Surveillance Tracking Systems 95–106
72. 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
73 Chan TK, Chin CS (2020) A comprehensive review of polyphonic sound event detection. IEEE Access 8:103339–103373
74. Imoto K, Tonami N, Koizumi Y, Yasuda M, Yamanishi R, Yamashita Y (2020) Sound event detection by multitask learning of sound events and scenes with soft scene labels. In ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 621–625. IEEE
75. Hebbar R, Bose D, Somandepalli K, Vijai V, Narayanan S (2023) A dataset for audio-visual sound event detection in movies. In ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE
76. Cheng S, Du J, Wang Q, Jiang Y, Nian Z, Niu S, Lee C-H, Gao Y, Zhang W (2023) Improving Sound Event Localization and Detection with Class-Dependent Sound Separation for Real-World Scenarios. In 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 2068–2073. IEEE
77. Lan C, Zhang L, Zhang Y, Lirong Fu, Sun C, Han Y, Zhang M (2022) Attention mechanism combined with residual recurrent neural network for sound event detection and localization. EURASIP J Audio Speech Music Process 2022(1):29
78. Min D, Nam H, Park Y-H (2023) Application of spectro-temporal receptive field on soft labeled sound event detection. Tech Rep Tech Rep DCASE2023 Chall
79. Gao L, Mao Q, Dong M (2024) On local temporal embedding for semi-supervised sound event detection. IEEE/ACM Trans Audio Speech Lang Process
80. Xie J, Zhu M (2022) Sliding-window based scale-frequency map for bird sound classification using 2D-and 3D-CNN. Expert Syst Appl 207:118054
81. Jung M, Chi S (2020) Human activity classification based on sound recognition and residual convolutional neural network. Autom Constr 114:103177
82. Sun Y, Maeda TM, Solis-Lemus C, Pimentel-Alarcon D, Burivalova Z (2021) Classification of animal sounds in a hyperdiverse rainforest using Convolutional Neural Networks. arXiv preprint arXiv:2111.14971
83. Messner E, Zöhrer M, Pernkopf F (2018) Heart sound segmentation—An event detection approach using deep recurrent neural networks. IEEE Trans Biomed Eng 65(9):1964–1974
84. Lee S, Kim H, Jang G-J (2023) Weakly supervised U-Net with limited upsampling for sound event detection. Appl Sci 13(11):6822
85. Ahmed A, Serrestou Y, Raoof K, Diouris J-F (2022) Empirical mode decomposition-based feature extraction for environmental sound classification. Sensors 22(20):7717
86. Kim C, Yang S (2022) Sound event detection system using Fix-Match for DCASE 2022 challenge Task 4. Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge

87. Kim S-J, Chung Y-J (2022) Multi-scale features for transformer model to improve the performance of sound event detection. Appl Sci 12(5):2626
88. Jin Ye, Wang M, Luo L, Zhao D, Liu Z (2022) Polyphonic sound event detection using temporal-frequency attention and feature space attention. Sensors 22(18):6818
89. Kong Q, Yong Xu, Sobieraj I, Wang W, Plumbley MD (2019) Sound event detection and time–frequency segmentation from weakly labelled data. IEEE/ACM Trans Audio Speech Lang Process 27(4):777–787
90. Kiktova E, Lojka M, Pleva M, Juhar J, Cizmar A (2013) Comparison of different feature types for acoustic event detection system. In Multimedia Communications, Services and Security: 6th International Conference, MCSS 2013, Krakow, Poland, June 6–7, 2013. Proceedings 6, pp. 288–297. Springer Berlin Heidelberg
91. Surampudi N, Srirangan M, Christopher J (2019) Enhanced feature extraction approaches for detection of sound events. In 2019 IEEE 9th International Conference on Advanced Computing (IACC), pp. 223–229. IEEE
92. Adavanne S, Pertilä P, Virtanen T (2017) Sound event detection using spatial features and convolutional recurrent neural network. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 771–775. IEEE
93. De Benito-Gorrón D, Ramos D, Toledano DT (2021) A multi-resolution CRNN-based approach for semi-supervised sound event detection in DCASE 2020 challenge. IEEE Access 9:89029–89042
94. Nguyen TNT, Watcharasupat K, Nguyen NK, Jones DL, Gan WS (2021) DCASE 2021 Task 3: Spectrotemporally-aligned features for polyphonic sound event localization and detection. arXiv preprint arXiv:2106.15190
95. Kim K, Ko H (2011) Discriminative training of GMM via log-likelihood ratio for abnormal acoustic event classification in vehicular environment. In 2011 First ACIS/JNU International Conference on Computers, Networks, Systems and Industrial Engineering, pp. 348–352. IEEE