Multimedia Tools and Applications (2024) 83:84699–84741
https://doi.org/10.1007/s11042-024-18740-9

Exploring current research trends in sound event detection: a systematic literature review

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

Received: 29 November 2022 / Revised: 16 January 2024 / Accepted: 24 February 2024 /
Published online: 29 April 2024

Abstract

Sound Event Detection (SED) plays a significant role in the present research, implemented in several areas such as Computer Science, Healthcare, Environmental Science, Security and Surveillance, etc. With the advancement of technology, SED can be deployed to mimic the human auditory system. In this paper, we have undertaken a Systematic Literature Review focused on sound event detection, presenting a comprehensive and well-structured analysis and in-depth discussions. This review is based on the authors' extensive knowledge and expertise in the field, and it compares various algorithms employed for sound event detection. The primary objective of this study is to offer valuable insights into datasets, feature extraction techniques, and execution models commonly used in SED, along with an examination of their corresponding accuracy, challenges, and limitations. Furthermore, the paper delves into identifying potential trends within the field, offering forward-looking information that can be invaluable for future research and development efforts in sound event detection. This systematic review aims to contribute to the continued advancement of SED technologies and applications by synthesizing existing knowledge and identifying emerging directions. It provides a foundation for researchers, practitioners, and stakeholders to make informed decisions and explore new possibilities within this evolving domain.

Keywords Sound event detection · Deep learning · Machine learning · Systematic literature review · Feature extraction

✉ Sallauddin Mohmmad
sallauddin.md@gmail.com

Suresh Kumar Sanampudi
sureshsanampudi@jntuh.ac.in

[1] School of Computer Science and Artificial Intelligence, SR University, Warangal, Telangana, India

[2] Research Scholar, JNTU, Hyderabad, India

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

1 Introduction

Sound event detection (SED) is a rapidly advancing field in current research dedicated to identifying and classifying diverse sounds within specified environments. Its applications span various domains: office spaces, parks, hospitals, forests, hotels, shopping malls, space research, and the automotive industry. SED is essential in recognizing auditory components such as speech, music, alarms, and specific sound events like a dog's bark or a gunshot [2, 3].SED's significance lies in its ability to detect momentary changes within an ongoing soundscape, making it invaluable for real-time monitoring and response systems. Whether monitoring noise levels in an office, detecting wildlife sounds in a forest, identifying hospital alarms, or recognizing security breaches in hotels, SED contributes to safety, security, and situational awareness [4, 5]. This field leverages machine learning and artificial intelligence techniques, particularly deep learning, to train models on extensive datasets for sound event classification. However, it faces challenges, including dealing with noisy data, overlapping sound events, and adapting to diverse acoustic environments. Nevertheless, SED continues to evolve, making it an essential tool for understanding and interacting with the acoustic world in various contexts [6].

The primary goal of the research on SED is to find a proper dataset that is adequately suitable for diverse environmental sounds. For instance, in classifying forest sounds, the dataset should encompass a spectrum of auditory elements such as birds, animals, insects, wind, rain, and more [10, 11]. Numerous datasets are available for sound classification, such as ESC-US, ESC-50, MAVD, Urban Sound Freefield1010, Dares G1, MIMII, Freiburg-106, AudioSet, SINGA: PURA, SONYC-UST-V2, FSD50K USM-SED, and the Million Song Dataset (2011) [23–25]. However, Creating a dataset exhibits distinct challenges. Firstly, a detailed capture of each sound event in the chosen environment ensures a diverse representation of sources. Secondly, dealing with background noise in the recorded sound samples integrated that irrelevant for required.

Furthermore, in situations illustrated by a low Signal-to-Noise Ratio, background noise poses a significant challenge to accurately determining sound events within a specific time frame. As a result, significant research efforts have been dedicated to advancing the effectiveness of current Sound Event Detection systems [26, 27]. Feature extraction is crucial in SED research, enabling more efficient identification and classification of sound events within audio signals. Numerous techniques exist for extracting features from sound data. Effective feature extraction methods are instrumental in converting raw audio data into an appropriate format for machine learning algorithms [35, 36]. Among the generally used techniques are Mel-Frequency Cepstral Coefficients (MFCCs), adept at capturing spectral characteristics; spectrograms, disclosing frequency content over time; chroma features, well-suited for music-related SED; Zero-Crossing Rate, valuable for detecting noisiness; energy-based features such as Root Mean Square (RMS) energy; statistical descriptors like mean and variance; temporal features encompassing zero-crossing rate and temporal centroid; wavelet transforms addressing spectral and temporal aspects; and deep learning-based features that harness neural networks to learn appropriate representations from raw audio autonomously. The selection of a particular feature extraction method for the specific SED task depended on the characteristics of the data [43, 45, 46].

On the other hand, to classify the featured data, recent researchers focused on various algorithms have been explored, spanning from traditional methods like the Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) and Non-negative Matrix Factorization (NMF) to well-established techniques such as Support Vector Machines (SVM) and Random Forest [52, 55, 57]. However, the prevailing directions indicate a noticeable shift towards utilizing neural network models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and other deep learning models. These neural architectures have achieved distinction for their capacity to extract complex features and representations from data, generating them particularly effective in diverse classification tasks [60, 64]. Furthermore, researchers have increasingly employed additional strategies such as Long Short-Term Memory (LSTM), Bidirectional Gated Recurrent Units (BiGRU), and Bidirectional LSTMs, models with enhanced performance by considering bidirectional context and capturing long-range dependences in sequential data [88, 90]. This paradigm shift reflects the evolution and widespread adoption of deep learning techniques across various domains, encompassing image processing, natural language processing, and time series analysis [86, 102].

1.1 Motivation

Numerous review-related research studies have been conducted on sound event detection. These reviews often aim to summarize the state of the art, identify research trends, highlight limitations of existing research, and outline the challenges that researchers face. These reviews provide valuable insights for researchers and anyone interested in Sound Event Detection. They help to consolidate knowledge, guide future research efforts, and facilitate collaboration among experts in the field.

Bansal et al. [111] comprehensively analyzed environmental sound classification. They explored the various datasets used in recent studies and the corresponding accuracy achieved with these datasets. The paper also delved into the different feature extraction methods employed by various researchers on selected data input. This review concluded that an algorithm's proper feature extraction combination would be essential for getting good results. The authors also discussed the limitations and challenges of using deep learning and machine learning approaches for sound classification.

Chan et al. [112] presented a comprehensive polyphonic sound event detection overview. The author approached this by categorizing concepts into neural and non-neural networking models, providing clear insights into feature extraction techniques, highlighting model limitations, and detailing the datasets used in each category. Notably, the review also dedicated a section to weakly labeled datasets, examining changes in accuracy levels and associated limitations when using such datasets. Overall, this review offers substantial information on polyphonic sound event detection.

Mesaros et al. [113] have reviewed sound event detection research trends and provided valuable knowledge. They initiated their review by addressing the general challenges of sound event detection. They subsequently delved into distinct areas, explaining machine learning methods, deep learning approaches, and various techniques for feature extraction. They introduced the CRNN model to illustrate its application in sound event detection using mel-spectrogram features as input.

Shreyas et al. [115] have presented a comprehensive analysis of sound classification in urban environments, attaching to the Systematic Literature Review (SLR) methodology. The authors of the reviewed papers extensively explored Deep Learning approaches, highlighting the advantages and disadvantages of each model concerning the datasets they employed. They meticulously examined various datasets encompassing multiple classes and samples related to urban sounds.

Abayomi-Alli et al. [116] encapsulated the current trends in sound event detection research—using the DCASE- Challenge 2016 dataset allowed for a deeper understanding of the nature of sound classification. In this review process, the authors included a section describing the need to select the feature extraction combination for a model. This review paper contributed significant information about implementing machine learning and deep learning models in SED.

Table 1 summarizes the comprehensive overview of the Sound Event Detection research studies that have been reviewed. This table categorizes the research by the environment under study, provides information about the number of papers referenced by the author, and indicates the years from which the papers were collected for the review.

1.2 Significant contributions of survey work

This paper aims to provide a systematic literature review on Sound Event Detection. The SLR systematically tracks our review of the study to collect and summarize the existing research. SLR also leads the integration of relevant studies of our research and supports the finding of domain-specific research questions to evaluate. In this paper, our research methodology followed the guidelines of Kitchenham et al. [1] to conduct the best review process for research. Kitchenham et al. also provided evidence-based, well-defined approaches to finding gaps in current research and suggested filling those gaps to move on to further investigation. The significant contributions of our research are given below:

1. Summarize the results and methodologies of previous studies in the field of SED to provide a clear overview of the current state of research.
2. Utilize the reviewed literature to formulate specific research questions that address the identified gaps, contribute to the advancement of SED knowledge, and provide recommendations for future research directions and strategies for filling the identified gaps in the SED field.
3. Adhere to established guidelines, such as those by Kitchenham et al. [1], to ensure a rigorous and evidence-based approach in the review process.
4. Contribute valuable insights and knowledge to the domain of Sound Event Detection through a well-defined comprehensive literature review.

The remainder of this paper is organized as follows: Section 2 outlines the Review Method, research inquiries, and the criteria for selection. Section 3 delves into the research question outcomes. Section 4 details the Evaluation Metrics, followed by Section 5, which explains the software resources used for SED research. Section 6 explores the potential future Scope, while Section 7 illustrates the Summary of the Review. Section 8 explores the Synthesis of our research, and finally, our conclusions and future work are presented in Section 9.

2 Review method

We designed the review questions based on PICOC criteria used by multiple researchers [110].

Table 1 Different research studies that have been reviewed on Sound Classifcation

Author	Environment	Number of References	References coverage Years	Observations
Bansal et al. [111]	Environmental Sound Classification	89	1999, 2006–2022	• Not included the limitations of feature extraction techniques with respect to the model • Not provided the future research scope of ESC • Not included the Evaluation Metrics for their research
Chan et al**.** [112]	Polyphonic Sound Event Detection	140	1989, 1990, 2003, 2005–2020	• Not included the preprocessing of sound sample • Reviewed the limited datasets
Nogueira et al. [114]	Sound Event Detection	50	1993, 2007, 2009, 2010–2020	• Not explained about preprocessing • Not included the Machine Learning and Baseline approaches • There is no comparative study among the executed models • Reviewed the limited datasets
Shreyas et al. [115]	Urban Environments sound classification	75	1967, 1987, 2012–2022	• Not explained about preprocessing • Not included the Machine Learning and Baseline approaches
Abayomi-Alli et al. [116]	Sound Event Detection	127	2006, 2009, 2010–2021	• Not included the limitations of feature extraction techniques with respect to the model • Limited discussion about Machine Learning models • Not provided the future research scope of ESC • Not included the Evaluation Metrics for their research

Population (P): This study focuses on Sound Event Detection within diverse environmental sets.
Intervention (I): The research explores using various elements, including datasets, feature extraction methods, classification techniques, and machine learning strategies.
Comparison (C): Our review process comprehensively assesses different research methodologies and their respective results and analyzes the strengths and weaknesses of each.
Outcomes (O): Our review critically analyzes existing methodologies, which contain considerable computational efficiency, scalability, robustness, and the potential for real-world deployment. The review also recommends future research directions, emphasizing areas where advancements can be made to enhance the field of Sound Event Detection within diverse environmental sets.
Context (C): The specific context for this investigation is not applicable (NA).

2.1 Review questions

To provide evidence for our systematic literature review of our studies on SED, we framed the following suitable review questions (RQ) related to our domain.

RQ1: What sound event detection datasets are accessible for research purposes, and what preprocessing steps should be applied to them?
The response to this question will present a catalog of datasets primarily related to SED and draft the import steps. In an essential case, it can also motivate us to prepare an efficient synthetic dataset for our research environment.
RQ2: What are the feature extraction methods available for sound events?
The answer to this question will provide the capability to gain knowledge on various feature extraction methods and libraries used to extract the features from sound events.
RQ3: What are the currently available Machine Learning and Neural Network-based models for sound event detection?
The answer to this question will provide knowledge of various Machine Learning algorithms and techniques involved in sound event detection, such as classification algorithms, regression algorithms, Neural Network related algorithms, etc.
RQ4: In recent research, which environments are typically selected for Sound Event Detection?
The answer to this question will provide knowledge of various environments that have been selected to conduct SED research so far.
RQ5: What are the limitations or challenges associated with current research methodologies?
This response addresses the limitations or challenges inherent in current research methods.

2.2 Search process

In order to enhance the comprehensiveness and stringency of our SLR, the search process was conducted on adequate resources such as IEEE Explore, Springer database, ACM, Science Direct, etc., to enrich and standardize SLR. Additionally, we widened the scope of our review by referencing papers published between 2009 and 2023, encompassing significant research in sound event detection, acoustic analysis, and speech processing. The included specific strings such as "Sound Event Detection," "Acoustic Event Detection," "Speech Processing," "Acoustic Classification," "Sound Classification," and "Environmental Sound Detection" were employed to retrieve relevant research papers.

2.3 Selection criteria

After gathering relevant documents from diverse warehouses, we systematically applied standards to decide which documents to include in our research and which ones to exclude. This careful process of inclusion and exclusion significantly improved the efficiency and effectiveness of our review.

