A Survey on Deep Learning Based Forest Environment Sound Classification at the Edge

DULANI MEEDENIYA, ISURU ARIYARATHNE, MEELAN BANDARA, and ROSHINIE JAYASUNDARA, Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka
CHARITH PERERA, School of Computer Science and Informatics, Cardiff University, United Kingdom

Forest ecosystems are of paramount importance to the sustainable existence of life on earth. Unique natural and artificial phenomena pose severe threats to the perseverance of such ecosystems. With the advancement of artificial intelligence technologies, the effectiveness of implementing forest monitoring systems based on acoustic surveillance has been established due to the practicality of the approach. It can be identified that with the support of transfer learning, deep learning algorithms outperform conventional machine learning algorithms for forest acoustic classification. Further, a clear requirement to move the conventional cloud-based sound classification to the edge is raised among the research community to ensure real-time identification of acoustic incidents. This article presents a comprehensive survey on the state-of-the-art forest sound classification approaches, publicly available datasets for forest acoustics, and the associated infrastructure. Further, we discuss the open challenges and future research aspects that govern forest acoustic classification.

CCS Concepts: • **Computing methodologies** → **Neural networks**; • **General and reference** → **Surveys and overviews;**

Additional Key Words and Phrases: Sound processing, edge computing, deep learning, artificial intelligence, Internet of Things

ACM Reference format:
Dulani Meedeniya, Isuru Ariyarathne, Meelan Bandara, Roshinie Jayasundara, and Charith Perera. 2023. A Survey on Deep Learning Based Forest Environment Sound Classification at the Edge. *ACM Comput. Surv.* 56, 3, Article 66 (October 2023), 36 pages.
https://doi.org/10.1145/3618104

1 INTRODUCTION

Forest monitoring systems are predominantly concerned with tracking deforestation and other environmental aspects to identify necessary actions to preserve forest ecosystems. With the advancements in the **Internet of Things (IoT)**, **Deep Learning (DL)** and wireless sensor networks, a surge of interest in sound classification based forest surveillance systems can be identified. Sound classification is a broad area of research that receives significant attention in numerous

Authors' addresses: D. Meedeniya, I. Ariyarathne, M. Bandara, and R. Jayasundara, Department of Computer Science and Engineering, University of Moratuwa, Moratuwa, 10400, Sri Lanka; e-mails: dulanim@cse.mrt.ac.lk, isuru.18@cse.mrt.ac.lk, meelan.18@cse.mrt.ac.lk, roshinie.18@cse.mrt.ac.lk; C. Perera, School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, UK; e-mail: pereraC@cardiff.ac.uk.

0360-0300/2023/10-ART66 $15.00
https://doi.org/10.1145/3618104

real-world applications, including healthcare [59], smart city management [106], music genre identification [7], and domestic audio identification [29]. Similarly, **Environment Sound Classification (ESC)** has evolved as an active research area where extensive contributions have been made. Such intelligent systems enable real-time monitoring of the forest environment by identifying illegal logging of trees, wildfires, poaching, weather changes, and more.

Over time, various attempts have been made for effective surveillance of large forest areas, including satellite image processing [131], video recording surveillance [49], and motion and vibration sensor data processing [89, 99]. However, these approaches experience limitations such as expensiveness, the requirement of advanced sensors, and high-power consumption [98, 117]. Acoustic surveillance, being a more feasible solution, provides a different sensory dimension on larger geographical boundaries [22]. Previously, environmental sound recognition was done by utilizing traditional **Machine Learning (ML)** approaches, which showed limited performance due to high error susceptibility and low ability in accurate data processing [24, 110]. Overcoming such challenges, DL approaches received high attention from the research community due to their ability to extract features from raw data, better self-learning capabilities, and precise results. IoT architecture also plays a major role in the practical implementation of a forest acoustic classification system and can be developed based on two major approaches. In the first approach, forest acoustics are recorded in a forest environment and the audio files are transferred to a cloud server for the processing to take place, whereas in the second approach, audio files are recorded and processed onsite using an edge device. Both approaches become challenging due to reasons like harsh weather conditions, animal disturbances, unavailability of efficient bandwidth, and resource constraints.

Deploying the classification model on the edge device itself reduces the communication latency, avoids flooding of raw data at the cloud level, improves security, and increases the quality of data due to near sensor processing [21, 36]. At the moment, the edge paradigm is still in its early stages, hence the number of available studies is limited. Therefore, deploying trained models on edge devices for real-time acoustic surveillance requires considerations that have not been sufficiently addressed in previous studies. The requirement of high computational power, memory constraints, and high energy consumption can be identified as the main challenges involved with edge computing.

Survey papers which explain the general practices, including feature extraction, model selection, training, and evaluation, for ESC can be identified [1, 3, 11, 75]. Further, studies that present the related work on implementing ML and DL models in edge devices, including the IoT aspects in a generic scope, are available [75, 117, 124]. But none of the mentioned surveys provide a solid comparison of datasets available for **Forest Sound Classification (FSC)**, **Convolutional Neural Network (CNN)**, and **Recurrent Neural Network (RNN)** based models used for ESC, hardware configurations used for ESC systems, utilized evaluation metrics, and optimization techniques used to deploy ML and DL models in edge devices. Thus, a survey paper that provides an overview of the full pipeline for FSC is not available. These limitations motivated us to complete this survey, primarily focusing on the state-of-the-art DL approaches for FSC implemented on resource-constrained edge devices.

This survey work is organized as follows. Section 1 provides an overview of forest monitoring systems. Section 2 emphasizes the scope and the motivation of the survey, including the evolution of related techniques. Section 3 includes a comprehensive study on sound pre-processing techniques, categorized under audio normalization, data augmentation, and feature extraction. Publicly available environment sound datasets and their limitations are addressed in Section 4. Section 5 and Section 6 focus on sound classification infrastructure that describes the evolution from ML to DL and the different DL models available for sound classification under CNN and RNN. Section 7 provides recommendations and discusses the open challenges and future research

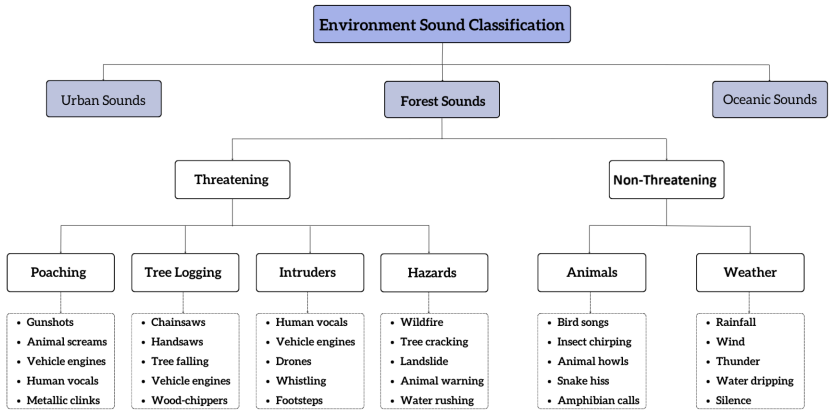

Fig. 1. Overview of ESC.

directions identified in this domain. Finally, Section 8 concludes this article with the intention of supporting future researchers working in the forest acoustics domain.

2 BACKGROUND

2.1 Real-World ESC

The classification of environmental sound events includes a number of areas, including ocean [76], forest, and urban [66] acoustics as shown in Figure 1. Scientists and conservationists can delve into the complex world of forest noises by utilizing cutting-edge methods for acoustic event recognition. This gives them the opportunity to learn important information about the risks, hazards, and non-threatening elements of the environment. The significance of forest acoustics is emphasized in this section, as well as its substantial contributions to ecological study and conservation activities.

Acoustic recognition systems can efficiently detect activities like poaching, illicit logging, and infiltration attempts in forest areas by using complex algorithms and models. DL algorithms used in gunshot detection systems make it possible to locate gunshots in tropical forests, which helps to stop poaching and other wildlife crimes [1]. Conservationists can fight deforestation operations using acoustic surveillance and monitoring techniques, which enable the detection and identification of noises connected to unlawful tree cutting [6, 79]. Additionally, the detection of noises associated with vehicle movements and drones is made easier with the use of forest acoustics, improving security measures and preserving the integrity of forested regions [45].

Forest acoustic recognition can be successfully utilized for locating and minimizing natural hazards. By identifying the noises connected to wildfires, acoustic event recognition enables early detection and prompt action [136]. Acoustic recognition techniques assist with climate assessment and the foretelling of possible threats like flooding by monitoring environmental noises like rainfall and wind patterns [12]. With the aid of these skills, researchers and forest authorities may gather crucial data for the creation of manageable plans that will lessen the effects of natural dangers on forest ecosystems.

Understanding non-threatening sounds, particularly animal vocalizations, is a component of forest acoustics. For the purposes of biodiversity research and conservation activities, DL algorithms can be utilized to distinguish and classify various animal sounds [132]. As a result, it is feasible to recognize various species, monitor their habitats, and defend both those ecosystems and the threatened wildlife that depends on them. Furthermore, sound categorization methods can be used to successfully identify the climate patterns present in forest environments [14].

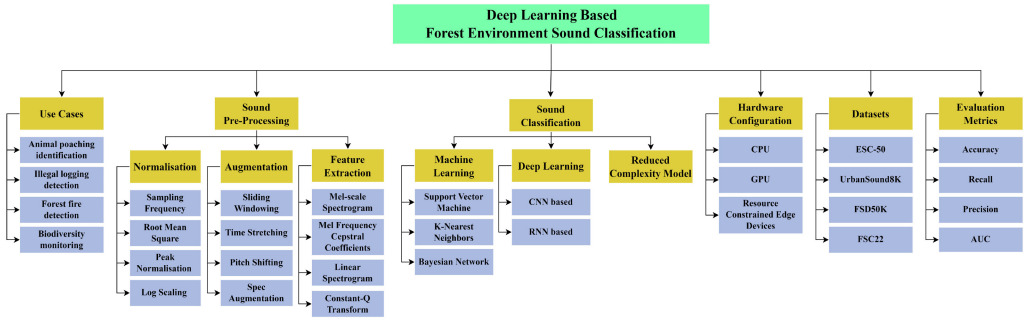

Fig. 2. Overall taxonomy.

Hence, forest acoustics becomes an essential instrument for preserving and safeguarding ecosystems. Its potential in threat detection, hazard mitigation, and sound classification of non-threatening nature offers priceless insights for ecological study and conservation activities. Forest acoustics equips researchers, environmentalists, and forest managers with the information and resources they need to efficiently monitor, manage, and conserve our priceless forests by utilizing cutting-edge acoustic detection techniques.

2.2 Scope and Motivation

Conducting a survey in the field of FSC offers a vital opportunity for researchers and environmentalists to actively contribute to the preservation and protection of such ecosystems. By employing advanced algorithms and models, especially DL techniques, forest acoustics holds the potential to accurately identify and categorize sound events linked to illegal human activities, leading to swift intervention and enhanced security measures. The insights gained through this survey will equip relevant parties with the information and resources necessary for efficient monitoring, management, and conservation of forests. Leveraging state-of-the-art acoustic detection techniques, researchers can gather rich data and develop actionable plans to mitigate the illegal activities by humans and natural hazards. Thus, the field of FSC emerges as a powerful tool that not only propels the advancement of ecological research but also facilitates impactful conservation activities.

Motivated by prior remarks, this survey explores the standard and novel techniques discussed in related literature, which can be used to develop forest acoustic classification systems. When developing such systems, acoustic data availability, data preprocessing techniques, feature engineering techniques, model selection, model evaluation, hardware configurations, and optimizations to support resource-constrained devices are of high importance. Figure 2 presents the overall taxonomy for FSC, which structures the preceding aspects in a hierarchical viewpoint. With the intention of identifying key areas to focus on in this survey, we carefully explored survey papers that discuss the preceding aspects and selected four highly related survey papers for further analysis. Among the selected papers, none specialized in FSC, but some works [1, 3, 11] provide a sound review of standard techniques used in ESC. Further, the work of Sharma et al. [113] provides a proper explanation of feature representation techniques that can be used for ESC. Table 1 presents a summary of the key contributions made by the considered survey papers.

