Proceedings of the International Conference on Trends in Material Science and Inventive Materials (ICTMIM-2025)
IEEE Xplore Part Number: CFP250J1-ART; ISBN: 979-8-3315-0148-8

2025 5th International Conference on Trends in Material Science and Inventive Materials (ICTMIM) | 979-8-3315-0148-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICTMIM65579.2025.10988199

A Systematic Literature Review on Sound Event Detection and Classification

S.Padmaja
Research Scholar
School of Computer Science and Artificial Intelligence ,
SR University,Warangal,
Telangana,India.
mahepadmaja@gmail.com

Dr. N. Sharmila Banu
Assistant Professor,
Department of Computer Science
School Of Computer Science and Artificial Intelligence,
SR University,Warangal,
Telangana,India
Sharmila.banu@sru.edu.in

Abstract—**Sound Event Detection (SED) has appeared as a fundamental study area due to its broad applicability, including environmental monitoring, healthcare systems, in smart cities, and in industrial automation. Accurate identification and categorization of sound events are essential for developing intelligent systems capable of understanding and responding to acoustic environments. This article represents a systematic literature review (SLR) to explore the advancements and challenges in SED, focusing on feature extraction techniques and classification models. Key challenges, like background noise and overlapping audio signals, are addressed by reviewing feature extraction methods, including Mel-frequency cepstral coefficients (MFCCs), spectrogram analysis, and wavelet transforms. These strategies are foundational for capturing discriminative sound patterns required for accurate classification, and this review exposed the role of advanced classifiers, including deep learning mechanisms like CNNs and RNNs and hybrid approaches that combine machine learning techniques for improved performance.**

Keywords—Sound Event Detection, MFCCs, Neural Networks, Machine Learning.

I. INTRODUCTION

Sound event detection and classification is crucial in understanding auditory environments and furnishing context and meaning to actions in our daily lives. Sounds are required for various applications in modern technology[1][2][3]. Accurately detecting and classifying sounds has become key for present research, improving safety measures and user experience in intelligent systems[4][5]. Different environments yield unique challenges for sound event detection like urban areas with high noise levels, natural environments feature various sound patterns. Applications of sound classification are vast and enclose supervision systems, healthcare monitoring, wildlife conservation, and human-computer interaction. In healthcare, sound classification can aid in noticing respiratory issues through cough or breathing sounds [6][7][8]. In wildlife conservation, it can help monitor animal populations by analyzing their vocalizations [9] [10].

The need for sound classification comes from the increasing demand for automated systems processing auditory data in real-time [11] [12] [13]. As the number of audio-enabled devices increases, intelligent systems can notice and interpret sound events accurately[14][15]. The indicated has directed to the development of various ML and neural network (NN) standards that can process audio signals effectively and make predictions[16][17].

Publicly available datasets, like AudioSet, UrbanSound8K, and ESC-50, have become measures for training and estimating ML and NN benchmarks. These datasets provide a various range of labeled sound events, allowing researchers to develop generalized models that perform well across different environments[18][19][20].

Saved Feature extraction and preprocessing mechanisms are necessary stages for effective sound categorization. Preprocessing methods like noise reduction, normalization, and segmentation help to improve the removal of irrelevant or redundant information. Feature extraction is used to convert raw audio signals into meaningful representations that capture relevant characteristics of the sound[21][22]. Standard features include Mel-frequency cepstral coefficients (MFCCs), chroma features, spectral contrast, and log-mel spectrograms[23][24]. These features allow ML and NN models to learn patterns effectively from audio data. Advanced feature extraction methods, often integrated into deep learning frameworks, allow for the automated discovery of high-level audio features like pitch, loudness, wavelength, speed and also improve classification performance by using advanced models [25] [26] [27].

Recent progress in machine learning and neural networks has brought better advances in sound classification. Standard approaches relied on feature engineering and classical ML algorithms, and modern procedures followed deep learning architectures like convolutional neural networks (CNNs) and RNNs, which have shown better performance in handling complex audio patterns[28][29]. Hybrid models that combine CNNs with attention mechanisms or transformer-based architectures further improve the ability to model temporal and spatial features in audio signals[30][31].

A. Motivation

The environmental sound categorization examination graph is constantly growing, instructed by creative mechanisms, different datasets, and research to improve performance. The collaborative efforts of investigators worldwide fuel this journey, each contributing unique understandings and refinements to the field. The following articles serve as beacons of inspiration, lighting pathways towards enhanced knowledge and utilization of environmental sound data.