Inclusion criteria 1: During our review, we included papers about the dataset, encompassing sound event detection and directly relevant to the evaluation.
Inclusion criteria 2. We included papers that explored implementing machine-learning techniques for Sound Event Detection.
Inclusion criteria 3: We plan to focus our upcoming research on environmental sounds. In alignment with this composition, the inclusion process included papers containing experimental results in environmental-based sound event detection and relevant articles for review.
Exclusion criteria: We excluded the articles that inadequately described the dataset, did not discuss the model's implementation, were irrelevant to our review, and exhibited poor writing quality.

2.4 Quality assessment

The inclusion and exclusion strategy of the papers has followed the quality assessment questions to ensure the document's quality. In addition, two subject experts from Sound Event Detection of our institution verified the documents and included those who interpreted their approach clearly with good analysis and validation.

The quality assessment questions are prepared based on the systemic literature review proposed by Kitchenham et al. [1]. The grading system consists of 0 or 1 for grading each question. Each document was verified with quality assessment questions and graded between 0 to 3. The paper graded with 2 or more values led to include the article. Otherwise, the document has been excluded from the list.

We created a set of quality assessment questions for our comprehensive study. Quality Assessment-1 focuses on internal validity, Quality Assessment-2 is designed to assess external validity, and Quality Assessment-3 specifically evaluates possible bias. Two active reviewers carefully examined the chosen documents during the paper selection process to determine the final enrollment of articles for analysis. We utilized the Quadratic Weighted Kappa score concept to measure the agreement between these two reviewers' final assessments.

We obtained the average result of 0.7102 from the above Quadratic Weighted Kappa score, which is a substantial agreement between the reviewers. Table 2 shows the quality assessment analysis to select the final set of papers. Table 3 presents the number of papers selected from different databases. Figure 1 shows the selection process of our final set of papers to review. Figure 2 shows the selected papers year wise count.

Table 2 Quality Assessment analysis

Number of Papers	Quality Assessment Score
72	3
42	2
69	1
27	0

Table 3 Final list of papers from various databases

Database	Paper Count
IEEE	52
ACM	11
DCASE	10
Elsevier	11
Springer	7
Others	31
Total	**122**

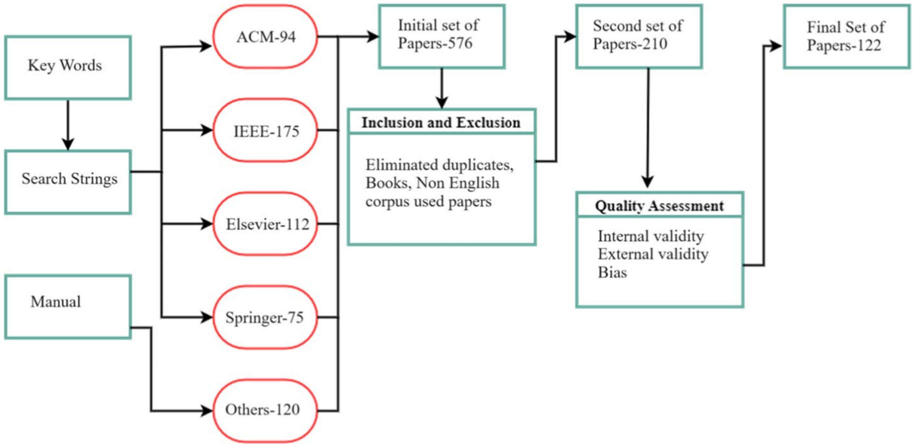

Fig. 1 Selection process

3 Results

3.1 *RQ1*, What sound event detection datasets are accessible for research purposes, and what preprocessing steps should be applied to them?

3.1.1 Dataset

The implementation of machine learning is essential for the automatic detection of sound events. Generally, the machine learning process involves training a model to identify a

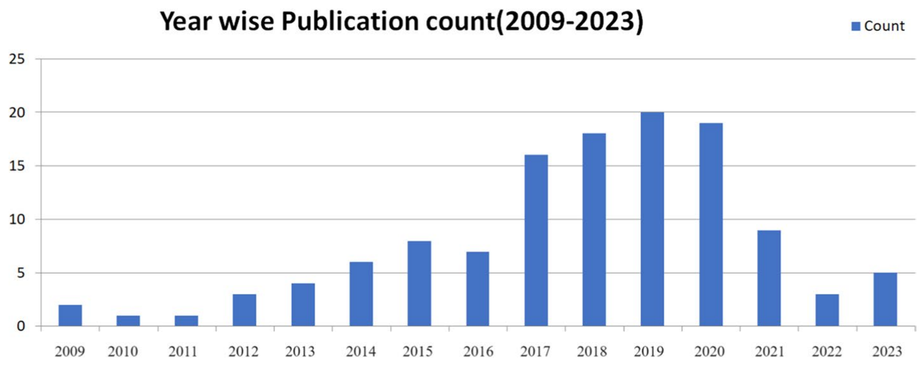

Fig. 2 Year wise papers count

various array of sounds, a task made possible by utilizing appropriate datasets. Sound characteristics can exhibit significant variations depending on factors such as the environment, background noise, and the specific sound events targeted. Different environments, such as urban areas, traffic, offices, natural landscapes like mountains and forests, underwater environments, and various other contexts, present distinct acoustic features.

Researchers need access to diverse datasets encompassing a wide range of real-world scenarios to create effective SED algorithms. This diversity helps algorithms recognize and distinguish between sound events in various contexts. SED has numerous applications in various fields, including health monitoring, robotics, security systems, wildlife monitoring, etc. By training algorithms on datasets that cover these different domains, researchers can develop solutions that can be readily applied in practical situations. Each environment presents its challenges. Urban environments may have complex mixtures of sounds, while underwater environments have unique acoustic properties. By training on a wide array of datasets, researchers can identify the specific challenges of each context and develop algorithms that address them effectively. Creating diverse datasets requires significant data collection, annotation, and duration effort.

Annamaria Mesaros et al. [2] have introduced a new dataset on sound event detection for residential and home environment sounds. They gathered the 587 min of the recorded set and categorized it into 15 numbers of classes. All these records have the sounds of residential areas and home environments. The cross-validation of this dataset has done with MFCC and GMM and achieved good results.

Hyungui et al. [3] provided a method for detecting sound events. The model incorporated the features of a RNN with LSTM units and 1DConvNet. The proposed model was evaluated by using DCASE 2017 Challenge Task-2 Dataset. The log amplitude and mel-spectrogram are used to extract the feature of the dataset. The error rate of the proposed system on the development dataset is 0.07 and F-Score is 96.26.

Karol J. Piczak [54] has provided sound event datasets named ESC to conduct the effect research on acoustic identifications. According to the author, the ESC dataset is categorized into ESC-10, ESC-50 and ESC-US. Here ESC-50 dataset consists of 2000 labeled records with equally balanced 50 classes. All 50 classes are grouped into five types, each assigned ten classes such as animal sounds, water sounds, Human (non-speech) sounds, domestic and urban sounds. The ESC-10 interprets the concept of assigning or selecting of 10 classes to the above-mentioned environment types from 50 classes. Here each environment type is assigned with 10 classes of sounds and 40clips for each class. On the other side ESC-US provide complex knowledge discovery rather than labeled ESC-50 dataset. It consists of 250000 records extracted from field recordings in a short clip of 5 s formats. According to the author, this dataset is more suitable for unsupervised learning.

Pablo Zinemanas et al. [55] provided a dataset, namely MAVD on Urban noise Monitoring. In this dataset author majorly focused on the traffic noise in urban areas, which has records of different vehicle sounds, component sounds, and vehicle action-related sounds. This dataset provided sufficient information for traffic sound to research sound event detection in urban areas. Authors have recorded residential areas, parks with residential neighborhoods, and residential with few shops. The authors prepared 4 h of recordings in this dataset with 21 classes.

Justin Salamon et al. [57] provided a dataset for effective research on sound event detection on urban area sounds. The dataset UrbanSound has implemented for the proposed model contains 27 h of audio records and ten numbers of different classes. The authors implemented MFCC to extract features and effectively classify the data using SVM. Dan Stowell et al. [58] provided a dataset freefield1010 of 7690 records sampled from the field-recording tag in the Free sound audio archive. The authors verified the records with MFCC features and GMM for classification.

Grootel et al. [59] provided a dataset DARE-G1 on everyday sounds of different environments such as streets, nature, home, public, buildings, vehicles, etc. The database consists of 120 fragments of 60-s recordings, resulting in about 1.3 GB of audio. The DARES-G1 collection is available on the DARE-Sounds website, http://www.daresounds.org. David S. Johnson et al. [60] have introduced a new form for dataset creation on sound event detection in federal learning. They have provided a vast dataset named DESED-FL and URBAN –FL, which consists of Domestic Environment Sound and urban sound records. STFT validated the evaluation of the dataset for preprocessing and executed with CNN-based algorithms.

Harsh Purohit et al. [61] have proposed a dataset named MIMII for industrial sounds to find the different kinds of machine and machine activity sounds. They collected 26,092 normal sounds recorded for different industrial machines such as valves, pumps, fans, and slide rails and 6,065 sound records for various abnormal sounds such as contamination, leakage, rotating unbalance and rail damage. In this MIMII dataset, each record has 10 s sound segments.

Lars Hertel et al. [62] have introduced a new dataset for sound event detection named *Freiburg-106,* which contains the human activities sound are recorded as samples. The total audio samples are 1,479 and the entire duration is 48 min. Jort F. Gemmeke et al. [63] collected a dataset of generic audio events, comprising an ontology of 632 audio event categories and a collection of 1,789,621 labeled 10-s excerpts from YouTube videos. As mentioned above, several datasets are available to research sound event detection. Table 4 presents the sample dataset available for sound event detection.

3.1.2 Preprocessing

Sound preprocessing techniques are methods used to enhance, clean, or transform audio data before further analysis or processing. Preprocessing is crucial because raw audio data can be complex, noisy, and high-dimensional, making it challenging to work with directly. The data can be transformed into a more suitable format for further processing or analysis by applying preprocessing techniques [4]. Common preprocessing included Sampling

Table 4 Sample datasets on sound

Reference	Dataset	Classes	Examples	Size(min)
[2]	TUT Sound Scenes 2016	18	954	78
[54]	ESC-US	-	250000	20833
[54]	ESC-50	50	2000	166
[55]	MAVD	21	-	240
[57]	Urban Sound	10	-	1620
[58]	Freefield1010	7	7690	1282
[59]	Dares G1	28	123	123
[61]	MIMII	-	32157	5359
[62]	*Freiburg-106*	22	1479	48
[63]	AudioSet	632	>2 M	>340 k
[64]	SINGA:PURA	14	6547	1092
[65]	SONYC-UST-V2	-	18510	3084
[66]	FSD50K	200	51,197	6000
[67]	USM-SED	27	20000	1660
[68]	Million Song Dataset (2011)	-	1 M	-
[70]	FSD early snapshot (2017)	398	23,519	7140
[71]	ToyADMOS	-	4000	10800
[72]	SONYC-UST	23	3068	511
[75]	TUT Urban Acoustic Scenes 2018	10	8640	1440
[76]	TAU Urban Acoustic Scenes 2019	10	14400	2400
[77]	TAU Urban Acoustic Scenes 2020	10	23040	3840
[78]	Litis Rouen Dataset	19	3026	1513
[79]	TUT Sound Events 2017	6	729	92

rate, NormalizationNoise removal, segmentation, and windowing [5, 99]. The Sampling rate conversion is changing the sampling rate of a digital audio signal while preserving its content. The unit of measurement for the sampling rate is Hertz (Hz), which represents the number of samples per second. 44.1 kHz (44100 Hz) and 48 kHz are the basic sampling rates to provide high-quality audio suitable for most music playback and digital audio applications [8].

The Normalization Scaling of the audio waveform to a standard range, often between -1 and 1, to ensure consistent volume levels across different audio samples. Noise Removal filters out background noise and unwanted artifacts from the audio signal using techniques like noise reduction filters. The Segmenting is used to split the audio into smaller chunks for analysis [10, 11]. Windowing is often applied to divide a more extended audio signal into shorter segments, which can then be processed individually. This is particularly useful when working with non-stationary signals, where the properties of the signal may change over time [17].

Hongning Zhu et al. [21] have proposed a model to classify the Sound Events using the Adopted CRNN on DCASE 2017 Task 3 dataset. In this dataset, audio is recorded using an iPhone XS smartphone. In this dataset, 100 records of 5-min sound samples are there. These sounds were recorded at a 44.1 kHz sampling rate and saved the sound samples on WAV format.

Komatsu et al. [26] have compared the NMF and Random forest models on DCASE-2016 Challenge dataset. They extracted the mel-spectrogram features with a 23 ms window size and a sampling frequency of 44.1 kHz.

Gómez et al. [44] have introduced a model to detect the drone sound by using deep learning. For that, they collected the synthetic dataset using smartphone to record the audio clips. The records are sampled at 44.1 kHz and with a mono-channel audio bit-rate of 64Kbps. All the audio is saved in WAV, and the sound sample length is 6 s each. They investigated with two drones, and to separate the sounds in the dataset, they labeled them Mambo and Bebop.