We have identified that none of the explored surveys discusses or compares techniques utilized for optimizing DL models to function with constrained resources, hardware configurations best suited to edge deployment, audio normalization techniques, and evaluation metrics. In addition, expositions including a sufficient amount of use cases for ML models, CNN models, RNN models, and transfer learning techniques used in FSC or ESC are not provided. These limitations motivated us to explore the techniques used in related studies which can be utilized to deploy edge DL

Table 1. Summary of Existing Surveys

Consideration	[11] (2022)	[1] (2022)	[113] (2020)	[3] (2020)	Our Article
Datasets	•	•	-	-	•
Audio Normalization	-	-	-	-	•
Data Augmentation	-	•	-	◦	•
Feature Extraction	•	•	•	◦	•
Traditional Machine Learning	◦	◦	-	-	•
CNN	◦	◦	-	◦	•
RNN	◦	◦	-	◦	•
Transfer Learning	-	-	-	◦	•
Hardware Configurations	-	-	-	-	•
Resource-Constrained Edge Devices	-	-	-	-	•
Evaluation Metrics	-	-	-	-	•

◦ Partially Discussed • Well Discussed.

models for FSC. This study stands out for comprehensively reviewing the entire pipeline of FSC. In contrast to existing survey papers that focus on specific aspects, our article provides a holistic analysis, covering all stages of the pipeline by exploring the state-of-the-art techniques presented in related studies. Through this comprehensive analysis, we not only bring a novel contribution to the field but also establish concrete future directions and highlight open challenges in the domain. Accordingly the contributions incorporated into this survey work can be summarized as follows:

- We present an overview and a comparison of sound pre-processing techniques used for audio normalization, data augmentation, and feature extraction.
- We review the publicly available environment sound datasets with their limitations, alongside the requirement for a real-world benchmark dataset.
- We explain the effectiveness of DL over ML for ESC, aided with scenarios from related literature.
- We present a comparative review of state-of-the-art CNN and RNN architectures utilized in related research work.
- We present a comparison of different approaches followed by researchers to deploy DL models in edge devices.
- We provide a review of hardware technologies used in edge device implementations.
- Finally, the survey provides recommended approaches, open challenges, and future research directions for forest acoustic surveillance.

2.3 Methodology

For this literature survey, we conducted an extensive exploration of research articles, intending to understand the current state-of-the-art techniques utilized in DL-based ESC for forest surveillance and capture aspects that need the attention of future research. Our search was performed using the advanced search feature supported by the Google Scholar platform due to the ability to search by the published source with different keywords. The search was primarily focused on five mainstream research sources—ACM, IEEE Xplore, Elsevier, MDPI, and Springer—due to the abundant related literature in these databases and the high quality maintained in the published papers. The data acquisition procedure employed for this survey is illustrated in Figure 3.

For the Google Scholar search query, we used keywords like "Environment Sound Classification," "Environment Sound Recognition," "Deep Learning," "Forest Sound," and names of different DL techniques such as "LSTM," "VGG," "GRU," and "Inception," where the source field is specified with the aforementioned source list. Further, a time filter was used to capture only the studies after 2018. Thus, our comparisons will provide an overview of the current state of the ESC domain.

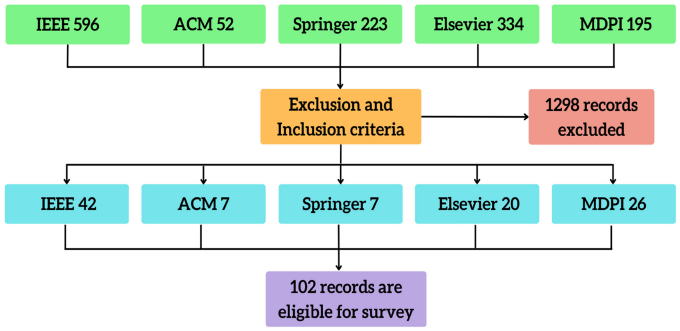

Fig. 3. The paper selection process for the survey.

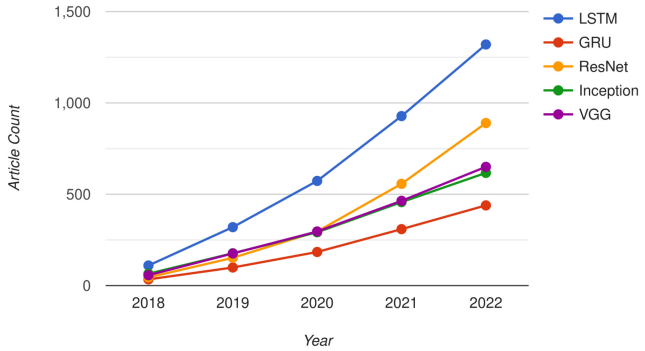

Fig. 4. Evolution of DL techniques.

Due to the noise inherent in the search query, many research papers that cannot be used for this survey were also returned. Hence, a screening phase was conducted to remove unusable papers such as duplicates and papers where keywords were only mentioned in the related works section or references. In addition, we have selected papers written in the English language only. After the complete selection process, 102 eligible research papers on ESC were identified for the survey, as summarized in Figure 3. These studies include acoustic processing using DL in different settings and environments, where we can explore possible techniques and infrastructure for FSC.

2.4 Evolution of DL Techniques

DL is a rapidly growing approach in data engineering. It allows the analysis of different data types and automated feature extraction from raw data [73]. We explored the progressive evolution of DL techniques utilized for the ESC domain during the past few years. Figure 4 presents a quantitative analysis of the usage of different DL models used in ESC for the period starting from the year 2018 to the end of 2022. The search query employed in Google Scholar to obtain the information was "(technique name)" + "environmental sound" + "classification" OR "recognition." However, all the obtained records for a given search may not be directly associated with the considered scope due to the inherent noise in the querying process. Hence, we assumed that this error would similarly affect all the queries when presenting the evolution of related techniques in Figure 4.

We have considered two RNN-based and three CNN-based models which are widely used in the ESC-based literature. At the time of writing, Google Scholar reported 1,230 and 3,270 search results when "RNN" and "CNN" were used as the technique name, respectively. Hence, it becomes apparent that there is a higher interest in CNN-based approaches for the ESC domain. Further,

Visual Geometry Group (VGG), Inception, and **Gated Recurrent Unit (GRU)** techniques have shown a similar and steady rise in usage since 2018. In 2020, a comparatively high rise in the usage of Residual Network (ResNet) architecture can also be identified. However, there has been a higher interest and rapid growth in the **Long Short-Term Memory (LSTM)** technique for ESC compared to other techniques over time. In addition, these results include studies that have used DL techniques for ESC on both cloud-based and edge-based implementations.

Furthermore, a scarcity of research work can be identified for edge-based ESC using DL. To analyze the edge-based research in a quantitative manner, the preceding query was extended with the keyword "edge" or "embedded." Google Scholar reported 521 and 1,360 search results when the new query was used with the technique names "RNN" and "CNN," respectively. This is considerably low when compared to the search results obtained with the original query. Thus, the requirement of conducting further research on the edge-based acoustic classification domain can be clearly identified.

3 SOUND PRE-PROCESSING

3.1 Audio Normalization

Audio samples are utilized to train the classifier, and later classification decisions are made on the real-time acquired audio samples. These audio samples contain different loudness levels and may contain different interferences. Therefore, for the classifier to generate efficient and accurate classification decisions, it requires normalized data samples [58]. Here, the same normalization techniques need to be followed in the real-time classification phase as well. Audio normalization is an effective pre-processing method to obtain more consistent data before passing it into the learning models. Generally, this process changes the overall volume by a constant amount to reach the amplitude to a target level. Since the same amount of gain is applied over the data, it does not affect the signal-to-noise ratio and relative dynamics. Thus, audio normalization does not change the sound or compress the data.

Table 2 presents the usage of audio normalization techniques including peak normalization, root mean square, **European Broadcast Union (EBU)** standard, pre-emphasis filters, sampling frequency, the bit depth of audio, and log scaling normalization for environment sound pre-processing. Accordingly, it can be seen that techniques such as sampling frequency and pre-emphasis filters are widely used for the audio normalization of forest sounds. Using higher sample frequency, a higher number of samples of audio, which are taken per second, generates more accurate representation of the original audio, which is helpful to preserve the nuances and details of the original audio. Pre-emphasis filters improve the overall quality and clarity of the audio by boosting the higher frequency of audio signals. It can be used to detect and correct issues in the audios.

Among many related studies, Shah et al. [111] have compared the usage of peak normalization, root mean square normalization, and the EBU standard as audio normalization techniques. They have comparatively shown that the EBU normalization technique performs with the highest accuracy of 95% when coupled with data augmentation techniques. In addition, pre-emphasis filters are used to normalize the audio (bit depth of audio) samples to balance the audio noise ratio and numerical complexities [102]. Sampling frequency and bit depth of audio samples also play a vital role. Andreadis et al. [6] have analyzed the tradeoff between computational cost and the quality of the input using 2 bit depths (8 and 32) sampling at 16 kHz. They have shown that the highest classification accuracy of 85.37% can be obtained with 32-bit input. In contrast, an accuracy of 85.03% is gained with 8-bit audio samples. The authors have further shown that the usage of 8-bit audio samples significantly reduces computational resources, whereas the inference time of the classifier is reduced from 1,089 ms to 232 ms when compared with the 32-bit approach.

Table 2. Usage of Audio Normalization Techniques in Forest Observatory Studies

Technique	Dataset	Studies	Total
Sampling Frequency	ESC	[6, 37, 43, 48, 77, 78, 95, 130]	27
	UrbanSound8k	[2, 26, 31, 43, 48, 52, 77, 95]	
	Bird sound (Xeno-canto)	[22, 137, 138]	
	DCASE2016	[52, 130]	
	AudioSet	[37, 43]	
	Forest sound	[46]	
	Animal sounds	[58]	
	BirdCLEF	[65]	
	Bird Sound (CLO-43DS)	[132]	
Pre-Emphasis Filters	UrbanSound8k	[52, 57, 102, 107, 133]	10
	ESC	[102, 107]	
	FSDKaggle18	[102, 107]	
	DCASE2016	[52]	
Peak Normalization	DCASE2016	[58]	2
	UrbanSound8k	[111]	
Root Mean Square	Chainsaw sound	[79]	2
	UrbanSound8k	[111]	
Bit Depth of Audio	Bird sound (Xeno-canto)	[22]	2
	UrbanSound8k	[77]	
European Broadcast Union Standard	UrbanSound8k	[111]	1
Log Scaling	Animal sounds	[22]	1

Moreover, the usage of log scaling and z-score normalization can also be identified in the literature. Cartwright et al. [18] have used log scaling over the raw audio files before the corresponding feature spectrograms are extracted. Further, they have compared the usage of per-channel energy normalization as an alternative to log scaling and provided evidence that such normalization increases the performance of the classifier by a considerable margin. Furthermore, the z-score is applied to normalize raw audio data in the work of Xie et al. [132]. The authors have clipped the audio frames sampled at 22.05 kHz to remove unnecessary parts and then applied z-score normalization before extracting Mel spectrograms.

3.2 Data Augmentation

Data augmentation is used to increase the dataset by forming new and slightly modified copies of the existing data. Thus, augmentation helps to synthetically increase the number of data points during sound data pre-processing [73, 83]. By generating a rich and sufficient dataset, data augmentation techniques support the generalization of the classifier and prevent overfitting in the training phase, which results in increased model performance [65]. Several techniques, such as noise injection that adds a random value to data, shifting time that shifts audio to left or right with a random second, and changing pitch and speed, are discussed in the related literature [37, 83, 85]. Thus, appropriate techniques to augment the dataset should be selected based on the given data pattern.

Most of the related studies have applied data augmentation in ESC tasks for both urban and remote areas. Das et al. [31] have used pitch shift, time stretch, and pitch shift combined with time stretch as audio data augmentation techniques on the UrbanSound8k dataset. Nanni et al. [86] have

used a single frame to create 10 more audio clips using techniques like signal speed scaling and **Pitch Shifting (PS)** by a random number in their work on animal sound processing. In addition, random time delays, **Time Stretching (TS)**, and PS have been used to augment environment sound data [95].

The sliding window is another technique used in both real-time acquisitions of sound and augmentation of training datasets. This technique generates sound frames of known length from audio streams because shorter sound signals can better capture acoustic phenomena [23]. It allows overlapping between successive frames, thus increasing the number of data frames. For instance, a sliding window of 4,000 ms with 50-ms hop length (3,950-ms overlap) can be used to create 21 audio frames each of 4,000-ms duration from a 5,000-ms audio signal [6]. Similarly, Mporas et al. [79] have used a sliding window of 20 ms with a 50% overlapping between successive audio frames. Although windowing presents the preceding advantages to the classification procedure, determining the best-suited window size to represent the relevant acoustic phenomena by the generated frames is challenging.