Karol J. Piczal.[3] This seminal work addresses the shortage of suitable datasets for environmental sound classification by introducing a precisely annotated collection of 2,000 short clips alongside a vast compilation of 250,000 unlabeled auditory excerpts. Piczak stresses the significance of dataset quality in extending research efforts by assessing human precision instead of baseline classifiers.

Honglie Chen et al.[7] by creating the VGG Sound dataset comprising over 200k video clips across 300 audio classes, Chen et al. explained the power of automated pipelines for dataset creation, eliminating manual annotation. Their facility of baseline measures set a new prototype for large-scale audio-visual datasets, allowing robust audio recognition research opportunities.

Bandara et al.[8] described a FSC22 dataset that manages a marked gap in forest sound classification analysis by providing an exhaustive collection of sound clips representing 27 different forest sounds. According to Bandara et al.'s analysis, selecting a suitable dataset and applying data augmentation strategies are essential for achieving high accuracy in sound classification.

Sena et al. [9] discussed a AUDIO-MC framework for audio classification, which has overwhelming context-specific limitations, by utilizing Mel Spectrograms and CNN architectures. Praciano et al.'s framework achieves impressive accuracy across diverse datasets, specifying a foundation for multi-context audio classification research.

Ruofei Ma et al.[12] introduced new mechanism for classifying environmental sounds by transforming sound spectrograms into RGB images. Their experiments show that this method performs well and provides a fresh understanding of practical ways to describe features for sound classification tasks using transfer learning and CNN architectures.

Md Tamzeed Islam [16] developed a unique mobile app to identify audio events. The app authorizes users to apply text-based semantic knowledge to built-in sound type. The app's flexibility and high precision in identifying different sounds show the powerful potential of mobile-based solutions for audio categorization.

B. Significant Contribution:

This analysis aims to provide a Systematic Literature Review (SLR) to analyze the methodologies, challenges in SED [32]. The review also focused on key challenges in processing complex audio signals, including noise, overlapping events, dynamic acoustic environments to accurate classification. The study provided a comprehensive overview of feature extraction techniques such as MFCC, spectrogram analysis, and wavelet transforms projected their effectiveness in capturing different sound patterns. This review provided the study of application of advanced classification models-Neural Networks models, and hybrid machine learning approaches to examine the robustness in sound classification. This review synthesized the strengths and limitations of existing techniques, identifying key research gaps and opportunities for future advancements.

II. Review Method

A. Review Questions

The systematic literature review on sound classification framed appropriate review questions to provide evidence for our studies.

RQ1: What Sound classification datasets are available for research purposes that explore the preprocessing techniques on datasets?

This answer will provide an overview of the datasets commonly used in sound classification and the preprocessing methods for audio data preparation.

RQ2: What approaches are available to extract features for classifying sounds?

This answer will describe different methods for extracting features and how well they work for identifying sounds.

Q3: How do ML and NN models improve Sound classification, and what factors affect their success and use in different areas?

This answer will outline where SED is used and why these models are essential for improving performance.

RQ4: What are the main challenges in classifying forest sounds?

This answer will focus on the challenges discussed in this research paper regarding sound classification.

B. Search Process

To identify relevant literature, we performed a complete search process involving the following steps:

1. Keyword Selection: Initially started by selecting keywords and phrases suitable to our research topic, including terms like "sound classification," "acoustic analysis," and "machine learning in sound classification."
2. Database Search: Preferred to PubMed, IEEE Xplore, Google Scholar, and Web of Science with our chosen keywords.
3. Filtering and Screening: After retrieving many search results, we used filters based on publication date, language, and relevance to narrow the list of articles. The rest of the articles by reviewing titles, abstracts, and keywords to provide alignment with our analysis goals.

C. Selection Criteria

Inclusion Criteria: Articles were included if they addressed topics relevant to sound or acoustic analysis category, machine learning, or deep learning techniques. To include the latest research in this field, preference was given to papers published within the last five years.

Exclusion Criteria: Papers unrelated to acoustic classification or machine learning applications were excluded from consideration. Papers lacking methodological clarity or presenting insufficient data were excluded to provide the reliability of the included literature. Researcher excluded documents in languages other than English due to language barriers.