Mesaros et al. [56] used the GMM to classify the TUT Sound Events 2016 dataset sound events. This dataset has prepared Roland Edirol R09 wave recorder, and the records are sampled at 44.1 kHz with 24-bit resolution. In this model, features are extracted by using MFCC. These features were calculated with a 40 ms frame with a hamming window 50% overlap and 40 mel bands.

Johnson et al. [60] introduced DESED-FL and URBAN-FL datasets on domestic and urban environments. The input audios are downsampled to 22.05 kHz. The STFT was implemented with a FFT of size 2048 input array. A mel-filter bank of 256 mel bands is then applied. Finally, 43 windows are stacked together, resulting in a feature representation (43 × 256x1) of one second.

Adavanne et al. [93] investigated sound classification using the Bi-CRNN model to classify the TUT-SED synthetic 2016 dataset that contains the 566 min with 16 various classes. All the data is divided into 5-s audio chunks. The five-second chunks were short-time Fourier transformed (STFT) with 50 ms frames and 50% overlap, and the results were given in the log magnitude of 40 mel per frame.

3.2 *RQ2*, what are the features extraction methods available for sound events?

The model's performance depends on feature extraction, which is essential for better results. Feature extraction techniques for sound events play a crucial role in various applications such as speech recognition, music analysis, and sound classification. These techniques help in representing audio data in a format that is suitable for algorithms to process and analyze [23, 24]. Feature extraction involves the consideration of various parameters, including waveform attributes, signal energy, pitch, and spectrum, among others. Many techniques are available for extracting features from sound data, including MFCC, Mel-Spectrogram, RMS, ZCR, LPCC, and more [27, 35, 36]. In their research, the researchers employed single and multi-feature extraction methods when analyzing sound data. In our study the feature extractions of sound explained with five categories that are time domain, frequency domain, wavelet domain, Image domain and Cepstral domain [39].

Time domain feature extraction in sound analysis involves quantifying various characteristics of a sound signal directly from its time waveform, such as its amplitude statistics (e.g., mean, RMS), temporal properties (e.g., duration, zero-crossing rate), and statistical attributes (e.g., variance, skewness). These features provide valuable information about the signal's energy, shape, and temporal behavior, enabling applications like speech recognition, music analysis, and sound classification [41, 49].

On the other hand the frequency domain feature extraction in sound processing involves transforming audio signals from the time domain (amplitude vs. time) into the frequency domain (amplitude vs. frequency), revealing critical information about the signal's spectral content. Techniques such as the FFT, MFCCs and spectrograms are used to extract features like pitch, timbre, and spectral characteristics [50, 52].

The Wavelet domain feature extraction for sound involves applying wavelet transforms to sound signals to decompose them into time-frequency subbands, extracting relevant features such as energy, spectral attributes, and statistics from these subbands, and using these features for tasks like sound classification or analysis. This technique provides a comprehensive representation that captures both temporal and spectral information, making it valuable in applications like speech recognition, music analysis, and environmental sound classification, where a more detailed understanding of sound characteristics is required for improved accuracy and robustness [81, 83].

The Image domain feature extraction from sound involves converting audio data into visual representations, such as spectrograms or waveform images, by segmenting the audio, extracting relevant features like frequency content or spectral patterns, converting these features into images, and potentially using machine learning techniques for analysis [86]. Figure 3 presets the various feature extraction methods for sound event detection.

Anusha Koduru et al. [79] have introduced a model to detect the sound and applied various feature extraction techniques such as MFCC, DWT, pitch, energy, and ZCR algorithms. The feature extracted data classified and results verified with SVM, LDA, and Decision Tree. As in this approach, the authors proved the Decision tree has better results with the above-mentioned extraction procedures.

Turgut Özseven et al. [81] introduced a new toolbox to extract the feature named Speech Acoustic (SPAC), which integrated with sound quality, LPCC, MFCC, pause rate, etc. This new toolbox is implemented on the Matlab platform with a GUI interface. In this toolbox, the overall accuracy on the different parameters was obtained 82.1% with SVM classification. A Dang et al. [82] have done their research for efficient classification with CNN with multi-feature extraction. They integrated the log Mel band energies and MFCC to extract the feature. They have implemented three different window sizes with 0.02,0.04, and 0.06 s. They concluded their classification with CNN and obtained the 85.9%.

Adam Glowacz et al. [83] have implemented a new method to extract the feature to find the fault diagnosis of commutator motor named MSAF-15-MULTIEXPANDED-8-GROUPS (Method of Selection of Amplitudes of Frequency Multi expanded 8 Groups). This approach followed a sequence of steps such as amplitude normalization, windowing, FFT, and the proposed method MSAF to extract the features. This MSAF-15-MULTIEXPANDED-8-GROUPS approach formed a feature vector of 1–15 frequency components. In their model amplitude of frequencies, Multi expanded 8 Groups on the bases of spectra of the sound signal. The authors defined the average feature vector to extract the feature in the neatest mean classifier as given below.

$$afv = \frac{1}{p}\sum_{i=1}^{p} y_i$$

p defines for number of essential frequency components and y_i represents the value of essential frequency component with i index.

Muqing Deng et al. [84] have done their research with the improved methodology of MFCC and CRNN to detect the sound. Fifth order bandpass procedure has been used to preprocess the sound data to remove the low-frequency artifacts. The feature of the sound was extracted with first-order MFCC (ΔMFCC) and second-order MFCC (Δ2MFCC), which the authors derived. Finally, the model classification was implemented with CRNN and achieved an accuracy of 98.34%.

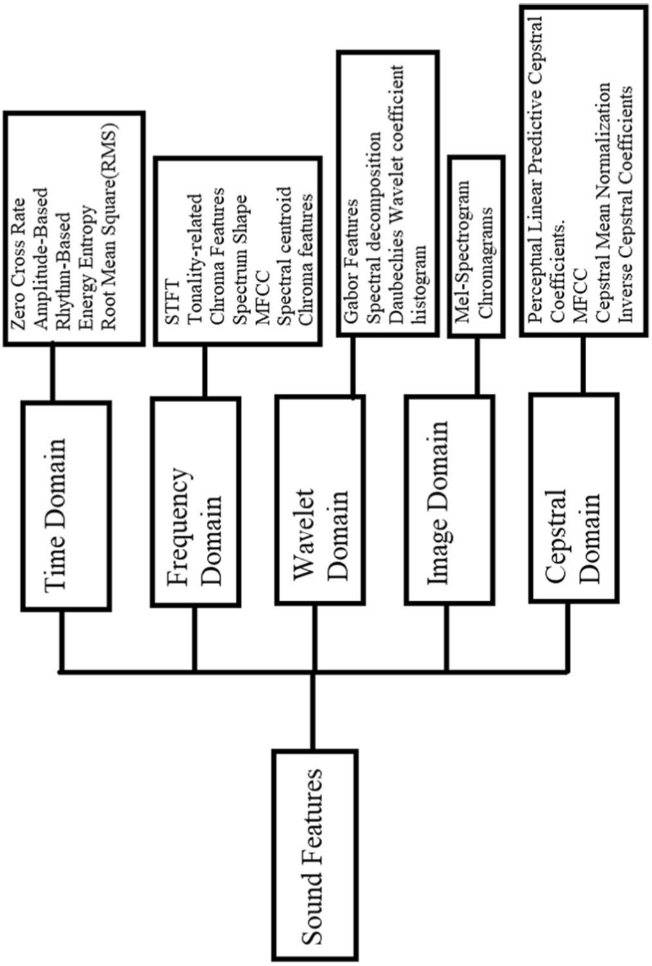

Fig. 3 Feature extraction techniques of SED

Sharath Adavanne et al. [90] have processed their research with a set of different feature extraction on urban sounds. Their feature extraction was extended with LSTM in RNN to gain better results. The features such as MFCC, Pitch, Mel Spectrogram, tdoa, and etc. were applied to the sound events of home and residential environmental sounds.

Based on the preceding discussion, the authors have employed various feature extraction methods to enhance their research results. They have systematically extracted multiple features from a single dataset, aiming to enrich their research findings. Among all papers, MFCC has been implemented in more papers with the combination of other feature extractions. Because the MFCCs are a compact representation of an audio signal. It contains information about rate changes of different spectrum bands and concisely describes the spectral envelope's shape. So, most researchers have preferred using the MFCC as a Feature Extraction technique.

After conducting a comprehensive analysis, multi-feature extraction processes become more complex but provide more accuracy than single-feature extractions, and it is based on the dataset only. In our future research, we will prefer to extract relevant features from the sound samples. Therefore, we will select the proper feature extraction combinations to extract the feature from sounds appropriate for the forest environment. Figure 4 shows Different feature extraction ratios among all articles. Table 5 presents the various feature extraction methods in different research.

3.3 *RQ3*, What are the currently available Machine Learning and Neural Network-based models for sound event detection?

Researchers presented numerous models and methods for sound event detection, each formulated to address the unique limitations of different environments and datasets. This broad

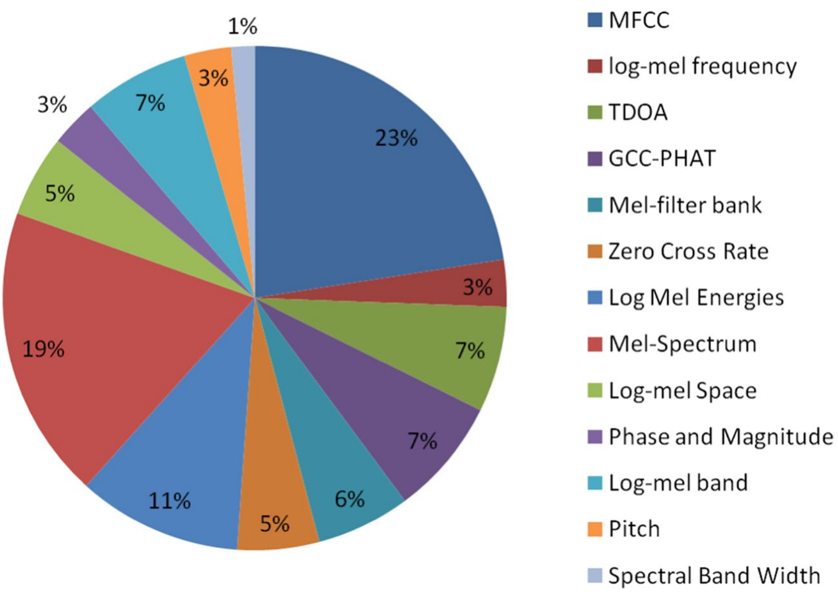

Fig. 4 Different feature extractions ratios among all papers

Table 5 Various feature extraction methods in different researches

Reference	Model	Features	Results	Limitations of features
[12]	CNN	MFCC + Mel-Filter Bank	Error rate:0.78	• MFCCs and Mel-Filter Bank features capture similar spectral information, which can lead to redundancy • They do not capture long-term temporal dependencies in the audio signal • Sensitive to noise
	Random Forest		Error rate:0.83	
[37]	CRNN	Spectro-temporal features + log mel-band energies	Accuracy:95.7%	• Produce High-Dimensional feature vectors that lead to increased computational complexity • The choice of window size and overlap in the STFT affects the time–frequency trade-off
[40]	CRNN	MFCCs + ZCR + DWT	Accuracy:64%	• Increased training times Combining multiple features can increase the risk of overfitting
[81]	ANN	MFCC + Pitch + Energy + ZCR + DWT	Accuracy:51.19%	• Pitch and energy may be computed over longer segments • Pitch extraction is well-suited for only tonal sounds with clear harmonic structures • ZCR is also sensitive to changes in signal amplitude
	KNN		Accuracy:64%	
	HMM		Accuracy:76%	
	SVM		Accuracy:77%	
	LDA		Accuracy:65%	
	Decision Tree		Accuracy:85%	
[83]	CNN	MFCC	Accuracy:84.4%	• MFCCs and Mel-Filter Bank features capture similar spectral information, which can lead to redundancy • Log-mel features are typically computed in fixed-size frames, which mean that temporal information is partially lost
		Log-Mel band	Accuracy:82.2%	
[87]	Coupled NMF	Mel-Spectrogram	Accuracy:81.6%	• Mel-spectrograms only capture magnitude information and discard phase information, • When multiple sound events overlap in time and frequency, it can be challenging to accurately represent and detect those using mel-spectrograms alone
[90]	CRNN BiCRNN	Log mel energies	Accuracy:64.7%	• Log mel energies features can be sensitive to background noise. Long mel energies features are computed using a fixed frame length, typically 20–40 ms This fixed frame length might not capture sound events that vary in duration

field encompasses algorithms such as HMM, NMF, SVM, Random Forests, DNN, CNN, RNN, CRNN, and numerous others [3, 28]. The selection of these algorithms is guided by their suitability for particular tasks and ability to adjust to varying data characteristics.

Moreover, researchers have fine-tuned their approaches by manipulating parameters such as frame size, frequency modulation, normalization methods, and filters. These elements are critical in preprocessing audio signals and extracting meaningful features, laying the foundation for effective sound event detection. In addition to these algorithmic choices, techniques like GRU, LSTM, BiGRU, and BiLSTM have been incorporated into models, enabling them to capture complex temporal and spatial dependencies in audio data [90, 91, 93].