Table 3 summarizes the usage of audio augmentation techniques including TS, spec augmentation, PS, signal speed scaling, random time delays, and **Sliding Windowing (SW)** in environment sound pre-processing. It can be observed that techniques such as SW, PS, and TS are widely used for sound data augmentation, whereas most of the studies incorporate multiple augmentation techniques to overcome the lack of audio data available for ESC and to prevent models from overfitting. SW generates multiple variations of an audio signal by applying transformations to small, overlapping segments of the signal rather than the entire signal at once. PS generates multiple variations of audio through raising and lowering the pitch and TS speeding and slowing down the audios to generate multiple audios. Both TS and PS can preserve the timbre of the original audio signal while still making it sound different. SW, PS, and TS augmentation methods are resource-intensive and efficient augmentation methods capable of providing more complex and varied transformations to the audio signal that make researchers use these techniques more frequently in data augmentation.

A comparison is presented by Wei et al. [127]. While using the aforementioned augmentation techniques, they introduce an efficient technique called *Mixed Frequency Masking*, which outperforms other augmentation methodologies. Moreover, Mushtaq et al. [83] have compared different sound augmentation techniques. They have shown that a combination of positive pitch shift, negative pitch shift, slow time stretch, fast time stretch, and silence trimming outperforms traditional techniques such as zoom range, width shift, fill mode, brightness range, rotation angle, height shift, shear range, and horizontal flip for the ESC-50 dataset. For the comparison, they have used a novel deep CNN and state-of-the-art transfer learning models including AlexNet, ResNet, DenseNet, and VGG.

Apart from expanding the available data points as described earlier, actions are needed to increase the quality of the available data as well. Mushtaq and Su [82] have described the use of silence trimming functions available in the librosa package [72] to increase the quality of the data points, as significant portions of available audio files contain irrelevant low decibel sounds. The effect of data efficiency on the pipeline efficiency of a classification system is discussed by Li et al. [67]. They have shown the impact of missing files and speed for data efficiency, and the importance of using smaller features to obtain maximum efficiencies.

3.3 Feature Extraction

A compact representation of audio signals is generated by applying feature extraction before the model training to reduce the computational complexities [35, 113]. In audio data, feature extraction mainly considers the time and frequency domain representation. We discuss the usage of related techniques such as linear spectrogram, Mel-scale spectrogram, **Mel Frequency Cepstral**

Table 3. Usage of Audio Data Augmentation Techniques in Environment Sound Processing

Technique	Dataset	Studies	Total
Sliding Windowing	UrbanSound8k	[21, 34, 52, 53, 94, 115, 133]	24
	ESC	[6, 21, 53, 53, 108, 115, 130]	
	Bird sound (Xeno-canto)	[22, 137, 138]	
	DCASE	[52, 130]	
	Chainsaw sound	[79]	
	Animal sounds	[68]	
	Forest sound	[89]	
	TUT sound events	[21]	
	Bird sound (CLO-43DS)	[132]	
Time Stretching	ESC	[48, 69, 83, 85, 95, 95, 121]	17
	UrbanSound8k	[31, 48, 69, 83, 95, 120]	
	Birds and cat sounds	[86]	
	AudioSet	[67]	
	DCASE	[121]	
Pitch Shifting	UrbanSound8k	[21, 31, 48, 69, 83, 95, 120]	16
	ESC	[21, 48, 69, 83, 85, 95]	
	Birds and cat sounds	[85, 86]	
	TUT sound events	[21]	
Spec Augmentation	ESC	[85, 121]	6
	Bird, Cat	[85]	
	DCASE	[121]	
	UrbanSound8k	[26]	
	BirdCLEF	[65]	
Random Time Delays	ESC	[85, 95, 95]	5
	Bird, Cat	[85]	
	UrbanSound8k	[95]	
Signal Speed Scaling	ESC	[85, 121]	4
	Bird, Cat	[85]	
	DCASE	[121]	

Coefficients (MFCC), **Continuous Wavelet Transform (CWT)**, and **Fast Fourier Transform (FFT)** spectrogram for feature extraction of environmental sounds.

Table 4 provides an overview of different feature extraction techniques used in related literature, including linear spectrogram, Mel-scale spectrogram, MFCC, CWT, **Constant-Q Transform (CQT)** spectrogram, and zero crossing rate for forest environment sound pre-processing. Accordingly, it can be seen that techniques such as MFCC and Mel-scale spectrogram are widely used for feature extraction of environment sounds due to the ability to perceive natural human hearing processes using the Mel scale.

The linear spectrogram is a simple feature extraction method that considers both time and frequency domain features of audio data. Fourier transform is used to generate the frequency domain features because a given audio signal is a combination created by the super-imposition of distinct audio signals with different frequencies and amplitudes [77]. Several feature extraction methods based on the Fourier extraction of an acoustic signal are available in the literature. Among them, a simple application of the Fourier spectrum is used by Segarceanu et al. [107]. They have calculated the sum of squared Fourier coefficients for a given analysis frame, which is termed the *power*

Table 4. Usage of Feature Extraction Techniques of Sound Data

Technique	Dataset	Studies	Total
Mel-Scale Spectrogram	ESC	[6, 43, 69, 82, 83, 85, 108, 115, 130]	26
	UrbanSound8k	[31, 43, 69, 82, 83, 115, 120]	
	AudioSet	[43, 50, 67]	
	DCASE	[58, 130]	
	YouTube-100M	[50]	
	BIRDZ, Cat sound	[85]	
	TUT sound	[69]	
	Bird sound (CLO-43DS)	[132]	
	Bird sound (Xeno-canto)	[138]	
Mel Frequency Cepstral Coefficients	UrbanSound8k	[31, 52, 53, 57, 102, 120]	23
	ESC	[6, 37, 53, 102, 130]	
	Forest sound (private)	[89, 107]	
	Environment sound (private)	[30, 64]	
	Chainsaw sound (private)	[46, 79]	
	DCASE	[52, 130]	
	Bird sound (Xeno-canto)	[22]	
	Animal sound	[128]	
	AudioSet	[37]	
	FSDKaggle2018	[102]	
Linear Spectrogram	ESC	[6, 53, 108, 115]	13
	UrbanSound8k	[31, 53, 111, 115]	
	Bird sound (Xeno-canto)	[137, 138]	
	Chainsaw sound (private)	[46]	
	Forest sound (private)	[89]	
	Own dataset	[61]	
Constant-Q Transform	ESC	[37, 53]	6
	UrbanSound8k	[31, 53]	
	Chainsaw sound (private)	[46]	
	AudioSet	[37]	
Spectrogram and Zero Crossing Rate	Chainsaw sound (private)	[79]	4
	Forest sound (private)	[89]	
	ESC	[82]	
	UrbanSound8k	[82]	
Continuous Wavelet Transform	Chainsaw sound (private)	[46]	3
	ESC	[53]	
	UrbanSound8k	[53]	

spectra of the signal and is used to represent the signal in the frequency domain. This study has used the power spectra obtained for the signal with dynamic time warping and feedforward networks to generate classification decisions. Moreover, Andreadis et al. [6] have used the linear spectrogram technique with a sub-framing size of 20 ms. They have calculated the spectrogram for 256 different frequency bands. In addition, a comparison between linear and log spectrograms is presented by Dennis et al. [33]. They have shown how less information available in linear spectrograms causes lower accuracies when classification decisions are made. The clarity and the details included in a linear spectrogram can be increased by converting the amplitudes to decibel scale.

The Mel scale is a perpetual scale developed based on the way that humans recognize the sound, and it is the result of a non-linear transformation of the frequency scale [125]. Generally, a spectrogram visualizes the frequencies of a signal varying with time. A Mel spectrogram logarithmically renders frequencies above a certain threshold. Among several studies, Hyder et al. [54] have presented a comparison between the effectiveness of feature extraction with linear-scaled, log-scaled, and Mel-scaled filter banks when classification is done with a CNN model. MFCCs are a set of features derived from the Mel spectrogram of a given audio signal. This has been considered the dominant feature extraction method for audio analysis [100]. Although a large set of MFCCs can be extracted from a given audio frame, generally the first 8 to 13 MFCCs are used in feature extraction, as they generate a robust and accurate representation of the considered audio [5]. Sharan and Moir [112] have presented an analysis of linear-scaled and log-scaled MFCCs for environmental sound analysis. They have shown that high accuracies can be obtained using noise-free audio data.

Several studies have compared different feature extraction techniques. In the work of Das et al. [31], a comparison between MFCC, Mel spectrogram, Chroma STFT, Chroma CQT, Chroma CENS, spectral contrast, and Tonnetz as feature extractors are presented. They have incorporated CNN and LSTM against the UrbanSound8k dataset in their comparative study and show that the best results are obtained with MFCCs. Another comparative study [134] presents a comparison between Fourier spectrogram, Mel-scale spectrogram, CQT, and MFCCs. Those authors have shown that raw Fourier transform, which can achieve an accuracy of 77.38%, can be increased to 91.61% when combined with a histograms of oriented gradients local descriptor. Moreover, the effectiveness of the Mel-scale spectrogram, **Log-Mel Spectrogram (LM)**, and MFCC are addressed by Peng et al. [94]. They have combined these feature extracting methods to generate Mel-MFCC, LM-MFCC, and T-M (Mel-LM-MFCC) and conclude that LM-MFCC feature fusion can achieve high performance against the ESC dataset using the GRU architecture. The tradeoff between the accuracy of the predictions and the computational and memory cost has become a vital factor when the classification model is implemented in a low-resourced edge device. Beneficially, MFCC can be identified as a proper feature extraction methodology that does not require substantial memory and computational cost [6]. Further, Andreadis et al. [6] have shown that the usage of MFCC reduces the inferencing time of the model when compared with other feature extraction methods such as linear spectrograms and Mel-scale spectrogram when used with CNNs.

The CWT scalogram is another feature extraction approach that provides better time localization and is suited for non-stationary signals [60, 119]. Copiaco et al. [29] have proposed an FFT method to extract feature coefficients from the CWT scalogram such that the computational complexities can be reduced. This feature extraction method has produced robust and high-accuracy results compared to MFCC and Mel spectrogram. Another novel technique has been introduced by Okawa et al. [88], showing that the bit representation of audio contains more properties than the integer-based waveform representations. To introduce the noise robustness to feature extraction, Huang et al. [52] have proposed a fusion between MFCC and Gammatone frequency cepstral coefficients. They have shown the effectiveness of the fusion of MFCC and Gammatone frequency cepstral coefficients against UrbanSound8k and DCASE2016 while achieving high accuracies for both datasets.

4 ENVIRONMENT SOUND DATASETS

4.1 Benchmark Dataset

A benchmark dataset consists of a variety of data, representing real-world scenarios. This will reduce the biases that can be introduced to the datasets. Moreover, such a dataset can be utilized as a general measure to determine the strengths and weaknesses of different methodologies with rigorous evaluation. Although there are several public datasets available, they are inconsistent

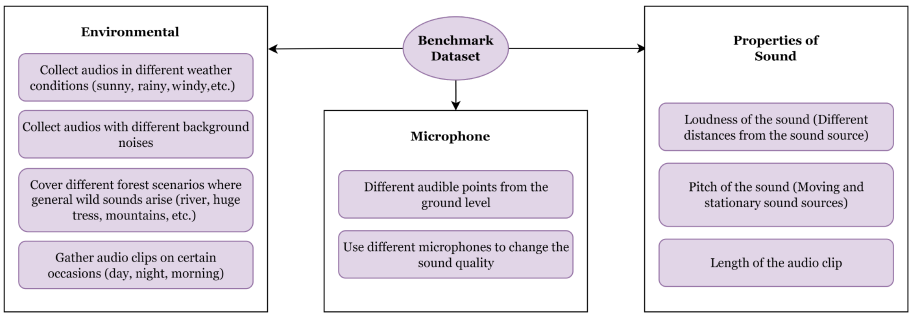

Fig. 5. Requirements for a benchmark dataset.

in some scenarios and lack diversity. Figure 5 shows the major features of benchmark datasets under three aspects, namely different environmental factors included in the sound, the importance of varying sound-related properties, and microphone-related factors [84]. Changes in different environmental conditions can introduce a variety of noise effects on audios, which is a major problem when creating a benchmark dataset. Thus, to improve the diversity of the dataset, it is essential to collect audios under different weather conditions covering common forest scenarios representing varying periods. Consequently, the properties of sounds are another considerable factor for a high-quality benchmark dataset. Therefore, it is required to include a recording of the same sound source with additional loudness levels, pitch levels, and lengths. Moreover, using different microphones and varying the position from the ground level to capture the sound can improve the generality of a benchmark dataset.