D. Quality Assessment

We achieved a detailed quality assessment to provide the reliability and validity of the selected analysis reports. Each paper was estimated using multiple criteria, including the strength of the approaches, relevance to our research topic, clarity of presentation, and the credibility of the results. We used the following scoring procedure for this quality assessment:

1. Methodological Rigor (0-5): This standard estimates the soundness of the research procedure used in the paper. Papers with well-defined research designs, clear data collection procedures, appropriate statistical analyses, and robust validation methods received higher scores.
2. Relevance to Research Topic (0-5): We estimated every paper's relevance to our investigation issue based on its alignment with our analysis goals and scope, and papers concentrating on acoustic classification, machine learning mechanisms, and related strategies received higher scores.
3. Clarity of Presentation (0-5): This benchmark predicts the transparency and coherence of the paper's presentation, including the organization of content, clarity of writing, and importance of visual aids. Papers with clear and brief definitions, structured sections, and visually attractive figures received higher scores.
4. Credibility of Findings (0-5): The credibility of the paper's results was predicted based on the strength of the proof submitted, the validity of the conclusions drawn, and the acknowledgment of potential limitations. Papers nourished robust proof, backed by proper statistical studies, and examined limitations transparently received higher scores.

III. RESULTS

A. RQ1: What Sound classifications datasets are available for research purposes, and explore the preprocessing techniques that generally using for datasets?

Datasets means it is a collection of structured data points used for analysis or machine learning tasks. These data points represent a single observation or example, with features or attributes that describe its characteristics. There are various types of datasets, including text datasets, image datasets, audio datasets, video datasets, tabular datasets, time series datasets, and spatial datasets. Text datasets consist of textual data, such as documents, articles, emails, or social media posts, used for tasks like natural language processing, sentiment analysis, or text classification.

In generally the time series datasets consist of data points collected over time, used for tasks like forecasting and trend analysis.

Spatial datasets, consisting of maps, satellite images, or GPS coordinates, are utilized for tasks like GIS, land cover classification, and route optimization.

Environmental sound classification datasets are utilized by researchers in audio processing, machine learning, and environmental science for developing and evaluating algorithms for automatic sound recognition, aiding in wildlife monitoring, urban noise pollution analysis, and smart city applications.

Sound preprocessing techniques are used to improve the quality of audio data before it is used in machine learning models. These methods include resampling, normalization, noise reduction, feature extraction, segmentation, augmentation, filtering, feature scaling, temporal context, and data augmentation. These techniques ensure consistency across recordings, reduce background noise, extract relevant features, divide longer recordings into smaller frames, generate additional training data, filter frequencies, scale features, incorporate temporal context, and increase data diversity. These methods enhance the performance of machine learning models in tasks like speech recognition, sound event detection, and audio classification. The table 1 shows the different datasets and their details related to sound classification.

Justin Salamon et al.[4] provided the dataset on urban sounds. This dataset has total 27 hours of audio records along with that 18.5 hours records have annotated sounds. The dataset derived has 10 classes. The UrbanSound dataset consists of 10 urban sound classes chosen for their frequency in noise complaints, with field recordings obtained from the Freesound repository. After manual filtering, 1314 recordings were retained, totaling over 27 hours of audio. Each recording was meticulously annotated with start and end times for occurrences of the target sound class, resulting in 3075 labeled occurrences. The dataset, is also available online for research in sound event detection. Additionally, an UrbanSound8K subset was created for sound source identification research, containing 8732 labeled slices of short audio snippets.

Imran et al. [10] FSDKaggle2018 dataset contains 3,710 training data that are clearly labeled with respective class. On the otherhand 5,763 samples are without labels and not belong to any of 41 classes. They performed preprocessing on 12 techniques. They utilized a single-block DenseNet architecture and applied batch-wise loss masking for label noise, and experimented with an CNN model.

Simiyu et al. [26] introduced a new dataset FSC22, which included with 2025 sound clips under 27 acoustic classes of forest environment. They have shown the comparisons of already existing dataset and their custom dataset. They have taken existing datasets like ESC-50, U8K, and FSD50K.They implemented data augmentation, feature extraction and CNN approach for classification with accuracy of 92.59%.

Korkmaz et al. [28] has included three datasets ESC-50, ESC-10, and AudioSet.There approach differs from one-shot and data augmentation techniques. They presented a unique mobile application for audio event detection, enabling users to input desired sound types and train audio clips. The performance is evaluated on an empirical dataset and in real-world scenarios and achieved a classifier accuracy of 60-90% for 6-10 class problems with 2-5 non-training examples.

Table 1. Details of various datasets on sound classification.