This multifaceted approach to sound event detection reflects the dynamic nature of the field, where researchers continually innovate and adapt established techniques to push the boundaries of what is possible. As computational capabilities grow and datasets become more extensive and diverse, the potential for more accurate and robust sound event detection models continues to expand.

3.3.1 Machine learning models

Machine learning algorithms commonly applied to sound event detection tasks encompass a variety of approaches. Support Vector Machines, Random Forests, k-nearest Neighbors, Gaussian Mixture Models, Hidden Markov Models, Naive Bayes, Decision Trees, Logistic Regression, Gradient Boosting, Principal Component Analysis, and Multiple Instance Learning are among them. These algorithms effectively classify and detect sound events, leveraging features derived from audio data or extracted representations to make classifications or identify temporal patterns. The choice of algorithm depends on factors like data complexity, desired interpretability, and the nature of the sound event detection problem at hand, providing diverse options for researchers and practitioners in the field.

Kawaguchi et al. [4] have proposed a system to detect sounds. The proposed system consists of front and back end portions to detect sounds in the environment of factories. In the front end, they implemented Blind Dereverberation (BD) and ASE algorithms to improve sound event detection in a noisy environment. The anomalous-sound-extraction (ASE) algorithm was implemented with NMF. The system's back end runs based on feature extraction and deep auto-encoders. *Kawaguchi et al.* [7] have incorporated the NMF and Semi-supervised NMF(SSNMF). The proposed model is mainly related to a non-negative matrix under approximation (NMU).

Selver Ezgi et al. [8] have defined a model for Multimedia event detection (MED), and they extracted MFCC features with varying numbers of coefficients in the first phase. In the second phase, extracted MFCC features with variable window and hop sizes to observe how they affect classification performance. To improve evaluation outcomes, the third phase applied the 5-fold cross-validation approach and the grid search strategy to change SVM parameters. SVM classifiers are implemented with a confusion matrix where occurrences in an actual class are represented in the rows.

Man-Wai et al. [9] defined a new model to balance detection accuracy with power consumption to extend battery life. Various audio characteristics, performance, and power efficiency were examined with SVM. This supports the idea of inherent complexity, which states that the polynomial SVMs' scoring model can be defined as a sorted order. The authors introduced an Intrinsic Complexity that reduced the use of CPU and did not affect the polyphonic SVM classification process.

HuyDat et al. [11] have offered a new classification method based on probability distance SVM and explored a parametric approach for describing audio signals. An SVM frame uses a sub-domain temporal envelope (STE) distribution and kernel technologies for subdomain probability distance (SPD). The authors derived the SPDSVM framework. In this model, sound features were extracted using Complex Mel-Gabor Filters and the temporal filter-based STE, which was an alternative to the MFCC frame-based features. Then applied, the Gamma modeling produced the two-way output, such as mapping SPD kernel (Linear SVM) and direct SPD kernel (Kernal SVM).

Phan et al. [13] have classified target variables on random forest regression, a collection of distinct regressions to detect the sound event. The model was learned using displacement information. Audio Event Classification (AEC) that operates on segmented AEs may be processed quickly with various pre-made classifiers and acoustic characteristics. The experimental model was implemented on UPC-TALP database.

Xianjun et al. [14] have a general goal of their work is to increase acoustic event detection accuracy .they approach a regression via classification (RvC) to the grouping with a random forest technique used to detect auditory events. In this paper, RFR-UW and RFR-W, as well as a suggested methodology based on AED, Random Forest Classification, are being examined for localization. This study investigated the identification of solo sound occurrences and proposed a new approach for localization based on random forest classification. According to their research, Compared to RFR systems, the RFC system has an equal error detection rate for the whole database, demonstrating its overall higher efficiency.

Stoller et al. [16] proposed a model to simplify the process of selecting the best method for sound event detection. They presented the Wave-U-Net model, which adopts 1D U-Net architecture. Against the previous system, Wave-U-Net avoided artifacts at the boundaries of the output window. Authors said that U-Net architecture on magnitude spectrograms had achieved new sound resources with state-of-the-art output division for music and speech reverberation.

Park et al. [17] have overcome the problem by presenting a simple strategy for restricting the spectral envelope obtained from linear prediction. They applied a blind scenario in which the instruments were unknown. The NMF's spectral bases are partitioned before iteration and then restricted to resemble the recovered envelopes to approach the mixed spectrogram. The spectral envelopes of each basis are calculated using LPC, and all envelopes belonging to the same group were averaged. These two techniques are based on the source-filter model with LPC but they acquire the spectra envelopes (or spectral bases of the NMF) without transforming them into time-domain signals.

Bisot et al. [19] have proposed a Non-negative Matrix Factorization model for detecting overlapping sound occurrences in real-time audio systems. Some works can be considered from the NMF, where isolated events are learned by spectral templates. The goal of supervised NMF is to combine the NMF as well as classification phases into a single bi-level optimization problem. In this model, both unsupervised and supervised NMF systems can compete with DNN-based approaches, according to the results. Finally, by using the Kullback-Leibler TD-NMF on segments with overlapping events, the potential of the Kullback-Leibler TD-NMF for highly polyphonic AED was proven.

3.3.2 Neural network models

Sound event detection has significantly improved by adopting deep learning and neural network models. CNNs are widely used for their ability to capture spectral features from audio data, often processing time-frequency representations like spectrograms or MFCCs. RNNs, including LSTM and GRU variants, excel at modeling temporal dependencies within audio sequences, making them valuable for SED. CRNN merges the benefits of both CNNs and RNNs, effectively capturing local and global context. Initially designed for natural language processing, transformer-based models are being adapted for SED tasks to capture long-range dependencies. Attention mechanisms enhance focus on critical audio components, while hybrid models combine different neural architectures. Data augmentation, pretraining, and ensemble methods further enhance SED performance. The choice of model depends on specific SED requirements, data availability, and computational resources, and the field continues to evolve with emerging techniques and architectures.

Hyungui et al. [3] have provided a method for detecting unusual sound events that combine a RNN and LSTM units and a 1D Convolutional Neural Network (1DConvNet). The system is evaluated by using DCASE 2017 Challenge Task 2 Dataset. They have implemented a deep learning model by RNN-LSTM without fading gradients, which can cause problems with long-term sequence learning. Returned features of the RNN-LSTM layer are fed into a fully connected layer with 128 hidden neurons. The results reveal that their strategy performs best when it comes to glass breaking, baby wailing, and gunshots.

Adavanne et al. [5] have implemented the concept of SED with a weakly labeled data set using CNN. The mel-band energy features are initially extracted from the sound and fed into stacked CNN and RNN. They have prepared strong and weak labeled data in this model. The proposed system originally mapped the strong labels, and then the strong ones were mapped to the weak ones.

Qiuqiang et al. [99] proposed a model integrated with CNN-Transformer, which is similar to CRNN. Their approach implemented threshold optimization like Mean Average Precision (MAP) for SED. This system categorized the weakly labeled data into the segment and clip-wise training data units. The authors derived the loss function on segment-wise trained data and clip-wise trained data. SED results were found in each segment, then aggregated and provided overall results. The authors implemented the improved architecture of LSTM called BiGRU with CNN.

Xia et al. [15] proposed a model to optimize the kernel size in a neural network approach to detect sound events. They have used DCASE Challenge 2017 Task 4 and DCASE Challenge 2018 Task 4 for the proposed model. According to the model, CNN with optimized kernel size provided better results on selected datasets.

Chan et al. [18] Proposed model integrated NMF with CNN. In this model, they proposed five layers of CNN out of this one layer as the input layer, and four layers are convolutional. The kernel size of this model is 5×5 with a padding size of 2×2 and strides 1×1. The dataset used for this model is DCASE 2018 task 4 test set. The best-performing model was trained using the combination of Weakly Labeled and Strongly Labeled Synthetic Data, which achieved an F1-score of 31%.

Adavanne et al. [19] have proposed a model with CRNN with Bi-LSTM. In this model, 19 h large TUT-SED 2016 dataset was used. It was observed that sound events were better recognized using binaural spatial features than monaural features.

Grondin et al. [28] have introduced a model to find the sound event localization and detection. They have implemented the CRNN to detect the sound and find the direction of arrival. MFCC and Mel-spectrogram features are extracted to find the sound events. TDOA features are extracted to recognize the localization of the sound system. Adavanne et al. [29] proposed model with CRNN to find sound events and localization DCASE 2019 challenge to promote SELD research. They have extracted the TDOA and Mel spectrogram features from the dataset.

Dang et al. [82] have proposed a model to detect the sound event using CNN approach. They have extracted multiple features from the dataset to gain better accuracy. In this research, they have used TUT Acoustic Scenes 2016 dataset. The proposed approach obtains a high accuracy of 85.9%.

Deng et al. [84] proposed a system for heart sound classification based on CRNN algorithm. They have introduced a new approach named improved Mel-frequency cepstrum coefficient features to extract the features. The Improved MFCC extracts the first and second-order difference features of MFCC. The PhysioNet Computational Cardiology (CinC) 2016 Challenge Database is used to test the effectiveness of the proposed algorithm.

Parascandolo et al. [89] The authors have presented a new approach to sound event detection. Authors build their proposal with RNN and Bi-LSTM. They have tested on real-life recordings with 61 classes from 10 different everyday contexts. Raw audio signals are then input into the proposed system. Database recordings of 10 to 30 min were acquired with a binaural microphone at 44.1 kHz sampling rate and 24-bit resolution. They have extracted Mel bands energies as a feature from the dataset. They compared the average scores for each context for FNN, BLSTM, LSTM, and BLSTM trained with the augmented data.

Cakır et al. [91] proposed a model with multi-label data classification with CRNN for polyphonic sound event detection. They have implemented three datasets. TUT SED 2009,TUT-SED 2016, CHiME-Home datasets and compared the results with FNN, CNN, RNN, and GMM baselines.

Liwei Lin et al. [96] build a model to detect sound events. For their model, they have used CNN for better classification. They have extracted multi-features such as MFCC and Mel spectrogram to analyze the data. According to the research, the proposed model secured the F-Score as 45.43% for the validation dataset and 42.7 for the test dataset.

Adavanne et al. [98] introduced an approach with CRNN to detect sound events. 3D-CNN layer was used to implement the inter- and intra-channel convolutions. The proposed C3RNN method recognizes overlapping sound events and performs better SED. According to the proposed model, the overall F-score improved by 7.5%, overall ER improved by 10%, and 15.6%.

Incze et al. [108] have implemented CNN system to find and classify the bird. Initially, preprocess the dataset with STFT and extract the Spectro temporal information from the dataset. The window size of 448 is used during the STFT because the used neural network needs a fixed size of 224×224 pixels input. The training on the Jet colormap shows higher accuracies by an average of 7.4%.

Chatterjee et al. [109] have proposed Transposed CRNN to detect sound events. In their model, they have implemented the TUT sound Event 2016 development dataset and extracted Instantaneous Frequency spectrogram features. They have compared the performance of the proposed model TCRNN with the CRNN for the joint Acoustic Scene Classification and SED task. They have achieved CRNN F1 score is 98.0 and Error Rate is 1.00. The Proposed TCRNN F1 score is 98.3, and Error Rate is 0.83.

According to the above review, the authors implemented machine learning algorithms with different feature extraction techniques to find the sound event classification. Most researchers have implemented the CRNN approach underlying LSTM or GRU to achieve good results. On the other hand, some researchers have implemented only either RNN or CNN. Machine learning approaches such as SVM, K-NN, Decision Tree, GMM-HMM, NMF, and Random Forest are not preferred in most research compared to neural network models. Figure 5 shows the paper count of my survey on sound event

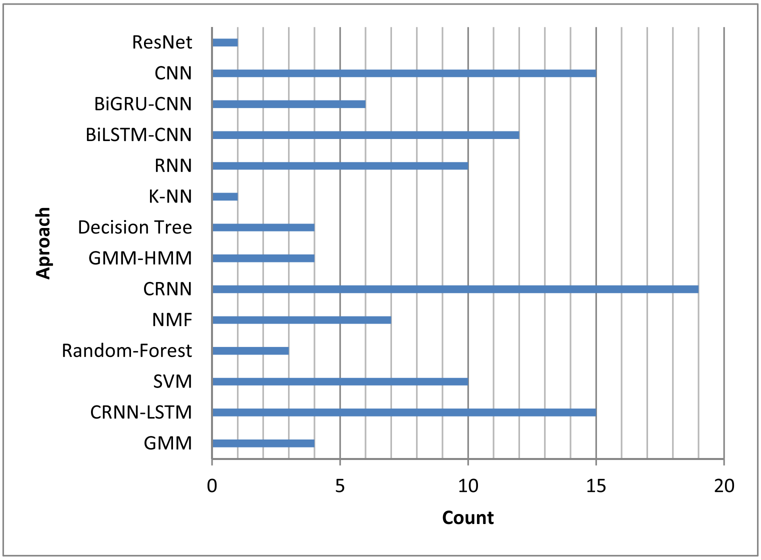

Fig. 5 Differnt approaches used to detect the sound event

detection with various approaches. Table 6 describes the different machine learning and neural network models discussed in our review.

3.4 *RQ4*, In recent research, which environments are typically selected for Sound Event Detection?