4.2 Public Datasets

Studies focused on forest acoustic monitoring generally develop new datasets by combining audio data available in public datasets and their privately collected audio dataset. This section provides an overview of publicly available datasets that govern the ESC domain and can be used for the preceding purpose as well:

***FreeSound*:** An audio-based public dataset in a collaborative platform. It consists of more than 500,000 audio clips. In addition, FreeSound provides an API to access audio data in different formats with complex search functionalities [41, 42]. Datasets such as UrbanSound8k and ESC-50 were created by extracting data from the FreeSound dataset. The dataset is available at https://annotator.freesound.org/ and https://labs.freesound.org.

***BBC Sound Effects library*:** A collection of 2,400 acoustic data captured in different geographical areas such as Europe, Asia, the Middle East, and South America [13]. All the audio files are easy to search due to the extensive metadata embedded and are available to download as mp3 or WAV files via the online platform of the library. Many studies have extracted forest acoustic surveillance data from this dataset to create their dataset [27, 112]. The data can be retrieved from https://www.epidemicsound.com/sound-effects.

***ESC-50*:** A collection of 2,000 short audio clips equally distributed among 50 classes, which are again divided mainly into five major classes: animal sounds, human (non-speech) sounds, urban noises, natural soundscapes and water sounds, and domestic sounds [97]. The ESC dataset is created by querying the Freesound database described earlier, considering classes of selected ESC taxonomy. Five-second-long recordings were extracted from the audio events, whereas shorter events with less than 5 seconds were padded with silence and converted to a unified format (44 KHz, single channel). The size of the dataset is approximately 600 MB and is

available at https://github.com/karolpiczak/ESC-50 [96]. Further, ESC-10, which is a less complex subset of ESC-50, is frequently used in related literature. It consists of 10 classes (dog, sneezing, clock ticking, crying baby, crowing rooster, rain, sea waves, fire crackling, helicopter, chainsaw) chosen from the main 50 classes.

***UrbanSound8k*:** A subset of the main UrbanSound dataset, which contains 27 hours of audio with 18.5 hours of annotated sound events [103]. The UrbanSound dataset is created by querying audios from the Freesound platform. The UrbanSound8k dataset consists of 8,732 audio clips and is categorized under 10 main classes (air conditioner, playing children, car horn, dog bark, engine handling, jackhammer, street music, siren, gunshot and drilling). The dataset contains nearly 1,000 audio clips per class. Audios were selected considering three main aspects: (1) containing sounds from urban environments, (2) considering recordings only from real field recordings, and (3) a sufficiently large dataset. Audios available in UrbanSound8k are 4 seconds or less in duration and are sufficient for classifying audios with considerable accuracy [27]. It should be noted that to generate publishable results for the UrbanSound8k dataset, studies must implement 10-fold cross validation with the pre-defined folds of the dataset [48]. The dataset is available at https://urbansounddataset.weebly.com/urbansound8k.html [104].

***AudioSet*:** A dataset consisting of human-annotated 10-second audio clips from YouTube [44]. The dataset is presented as a CSV file, which contains (1) a YouTube video identifier, (2) start time, (3) end time, and (4) labels of sound categories present in the audio. There are more than 2 million sound clips distributed among 527 unique classes. The available data points are divided as a balanced and unbalanced training set and evaluation set. This dataset is available at http://research.google.com/audioset/ontology/index.html [8].

***FreeSound Dataset 50K (FSD50K)*:** A subset of the AudioSet dataset consisting of 51,197 audio clips distributed among 200 classes based on the AudioSet ontology labeled by humans [40]. Most of the audio clips in the dataset are produced by physical sound sources and different production mechanisms. The dataset is available at https://annotator.freesound.org/fsd/release/FSD50K/ [39].

***SONYC-UST*:** A dataset with 30,68 audio records captured using the **Sounds of New York City (SONYC)** sensor network [19]. All the audio samples are classified into 8 main classes and extended to 23 fine-grained classes. An improved version of SONYC-UST named *SONYC-UST-V2* [16] is also publicly available with 18,510 audio recordings captured from the same sensor network. This has been used for the development and evaluation of machine listening models for realistic urban noise monitoring [18]. The dataset can be retrieved from https://doi.org/10.5281/zenodo.3966543 [17].

FSC22 dataset: A dataset that contains 2,025 audio samples, each with a duration of 5 seconds, resulting in 2.81 hours of forest acoustics. The audios are retrieved from the FreeSound Platform and distributed among 27 subclasses. Each subclass contains 75 manually selected audio samples. The dataset is available at https://dx.doi.org/10.21227/40ds-0z76 [10].

Table 5 shows a comparison of the aforementioned publicly available ESC datasets. The summary of ESC-10 and SONYC-UST is not included, as they are direct subsets of ESC-50 and SONYC-UST-V2, respectively. FSC22, which is a recently released dataset, is not included because no studies have been published based on it [10]. Further, the FreeSound database and the BBC Sound Effects Library are not included, as they are platforms, where audio data can be retrieved to generate related datasets.

Accordingly, these datasets contain acoustic events covering different phenomena that can be observed naturally or artificially in forest ecosystems. However, they lack sounds like specific

Table 5. Overview of Public ESC Datasets

Dataset	Source	Total Clips	Clip Length	Classes	Used Classes	Related Studies
ESC-50 [96]	Free-sound	2,000	5s	50	22	[4, 6, 21, 25, 37, 43, 48, 53, 77, 78, 82, 83, 87, 91, 95, 102, 115, 130]
Urban-Sound8k [104]	Free-sound	8,732	4s or less	10	3	[2, 21, 25, 26, 31, 34, 43, 48, 52, 53, 57, 69, 77, 82, 83, 85, 87, 91, 94, 95, 102, 106, 111, 115, 120, 133, 139]
AudioSet [8]	YouTube	2M	10s	527	17	[3, 37, 43, 50, 61, 63, 67, 135]
FSD50K [39]	Free-sound	51,197	0.3–30s	200	13	[40]
SONYC-UST-V2 [17]	SONYC acoustic	18,510	10s	23	9	[4, 16, 18]

animal sounds, engine sounds, and forest fires, which are of high significance. Thus, researchers have used datasets like BIRDZ [85, 86], Xeno-canto Archive [137, 138, 141], TUT Sound Events [21, 69, 135], and YouTube-100M [50] to generate combined datasets to better suit their classification requirements.

4.3 Challenges in Public Datasets

Generally, public datasets are generated considering a real-world application domain like ESC. These datasets contain large volumes of audio data points according to common taxonomies [103] while covering a broad scope of scenarios. However, these saturated datasets cannot be directly used for the training and validation of classifier models in specific scenarios such as forest acoustic monitoring and surveillance. Therefore, a significant amount of resources needs to be utilized to extract data from public datasets and to annotate the data points according to a suitable taxonomy.

Public datasets are mostly created using two types of audio clips: (1) clips recorded by microphones capturing the required scenario and (2) clips extracted from videos that contain the considered scenario captured by different types of devices. Inherently, these audio clips are clearer and without background noise when compared to the actual real-world scenarios. Further, the effect of aspects such as different weather conditions, the varying distance between non-stationary sound sources and the microphone, and different geographical factors are not properly implied in the publicly available datasets. Due to the preceding reasons, it becomes inconvenient to directly use public datasets for forest acoustic monitoring and surveillance. Different approaches such as synthetic data generation have been followed to incorporate the aforementioned factors into the dataset that are used for training and validation to obtain more accurate and efficient decisions.

4.4 Synthetic Datasets

Collecting dedicated audio samples of real-world scenarios is challenging due to the high resource requirement. Available public ESC datasets cover a broad spectrum and are hard to use for a given sound classification application. For instance, the ESC-50 or UrbanSound8k dataset cannot be directly used for forest sound monitoring and surveillance. As a solution, synthetic sound generation techniques are used to create sound clips with sufficient realisticity.

Among several studies, Mun et al. [80] have proposed a **Generative Adversarial Network (GAN)**-based technique to synthetically develop larger datasets. They have used GANs for each class and iteratively generate new data points for each class. In addition, they have utilized a **Support Vector Machine (SVM)** hyperplane for each class to select only the suitable samples to

be merged into the original dataset. The overall accuracy of both the classification models has obtained significant improvements with the combined dataset when compared to the results obtained with the original dataset. In addition, GAN-based techniques are used to overcome the data scarcity issue [69].

Moreover, Elliott et al. [37] have used **Vector-Quantized Variational Autoencoders (VQ-VAE)** to generate synthetic data for ESC. The authors have experimented with the VQ-VAE over the ESC-50 dataset and compared it with the other optimization techniques such as curve tokenization and amplitude reshaping. They have shown that VQ-VAE optimization achieves the lowest performances due to the heterogeneity of the ESC-50 dataset. Furthermore, a benchmark on the sound event detection system [122] has used the DESED (Domestic Environment Sound Event Detection) synthetic soundscape evaluation set. It consists of audio clips of 10 seconds, which were created from the Scraper [105] library. Scraper automatically mixes a selected set of foreground and background audios randomly with the user's requirement. Moreover, it controls the signal-to-noise ratio and controls several other key characteristics. These synthetic soundscapes are annotated automatically by Scraper. Serizel et al. [109] have shown that DESED relies on biases which restrict reaching generalization to real case conditions, even though the performance was improved [61].

5 SOUND CLASSIFICATION INFRASTRUCTURE

5.1 ML vs DL

Present ESC classification studies are mainly based on ML and DL. Generally, ML uses algorithms to learn data and make predictions without being explicitly programmed [81]. In contrast, DL uses a complex structure of algorithms modeled as of the human brain, to learn and classify the data. Most of the ML-based studies have used SVM and KNN for ESC with private datasets [46, 79]. Generally, ML algorithms endure the selectivity invariance problem, as they have limited ability to process data in their original format [24]. With the advancement of DL techniques, DL models such as CNN, LSTM, VGG, and MobileNet have been utilized for ESC [100]. However, by nature, DL techniques require large volumes of data to train models and generate accurate and efficient decisions [24]. Considering the comparative studies, Chen et al. [25] have compared the performance of ML and DL approaches for ESC. Based on their results, SVM showed an accuracy of 63.3%, whereas CNN and RNN showed maximum accuracies of 85.5% and 91.1%, respectively, with the UrbanSound8k dataset, indicating the superiority of DL algorithms over ML algorithms. We have analyzed more than 100 unique implementations based on the two major architectures (CNN and RNN) for ESC and identified the best-performing models against the most frequent datasets. Among CNN-based models, DenseNet has shown better performance [82, 83, 91]. Moreover, among RNN-based models, LSTM has provided better results considering the sequencing nature of audio data [31, 57].

Accordingly, ML and DL present unique advantages and disadvantages according to the domain in which they are deployed. Thus, both architectures have been extensively used in different studies that cater to unique requirements. Figure 6 shows the frequency of using ML models such as KNN, SVM, and the Bayesian network (BN in Figure 6), and DL models such as GRU, AlexNet, LSTM, VGG, ResNet, and DenseNet based on standard audio datasets such as AudioSet, DCASE (Detection and Classification of Acoustic Scenes and Events), ESC-10, ESC-50, and UrbanSound8k for ESC. It is observable that techniques such as SVM, AlexNet, ResNet, and DenseNet have been used widely for ESC in general. Since DL techniques support classifying complex audio data with implicit feature engineering, they outperform ML approaches [115]. Therefore, this survey focuses on studies based on DL techniques for forest surveillance through sound classification, and Section 6 describes them in detail.

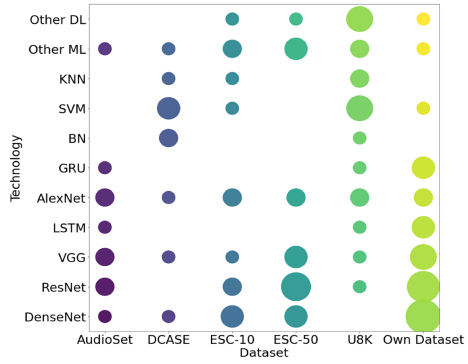

Fig. 6. Usage of ML and DL models against audio datasets.