Reference	Dataset	Classes	No. of Samples	Duration of Each sample
[1]	ESC-50	50	250000	5 seconds
[17]	SONYC-UST-V2	10	18510	30seconds
[7]	VggSound dataset	300	200,000	10s/clip, 300 clips/class
[18]	FSC22 dataset	27	2025	5 seconds of each audio clip
[19]	FSDKaggle2018 dataset	41	3,710 verified labels, 5,763 non-verified labels	300ms to 30 seconds for each audio clip
[20]	FSD50k	200	51000	10sec to 20sec

B. RQ2: What are the feature extractions seen in classification of sounds?

Feature extraction is the technique of converting raw data into a set of features that capture related information for a particular task or analysis. Audio categorization converts raw audio signals into numerical features, and this can be input to machine learning algorithms for classification, clustering, or regression. Standard methods for feature extraction have MFCC, spectrogram features, time-domain characteristics, pitch and harmonic features, temporal features, and wavelet transform features. The selection of feature extraction mechanism depends on the particular features of the audio data and the needs of the classification work. Feature selection or dimensionality reduction techniques are typically utilized to refine the feature set before feeding it into a classification algorithm. Overall, feature extraction is essential in modifying raw audio data into a format appropriate for machine learning-based audio classification tasks. Table 2 illustrate the using of various features in different research and respective observations.

Stowell et al.[5] found that feature type significantly influenced recognition performance, with MFCCs outperforming Mel spectra and learned characteristics. For the largest dataset, lifeclef2014, feature learning led to categorization performance up to 85.4% AUC, while raw Mel spectra and MFCCs performed at 82.2% and 69.3%, respectively. Switching to learned features and raw Mel spectral features also improved recognition performance. Mean-and-standard-deviation summarization constantly provided the outstanding achievement over maximum or modulation coefficients.

Mesaros et al. [6] found that feature type significantly influenced recognition performance, with MFCCs outperforming Mel spectra and learned features. For the largest dataset, lifeclef2014, feature learning directed to categorization achievement up to 85.4% AUC, while raw Mel spectra and MFCCs performed at 82.2% and 69.3%, respectively. Switching to learned features and raw Mel spectral features also improved recognition performance. Mean-and-standard-deviation summarization constantly provided the strongest performance over maximum or modulation coefficients.

Bandara et al.[8] used Mel spectrograms and MFCC as feature extraction methods in audio categorization. The librosa, Feature library provides these spectrograms. Every audio file is sampled into overlapping frames, and cepstral coefficients are calculated for each frame model. The resulting two-dimensional array of shapes is converted to the decibel scale. ML-based categorization mechanism typically uses one-dimensional features, so the dimensionality of the produced spectrograms is reduced before using the XGBoost model. For DL-based categorization, an image-like presentation of characteristics in accordance with RGB mode is essential. Three spectrograms are created for each audio sample, with windowing lengths of 93 ms, 46 ms, and 23 ms, using the sample rate parameter at 22,050 Hz and the FFT parameter at 2048, 1024, and 512, correspondingly.

Lucas et al.[9] conducted experiment on audio signal preprocessing to transforming digital waveforms into more informative representations like images. Spectrograms and mel spectrograms offer enhanced representations, with parameters like channels number, window length, hop length, and sample rate playing pivotal roles. These parameters dictate the resulting image, significantly impacting classification. However, converting spectrogram images back to raw audio can result in incomprehensible output, underscoring the importance of scrutinizing spectrogram images for optimal classifier performance. Therefore, careful analysis of spectrogram images is essential for accurate classification. $fm = 1125 \ln(1+ fa /700)$ where fm is the Mel frequency and fa is the actual frequency.

Theodoros et al. [11] used the Essential audio analysis library to extract MFCC from audio slices. MFCCs are generally utilized in environmental sound investigation and are used as a baseline for benchmarking new techniques. The features are extracted per-frame utilizing a window dimension of 23.2 ms and 50% frame overlap. The first 25 MFCC coefficients are kept, and the per-frame values are concluded with statistics, ensuring in a feature vector of measurement 225 per slice.

Ruofei et al.[12] discussed the importance of extracting audio signal features for ESC systems, focusing on MFCC, LMS, and scalogram as input characteristics. The authors use the librosa library to extract these features, resizing them to 299x299, the default input form of the Xception standard. They demonstrate the effectiveness of these features in recognizing sounds from washing machines and human laughter, highlighting the need for improved re-description methods for ESC systems.

Gourisaria et al.[29] extracted features from every clip, including zero-crossing rate and MFCC, which are useful in speech processing and harmonic content analysis. MFCCs were determined with the librosa package, and feature vectors were used to input KNN, and SVM.

Table 2. Observations on feature extraction techniques in different research.