Sound event detection research typically begins with researchers selecting distinct environment types for data gathering. These categories encompass residential areas, where researchers explore typical home sounds; environmental sounds, including natural sounds from forests or wildlife; urban environments with a variety of sounds like traffic and human activities; and even applications in medical research, such as detecting heartbeat sounds. The broad areas to conduct the SED classifications are office spaces, parks, hospitals, forests, Hotels, shopping malls, space research, motors, and many diverse areas. Meanwhile, SED focuses on detecting momentary changes within the ongoing soundscape, such as recognizing a dog's bark, the sharp sound of a gunshot, a door being knocked, or the constant hum of an engine. In essence, these chosen environments and categories serve as important starting points for researchers in their pursuit to create sound event detection systems with a wide range of practical applications.

Xianjun et al. [37] introduced a new approach for acoustic event detection in the home and street areas. Several types of research are conducted on urban sound event detection to classify the various sounds. Bird sound event detection system to discover bird category, extinction, and survival of birds. Ornithology of researchers mostly prefers to use the SED system to identify the state of birds in urban and forest areas.

William E et al. [44] have investigated drone sound detection that helps to detect the drone sound at country borders to avoid illegal entry. The authors have used a synthetic dataset recorded in different environments. Sobieraj et al. [48] have examined bird sound

Table 6 Note: This data is mandatory. Please provide

Reference	Dataset	Feature Extraction	Model	Results	Environment
[2]	TUT Acoustic Scenes 2016	MFCC	GMM	Accuracy:72.5%	Residential area and Home
[3]	TUT Rare Sound Events 2017’	Log-amplitude mel-spectrogram	1D-CRNN	Accuracy:93.1%	Gunshots and Home
[9]	Synthetic Dataset	MFCC,LPCC and Mel filter-bank spectrum	SVM	Accuracy:87%	Traffic
[11]	Synthetic Dataset	MFCC	SVM	Accuracy:96%	Office,traffic
[14]	UPC-TALP database	log-spectral, zero-crossing rate and spectral bandwidth	random forest	Accuracy:91.56%	meeting room environment
[18]	DCASE 2018 task 4	STFT	NMF-CNN	F1-Score: 75.2% Error Rate:0.17	Audio tracks
[19]	DCASE 2016	Mel spectrum	Task Driven NMF	Accuracy:82%	real life audio
[20]	TUT Acoustic Scenes 2016	Mel spectrum	CNN	Accuracy:78%	residential area and city
[28]	DCASE 2019 development	Cross-Spectrum Phase, Cross-Spectrum Log Amplitude,GCC-PHAT	CRNN	F1-Score: 92.2% Error Rate:0.14	environmental sounds
[30]	Synthetic Dataset	STFT, Mel-spectrogram	CRNN	F1-Score: 91% Error Rate:0.14	environmental sounds
[31]	DCASE 2019 development	log mel space and GCC-PHAT	CNN	F1-Score: 93% Error Rate:0.13	environmental sounds
[32]	TUT Sound Events 2018	Phase and magnitude spectrum	CRNN	F1-Score: 93% Error Rate:0.13	Music,Real life
[33]	DCASE-2019	Log-Mel Feature, Phase Feature	CNN	F1-Score: 96.9% Error Rate:0.13	Urban Sounds
[37]	TUT Sound Event 2016	short and the log mel-band energies	CRNN	F1-Score: 48.32% Error Rate: 0.59	Home,City and residential
[38]	MedleyDB dataset	MFCC	CRNN	Accuracy:82%	Music
[39]	MAPS dataset	NMF vector	NMF and CRNN	Accuracy:84%	Music

Table 6 (continued)

Reference	Dataset	Feature Extraction	Model	Results	Environment
[40]	MedleyDB datase	Pitch,MFCC&mel-spectogram	CNN	Accuracy:84%	Music
[41]	MedleyDB Dataset& MIR-1 K dataset	STFT,Pitch	CRNN	Accuracy:82%	Music
[42]	HST datasets	Log Mel-band energies	CRNN	Accuracy:97.4%	High-Speed Train Bogie
[48]	Warblr & freefield1010	NMF vector	NMF	Accuracy:80.1%	Bird Audio Detection
[49]	ADC2004, MIREX 05 and MedleyDB datasets	temporal features	CRNN	Accuracy:86.1%	Music Tracks
[53]	DCASE 2018 Task 5 database	Mel-spectrogram& NMF based features	CNN	Accuracy:95%	Urban Sounds
[55]	MAVD-traffic	MFCC	CNN	F1-Score: 63.1% Error Rate: 0.59	Traffic
[69]	HS data	wavelet packet transform and MFCC	SVM and FCN	Accuracy: 96.37%	Heart sound detection
[73, 74]	TUT Sound Events 2016	log mel-band energies	CRNN	F1-Score: 42.8% Error Rate: 0.64	Home,Residential Area
[79]	TUT Acoustic Scenes 2017	log mel-band energies	GMM-HMM	F1-Score: 72.7% Error Rate: 0.53	Environmental,Urban
[81]	RAVDESS	MFCC + Pitch + Mel Energy + ZCR + DWT	SVM Decision tree	Accuracy:70% Accuracy:85%	Speech
[82]	EMO-DB, EMOVA, eNTERFACE05 and SAVEE	MFCC	SVM K-NN	Accuracy:69.5 Accuracy:58.7	Speech
[85]	PhysioNet/CinC challenge 2016	improved MFCC	CRNN	Accuracy:97.34%	heart sound signals
[86]	CLEAR 2007 and DARES-G1 database	MFCC	GMM-HMM	Accuracy:92%	Urban and environmental sounds
[87]	real-life recordings	mel spectrum	NMF	Accuracy:57.8%	real-life recordings
[89]	Synthetic Dataset	MFCC	CNN	Accuracy:64%	Urban areas

Table 6 (continued)

Reference	Dataset	Feature Extraction	Model	Results	Environment
[91]	TUT sound events detection 2016 database	Log mel-band energy, tdoa and pitch	RNN-LSTM	Error Rate:0.82	Residential area and Home
[92]	CHiME-Home dataset and *TUT-SED 2016*	mel band energies and MFCC	CRNN	F1 Score:68.3% Error Rate: 0.3	Residential area
[93]	CHiME-Home dataset and *TUT-SED 2016*	MFCC	CNN-BLSTM	F1-Score: 74.0% Error Rate: 0.56	Residential area
[97]	ESC-50 dataset, UrbanSound8K	MFCC	CNN	Accuracy:81%	Public and Urban sounds
[98]	DCASE 2019 Task 4 Dataset	log mel-bank magnitudes	CNN	F1-Score: 69:06%	Home
[100]	DCASE 2017 Task 4 dataset	*STFT*	CNN-biGRU	F1-Score: 68.4% Error Rate: 0.68	Traffic
[101]	DCASE 2018 Task 4	*MFCC*	CNN	F1-Score: 67.3%	Residential area and Home
[104]	everyday environments	*MFCC and* Log Mel-band energies	DNN	Accuracy:63.8%	Environmental sounds
[105]	Urban-Sound8K dataset	Log Mel-band energies	CNN	Accuracy:97.03%	Urban sounds
[108]	BirdCLEF 2017	MFCC	CNN	Accuracy:68.6%	Bird Sound

detection to classify the different kinds of birds and the authors used the Warblr dataset in DCASE 2017 challenge. Deforestation is a challenging task for all governments to save forest lands. By using SED, identify the tree-cutting sound in the forest that leads to reducing deforestation.

Muqing Deng et al. [84] have done their research with the improved methodology of MFCC and CRNN to detect heart sounds. For that, they have used the PhysioNet dataset. Fifth order bandpass procedure has been used to preprocess the sound data to remove the low-frequency artifacts. Finally, the model classification was implemented with CRNN and achieved an accuracy of 98.34%. Mondal et al. [94] have done their research on biomedicine. Apart from that, they have investigated lung sound event detection to diagnose the lungs related diseases. Some of the researchers selected the home environment sounds. This home environment sound event detection system can provide security from thieves' illegal entry.

Sharma G et al. [107] introduced an investigation on tree cutting in the forest. The author used sound event detection methodology to classify the tree cutting sound in a noisy environment, for they implemented CRNN to identify the tree cutting event in the forest. Several countries are investing in border security systems to avoid illegal entries. But presently, intruders are finding many paths to enter. Apart from that, drone flying is one unlawful action from intruders. Several types of research have been done to find out illegal drone entries. Drone sound detection is one of the best solutions to find the drones that also come under sound event detection. Speech sound classification is also a significant challenge for researchers.

In the field of sound event detection, various types of research are conducted. Our upcoming work involves detecting sound in forest environments, explicitly identifying the different sounds of tree-cutting, which can be made with instruments such as an axe, machine saw, and chain saw. We will create a synthetic dataset of tree-cutting sounds to enhance our research. We have observed a need for more investigation in the forest environment and aim to contribute to this area with our future research. Figure 6 shows an implementation of sound event detection in distinct environments.

3.5 *RQ5*, What are the challenges/limitations in the current research?

Sound event detection research aims to automatically identify specific sounds or events within audio data and improve detection accuracy. However, it faces several challenges and limitations, including the variability of sound events, the need for extensive and accurately annotated datasets, scalability issues, the robustness of algorithms to noise, limited availability of data for certain events, the challenge of generalization to new events, real-time processing requirements, ethical and privacy concerns, defining appropriate evaluation metrics, and the interdisciplinary nature of the field. Researchers are actively working to address these challenges and advance the state-of-the-art in sound event detection through advancements in machine learning, signal processing, and access to diverse datasets.

Hyungui et al. [3] have provided a method for detecting unusual sound events that combines a CNN with long short-term memory units. The system is evaluated by using DCASE 2017 Challenge Task 2 Dataset. LSTM is also challenging because it will take longer to train and require more memory and restrict execution time. The use of LSTM with CNN will become more complex to implement. Basically, the sequential behavior of LSTM prevents the use of parallelization.

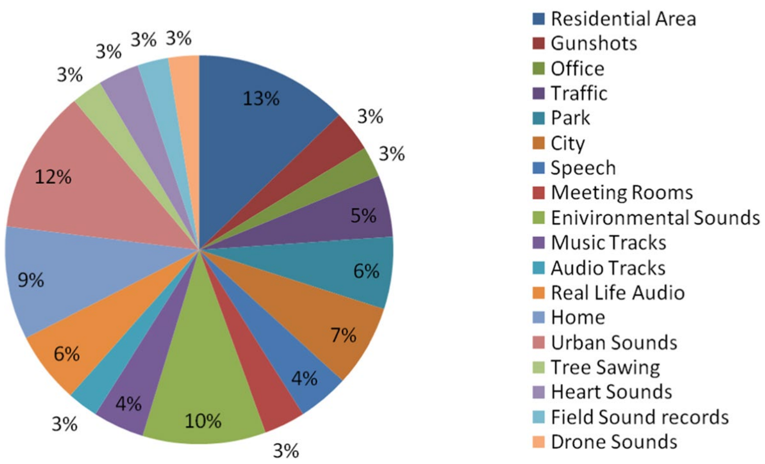

Fig. 6 Implementation of sound event detection in distinct environments

Xianjun et al. [37] introduced a new approach for acoustic event detection with CRNN. They introduced a new augmentation process called AC-GAN. Features are extracted by using MFCC and Mel-Spectrogram. Virtanen et al. [85] have approached their research on two levels. One is to identify the contexts using GMM and sound events modeled using HMM. Initially, audio contexts of signals are detected to convince the sound event detection using HMM. In the first level using GMM, the monophonic sound sequences are provided with the help of Viterbi decoding. The GMM-HMM achieved less accuracy in sound event detection compared to other models. MFCC will not be a good feature extraction process in the GMM-HMM based models. The window size increases then, which leads to reduce accuracy. This approach also needs a large set of the training dataset.

Heittola et al. [86] have implemented their idea of sound event detection using supervised NMF. The goal of supervised NMF is to combine the NMF and classification phases into a single bi-level optimization problem. The basic plan is to learn discriminative non-negative spectral template dictionaries to lower classification costs. There are some challenges to implementing the NMF, such as every training dataset, the models related to NMF need to initiate the clustering, and perfect clusters only derive from several inputs. The training dataset should not contain the overlap events.

Parascandolo et al. [89] have implemented a model using spatial and harmonic features to detect sound events. STFT features are extracted to estimate the pitch and periodicity. Likewise, TDOA features and mel band features are extracted. In this model, they used the LSTM and CNN combination to acquire better results.

Cakır et al. [91] have introduced a new way of sound event detection using LSTM-RNN by implementing different harmonic features. They utilized the TUT sound events detection 2016 database to detect mono-channel audio and multi-channel audios. In this model, they extracted several kinds of features and implemented the RNN-LSTM to classify the data accurately. According to the author's stance, they achieved good results on residential area sounds with mel and TDOA feature extraction.

Virtanen et al. [93] compared the low-level features with high-level feature extraction and proved that low level features provided better results than high-level features. The authors verified their model with TUT-SED 2009 and TUT-SED 2016 datasets. High-level features like log mel-band, TDOA, and dominant frequencies are concatenated for feature extraction implemented on CRNN. This approach compared low-level features like GCC-PHAT, and auto-correlation with CRNN. With these two approaches, low-level feature extractions gain better results. But CRNN architecture becomes complicated to classify the sound.