5.2 Transfer Learning

A system's ability to use previously learned skills and knowledge to analyze a novel task can be defined as transfer learning [73, 92]. With this approach, a model pre-trained on a general dataset can be specialized to a considered domain by training the model again over a problem-specific, comparatively smaller dataset [55]. DL algorithms require large volumes of data, which is not available for many environmental scenarios, and the creation of such datasets requires more effort. Therefore, learning schemas such as transfer learning play a vital role in this domain [87, 115].

For instance, Mushtaq et al. [83] have addressed the usage of ResNet, DenseNet, SqueezeNet, AlexNet, and VGG, which use transfer learning to obtain the required accuracies and efficiencies. They have shown that a combination of ResNet and DenseNet generates the best performance when coupled with meaningful data augmentation techniques. In addition, a comparison among DL techniques such as ResNet, DenseNet, and ESResNet is presented by Nasiri and Hu [87]. They have modified the ResNet architecture and introduced a new classification model called *SoundCLR* using transfer learning. SoundCLR achieved 93.6%, 99.75%, and 88.01% accuracies for ESC-50, ESC-10, and UrbanSound8k, respectively. Guzhov et al. [48] have studied the effect of using pre-trained weights learned from the ImageNet dataset for sound classification based on ESC-50, ESC-10, and UrbanSound8k datasets. Their best-performing model—ESResNet-Attention—shows a significant improvement in accuracy when the pre-trained weights were used, revealing the importance of using transfer learning for ESC. Further, Bhat et al. [15] have proposed a new audio classification model called *YAMNet* based on MobileNetV1 architecture. YAMNet achieves a classification accuracy of 98.4% over a subset of the ESC-50 dataset.

Although transfer learning approaches present vital advantages to practically implementing a DL model, they inherit severe challenges. For instance, the dataset used to pre-train the model can contain certain biases and even backdoor data insertions to generate erroneous decisions [55]. Therefore, when using transfer learning for ESC systems, proper precautions need to be taken to guarantee the interpretability of the decisions.

5.3 Hardware Configurations

Selecting an optimal hardware specification for DL tasks is critical, as it affects the performance and accuracy of the application. However, this decision predominantly depends on the factors such as the task carried out, computational requirement, cost, power usage, and size of the data. The studies explored in this survey were analyzed to present different usages of hardware configurations including the CPU, the **Graphics Processing Unit (GPU)**, and resource-constrained edge devices. Table 6 states the studies on ESC that have used CPU and GPU-based platforms, or

Table 6. Hardware Configurations Used in the Selected Studies on ESC

Implementation Type	Related Studies
GPU	[4, 22, 31, 43, 50, 61, 62, 65, 67, 69, 82, 83, 85, 86, 91, 93, 94, 102, 115, 116, 120, 121]
CPU	[28, 29, 43, 101]
Resource-constrained edge devices	[6, 9, 15, 20, 21, 32, 36, 37, 51, 56, 78, 87, 89, 111, 123, 126, 130, 140]

resource-constrained edge devices. It can be seen that many sound processing studies have used GPUs as their implementation platform.

Among the considered studies, only four studies have mentioned the usage of CPU for different classification tasks. For instance, the work presented by Copiaco et al. [29] has used an Intel Core i7-9850H CPU with a 2.60-GHz processor only for the feature extraction task but not for the model training. Gazneli et al. [43] have presented a comparative study on model inference time on the NVIDIA Tesla V100 GPU and Intel Xeon CPU. They have shown promising performance on the GPU compared to the CPU. Hence, we can see that all the processors are not compatible with handling a given DL task. Generally, CPUs are reasonable for inferencing tasks but not for data-intensive computation tasks. Hence, most of the researchers have utilized GPUs for DL-based sound classification. GPUs are more appropriate for DL tasks, as they can take the advantage of parallel processing that provides high-speed processing capability and improves productivity. From the identified literature, NVIDIA is the widely used GPU for DL tasks. A variety of NVIDIA GPUs are available with different specifications such as core count, clock speed, memory size, and power consumption. In the existing literature, different types of GPU usages were identified, including NVIDIA GeForce GTX 1080 [65, 83, 85, 86, 102], NVIDIA GeForce RTX 2080 [37, 91, 120], NVIDIA Tesla V100 [37, 43, 69], NVIDIA GeForce 2070 SUPER [22], and NVIDIA Tesla K40 [116]. Nevertheless, the hardware selection mainly depends on the requirement and the budget.

Furthermore, deploying DL models in resource-constrained edge platforms is an evolving approach due to the high interest in edge computing with IoT. Most of the time, researchers and developers use edge-based implementations to deploy models in resource-constrained environments. We have explored 19 such studies, as stated in Table 6. Section 5.4 discusses the available edge computing platforms which are experimented with in the prior studies. DL-based edge computing is challenging, and techniques to optimize the performance in these systems are further discussed in Section 6.3.

5.4 Resource-Constrained Edge Devices

Edge intelligence is an evolving research area that enables real-time data analysis with the execution of ML or DL algorithms. It has extended the potential of the IoT by bringing computing services close to the physical location of the data sources. For the implementation of forest monitoring systems, significant efforts have already been invested in IoT combined with cloud computing [70, 79, 132]. Although it could be a straightforward solution, from a practical perspective, cloud processing will not be appropriate for every circumstance. It comes with some bottlenecks including latency in critical data transferring, higher communication costs, reduced bandwidth, privacy issues, and reliability problems [20, 75, 78]. Hence, incorporating intelligence on IoT edge devices, which has been the current trend, will reduce the bottlenecks to some extent. Unfortunately. IoT devices have limited computing power and less memory, thus deploying an efficient DL model in an edge device could be challenging. Therefore, an optimal choice of

Table 7. Usage of Hardware Components in DL-Based Edge Computing

Paper	Year	Application	Edge Device	Hardware Specification	Model	Acc	Dataset	Remarks
[78]	2023	Environment sound classification	Sony Sprensense (ARM Cortex M4)	192 kHz, 1.5MB SRAM	ACDNet	81.5%	ESC-50	High performance, low power, high-quality audio input
[51]	2022	Bird species recognition	NVIDIA Jetson Nano	4 GB of RAM, 128-core GPU	Efficient Net B3	95%	Xeno-canto archive	High performance, fast inference time
[32]	2022	Speech background sound classification	NVIDIA Jetson AGX Xavier	32 GB of RAM, 512-core Volta GPU	CNN	95.2%	Noisy speech dataset	High performance, fast inference time
[37]	2021	Office sound detection	Arduino Nano 33 BLE Sense	256 kB of SRAM, 64 MHz	BERT Trans-former	-	Office sounds dataset	Less power and memory consumption at the edge
[6]	2021	Illegal logging monitoring	32-bit ARM Cortex MF4 chipset	256 kB of SRAM., 1 MB of flash memory, 64 MHz	CNN	85%	ESC-50	Low power consumption, better memory management, better performance
[126]	2021	Environment sound recognition	MZ7035FA (Xilinx Zynq 7035 ARM+FPGA)	1 GB of DDR3, 256 Mbit of flash memory	CNN	88.3%	UrbanSound8k	High cost, better performance
[140]	2021	Bat species identification	Google Coral TPU (quad Cortex-A53, Cortex-M4F)	4 GB of RAM, Google Edge TPU coprocessor	CNN	97.3%	Custom dataset	High performance, high cost
[123]	2021	Urban sound recognition	Zynq Z-7020 FPGA (ARM Cortex-A9)	650 MHz, 512 MB of DDR3, 128 Mbit of flash memory	1D CNN	75.2%	ESC-10	Low cost with a limited amount of resources
[130]	2021	Office sound classification	Raspberry Pi Zero (ARMv6 CPU)	1 GHz, 512 MB of RAM	BERT Trans-former	81.2%	Custom dataset	Low cost and low power consumption
[21]	2020	Environment sound classification	STM3276RG Nucleo (ARM Cortex M4)	34.3 kB of RAM usage, 80 MHz	VGGish	68%	UrbanSound8k	Efficient processing capability and better power management
[111]	2019	Urban sound monitoring	Raspberry Pi 4 (quad-core Cortex-A72)	1.5 GHz, 4 GB of RAM	2D CNN	95%	UrbanSound8k	High accuracy, high cost, high power consumption,
[56]	2017	Illegal logging detection	Raspberry Pi 3 Model B (quad-core ARM Cortex-A53)	1.2 GHz, 1 GB of RAM	CNN	92%	Custom dataset	High cost, high power consumption, better performance

hardware needs to be decided based on model accuracy, throughput, implementation cost, and power consumption.

With the recent advancements in embedded technology, several computationally powerful hardware components specialized in handling ML tasks have been introduced. For deploying a complex algorithm on an embedded platform, DL developers had to deal with the proper choice of hardware that fits the model design and memory constraints [75]. From the performance and accuracy perspective, cloud-tested models tend to deviate when they are deployed on an edge device. Thus, to achieve better performances, an efficient algorithm has to be coupled with an optimal hardware choice. There are several hardware devices specialized for ML tasks, including **Microcontroller Units (MCUs)**, Google's Edge TPU (Tensor Processing Unit), NVIDIA's Jetson, and **Field-Programmable Gate Arrays (FPGAs)** [117]. Analyzing related work helps to understand the performance of different models in diverse processing platforms and workarounds for increasing performance. Table 7 includes the choice of hardware for implementing IoT devices that enabled edge computing and the model behavior in different platforms.

As reported in Table 7, general-purpose microcontrollers usually come with limited computational resources such as small RAM sizes and slow clock speeds. Consequently, classification models with sufficient accuracy are often too large for the edge device. The basic specifications of

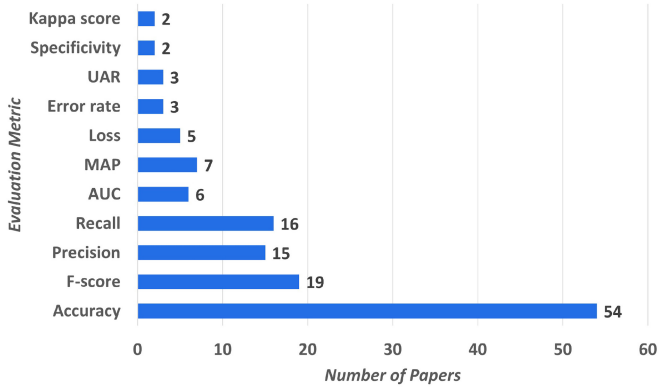

Fig. 7. Usage of evaluation metrics in ESC.

some common energy-efficient MCUs like STM3276RG have 32-bit ARM Cortex M4F CPUs and typically run on clock speeds below 200 MHz. Hence, deploying DL models on such MCUs will require extensive model optimization techniques. In the preceding studies, they have experimented with different model compression techniques like pruning [78], quantization [6, 21, 37, 78], and knowledge distillation [21, 78]. The goal of utilizing those techniques is to improve energy efficiency by reducing computing complexity, data volume, and hardware resources required during the execution of the networks [71]. Moreover, Raspberry Pi is a powerful MCU with a variety of models capable of running complex tasks. It shows significant performance and high accuracy in DL tasks when compared with the other MCUs. But the limitation is that it is relatively expensive and also consumes high power. In addition to the reported MCUs, FPGAs are power-efficient edge devices, highly suitable for computationally intensive algorithms like CNNs [123]. With the technical advancement, high-powered specialized edge devices with GPU support are introduced, addressing the scarcity of computational capability at the edge. For instance, Google's Coral TPU and NVIDIA's Jetson are powerful edge devices that are capable of delivering high accuracy while achieving high performance. They also offer fast inferencing time for DL algorithms with the tradeoffs of high-power consumption and inflated cost. Hence, to achieve the maximum benefits, a proper choice of hardware should be made according to the requirement.

5.5 Evaluation Metrics

Performance analysis of a classification system plays a vital role and supports fine-tuning the model. Generally, a confusion matrix provides a summary of prediction results on a classification problem [73]. The number of correct and incorrect predictions are summarized with count values and split by each class. The metrics true positive, true negative, false positive, and false negative are used when evaluating the model [118]. These parameters are used to derive the complex evaluation metrics. Accuracy [53, 68, 70, 79, 95] is the most used metric among other performance measures including the F1-score [38, 64, 74, 109, 132], precision [22, 34, 58, 122, 137], recall [36, 69, 82, 130, 141], **Area Under the Curve (AUC)** [25, 50, 61, 93, 135], and **Mean Average Precision (MAP)** [43, 63, 65, 67, 127].