Reference	Features Extracted	Model Used	Observations
[3]	Zero-crossing rate, MFCC	k-nearest neighbors, random forest, SVM	Addressed the lack of appropriate and publicly available datasets ESC for environmental sound classification.
[8]	Mel Spectrograms, MFCC	XGBoost, DL-based classification	Introduced the FSC22 dataset for forest environmental sound categorization, achieving a maximum classification precision of 92.59% using CNN-based approach. Compared existing datasets and implemented data augmentation and feature extraction techniques.
[9]	Mel Spectrograms	CNN	Proposed the AUDIO-MC framework for automatic audio classification, achieving over 80% accuracy across analyzed contexts. Converted audio files into Mel Spectrograms and used a CNN design for classification.
[12]	MFCC, LMS, Scalogram	CNN based on Xception model	Proposed a new CNN standard for environmental sound categorization using transfer learning. Trained the CNN architecture using the Xception model and achieved better performance on ESC dataset.
[30]	MFCC	SVM	Provided the UrbanSound dataset containing 27 hours with 10 classes. Created an UrbanSound8K subset for sound source identification research.
[31]	MFCC, Mel Spectrogram	CNN	Presented three tasks for sound event detection (SED), focusing on TIMIT speech database , training/testing on sound events, and overlapping events in real-life audio. Utilized deep learning for feature extraction and observed significant performance differences between top systems and baseline classifiers.

C. RQ3: How do Machine Learning and Neural Network-based models enhance Sound Event Detection (SED) technology, and what are the key factors influencing their effectiveness and adoption across diverse domains?

Machine learning has reconstituted sound classification by automating feature extraction, model training, and performance optimization. Its algorithms, including neural networks, SVMs, and random forests, can efficiently classify various sounds like speech, environmental sounds, and music genres. Advancements in computational power have enabled real-time processing of audio data, making it applicable in domains like virtual assistants and smart home devices. Machine learning models can adapt to changes in sound environments through retraining or fine-tuning, ensuring robust performance over time. Deep learning techniques excel at learning hierarchical representations of audio data, improving accuracy with additional labeled data.

Annamaria Mesaros et al.[6] focused on the use of log-mel spectrogram characteristics and CNN to extract spectrograms from raw audio signals. The systems were preprocessed by two participants, normalized by amplitude, and background noise removed. Data was augmented through pitch shifting and dynamic range compression. The standards utilized different window sizes, with a limit of 40 to 128 mel filterbanks. The system utilized a multi-level and multi-scale Convolution Neural Network for optimal parameter values.
Bandara et al.[8] introduced a model using XGBoost and achieved an average classification precision between 48.14% and 62.17% on the FSC22 dataset, with its highest accuracy reaching 62.71% when using the MFCC feature extraction method. The representation was most accurate in identifying "Silence" and "Bird Chirping" sounds, while it had the lowest accuracy with the "Axe" and "Generator" sounds.

Gantert et al.[21] using the ESC-50 dataset to showed the less changeability in model validation but a significant performance by the random forest ensemble (44.3%) correlated to Support Vector Machine (33.6%) and k-Nearest Neighbor (32.2%). The simplest model showed a drop in accuracy, suggesting more complex feature dependencies in the larger dataset. The SVM classifier got good results for animal sounds compared to random forest ensemble.

Vashishtha et al.[22]compared five different classification algorithms: decision tree, k-Nearest Neighbor , random forest, SVM. The Results showed a statistically significant difference in performance.Performance remains stable from 10 to 6 seconds, then starts decreasing gradually. The top performing classifier (SVM) has no symbolic variation between performance using 6s and 4s slices, supporting the alternative of 4s slices for Urban Sound8K. Additional insights can be obtained by examining the accuracy of the SVM for each class individually.

Wolf-Monheim et al.[23] utilized LMS, scalogram, and Mel-Frequency Cepstral Coefficients characteristics of sound signals and transfer learning mechanism to enhance the accurateness of the ESC standard. A deep CNN design was proposed, using RGB images from original LMS, scalogram, and MFCC characteristics. This model achieved high classification exactness for the ESC-50 dataset. Further research is needed to extract more features from environmental sound data and design deeper CNN-based models.

Bertocco et al.[24] focused on the use of K-means to classify the sounds of birds by using of unsupervised learning models .This research implemented with local optimization algorithm, to minimize the impact of data presentation order. To reduce the impact of random data variation, the experiments were carried out in two phases. In the first stage, data points were randomly sampled and shuffled, then processed with PCA whitening before starting the k-means clustering. In the second stage, k-means was trained further by using all data points. The researchers also experimented with a two-layer version of the feature-learning mechanism, which captured details over different time scales by applying spherical k-means to the dataset. This research also conducted a human categorization test to establish a baseline precision for the FSC22 dataset, followed by machine learning and deep learning classification to compare with similar studies.