Heittola et al. [103] proposed a model to detect overlapping sound events using multi-label deep neural networks. The spectral features are extracted to perform optimal classification. The number of hidden layers are 2, the number of filters are 800, the context window is five frames, and the median filter is 10 frame window. The authors proved that the neural networks achieved 19% more accuracy than Hidden Markov models.

Piczak et al. [102] Proposed a model to classify the environmental sounds by using CNN. In this model, two convolutional layers with max-pooling and 2 fully connected layers are trained on a low-level representation of audio data. The neural network models need more data to provide optimal results. But CNN does not encode the position and orientation of the object. The CNN models with the convolution filters research models can extract the features independently, they won't extract the features sequentially.

The neural network models take more time to take the train if they define with more layers.

Kao et al. [105] introduced a new model based on Faster-CRNN to analyze the process event level along with frame level named Region-CRNN. According to the existing models, the predictions have to be done initially at the frame level, and then reaming processes used to find event level prediction.64-dimensional log filter bank energies (LFBEs) are calculated for each frame. Then they aggregated to prepare the input spectrogram. The error rate was evaluated and compared with CNN,CRNN_GUU and CRNN_BiGRU on DCASE 2017 dataset and proved that CRNN_BIGRU achieved better results. BiGRU creates complexity in implementation.

Cakir et al. [106] researched bird sound classification using CRNN on the dataset TUT-SED 2009. In this model, they implemented 3 layers of CRNN with filter size 256 of type GRU. They proved that CRNN would provide a better result than CNN. But CRNN Architecture becomes complicated to classify. CRNN is not a simplified model for gain accuracy. Table 7 presents the various machine network models and limitations according to my review's selected papers.

4 Evaluation metrics

The computational metrics for Sound Event Detection should yield satisfying results from an application perspective. Typically, SED is implemented using Machine Learning and Neural Network models. These models validate classification, learning, training, and verification through testing data samples to derive results [96, 117].

Machine learning or neural network-based algorithms are mostly preferred to classify the sound in two types of evaluation metrics measure their evaluation are implementing they are segment-based evaluation and event-based evaluation metrics. The segment-based evaluation metrics mostly concentrate on time intervals with active sound events, while the event-based evaluation evaluates the system's ability to identify individual instances of sound events. Figure 7 shows the segment based and Event based metric evaluation process.

Table 7 Machine learning models with limitations

Reference	Feature Extraction	Model	Limitations
[2, 78]	MFCC	GMM	Each event class has to be modeled by an individual HMM Each dataset class must be process with individual HMM GMM will not work properly when the window size is large. MFCC is basic feature extraction and many not suitable for proper classification on large dataset
[3, 29, 30]	Log-amplitude, mel-spectrogram	1D-CRNN-LSTM	CRNN not simplified model for gain accuracy LSTM creates complexity in implementation Parallelization occurs with LSTM
[8, 9, 11]	MFCC, LPCC and Mel filter-bank spectrum	SVM	MFCC is basic feature extraction and many not suitable for proper classification on large dataset SVM not suitable for large dataset SVM not provide better accuracy for overlapping target classes SVM points data below or above of hyperlane so that SVM no probabilistic clarification
[12–14]	MFCC	random forest	MFCC is basic feature extraction and many not suitable for proper classification on large dataset Large number of trees limits the prediction rate Need provide more data sample to train
[17–19, 22]	MFCC, Mel spectrum	NMF	NMF hard to produce more accuracy because it is NP-hard Another issue with NMF is that there is not guaranteed to be a single unique decomposition Single unique decompositions not possible by NMF Perform the clustering for every newly created training dataset Need to set proper threshold value
[28, 30]	Cross-Spectrum Phase, Cross-Spectrum Log Amplitude,GCC-PHAT	CRNN	Architecture becomes complicate to classification CRNN not simplified model for gain accuracy Additional audio feature extractions make complex and little improvement
	STFT, Mel-spectrogram	CRNN	
[20, 31, 55]	log mel space and GCC-PHAT	CNN	Long temporal context information not provided by CNN
	MFCC	CNN	Lots of training data need to prepare to gain accuracy CNN becomes slow when implement the maxpool Training process takes long time with multiple layers

Table 7 (continued)

Reference	Feature Extraction	Model	Limitations
[90, 91]	spectral features Log mel-band energy, tdoa and pitch	RNN-BLSTM RNN-LSTM	Gradient vanishing and exploding problems Training process of RNN is very complex RNN becomes slow when use activation functions such as tanh and relu for long sequences Since BiLSTM has double LSTM cells so it is costly

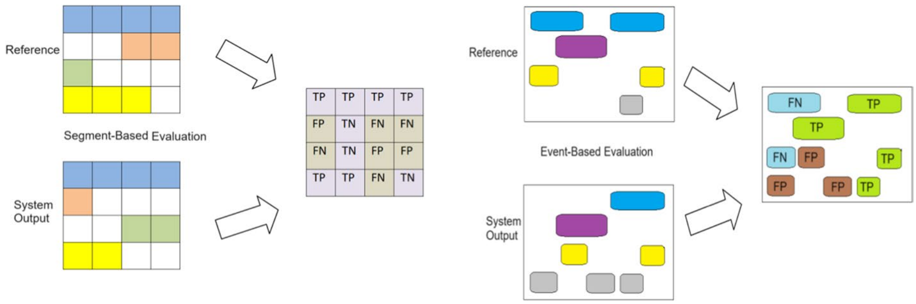

Fig. 7 Segment based and Event based metric evaluation process

During segment-based evaluation process, the sound event start and end point timing are adjusted to match a set evaluation grid. Sometimes, the duration of sound events may be extended to ensure that activity indicators cover the entire period when the sound is active, even if it's only for a short time. To compare reference annotations with system output, segment-level binary activity indicators are used to measure correctly and incorrectly detected events, including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These metrics can be used to calculate different performance measures.

Common metrics for the classification preferred are the precision (P), recall (R), F-score and error rate (ER). These metrics are evaluated based on the counts of correct and erroneous detections as follows

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F = \frac{2PR}{P + R}$$

Assessing and quantifying individual errors can be done using Precision (P), Recall (R), and the F-score. However, the Error Rate (ER) approach looks at things differently. It treats the occurrence of both a false positive and a false negative as a single substitution error, which is denoted as 'S.' This approach originated from evaluating speech recognition systems and is based on the concept of edit distance. This distance measures the difference between two strings (such as words) by calculating the minimum operations required to transform one string into another. Any false positives that are not accounted for within 'S' are considered as insertions ('I'), while unaccounted false negatives are labeled as deletions ('D'). The ER is calculated by adding up the total error count relative to the reference events ('N') and expressing it as a ratio.

$$ER = \frac{S + D + I}{N}$$

Working with these metrics presents a significant challenge as it involves balancing accurate detection with the possibility of missing detections. To overcome this challenge, one needs to select an ideal operating point where the trade-off between accuracy and potential missed detections is considered acceptable. On the other hand, metrics based on the Receiver Operating Characteristic (ROC) curve provide a more comprehensive view of performance across different operating points. This is achieved by plotting the True Positive Rate (TPR) or sensitivity/recall (calculated as described in Eq. (3)) against the False Positive Rate (FPR), which is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

A single metric describing this curve is the area under the curve (AUC), which is a metric commonly used in statistics and machine learning to evaluate the performance of binary classification models, such as logistic regression, support vector machines, and random forests. It quantifies the ability of a model to distinguish between the positive and negative classes.

Sensitivity and specificity represent the true positive rate and true negative rate, respectively. They are used as a pair, to illustrate the trade-off between the two measured components.

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

Accuracy, on the other hand, measures how often the classifier accurately predicts outcomes, calculated as the ratio of correctly predicted system outputs to the total number of outputs.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

5 Software resources

Several software libraries and frameworks were used for sound event detection and classification. These software tools provided developers and researchers with the necessary tools and resources to work on audio analysis tasks. Here are some of the existing software tools commonly used in the field:

Librosa: Librosa is a Python library for music and audio analysis. It provides functions for extracting various audio features, such as MFCCs and chroma features, making it useful for sound event classification tasks. It can be combined with machine learning libraries like scikit-learn for classification.

TensorFlow and Keras: TensorFlow, an open-source machine learning framework, and Keras, a high-level neural networks API, are commonly used for building and training

deep learning models for sound event detection and classification. TensorFlow's audio processing library, tf.signal, can be used for feature extraction.
PyTorch: PyTorch, another popular deep learning framework, is used for building custom neural network architectures for audio analysis. It provides flexibility and is favored by many researchers in the field.
Scikit-learn: Scikit-learn is a versatile machine-learning library for Python. While it's not specialized for audio, it can be used for sound event classification combined with feature extraction libraries like Librosa.
OpenSMILE: OpenSMILE is an audio feature extraction toolkit that provides a wide range of low-level and high-level audio features. It extracts features from audio data before feeding them into machine learning models.
PyAudio and PortAudio: PyAudio is a Python wrapper for the PortAudio audio I/O library. These libraries are used for real-time audio capture and playback, making them useful for applications like sound event detection and live audio processing.
Audacity: While primarily an audio editing tool, Audacity can be used for basic audio analysis tasks. It allows users to visualize and manually label audio segments, which can be useful for creating labeled datasets for training machine learning models.
MIR Toolbox: The Music Information Retrieval (MIR) Toolbox is a MATLAB-based software package for audio analysis, including feature extraction and audio classification tasks.
VGGish: VGGish is a pre-trained deep learning model developed by Google that can be used for audio classification tasks. It's particularly suitable for tasks involving audio embeddings and transfer learning.
YAMNet: YAMNet is another pre-trained deep learning model developed by Google. It's specifically designed for sound event classification and can classify audio events.
Essentia: Essentia is an open-source library for audio analysis and audio-based music information retrieval. It provides a variety of audio feature extraction functions.
Praat: Praat is a software for analyzing speech and other audio signals. It's commonly used in linguistics and phonetics research but can also be applied to sound event analysis.

6 Future scope and applications

The future scope of sound detection is vast and continues to expand as technology advances and our understanding of sound deepens. Here are several areas that hold significant potential for research and development in the field of sound:

Sonic Arts and Music Technology: Pushing the boundaries of sound synthesis, sound design, and music production techniques. Exploring new possibilities for interactive and generative music using AI and machine learning [38, 39].
Health and Wellness Applications: SED can be implemented for investigating the therapeutic effects of sound, such as using sound for stress reduction, relaxation, and improving sleep quality. SED can be used for ultrasound imaging techniques for medical diagnosis and treatment, including high-resolution imaging and targeted drug delivery. And also implement research on non-invasive methods for imaging and monitoring brain activity [69, 94].
Speech and Language Processing: SED can be executed for speech recognition and natural language processing technologies for improved human-computer interaction

and accessibility. SED can also enable the research for cross-lingual and accent-robust speech processing [81, 82].

Bioacoustics and Environmental Monitoring: Using a sound classification system, we monitor and study wildlife behavior, habitat health, and biodiversity. Properly developing a sound event detection model can detect natural disasters like earthquakes and tsunamis [86, 87].

Automotive and Transportation: Developing advanced sound event detection system designs for electric and autonomous vehicles can enhance safety and user experience. And also researching soundscapes that improve the overall auditory environment within vehicles and transportation hubs [44].

Virtual and Augmented Reality Audio: Advancing spatial audio technologies to create more immersive and realistic virtual and augmented reality experiences and researching personalized audio experiences that adapt to the user's movements and environment. It can Explore innovative ways to use sound as an input and feedback modality in various computing contexts, such as gesture recognition and touch less interfaces [90, 91, 118].

Noise Pollution Mitigation: This research can enable innovative methods to reduce noise pollution in urban environments, including quieter transportation systems and noise barriers. This research helps to design buildings, malls, and spaces with optimal acoustics for various purposes, such as concert halls, classrooms, offices, and homes [92, 93].

Multimodal Approaches: Large language models can be integrated into multimodal systems, combining text and audio processing. For example, a model could analyze transcriptions of spoken words alongside the corresponding audio signals to improve the accuracy of sound detection systems. The example applications are Duolingo or Rosetta Stone with integrated speech recognition, Otter.ai's live transcription service.

Contextual Understanding: Large Language excel at comprehending context in written text. Integrating this contextual understanding into sound detection algorithms could enhance their performance, especially in scenarios where the context influences the meaning of sounds. Example applications are YouTube Live automatic captions, Rev.ai for conference call transcription

Human–Machine Interaction: Large Language models are used in human-machine interaction systems. Integrating sound detection capabilities with natural language understanding can lead to more sophisticated systems that respond not only to textual input but also to spoken commands or queries. Example applications are Otter.ai's live transcription service.

7 Summary of review

Our research employed a systematic literature review methodology to conduct an extensive survey on sound event detection. With this framework, we formulated five specific review questions. We meticulously provided comprehensive responses for each of these questions within our review report. Our investigation encompassed a thorough examination of 122 articles in order to address these five key review questions. The selection of these papers was conducted with a keen focus on ensuring their quality, thereby enhancing the reliability of the gathered information.

Following providing answers to the research questions, we compiled this section to briefly overview our findings. Within this summary, we highlighted the dataset used in each model, the underlying algorithms, and the methods employed for feature extraction as the primary focal points of consideration. Table 8 summarizes various Sound Event Detection models, highlighting their respective datasets for reference and comparison.