Moreover, the error rate [58, 83], unweighted average recall [34, 59], weighted average recall [117], specificity [22, 132], true-positive rate [46], identification rate [107], Matthew's correlation coefficient [82], false discovery rate [82], Fowlkes-Mallows index [82], miss rate [82], kappa score [69, 82], mean reciprocal rank [65], dPrime [50], and Youden's index, likelihoods, and discriminant power [118] can be identified in the literature. Figure 7 shows the number of related

studies that have utilized a given evaluation metric for the top frequent metrics identified from the considered studies.

6 DL MODELS FOR AUDIO CLASSIFICATION

DL gives birth to many state-of-the-art audio processing systems [77, 100]. This section explores ESC studies implemented based on CNN or RNN, which are the main architectures used in DL. We also describe the studies that are specifically developed for resource-constrained edge environments. We represented the accuracy of each study with a gradient-based color code, which provides additional depth and nuance to the representation. This visual differentiation aids in understanding the performance obtained by the related studies and identifying the outliers. Therefore, by leveraging gradient-based color codes with varying levels of intensity, we can interpret the significance of the related studies.

6.1 Related Work on CNN-Based Models

The CNN is a major DL architecture utilized in different applications to derive classification decisions [29]. Usually, CNN takes input and assigns meaningful and learnable weights to develop the final classifier model [73, 108]. CNN is generally developed with multiple convolution layers, interleaved with pooling layers, followed by a dense layer [100]. Table 8 and Table 9 present state-of-the-art approaches for ESC using the existing CNN models and novel CNN models, respectively.

Studies stated in Table 8 and Table 9 have considered sounds observable in an environment in general. It can be observed that feature extraction techniques such as LM [26, 61, 91], MFCC [50, 52, 57, 91], **Short-Time Fourier Transform (STFT)** [48, 135, 141] are widely used, whereas some studies focus on combined feature extraction to obtain higher accuracies [52, 141]. Most of the studies have used PS [26, 48, 91], TS [26, 48, 91], noise addition [52], mixup [26], FFT [57, 135], and framing and overlap [57] for data augmentation to generate higher accuracies. Therefore, it can be seen that many studies have addressed ESC with CNN-based models together with data pre-processing techniques.

Among several related studies, a comparison of the effectiveness of using Xception, MobileNetV2, and DenseNet for ESC against a baseline CNN model is presented by Chhikara [26]. The author has discussed usage of the cyclic learning rate and the Adam optimizer with decoupled weight decay to optimize the model performance. The Xception model achieves the highest overall accuracy of 81% with the UrbanSound8k dataset, whereas the baseline CNN model, DenseNet, and MobileNetV2 achieved 79%, 75.9%, and 73.3%, respectively. Further, the study emphasized that these models achieved average accuracies due to data overfitting.

In another study, Palanisamy et al. [91] have addressed the use of Inception, ResNet, and DenseNet to identify acoustic phenomena in the environment using GTZAN, ESC-50, and UrbanSound8k datasets. They have shown that the CNN models with pre-trained weights performed significantly over CNN models with random weights. In addition, the authors have developed an ensemble based on the three DL models. Except for the pre-trained single DenseNet with the GTZAN dataset, all other combinations have shown that the ensemble models increase the accuracy by a significant margin. Finally, the authors have compared the accuracy of DenseNet with random weights, pre-trained weights, and pre-trained ensembles against the state-of-the-art classifiers. The results show that the pre-trained ensemble version of DenseNet outperforms other state-of-the-art models for the datasets ESC-50 and UrbanSound8k while achieving reasonable accuracies for the GTZAN dataset.

Moreover, Nanni et al. [85] have presented a deep and empirical analysis of the usage of CNN in ESC using the datasets BIRDZ, CAT, and ESC-50. Additionally, they have shown the effect of different data augmentation and feature extraction methodologies to achieve high performance. In

Table 8. Overview of Studies Using Existing CNN Models for ESC

Study	Year	Dataset	Model	Accuracy	Speciality
[26]	2021	UrbanSound8k	Xception	0.81	Endured data overfitting, thus showed average performance
			DenseNet	0.75	
			MobileNetV2	0.73	
[115]	2021	ESC-10	ResNet	0.78	Showed the effectiveness of self-supervised transfer learning for audio classification
		UrbanSound8k		0.76	
[83]	2021	ESC-50	DenseNet	0.97	Presented a comparative study on audio data augmentation
			ResNet	0.96	
			VGG	0.96	
			AlexNet	0.88	
		UrbanSound8k	ResNet	0.99	
			DenseNet	0.99	
			VGG	0.99	
			AlexNet	0.93	
[48]	2021	ESC-10	ESResNet(Pre-trained)	0.97	Presented the importance of multiple input channels, the evaluation procedure, and the usage of pre-trained weights from ImageNet
			ESResNet(Pre-trained)	0.97	
			ESResNet	0.94	
		ESC-50	ESResNet(Pre-trained)	0.91	
			ESResNet	0.83	
		UrbanSound8k	ESResNet(Pre-trained)	0.85	
			ESResNet	0.82	
[121]	2021	ESC-10	ResNet	0.91	Proposed a self-supervised learning-based classifier for ESC
			AlexNet	0.78	
			GoogleNet	0.63	
[82]	2020	ESC-10	DenseNet	0.99	Introduced novel audio feature extraction techniques and augmentation methods
		UrbanSound8k		0.97	
[52]	2020	UrbanSound8k	DCASE-2016	0.85	Also proposed a novel CNN model, D-2-DenseNet
		DenseNet		0.84	
[127]	2020	FSDKaggle2018	ResNet	0.93	Proposed a simple and effective augmentation method named *Mixed Frequency Masking*, which is not sensitive to parameters
[91]	2020	ESC-50	DenseNet	0.92	Used transfer learning and ensemble models; integrated gradients are used to realize the learning of the spectrogram shapes by the CNN
		GTZAN		0.90	
		UrbanSound8k		0.87	
[2]	2019	UrbanSound8k	GoogleNet	0.93	Presented a comparison between 1D CNN and 2D CNN
			VGG	0.70	
[61]	2018	AudioSet	VGG (CNN8)	0.89	Showed the effectiveness of CNN models for all DCASE tasks; task 2 results are shown here
			AlexNet (CNN4)	0.85	
[141]	2018	Private dataset	MobileNet	0.70	Showed the importance of noise reduction in sound pre-processing and the usage of different color maps for spectrogram images
[50]	2017	YouTube-100M	ResNet-50	0.92	Used transfer learning to experiment with large-scale datasets; showed that regularization can reduce the gap between the models trained on smaller datasets
			Inception V3	0.91	
			VGG	0.91	
			AlexNet	0.89	

the study, a comparison is performed among AlexNet, GoogleNet, VGG, ResNet, and Inception, and the authors have used the sum rule to create ensemble models. For each considered combination of datasets, feature extraction methodologies, and data augmentation techniques, ensemble versions outperformed by a significant margin.

Table 9. Overview of Studies with Novel CNN Models for ESC

Study	Year	Dataset	Model	Accuracy	Speciality
[102]	2021	UrbanSound8k	CFClean	0.94	Showed the importance of envelope function, segmentation, and normalization; discussed the inefficiency of regularization for the problem
			CF	0.85	
		ESC-50	CFClean	0.87	
			CF	0.45	
[85]	2021	ESC-50	FusionGlobal	0.88	Discussed the effectiveness of using ensembles of CNNs with data augmentation
			FusionGlobal-CO	0.88	
		BIRDZ	FusionGlobal-CO	0.97	
			FusionGlobal	0.96	
[69]	2021	UrbanSound8k	CNN based	0.98	Used conditional augmentation while training
		ESC-10	CNN based	0.96	
		TUT Acoustic	CNN based	0.73	
[83]	2021	ESC-50	CNN7	0.96	Presented a comparative study on audio data augmentation
			CNN9	0.95	
		UrbanSound8k	CNN7	0.95	
			CNN9	0.87	
[111]	2019	UrbanSound8k	2D CNN	0.95	ESC in real time on an embedded system; developed a fusion method with normalization and augmentation
[2]	2019	UrbanSound8k	1D CNN	0.89	Compared 1D CNN and 2D CNN

6.2 Related Work on RNN-Based Models

RNN models the time-bound dependencies in input data and thus can overcome the limited context size of CNN [73, 100]. In addition, vanishing or exploding gradients can be identified as major issues in RNNs, and many alterations to RNNs have been introduced to resolve this issue. LSTM [47], which uses a gating mechanism and memory cells to solve the gradient issue, can be identified as an improvement under RNN. LSTM has been effectively employed in ESC. Table 10 summarizes the related studies on RNN-based models. Similar to CNN-based models, these studies have used MFCC [57, 128] and Mel spectrogram [135] and combined extractors [25, 31] for feature extraction.

Among several studies, Wu and Lee [135] have implemented an LSTM architecture with three LSTM layers each having 2,048 units for classifying audio files available in the AudioSet [57] database, which produced a classification accuracy of 86.6%. Moreover, Das et al. [31] have compared the usage of different feature extraction methodologies with CNN and LSTM to generate classifications based on the UrbanSound8k dataset. They have shown that LSTM with MFCC and Chroma STFT produces a state-of-the-art accuracy of 98.81%. In another study, Banuroopaa et al. [57] have used two stacked LSTM layers accompanied by two time distributed layers, a flattening layer, and a dense layer to produce classification decisions based on the UrbanSound8k dataset. They have used MFCC for feature extraction and enhanced the LSTM model with different activation functions such as Softmax to generate an overall accuracy of 98.8%.

Moreover, the GRU architecture was introduced to capture dependencies of different time scales at each recurrent unit. GRU is similar to LSTM in many aspects, but it does not comprise separate memory cells [28]. For instance, GRU has two gates that are reset and updated, whereas LSTM has three gates (the input gate, output gate, and forget gate). GRU is less complex than LSTM

Table 10. Overview of Studies Based on RNN for ESC

Study	Year	Dataset	Model	Accuracy	Speciality
[57]	2022	UrbanSound8k	LSTM	0.98	Presented a novel audio fingerprinting method based on MFCC for sound classification
[31]	2020	UrbanSound8k	LSTM	0.98	Compared CNN and LSTM and showed that LSTM performs better
[94]	2020	UrbanSound8k	GRU-AWS	0.94	Analyzed the usage of the Attention Weight Similar (AWS) model for improvements
[25]	2019	UrbanSound8k	LSTM	0.91	Compared CNN, RNN, and SVM for audio classification with combined feature extraction
[133]	2019	UrbanSound8k	GRU	0.83	Introduced GRU for Audio Scene Classification (ASC); compared RNN and GRU
[135]	2018	AudioSet	GRU	0.87	Compared CNN, RNN, and MLP, and considered model complexity reduction
			LSTM	0.86	
[30]	2017	LITIS Rouen	DGRU	0.94	More experiments are needed to identify the performance of Deep GRU for tasks such as DCASE
			GRU	0.92	
			Baseline	0.91	
			LSTM	0.89	

because it has fewer gates. If the dataset is small, then GRU is preferred; otherwise, LSTM is preferred. GRU uses fewer training parameters and therefore uses less memory and executes faster than LSTM, whereas LSTM is more accurate on a larger dataset. The study by Wu and King [129] has presented a comparison between LSTM and GRU. They have shown that both models provide similar performance, but GRU requires only a few parameters. They have proposed a simplified version of LSTM, which outperforms both conventional LSTM and GRU architectures. Another similar study by Yang et al. [133] has used GRU for ESC with the UrbanSound8k dataset, where feature extraction is done by MFCC. Similarly, Peng et al. [94] have tested different feature extraction methods for audio tagging using GRU. They have shown that a combination of MFCC and LM performs better with GRU at an accuracy of 92%. Additionally, they have empirically shown that by introducing attention weight similar models, the accuracy can be further improved to 94.3%. Furthermore, a comparison between LSTM and GRU is presented by Dang and Tran [30]. GRU achieves an accuracy of 92.85% while outperforming LSTM with an accuracy of 89.04%. In their work, an improved version of deep GRU has shown an accuracy of 94.92% for the same audio tagging requirement over the LITIS Rouen dataset.

6.3 Edge-Based Reduced Complexity Models

Audio classification using DL can achieve high performance and accuracy when the models are deployed in resource-rich cloud environments. However, in real-world scenarios such as forest acoustic surveillance, it is practical to implement the classifier model at the edge device itself, due to the challenges in transferring audio data in real time to a cloud server with the low networking infrastructure available in such environments. This section explores related studies on edge-based classification models.