Neural networks model :Neural networks are widely utilized in sound classification due to their capacity to learn intricate patterns and features from raw audio data. Table 3 shows the classification of sound using various ML and NN models and the respective outcome.

Chene et al.[7] combined two datasets with dissimilar sound vocabularies, using common categories at training and selecting the intersection of AudioSet and VGGSound to form a single testset called AStest. This results in about 15,000 clips in AStest. The study also looks at how well audio can be identified using VGG and ResNet networks, with or without NetVLAD aggregation, using the new VGGSound dataset. The VGGSound dataset has more than 200,000 videos and 300 categories for unconstrained conditions. The study compares different Convolution Neural Network architectures and aggregation mechanisms for audio recognition on VGGSound.

Jeong et al.[10] used several techniques, such as the single-block DenseNet architecture, Squeeze-and-Excitation Network. These methods still need more testing to confirm how well they work. The waveform-based model did not perform as well as the logmel-based model. The authors also introduced a single-block DenseNet model and combined models with different low-level components and sampling rates, achieving top-level results.

Zaman et al. [25] implemented the DenseNet by choosing as a robust network for image recognition, with the aim of extracting relevant features. The AUDIO-MC model achieved 93.21% and 98.10% accuracy on ESC-50 and URBAN datasets, correspondingly, for environment sound classification.

Simiyu et al.[26] optimized the CNN for single-stream audio recordings, achieving the best performance in automatic vocal categorization of marmoset calls to date. The study compares the performance of separate finding and classification methods from Sharma et al., 2017 with our own feedforward deep CNN. The study established that adding LSTM layers to the model did not enhanced the task's performance, despite the fully connected RNN and LSTM-based neural network.

Table 3. Machine learning and Neural Networks models implemented of sound classification.

Reference	Dataset	Model	Accuracy
[3]	ESC-50	k-NN, random forest ensemble and SVM .	Accuracy of k-NN is 66.7% and accuracy of SVM is 67.5%
[4]	Urban Sound Research	SVM	Support Vector Method model accuracy: 88%
[6]	SED in the DCASE 2017 Challenge	CNN	Accuracy:95%
[8]	FSC22	CNN	Accuracy: 97%
[9]	AUDIO -MC	CNN	Accuracy: 94.94%
[10]	FSDKaggle2018	CRNN	Accuracy:95%
[11]	ARCHEO	SVM and CNN	Worked on only data augmentation
[12]	ESC -10	*Xception; CNN*	75.3% on ESC-50.

D. RQ4: What are some key challenges in the sound classification?

Some key challenges in forest sound classification include:
1. Background Noise: Forest environments often have diverse and unpredictable background noise, such as wind rustling through trees, animal calls, and other environmental sounds, which can interfere with the classification of target sounds.
2. Species Diversity: Forests are habitats to a wide variety of species, each with its own distinct vocalizations. Identifying and classifying these different species' sounds accurately can be challenging, especially when dealing with overlapping or similar calls.
3. Acoustic Variability: Sound produced by animals can differ greatly based on factors like distance from the recording device, environmental conditions, and individual variability within species. This variability adds complexity to the classification task
4. Data Imbalance: In real-world forest environments, particular species may be more general or vocal, leading to imbalanced datasets where a few classes have largely more examples than others. This can affect the performance and generalization ability of classification models.
Saved
5. Seasonal and Temporal Variations: Forests' acoustic geographies can vary seasonally and throughout the day, influenced by breeding seasons, migration patterns, and diurnal/nocturnal activity. Accounting for these temporal variations is essential for robust classification.
6. Cross-Species Similarities: Some animal species may produce sounds acoustically similar to those of other species. Differentiating between these similar sounds and accurately categorizing them can be challenging, requiring fine-grained spectral and temporal characteristics analysis.
Handling these challenges requires a combination of advanced signal processing tools, feature engineering, machine learning algorithms tailored to manage sparse and imbalanced data, and domain expertise in ecology and acoustics. Additionally, continuous advancements in sensor technology and data collection methods can help improve the quality and quantity of data available for forest sound classification tasks. The table 4 illustrates the various challenges and mitigations in the present research scenario on sound classification.

Table 4. Challenges and mitigations in the present research scenario on sound classification.