8 Synthesis

In our Systematic Literature Review on sound event detection, we initially collected 576 papers from various publishers searching with keywords related to our research area from different databases. We short-listed the 210 papers based on the inclusion and exclusion criteria. We have allotted two reviewers to assess the quality of short-listed papers to filter the best papers from 210. Finally, we have chosen 122 quality papers to write the review on SED. According to our observation of sound event detection from 2009 to 2023, we have concluded the review points as follows:

From our study we observed that there are several datasets available to research different environments. The major domain areas are residential area, office, traffic, park, city, real life, urban sounds, heart sounds, and drone sounds, etc. From these domains we currently have SED datasets, such as the CHiME-Home dataset, TUT-SED, ESC-10 &ESC-50, DARES-G1 database, MAVD-traffic, freefield1010, etc. The researchers used the above datasets based on their selected environment for their research. Apart from that, some researchers prepared the synthetic dataset for their research to enrich the investigation. In our future research, we will prepare a synthetic dataset to detect the sound in the forest environment during tree-cutting events.

The performance of the model and implementation depended on various feature extraction techniques. In the sound event detection system, feature extractions are classified into 5 types. They are the time domain, frequency domain, wavelet domain, image domain, and Cepstral domain. The authors integrated the multiple feature extractions to gain accuracy in the model. The frequency and cepstral domains played a significant role in sound event detection. Some of the research models, such as [2, 8, 9, 11, 12, 17, 55], have used MFCC as a feature extraction method. This method provides frequencies with respect to the time of sounds as a feature that becomes an input to classification algorithms. From our observations, only the MFCC feature extraction method is insufficient for better classification. Because MFCC is to convert audio in the time domain into the frequency domain but has poor robustness to noise signals, as noise signals change all MFCCs if at least one frequency band is tilted. Another critical issue with MFCCs is that these are derived only from the power spectrum of a speech signal, ignoring the spectrum phase. The sound event detection research models must apply proper feature extractions according to the sound environment and a proper classification algorithm. Proper feature extractions lead to producing better results from classifiers. Some of the researchers [9, 25, 83, 92, 104] implemented the MFCC and Mel spectrogram to extract sound feature from dataset. Mel spectrogram is one feature extraction used to visualize the sound feature. This model generates frequency and image domain features to classify sound events. The major limitation of Mel Spectrogram is that t difficult to separate simultaneous sounds in spectrogram representations.

Similarly, some models have implemented integrated feature extraction methods by combining multiple feature extractions such as ZCR, MFCC, spectro-temporal features, log mel-band energies, Mel spectrograms. The researchers showed that extraction of multi-features from data provides better results [9, 28, 31, 34, 50, 91, 96] though these

Table 8 Summary of various SED model

Dataset	Reference	Time Domain	Frequency Domain	Wavelet Domain	Image Domain	Cepstral Domain	Model	Results
TUT Acoustic Scenes 2016	[2]		✓			✓	GMM	Accuracy:72.5%
	[20]				✓		CNN	Accuracy:78%
	[37]	✓					CRNN	F1-Score: 48.32%
	[82]		✓			✓	CNN	Accuracy: 85.9%
TUT Rare Sound Events 2017	[3]	✓			✓		CRNN	Accuracy:93.1%
	[80]				✓		CRNN	F1-Score: 89.9%
TUT Acoustic Scenes 2017	[12]		✓			✓	Random Forest	Error rate:0.83
	[78]		✓				GMM-HMM	F1-Score: 72.7%
UPC-TALP	[13]		✓			✓	Random Forest	Accuracy:64.6
	[14]	✓	✓	✓		✓	Random Forest	Accuracy:91.56%
DCASE 2018 task 4	[18]	✓	✓				NMF-CNN	F1-Score: 75.2%
	[101]		✓			✓	CNN	F1-Score: 67.3%
DCASE 2019 development	[28]	✓	✓	✓			CRNN	F1-Score: 92.2%
	[31]	✓		✓			CNN	F1-Score: 93%
	[33]	✓		✓			CNN	F1-Score: 96.9%
TUT Sound Events 2018	[32]			✓	✓		CRNN	F1-Score: 93%
DCASE2020 SELD dataset	[34]		✓	✓	✓	✓	CNN	F1-Score: 71.2%
MedleyDB dataset	[38]		✓			✓	CRNN	Accuracy:82%
	[41]	✓	✓				CRNN	Accuracy:82%
	[49]	✓	✓				CRNN	Accuracy:86.1%
freefield1010	[48]	✓	✓			✓	NMF	Accuracy:80.1%
	[58]		✓				CNN	
TUT Rare Sound Events dataset	[47]		✓			✓	CRNN	F1-Score:52%
	[50]	✓	✓	✓			SVM	F1-Score:70%
RAVDESS	[51]		✓			✓	CNN	F1-Score:70%
	[81]	✓	✓	✓		✓	Decision Tree	Accuracy:85%

Table 8 (continued)

Dataset	Reference	Time Domain	Frequency Domain	Wavelet Domain	Image Domain	Cepstral Domain	Model	Results
real-life recordings	[87]				✓		NMF	Accuracy:57.8%
	[90]		✓			✓	Bi-CRNN	Accuracy:64.7%
ESC-10 &ESC-50	[95]		✓			✓	CNN	Accuracy:94.9%
	[97]		✓			✓	CNN	Accuracy:81%
DCASE 2017	[96]		✓			✓	CNN	F1-Score: 52.6%
	[100]	✓	✓				BiGRU-CNN	F1-Score: 68.4%
TUT-SED 2017 development dataset	[99]	✓	✓	✓			3D-CNN	F1-Score: 67.5%
Urban-Sound8K dataset	[57]		✓			✓	SVM	Accuracy:78%
	[97]		✓			✓	CNN	Accuracy:81%
	[105]	✓	✓				CNN	Accuracy:97.03%
CHiME-Home dataset and *TUT-SED 2016*	[92]	✓	✓			✓	CRNN	F1 Score:68.3%
	[93]		✓			✓	BiCRNN	F1-Score: 74.0%

integrated feature extraction method providing good results but not suitable to heavy noise environment. According to our study, researchers are selected feature extraction methods randomly instead of that selecting feature extraction methods with respect to environment. These multiple feature extractions methods are inappropriate for every environment. Researchers selected the features randomly without concerning the environment which leads to create bios. Our observation on feature extraction techniques is that selection of feature extraction techniques must be based on environment and noise level.

To classify the input data, researchers preferred to use Machine Learning and Neural Network models. According to the analysis of several models, the Neural Network algorithms performed well compared to Machine Learning algorithms. We studied Some models [2, 8, 12, 17, 81] that used SVM, GMM-HMM, and NMF to classify the input sounds. But their results are not reached the expected range like neural network models. Because all machine learning models miss the feature connectivity with respect to time. In our investigation, the researchers mostly implemented CNN, RNN, and CRNN to train the extracted features [119–122]. CNN works based on analogous architecture, such as image context. That image generated based on frequency and time in spectrogram feature. This spectrogram feature becomes input to the CNN. The objects are separately analyzed and classified with CNN. Other side, the RNN models store the feature vectors with respect to the time and will connect to the previous features.

In the present days of research, the Sound Event Detection system has massive demand in many areas to find something interrelated to sound or speech. Presently, sound event detection is deployed in Physics and Astronomy, Computer Science, Medicine, Neuroscience, Earth and Planetary, Psychology, Biochemistry, Environmental Science, etc. Moreover, research implemented Sound Event Detection on information retrieval systems like Google, Mobile applications like Sound Meter & Noise Detector. In addition, several echo systems, such as Amazon Alexa Gaurd, are deploying sound event detection to find the various sounds.

9 Conclusion

According to our review standards, We referred to 122 documents sourced from various databases by our review criteria. We comprehensively examined numerous algorithms and diverse procedures for extracting features in the context of sound event detection. Furthermore, this paper provides a detailed exposition of various parameters and metrics for assessing sound events. This research domain presents substantial challenges when it comes to deploying models. Despite numerous researchers' dedicated efforts to construct a resilient Sound Event Detection system, achieving the expected accuracy still needs to be achieved. Most research endeavors rely on well-known datasets such as DCASE, TUT, CHiME-Home, Urban-Sound8K, among others. However, it's worth noting that these datasets may only be universally applicable to some environmental contexts. The researchers who achieved commendable results from these datasets did so due to thorough preparation and minimal noise interference. As a result, researchers are encouraged to develop robust synthetic datasets tailored to their specific environments. In the current study, we advocate using neural network models for sound classification, leveraging multi-feature extraction techniques. Unfortunately, many researchers have randomly selected multiple feature extraction methods without considering their suitability for specific datasets and environmental conditions. This approach consistently leads to biased learning outcomes.

Our proposed research focuses on detecting Sound Event Detection in forest environments, explicitly targeting tree-cutting events. Initially, our approach involves the creation of a comprehensive tree-cutting dataset encompassing various classes related to tree-cutting activities. Subsequently, we will employ appropriate feature extraction techniques to extract relevant features from the sounds associated with tree-cutting within the forest environment. Finally, our study will leverage a suitable Machine Learning algorithm to effectively classify these sounds from a noisy background.

This study contributes by providing insights into a range of aspects, including developing datasets across different environments, utilizing various feature extraction methods, and applying classification techniques to identify sounds. Nevertheless, it's essential to acknowledge that some specific functional details related to this subject matter may need to be covered comprehensively. However, our research presents valuable content to conduct a Systematic Literature Review.

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. Kitchenham B, Brereton OP, Budgen D, Turner M, Bailey J, Linkman S (2009) Systematic literature reviews in software engineering–a systematic literature review. Inf Softw Technol 51(1): 7–15
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, Virtanen T (2017) Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network. arXiv preprint arXiv:1710.02998
6. Archontis P, Mesaros A, Adavanne S, Heittola T, Virtanen T (2020) Overview and evaluation of sound event localization and detection in DCASE2019. IEEE/ACM transactions on audio, speech, and language processing, 29 pp 684–698
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. Mak M-W, Kung S-Y (2012) Low-power SVM classifiers for sound event classification on mobile devices. In: 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1985–1988
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. Yu C-Y, Liu H, Qi Z-M (2017) Sound event detection using deep random forest. Detection and Classification of Acoustic Scenes and Events
13. Phan H, Maaß M, Mazur R, Mertins A (2014) Random regression forests for acoustic event detection and classification. IEEE/ACM Trans Audio Speech Lang Process 23(1):20–31
14. Xia X, Togneri R, Sohel F, Huang D (2017) Random forest classification based acoustic event detection. IEEE International Conference on Multimedia and Expo (ICME), pp 163–168
15. 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(2018):1–13
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. Park J, Shin J, Lee K (2018) Separation of instrument sounds using non-negative matrix factorization with spectral envelope constraints. arXiv preprint arXiv:1801.04081
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. Bisot V, Essid S, Richard G (2017) Overlapping sound event detection with supervised nonnegative matrix factorization. IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 31–35
20. 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
21. Wei W, Zhu H, Benetos E, Wang Y (2020) A-crnn: A domain adaptation model for sound event detection. In: ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 276–280
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. Turpault N, Serizel R, Wisdom S, Erdogan H, Hershey JR, Fonseca E, Seetharaman P, Salamon J (2021) Sound event detection and separation: a benchmark on desed synthetic soundscapes. In: ICASSP 2021–2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 840–844
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. Kong Q, Cao Y, Iqbal T, Xu Y, Wang W, Plumbley MD (2019) Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems. arXiv preprint arXiv:1904.03476
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 Top 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. Basaran D, Essid S, Peeters G (2018) Main melody extraction with source-filter NMF and CRNN. In: 19th International Society for Music Information Retreival. 2018
39. Boulanger-Lewandowski N, Mysore GJ, Hoffman M (2014) Exploiting long-term temporal dependencies in NMF using recurrent neural networks with application to source separation. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 6969–6973. IEEE
40. Liu S, Guo L, Wiggins GA (2018) A parallel fusion approach to piano music transcription based on convolutional neural network. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 391–395. IEEE
41. Hsieh T-H, Su L, Yang Y-H (2019) A streamlined encoder/decoder architecture for melody extraction. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 156–160. IEEE
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 S-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. Eco 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. Roma G, Nogueira W, Herrera P (2013) Recurrence quantification analysis features for environmental sound recognition. In: 2013 IEEE workshop on applications of signal processing to audio and acoustics, pp 1–4. IEEE
48. Sobieraj I, Kong Q, Plumbley MD (2017) Masked non-negative matrix factorization for bird detection using weakly labeled data. In: 2017 25th European signal processing conference (EUSIPCO), pp 1769–1773. IEEE
49. Yu S, Yi Yu, Chen Xi, Li W (2021) HANME: hierarchical attention network for singing melody extraction. IEEE Signal Process Lett 28:1006–1010
50. 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
51. Gumelar AB, Kurniawan A, Sooai AG, Purnomo MH, Yuniarno ME, Sugiarto I, Widodo A, Kristanto AA, Fahrudin TM (2019) Human voice emotion identification using prosodic and spectral feature extraction based on deep neural networks. In: 2019 IEEE 7th international conference on serious games and applications for health (SeGAH), pp 1–8. IEEE
52. Jain U, Nathani K, Ruban N, Raj ANJ, Zhuang Z, Mahesh VGV (2018) Cubic SVM classifier based feature extraction and emotion detection from speech signals. In: 2018 international conference on sensor networks and signal processing (SNSP), pp 386–391. IEEE
53. Lee S, Pang H-S (2020) Feature extraction based on the non-negative matrix factorization of convolutional neural networks for monitoring domestic activity with acoustic signals. IEEE Access 8:122384–122395
54. Piczak KJ (2015) ESC: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on multimedia, pp 1015–1018