Although CNN models perform better when compared with RNN and **Multi-Layer Perceptron (MLP)** for the large-scale audio classification domain, it is not practical to use them in resource-constrained edge devices due to the layered complexity of CNN models. Several studies have proposed approaches to reduce such model complexities, as summarized in Table 11. These studies have used evaluation metrics including MAP, and AUC.

Among the studies, Gazneli et al. [43] have introduced two **End-to-End Audio Transformer (EAT)** models named *EAT-S* and *EAT-M*, where the complexities are small and medium,

Table 11. Comparison of ESC with Reduced Complexity Models

Study	Year	Model	Param. in Mil.	Metric	ESC-10	ESC-50	UrbanSound8k	AudioSet
[43]	2022	EAT-S	5.3	MAP	-	0.95	0.88	0.40
		EAT-M	25.5		-	0.96	0.90	0.42
[77]	2022	ACDNet (Pruning)	0.13	Acc.	0.92	0.75	0.70	
		AclNet (Pruning)	0.13		0.90	0.72	0.67	-
		ACDNet (XNOR-NET)	1.05		0.82	0.31	-	-
		AclNet (XNOR-NET)	1.05		0.80	0.31	-	-
		ACDNet (Baseline)	4.74		0.96	0.87	0.84	-
		AclNet (XNOR-NET)	10.63		0.95	0.85	0.79	-
[135]	2018	AlexNet-BN with GAP	2.59	AUC	-	-	-	0.91
		AlexNet-BN with 64 FC	3.07		-	-	-	0.84
		ResNet-50	24.58		-	-	-	0.91
		AlexNet	56.09		-	-	-	0.89
		AlexNet-BN	56.11		-	-	-	0.92

respectively. The study has compared the performance of both the models for three datasets—UrbanSound8k, AudioSet, and ESC-50—against other state-of-the-art models. With larger network sizes, EAT-M performs better with a close margin to the EAT-S model. Due to the high accuracies achieved for EAT-S models that resemble the MobileNet architecture, it becomes a viable candidate for edge deployment in resource-constricted environments. Further, they have analyzed different data augmentation approaches to enhance the functionality of the proposed architecture.

Moreover, pruning and XNOR-NET techniques have been used to reduce the model complexity of ACDNet and AclNet, which can be deployed in resource-constrainted edge devices [77]. They have shown that derivations can be made from ACDNet and AclNet, which produce comparable accuracies with suitable size parameters to be deployed in an edge environment, by using XNOR-NET. However, when the number of classes considered is increased, the accuracy of the XNOR-NET version of the baseline models takes a significant drop. The derivations made using pruning combined with proper quantization techniques also satisfy the size requirement for edge, and their drop of accuracy was significantly low compared to that of XNOR-NET versions.

Another study by Wu and Lee [135] has presented a model that reduces the complexity using global average pooling and BN (bottleneck) layers as AlexNet-BN. Their model achieved an accuracy of 92.7% with 56.11 million parameters without any optimizations, which is closely followed by ResNet-50 at an accuracy of 91.4% with 24.58 million parameters. Additionally, they have shown that the parameters of AlexNet-BN can be reduced with global average pooling by a 1/22 factor, which results in an accuracy of 91.6% at 2.59 million parameters. Further, they have validated the performance of the proposed architecture against the TUT Acoustic Scenes 2016 database, where AlexNet-BN, AlexNet-BN with global average pooling, LSTM and three-layer MLP registers have shown the accuracy of 87.4%, 85.9%, 82.8%, and 78.2%, respectively.

Among other studies on edge-based ESC, CNN approaches discussed by Copiaco et al. [29] have considered domestic sound classification using a private dataset. They have compared different parameters of AlexNet, ResNet, and Xception and proposed an improved version of AlexNet called *MAlexNet-33*. This model has outperformed edge-suitable neural networks such as MobileNetV2, SqueezeNet, NasNet, and ShuffleNet from the F1-score while achieving the lowest average execution time. From another point of view, a BERT (Bidirectional Encoder Representations from Transformers)-based transformer proposed by Elliott et al. [37] has outperformed a CNN using

99.85% fewer parameters, which can execute on a microcontroller. Another specially designed network to identify urban and noise sounds to be used in hearing aids is proposed by Ting et al. [120]. The proposed model has a significantly low number of parameters and can be easily deployed in edge cases. With a reduced number of parameters, the UrbanSound8k and HANS datasets have shown an accuracy of 83.3% and 75.27%, respectively, using an inception block with dense connectivity. Furthermore, an approach to identifying tree-cutting sound events using edge devices with low power and a memory-constrained setting is presented by Andreadis et al. [6]. They have shown an accuracy of 85% using a CNN-based model with the ESC-50 dataset.

7 DISCUSSION

7.1 Study Contribution and Lessons Learned

In this survey, we have explored the trends of recent studies on DL-based ESC at the edge. With a discussion of the main sound processing approaches, we presented a comprehensive study on sound pre-processing techniques including normalization, augmentation, and feature extraction of sound data. Moreover, this study reviewed publicly available environmental sound datasets with their limitations and emphasized the requirement for a real-world benchmark dataset. Importantly, we explored the literature that has applied DL techniques in ESC and analyzed their advantages and disadvantages. Therefore, this survey directs the researchers and developers in ESC to identify the recommended approaches, open challenges, and future research directions mainly in forest acoustic surveillance applications.

Accordingly, the practical implementation of forest monitoring systems that are based on acoustic analysis introduces unique challenges according to the architecture of the system. Initially, a suitable audio dataset for ESC should be identified. If the already existing datasets are not sufficient or not suitable, audio samples need to be recorded or synthetically generated with the available resources [37, 80, 122]. The next challenge would be the selection of the model architecture to develop the classifier. Generally, the domain knowledge and the level of feature engineering expertise required for building an ML model are comparatively high compared to the DL approach [55]. Following the DL approach, a variety of state-of-the-art models that are based on CNN and RNN can be used.

The next process would be deciding whether to use transfer learning or complete end-to-end model development according to the time and resource availability. Importantly, different optimization techniques to fine-tune the classifier and its deployment in the real world need to be explored. Finally, considering the application domain, the model can be deployed in resource-rich cloud environments with the support of configured GPUs or in an edge device under constrained power and computational resources. The aforementioned process of design, development, and deployment of an acoustic surveillance solution is presented in Figure 8, where the associated state-of-the-art techniques and methodologies governing all of the preceding aspects are discussed in this survey. This can be referred to as guidance by researchers and developers when making decisions on the development of ESC solutions.

Furthermore, Table 12 summarizes the models used and classification accuracies obtained using main public datasets for ESC. Here the terms CEL, SCL, and HL denote cross-entropy loss, supervised contrastive loss, and hybrid loss, respectively. It can be seen that some studies have obtained different accuracy for the same dataset by applying the same DL model. The main reason for these differences could be the use of rich pre-processing techniques. Consider the classifiers presented by Mushtaq and Su [82], Mushtaq et al. [83], and Nasiri and Hu [87], which are based on DenseNet, VGG, and ResNet, respectively. These studies outperform the results obtained by Chhikara [26], Abdoli et al. [2], and Nanni et al. [85], respectively, for the same model and dataset by a

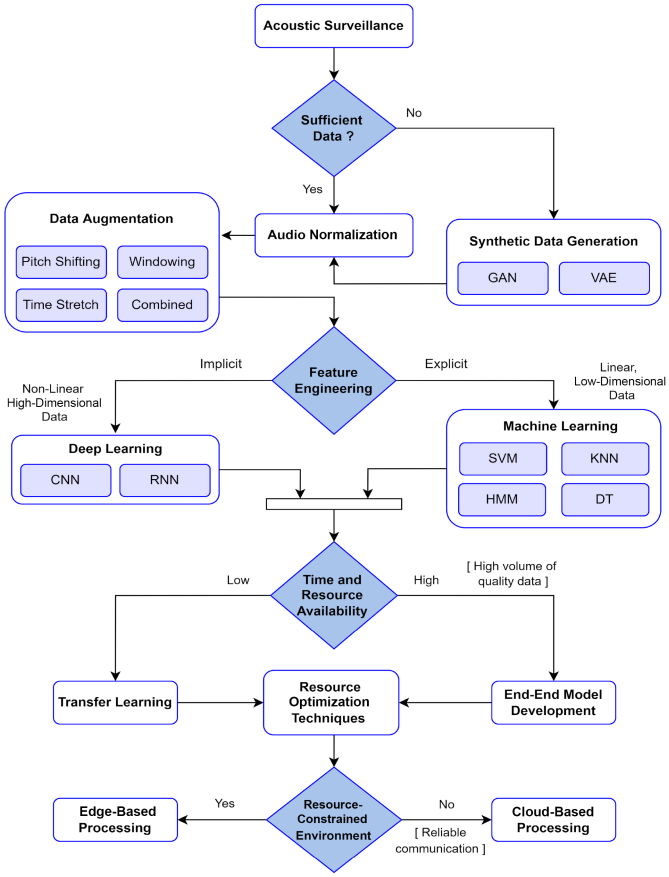

Fig. 8. Recommended techniques to use in an acoustic monitoring system.

significant margin. For instance, considering the results of the ESC with UrbanSound8k dataset using DenseNet, the obtained highest accuracy is 97.18% [82] and the lowest accuracy is 75.9% [26]. Therefore, it is interesting to analyze the reasons for the differences in the results.

Accordingly, Mushtaq and Su [82] have introduced two novel feature aggregation techniques and employ fine-tuning techniques over the transfer learning based models. Moreover, the studies by Mushtaq and Su [82] and Mushtaq et al. [83] have used novel data augmentation techniques to achieve significantly higher classification accuracies. In addition, the work presented by Nasiri and Hu [87] has used a hybrid loss function to train DL models instead of singly relying on the cross-entropy loss function or contrastive loss function. Therefore, the usage of such advanced techniques for feature extraction, data augmentation, and model training aspects of an ESC system can be identified as the foremost reason for state-of-the-art classifiers [82, 83, 87] to outperform the traditionally developed classifiers [2, 26, 85] by a significant margin.

7.2 Open Challenges and Future Research Directions

7.2.1 Requirement of a Suitable Dataset. A standard benchmark dataset is a fundamental requirement when building a classification system, and the same specificity is required to be implied in the considered dataset. Considering the domain of ESC, standardized datasets with proper

Table 12. Approaches Used with Main Public Datasets for ESC

ESC-50			UrbanSound8k			UrbanSound8k		
Study	**CNN Based**	**ACC**	**Study**	**CNN Based**	**ACC**	**Study**	**RNN Based**	**ACC**
[82]	DenseNet	0.98	[82]	DenseNet	0.97	[57]	LSTM	0.98
	AlexNet	0.88		AlexNet	0.93			
	ResNet	0.96		ResNet	0.99			
	VGG	0.96		VGG	0.99			
[83]	DenseNet	0.97	[83]	ResNet	0.99	[139]	APNet	0.65
	ResNet	0.96		DenseNet	0.99		APNet (R prot.)	0.68
	VGG	0.96		VGG	0.99		APNet (R. ch.)	0.69
	CNN7	0.96		CNN7	0.95		OpenL3	0.67
	CNN9	0.95		AlexNet	0.93			
	AlexNet	0.88		CNN9	0.87			
[91]	DenseNet	0.92	[91]	DenseNet	0.87	[94]	GRU-AWS	0.94
[87]	ResNet-50 (CEL)	0.92	[87]	ResNet-50 (CEL)	0.91			
	ResNet-50 (SCL)	0.92		ResNet-50 (SCL)	0.86			
	ResNet-50 (HL)	0.93		ResNet-50 (HL)	0.86			
	DenseNet	0.91		DenseNet	0.85			
[4]	CNN based	0.89	[69]	CNN based	0.98	[133]	GRU	0.83
[85]	FusionGlobal	0.88	[111]	2D CNN	0.95	[21]	GRU	0.68
	FusionGlobal-CO	0.88						
[102]	CFClean	0.87	[102]	CFClean	0.94			
	CF	0.45		CF	0.85			
[78]	ACDNet	0.87	[2]	GoogleNet	0.93			
				1D CNN	0.89			
				VGG	0.70			
[77]	ACDNet	0.87	[77]	ACDNet	0.84			
[6]	CNN based	0.85	[26]	Xception	0.81			
				MobileNetV2	0.73			
				DenseNet	0.75			
[43]	ResNet	0.83	[43]	ResNet	0.82			
[37]	Transformer based	0.81	[52]	DenseNet	0.84			
[115]	ResNet	0.78	[115]	ResNet	0.76			
[53]	Conv-3	0.54	[53]	Conv-3	0.74			
	Conv-5	0.50		Conv-5	0.69			
[95]	CNN based	0.44	[95]	CNN based	0.78			
			[120]	ResNet	0.73			
				Inception	0.75			
				DenseNet	0.76			
Study	**Edge Based**	**ACC**	**Study**	**Edge Based**	**ACC**			
[43]	EAT-S	0.95	[43]	EAT-S	0.88			
	EAT-M	0.96		EAT-M	0.90			
[77]	ACDNet (Baseline)	0.87	[77]	ACDNet (Baseline)	0.84			
	ACDNet (Pruning)	0.75		ACDNet (Pruning)	0.70			
	AclNet (Baseline)	0.85		AclNet (Baseline)	0.79			
	AclNet (Pruning)	0.72		AclNet (Pruning)	0.67			

taxonomies are abundantly identified as described in Section 4.3. However, in forest acoustic monitoring and surveillance, such benchmark datasets with proper taxonomies are not publicly available. Most of the related studies have created their datasets using extracted data from the available ESC datasets and recorded audio samples capturing specific scenarios. Such datasets are specifically used for the considered study and do not generalize for other research scopes. Therefore, it is challenging to create an appropriate dataset for FSC, as the datasets need large

volumes of unbiased data, and all the data is required to be manually tagged by inspection. Hence, the requirement for a proper dataset with a common taxonomy for forest acoustics can be identified as an open challenge that requires future research considerations.