Risk	Description	Mitigation
Model Inaccuracy	The model may misidentify sounds, leading to miscommunication.	Continuously evaluate and fine-tune the model to improve accuracy. - Regularly update the training data with new and diverse audio recordings to enhance the system's performance.
Data Bias	If there is lack of diverse training data, the model may exhibit bias in recognizing certain animal species.	Ensure a balanced and diverse dataset for training to minimize bias. - Collaborate with experts and researchers to curate high-quality and representative data.
Model Development Complexity	The complexities in deploying and developing the model could lead to project delays or failures.	Engage experienced machine learning engineers and data scientists to simplify the development process. - Maintain clear documentation for model architecture and deployment procedures.
Environmental Conditions	Changes in environmental conditions, such as weather or habitat alterations could affect the sounds captured, potentially impacting animal recognition.	Develop models and data collection methods that account for environmental variations. - Implement real-time environmental monitoring to enhance sound capture reliability.
Resulting Actions	To ensure model accuracy, maintain diverse and updated training data.	To minimize bias, curate balanced and high-quality training data.

IV. CONCLUSION

Our research conducted a Systematic Literature Review (SLR) on Sound Event Detection and Classification, providing a comprehensive analysis of existing methodologies, datasets, and challenges in the field. We examined various publicly available sound event datasets. Additionally, we explored key feature extraction techniques, such as MFCC, Mel-Spectrogram and etc. while discussing their constraints in

improving classification performance. This review provided the overall knowledge on using of machine learning and Neural Networks models and their role in classification of the sound events. Our review process also described the challenges related to the sound classification research. Our future research aims to address these challenges by exploring advanced feature methods, optimizing deep learning architectures, and integrating hybrid machine learning models to enhance classification accuracy and robustness in different sound environments.

REFERENCES

[1] Cunningham, Stuart, Harrison Ridley, Jonathan Weinel, and Richard Picking. "Supervised machine learning for audio emotion recognition: Enhancing film sound design using audio features, regression models and artificial neural networks." Personal and Ubiquitous Computing 25, no. 4 (2021): 637-650.

[2] Narasimhan, Revathy, Xiaoli Z. Fern, and Raviv Raich. "Simultaneous segmentation and classification of bird song using CNN." In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 146-150. IEEE, 2017.

[3] Piczak, Karol J. "ESC: Dataset for environmental sound classification." In Proceedings of the 23rd ACM international conference on Multimedia, pp. 1015-1018. 2015.

[4] Salamon, Justin, Christopher Jacoby, and Juan Pablo Bello. "A dataset and taxonomy for urban sound research." In Proceedings of the 22nd ACM international conference on Multimedia, pp. 1041-1044. 2014.

[5] Stowell, Dan, and Mark D. Plumbley. "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning." PeerJ 2 (2014): e488.

[6] Mesaros, Annamaria, Aleksandr Diment, Benjamin Elizalde, Toni Heittola, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. "Sound event detection in the DCASE 2017 challenge." IEEE/ACM Transactions on Audio, Speech, and Language Processing 27, no. 6 (2019): 992-1006.

[7] Chen, Honglie, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. "Vggsound: A large- scale audio-visual dataset." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 721-725. IEEE, 2020.

[8] Bandara, Meelan, Roshinie Jayasundara, Isuru Ariyarathne, Dulani Meedeniya, and Charith Perera. "Forest sound classification dataset: Fsc22." Sensors 23, no. 4 (2023): 2032.

[9] Sena, Lucas B., Francisco DBS Praciano, Iago C. Chaves, Felipe T. Brito, Eduardo Rodrigues Duarte Neto, José Maria Monteiro, and Javam C. Machado. "AUDIO-MC: A General Framework for Multi-context Audio Classification." In ICEIS (1), pp. 374-383. 2022.

[10] Jeong, Il-Young, and Hyungui Lim. "Audio tagging system using densely connected convolutional networks." In DCASE, pp. 197-201. 2018.

[11] Psallidas, Theodoros, Alexander Mitsou, George Pikramenos, Evaggelos Spyrou, and Theodore Giannakopoulos. "ARCHEO: A Dataset for Sound Event Detection in Areas of Touristic Interest." In 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA, pp. 1-6. IEEE, 2020.

[12] Lu, Jianrui, Ruofei Ma, Gongliang Liu, and Zhiliang Qin. "Deep convolutional neural network with transfer learning for environmental sound classification." In 2021 International Conference on Computer, Control and Robotics (ICCCR), pp. 242-245. IEEE, 2021.

[13] Ibraheam, Mai, Fayez Gebali, Kin Fun Li, and Leonard Sielecki. "Animal species recognition using deep learning." In Advanced Information Networking and Applications: Proceedings of the 34th International Conference on Advanced Information Networking and Applications (AINA-2020), pp. 523-532. Springer International Publishing, 2020.