55. Zinemanas P, Cancela P, Rocamora M (2019) MAVD: a dataset for sound event detection in urban environments. Detection and classification of acoustic scenes and events, DCASE 2019, New York, NY, USA, 25–26 Oct, page 263–267
56. Mesaros A, Heittola T, Virtanen T (2016) August. TUT database for acoustic scene classification and sound event detection. In: 2016 24th European Signal Processing Conference (EUSIPCO), pp 1128–1132). IEEE
57. Salamon J, Jacoby C, Bello JP (2014) A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM international conference on Multimedia, pp 1041–1044
58. Stowell D, Plumbley MD (2013) An open dataset for research on audio field recording archives: freefield1010. arXiv preprint arXiv:1309.5275
59. Vozáriková E, Juhár J, Čižmár A (2011) Acoustic events detection using MFCC and MPEG-7 descriptors. In: International conference on multimedia communications, services and security, pp 191–197. Springer, Berlin, Heidelberg
60. Johnson DS, Lorenz W, Taenzer M, Mimilakis S, Grollmisch S, Abeßer J, Lukashevich H (2021) Desed-fl and urban-fl: Federated learning datasets for sound event detection. In: 2021 29th European signal processing conference (EUSIPCO), pp 556–560. IEEE
61. Purohit H, Tanabe R, Ichige K, Endo T, Nikaido Y, Suefusa K, Kawaguchi Y (2019) MIMII dataset: sound dataset for malfunctioning industrial machine investigation and inspection. arXiv preprint arXiv:1909.09347
62. Hertel L, Phan H, Mertins A (2016) Comparing time and frequency domain for audio event recognition using deep learning. In: 2016 International Joint Conference on Neural Networks (Ijcnn), pp 3407–3411. IEEE
63. Gemmeke JF, Ellis DP, 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
64. Ooi K, Watcharasupat KN, Peksi S, Karnapi FA, Ong ZT, Chua D, Leow HW, Kwok LL, Ng XL, Loh ZA, Gan WS (2021) A strongly-labelled polyphonic dataset of urban sounds with spatiotemporal context. arXiv preprint arXiv:2111.02006
65. Cartwright M, Cramer J, Mendez AEM, Wang Y, Wu HH, Lostanlen V, Fuentes M, Dove G, Mydlarz C, Salamon J, Nov O (2020) SONYC-UST-V2: An urban sound tagging dataset with spatiotemporal context. arXiv preprint arXiv:2009.05188
66. Fonseca E, Favory X, Pons J, Font F, Serra X (2020) FSD50k: an open dataset of human-labeled sound events. arXiv preprint arXiv:2010.00475
67. Abeßer J (2021) USM-SED-A dataset for polyphonic sound event detection in urban sound monitoring scenarios. arXiv preprint arXiv:2105.02592
68. McFee B, Bertin-Mahieux T, Ellis DP, Lanckriet GR (2012) The million song dataset challenge. In: Proceedings of the 21st International Conference on World Wide Web, pp 909–916
69. Gao S, Zheng Y, Guo X (2020) Gated recurrent unit-based heart sound analysis for heart failure screening. Biomed Eng Online 19(1):1–17
70. Fonseca E, Pons Puig J, Favory X, Font Corbera F, Bogdanov D, Ferraro A, Oramas S, Porter A, Serra X (2017) Freesound datasets: a platform for the creation of open audio datasets. In: Hu X, Cunningham SJ, Turnbull D, Duan Z (eds) Proceedings of the 18th ISMIR Conference; 2017 oct 23–27; Suzhou, China.[Canada]: International Society for Music Information Retrieval, pp 486–93. International Society for Music Information Retrieval (ISMIR)
71. Koizumi Y, Saito S, Uematsu H, Harada N, Imoto K (2019) ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection. In: 2019 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA), pp 313–317. IEEE
72. Cartwright M, Mendez AEM, Cramer J, Lostanlen V, Dove G, Wu HH, Salamon J, Nov O, Bello J (2019) SONYC Urban Sound Tagging (SONYC-UST): A multilabel dataset from an urban acoustic sensor network
73. Li Y, Liu M, Drossos K, Virtanen T (2020) Sound event detection via dilated convolutional recurrent neural networks. In: ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 286–290. IEEE
74. Mesaros A, Heittola T, Virtanen T (2018) A multi-device dataset for urban acoustic scene classification. arXiv preprint arXiv:1807.09840
75. Wan M, Wang R, Wang B, Bai J, Chen C, Fu Z, Chen J, Zhang X, Rahardja S (2019) Ciaic-ASC system for DCASE 2019 challenge task1. Tech. Rep., DCASE2019 Challenge
76. Heittola T, Mesaros A, Virtanen T (2020) Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions. arXiv preprint arXiv:2005.14623

77. Rakotomamonjy A, Gasso G (2014) Histogram of gradients of time–frequency representations for audio scene classification. IEEE/ACM Trans Audio Speech Lang Process 23(1):142–153
78. Mesaros A, Heittola T, Diment A, Elizalde B, Shah A, Vincent E, Raj B, Virtanen T (2017) DCASE 2017 challenge setup: Tasks, datasets and baseline system. In: DCASE 2017-workshop on detection and classification of acoustic scenes and events
79. Koduru A, Valiveti HB, Budati AK (2020) Feature extraction algorithms to improve the speech emotion recognition rate. Int J Speech Technol 23(1):45–55
80. Zhang Keming, Cai Yuanwen, Ren Yuan, Ye Ruida, He Liang (2020) MTF-CRNN: multiscale time-frequency convolutional recurrent neural network for sound event detection. IEEE Access 8:147337–147348
81. Özseven T, Düğenci M (2018) SPeech ACoustic (SPAC): A novel tool for speech feature extraction and classification. Appl Acoust 136:1–8
82. Dang A, Vu TH, Wang JC (2018) Acoustic scene classification using convolutional neural networks and multi-scale multi-feature extraction. In: 2018 IEEE international conference on consumer electronics (ICCE), pp. 1–4. IEEE
83. Glowacz Adam (2018) Acoustic-based fault diagnosis of commutator motor. Electronics 7(11):299
84. Deng M, Meng T, Cao J, Wang S, Zhang J, Fan H (2020) Heart sound classification based on improved MFCC features and convolutional recurrent neural networks. Neural Netw 130:22–32
85. Heittola T, Mesaros A, Eronen A, Virtanen T (2013) Context-dependent sound event detection. EURASIP J Audio Speech Music Process 2013(1):1–13
86. Mesaros A, Heittola T, Dikmen O, Virtanen T (2015) Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 151–155. IEEE
87. Ohishi Y, Mochihashi D, Matsui T, Nakano M, Kameoka H, Izumitani T, Kashino K (2013) Bayesian semi-supervised audio event transcription based on Markov Indian buffet process. In: 2013 IEEE international conference on acoustics, speech and signal processing, pp 3163–3167. IEEE
88. Cakir E, Heittola T, Huttunen H, Virtanen T (2015) Multi-label vs. combined single-label sound event detection with deep neural networks. In: 2015 23rd European signal processing conference (EUSIPCO), pp. 2551–2555. IEEE
89. Parascandolo G, Huttunen H, Virtanen T (2016) Recurrent neural networks for polyphonic sound event detection in real life recordings. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 6440–6444. IEEE
90. Adavanne S, Parascandolo G, Pertilä P, Heittola T, Virtanen T (2017) Sound event detection in multichannel audio using spatial and harmonic features. arXiv preprint arXiv:1706.02293
91. Cakır E, Parascandolo G, Heittola T, Huttunen H, Virtanen T (2017) Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Trans Audio Speech Lang Process 25(6):1291–1303
92. Jung S, Park J, Lee S (2019) Polyphonic sound event detection using convolutional bidirectional lstm and synthetic data-based transfer learning. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 885–889. IEEE
93. 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
94. Mondal Ashok, Banerjee Poulami, Tang Hong (2018) A novel feature extraction technique for pulmonary sound analysis based on EMD. Comput Methods Programs Biomed 159:199–209
95. Mushtaq Zohaib, Shun-Feng Su (2020) Environmental sound classification using a regularized deep convolutional neural network with data augmentation. Appl Acoust 167:107389
96. Lin L, Wang X, Liu H, Qian Y (2019) Guided learning convolution system for dcase 2019 task 4. arXiv preprint arXiv:1909.06178
97. Altinors Ayhan, Yol Ferhat, Yaman Orhan (2021) A sound based method for fault detection with statistical feature extraction in UAV motors. Appl Acoust 183:108325
98. Adavanne S, Politis A, Virtanen T (2018) Multichannel sound event detection using 3D convolutional neural networks for learning inter-channel features. In: 2018 international joint conference on neural networks (IJCNN), pp 1–7. IEEE
99. Kong Q, Xu Y, 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
100. Lin L, Wang X, Liu H, Qian Y (2020) Guided learning for weakly-labeled semi-supervised sound event detection. In: ICASSP 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP, pp 626–630. IEEE

101. Alías F, Socoró JC, Sevillano X (2016) A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. Appl Sci 6(5):143
102. Piczak KJ (2015) Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th international workshop on machine learning for signal processing (MLSP), pp 1–6. IEEE
103. Cakir E, Heittola T, Huttunen H, Virtanen T (2015) Polyphonic sound event detection using multi label deep neural networks. In: 2015 international joint conference on neural networks (IJCNN), pp 1–7. IEEE
104. Madhu A, Kumaraswamy S (2019) Data augmentation using generative adversarial network for environmental sound classification. In: 2019 27th European signal processing conference (EUSIPCO), pp 1–5. IEEE
105. Kao CC, Wang W, Sun M, Wang C (2018) R-CRNN: Region-based convolutional recurrent neural network for audio event detection. arXiv preprint arXiv:1808.06627
106. Cakir E, Adavanne S, Parascandolo G, Drossos K, Virtanen T (2017) Convolutional recurrent neural networks for bird audio detection. In: 2017 25th European signal processing conference (EUSIPCO), pp 1744–1748. IEEE
107. Sharma G (2018) Acoustic signal classification for deforestation monitoring: tree cutting problem. J Comput Sci Syst Biol 11:178–184
108. Incze A, Jancsó H-B, Szilágyi Z, Farkas A, Sulyok C (2018) Bird sound recognition using a convolutional neural network. In: 2018 IEEE 16th international symposium on intelligent systems and informatics (SISY), pp 000295–000300. IEEE
109. Chatterjee CC, Mulimani M, Koolagudi SG (2020) Polyphonic sound event detection using transposed convolutional recurrent neural network. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 661–665. IEEE
110. Riaz M, Mendes E, Tempero E (2009) A systematic review of software maintainability prediction and metrics. 2009 3rd international symposium on empirical software engineering and measurement, pp 367–377. https://doi.org/10.1109/ESEM.2009.5314233
111. Bansal A, Garg NK (2022) Environmental sound classification: a descriptive review of the literature. Intell Syst Appl 200115
112. Chan TK, Chin CS (2020) A comprehensive review of polyphonic sound event detection. IEEE Access 8:103339–103373
113. Mesaros Annamaria, Heittola Toni, Virtanen Tuomas, Plumbley Mark D (2021) Sound event detection: a tutorial. IEEE Signal Process Mag 38(5):67–83
114. Nogueira AFR, Oliveira HS, Machado JJM, Tavares JMRS (2022) Sound classification and processing of urban environments: a systematic literature review. Sensors 22(22):8608
115. 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
116. Abayomi-Alli Olusola O, Damaševičius Robertas, Qazi Atika, Adedoyin-Olowe Mariam, Misra Sanjay (2022) Data augmentation and deep learning methods in sound classification: a systematic review. Electronics 11(22):3795
117. Mesaros Annamaria, Heittola Toni, Virtanen Tuomas (2016) Metrics for polyphonic sound event detection. Appl Sci 6(6):162
118. Xiao Y, Khandelwal T, Das RK (2023) FMSG submission for DCASE 2023 challenge task 4 on sound event detection with weak labels and synthetic soundscapes. Proc. DCASE Challenge
119. Martín-Morató I, Harju M, Ahokas P, Mesaros A (2023) Training sound event detection with soft labels from crowdsourced annotations. In: ICASSP 2023–2023 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1–5. IEEE
120. Cai X, Gan Y, Wu M, Wu J (2023) Weak supervised sound event detection based on Puzzle CAM. IEEE Access
121. Xu L, Wang L, Bi S, Liu H, Wang J (2023) Semi-Supervised sound event detection with pre-trained model. In: ICASSP 2023–2023 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1–5. IEEE
122. Wang Qing, Jun Du, Hua-Xin Wu, Pan Jia, Ma Feng, Lee Chin-Hui (2023) A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection. IEEE/ACM Trans Audio Speech Lang Process 31:1251–1264

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.