[14] Rubert, J., N. Sebastià, J. M. Soriano, C. Soler, and J. Mañes. "One-year monitoring of aflatoxins and ochratoxin A in tiger-nuts and their beverages." Food chemistry 127, no. 2 (2011): 822-826.

[15] Ranparia, Devsmit, Gunjeet Singh, Anmol Rattan, Harpreet Singh, and Nitin Auluck. "Machine learning-based acoustic repellent system for

[16] "Protecting crops against wild animal attacks." In 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), pp. 534-539. IEEE, 2020.

[17] Islam, Md Tamzeed, and Shahriar Nirjon. "Soundsemantics: exploiting semantic knowledge in text for embedded acoustic event classification." In Proceedings of the 18th International Conference on Information Processing in Sensor Networks, pp. 217-228. 2019.

[18] Cartwright, Mark, Jason Cramer, Ana Elisa Mendez Mendez, Yu Wang, Ho-Hsiang Wu, Vincent Lostanlen, Magdalena Fuentes et al. "SONYC-UST-V2: An urban sound tagging dataset with spatiotemporal context." *arXiv preprint arXiv:2009.05188* (2020).

[19] Qurthobi, Ahmad, Robertas Damasevicius, Vytautas Barzdaitis, and Rytis Maskeliunas. "Robust Forest Sound Classification Using Pareto-Mordukhovich Optimized MFCC in Environmental Monitoring." IEEE Access (2025).

[20] Imran, Mohammed Safwat, Afia Fahmida Rahman, Sifat Tanvir, Hamim Hassan Kadir, Junaid Iqbal, and Moin Mostakim. "An analysis of audio classification techniques using deep learning architectures." In 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 805-812. IEEE, 2021.

[21] Fonseca, Eduardo, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. "Fsd50k: an open dataset of human-labeled sound events." IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2021): 829-852.

[22] Gantert, Luana, Matteo Sammarco, Marcin Detyniecki, and Miguel Elias M. Campista. "Super Learner Ensemble for Sound Classification using Spectral Features." In 2022 IEEE Latin-American Conference on Communications (LATINCOM), pp. 1-6. IEEE, 2022.

[23] Vashishtha, Srishti, Rachna Narula, and Poonam Chaudhary. "Classification of Musical Instruments' Sound using kNN and CNN." In 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1196-1200. IEEE, 2024.

[24] Wolf-Monheim, Friedrich. "Spectral and Rhythm Features for Audio Classification with Deep Convolutional Neural Networks." arXiv preprint arXiv:2410.06927 (2024).

[25] Bertocco, Matteo, Stefano Parrino, Giacomo Peruzzi, and Alessandro Pozzebon. "Estimating volumetric water content in soil for IoUT contexts by exploiting RSSI-based augmented sensors via machine learning." Sensors 23, no. 4 (2023): 2033.

[26] Zaman, Khalid, Melike Sah, Cem Direkoglu, and Masashi Unoki. "A survey of audio classification using deep learning." IEEE Access (2023).

[27] Simiyu, Daniel, Henry Muchiri, Allan Vikiru, Julius Butime, and Zainabu Muti. "A Forest Acoustics–Temporal Frequency Convolution Neural Network Model for Detecting Illegal Logging Activities in Forest." In 2024 5th International Conference on Smart Sensors and Application (ICSSA), pp. 1-6. IEEE, 2024.

[28] Adapa, Sainath. "Urban sound tagging using convolutional neural networks." *arXiv preprint arXiv:1909.12699* (2019).

[29] Korkmaz, Yunus. "SS-ESC: a spectral subtraction denoising based deep network model on environmental sound classification." Signal, Image and Video Processing 19, no. 1 (2025): 1-13.

[30] Gourisaria, Mahendra Kumar, Rakshit Agrawal, Manoj Sahni, and Pradeep Kumar Singh. "Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques." Discover Internet of Things 4, no. 1 (2024): 1.

[31] Avadhani, Mita, Anupama P. Bidargaddi, and S. Thushara. "Multi-Class Urban Sound Classification with Deep Learning Architectures." In 2024 5th International Conference for Emerging Technology (INCET), pp. 1-7. IEEE, 2024.

[32] Ustubioglu, Arda, Beste Ustubioglu, and Guzin Ulutas. "Mel spectrogram-based audio forgery detection using CNN." Signal, Image and Video Processing 17, no. 5 (2023): 2211-2219.

Mohmmad, Sallauddin, and Suresh Kumar Sanampudi. "Exploring current research trends in sound event detection: a systematic literature review." *Multimedia Tools and Applications* (2024): 1-43.