7.2.2 Demand for Resources and Power. Deploying complex classification algorithms on the edge has been impeded because edge-based sound classification systems need to satisfy requirements including computational capability, memory management, power consumption, and cost. Thus, strategies and techniques to implement sustainable and reliable systems which can satisfy the preceding non-functional requirements become important. In particular, sufficient computational power is critical to provide an accurate classification decision by processing the data in real time. Therefore, an optimal selection of hardware specifications should be decided by considering aspects such as computational capacity, accuracy, and throughput. Importantly, the model size directly affects the complexity of the algorithm [90]. Several strategies, such as knowledge distillation and quantization, have been suggested, which compress the model, allowing it to execute on a constrained platform [75]. However, it may cause a loss of accuracy in return. In addition, neural architecture search based implementations can be applied, which automatically find the optimal architectures with low losses [114]. Hence, in future investigations, more focus should be given to designing models with a reduced size while preserving performance and accuracy. Moreover, edge computing consumes considerably higher power, which is challenging to accommodate in a forest environment. Solar energy harvesting might not be the best option during rainy or snowy seasons. Consequently, future work is required to develop practical systems that can operate with minimal power and resource consumption.

7.2.3 Real-Time Notification System. At present, one of the fundamental requirements of a forest acoustic monitoring system is to identify and prevent illegal activities from being carried out in forest environments. Therefore, monitoring systems are required to provide accurate updates on such activities as early as possible, hence proper strategies need to be implemented to ensure real-time operation. When considering cloud-based approaches, the recorded audio data needs to be transferred to the cloud over a network [56]. This transmission needs to be done in real time without any losses, thus accurate and instant information can be given to the relevant authorities. Yet, data transmission becomes challenging due to the low infrastructure available in forest environments. However, when considering edge-based approaches, the classification decision is derived at the edge itself and only the alerts need to be transferred over a network to the relevant authorities [6]. Generally, alerts are not frequently generated. Even when alerts are generated, the size of the alerts is comparatively small compared to the size of audio frames that needs to be transferred over the network. Hence, ensuring real-time function in an edge-based approach can be comparatively less challenging. Yet, the amount of research conducted dedicated to ensuring this real-time operation is significantly low in the literature. Therefore, addressing real-time operations can be identified as a new research direction and an open challenge for the forest acoustic monitoring and surveillance domain.

7.2.4 Explainable DL Model. Explainable AI is an evolving area of research that focuses on producing explainable DL models that can be understandable by humans while maintaining a high level of learning performance. It helps improve the trust and manage the emerging generation of partners effectively. Interpretability of an explainable system indicates the extent to which a cause and effect can be observed within a system. The term *explainability* indicates the extent to which the internal mechanics of an ML or DL system can be explained in human terms. Using interpretability and explainability, the users can understand the impact of different inputs on the generated classification decisions. Since the forest acoustic observation systems are deployed in

forest reserves to monitor abnormal activities such as illegal tree logging and poaching, such applications need to function in real time with high interpretability and explainability. Currently, to the authors' knowledge, there are very limited state-of-the-art models for ESC that have addressed explainability considerations in their studies. As a notable effort, Zinemanas et al. [139] have presented a novel approach to ensure the interpretability of DL models named the **Audio Prototype Network (APNet)**. They have implemented a similarity measuring and weight assigning layer in their network to increase the interpretability of the proposed model. Thus, further research is required to saturate the interpretability and explainability aspects of forest acoustic monitoring systems.

8 CONCLUSION

Forest ecosystems play a vital role in the sustainable existence of life on earth. Acoustic surveillance systems support authorities to manage artificial and natural scenarios that cause unfavorable effects on such environments. This survey article explored the main approaches to developing FSC systems at the edge, including sound pre-processing techniques, available standard datasets, state-of-the-art DL models, and different hardware configurations and evaluation metrics. Finally, we provided recommended techniques for the ESC domain and discussed the challenges with future research directions, which will be helpful for researchers and developers of this domain.

REFERENCES

[1] Olusola O. Abayomi-Alli, Robertas Damaševičius, Atika Qazi, Mariam Adedoyin-Olowe, and Sanjay Misra. 2022. Data augmentation and deep learning methods in sound classification: A systematic review. *Electronics* 11, 22 (2022), 3795.

[2] Sajjad Abdoli, Patrick Cardinal, and Alessandro Lameiras Koerich. 2019. End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Systems with Applications* 136 (2019), 252–263.

[3] Jakob Abeßer. 2020. A review of deep learning based methods for acoustic scene classification. *Applied Sciences* 10, 6 (2020), 1–16.

[4] Sainath Adapa. 2019. Urban sound tagging using convolutional neural networks. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE'19)*. 5–9.

[5] Shalbbya Ali, Safdar Tanweer, Syed Sibtain Khalid, and Naseem Rao. 2021. Mel frequency cepstral coefficient: A review. In *Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development (ICIDSSD'21)*. 92–101.

[6] Alessandro Andreadis, Giovanni Giambene, and Riccardo Zambon. 2021. Monitoring illegal tree cutting through ultra-low-power smart IoT devices. *Sensors* 21, 22 (2021), 7593.

[7] Paolo Annesi, Roberto Basili, Raffaele Gitto, Alessandro Moschitti, and Riccardo Petitti. 2007. Audio feature engineering for automatic music genre classification. In *Proceedings of Large Scale Semantic Access to Content (Text, Image, Video, and Sound) (RIAO'07)*. 702–711.

[8] AudioSet. 2021. AudioSet—A Large-Scale Dataset of Manually Annotated Audio Events. Retrieved July 20, 2022 from https://research.google.com/audioset/download.html

[9] Miroslav Babiš, Maroš Ďuríček, Valéria Harvanová, and Martin Vojtko. 2011. Forest Guardian—Monitoring system for detecting logging activities based on sound recognition. *Researching Solutions in Artificial Intelligence, Computer Graphics and Multimedia, IIT.SRC* 2011 (2011), 1–6.

[10] Meelan Bandara, Roshinie Jayasundara, Isuru Ariyarathne, Dulani Meedeniya, and Charith Perera. 2023. Forest sound classification dataset: FSC22. *Sensors* 23 (2023), 1–22.

[11] Anam Bansal and Naresh Kumar Garg. 2022. Environmental sound classification: A descriptive review of the literature. *Intelligent Systems with Applications* 16 (2022), 200115.

[12] Bipendra Basnyat, Nirmalya Roy, Aryya Gangopadhyay, and Adrienne Raglin. 2022. Environmental sound classification for flood event detection. In *Proceedings of the 18th International Conference on Intelligent Environments*. IEEE, Los Alamitos, CA, 1–8.

[13] BBC. 2020. BBC Sound Effects. Retrieved July 20, 2022 from https://sound-effects.bbcrewind.co.uk

[14] Carol Bedoya, Claudia Isaza, Juan M. Daza, and José D. López. 2017. Automatic identification of rainfall in acoustic recordings. *Ecological Indicators* 75 (2017), 95–100.

[15] K. Manasvi Bhat, Manan Bhandari, ChangSeok Oh, Sujin Kim, and Jeeho Yoo. 2020. Transfer learning based automatic model creation tool for resource constraint devices. *arXiv abs/2012.10056* (2020).

[16] Mark Cartwright, Jason Cramer, Ana Elisa Méndez Méndez, Yu Wang, Ho-Hsiang Wu, Vincent Lostanlen, Magdalena Fuentes, Graham Dove, Charlie Mydlarz, Justin Salamon, Oded Nov, and Juan Pablo Bello. 2020. SONYC-UST-V2: An urban sound tagging dataset with spatiotemporal context. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE'20)*. 1–5.

[17] Mark Cartwright, Jason Cramer, Ana Elisa Mendez Mendez, Yu Wang, Ho-Hsiang Wu, Vincent Lostanlen, Magdalena Fuentes, Graham Dove, Charlie Mydlarz, Justin Salamon, Oded Nov, and Juan Pablo Bello. 2020. SONYC Urban Sound Tagging (SONYC-UST): A Multilabel Dataset from an Urban Acoustic Sensor Network. Retrieved July 20, 2022 from https://doi.org/10.5281/zenodo.3966543

[18] Mark Cartwright, Jason Cramer, Justin Salamon, and Juan Pablo Bello. 2019. Tricycle: Audio representation learning from sensor network data using self-supervision. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'19)*. IEEE, Los Alamitos, CA, 278–282.

[19] M. Cartwright, A. Mendez, J. Cramer, V. Lostanlen, G. Dove, H. Wu, J. Salamon, O. Nov, and J. Bello. 2019. SONYC Urban Sound Tagging (SONYC-UST): A multilabel dataset from an urban acoustic sensor network. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*. 35–39.

[20] Gianmarco Cerutti, Renzo Andri, Lukas Cavigelli, Elisabetta Farella, Michele Magno, and Luca Benini. 2020. Sound event detection with binary neural networks on tightly power-constrained IoT devices. In *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*. ACM, New York, NY, 19–24.

[21] Gianmarco Cerutti, Rahul Prasad, Alessio Brutti, and Elisabetta Farella. 2020. Compact recurrent neural networks for acoustic event detection on low-energy low-complexity platforms. *IEEE Journal of Selected Topics in Signal Processing* 14, 4 (2020), 654–664.

[22] C. Chalmers, P. Fergus, S. Wich, and S. N. Longmore. 2021. Modelling animal biodiversity using acoustic monitoring and deep learning. In *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN'21)*. IEEE, Los Alamitos, CA, 1–7.

[23] Chuan-Yu Chang and Yi-Ping Chang. 2013. Application of abnormal sound recognition system for indoor environment. In *Proceedings of the 9th International Conference on Information, Communications, and Signal Processing*. IEEE, Los Alamitos, CA, 1–5.

[24] Nitin Kumar Chauhan and Krishna Singh. 2018. A review on conventional machine learning vs deep learning. In *Proceedings of the International Conference on Computing, Power, and Communication Technologies*. IEEE, Los Alamitos, CA, 347–352.

[25] Chai Chen, Yuxuan Liu, Haoran Sun, and Moyan Zhou. 2019. Audio Feature Extraction and Classification for Urban Sound. Retrieved September 11, 2023 from https://github.com/yuxuan3713/ECE-228-Project

[26] Jasmine Chhikara. 2021. Transfer learning models based environment audio classification. *International Journal of Emerging Technologies in Engineering Research* 9 (2021), 1–8.

[27] Selina Chu, Shrikanth Narayanan, and C.-C. Jay Kuo. 2009. Environmental sound recognition with time–frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 6 (2009), 1142–1158.

[28] Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proceedings of the NIPS Workshop on Deep Learning*. 1–9.

[29] Abigail Copiaco, Christian Ritz, Nidhal Abdulaziz, and Stefano Fasciani. 2021. A study of features and deep neural network architectures and hyper-parameters for domestic audio classification. *Applied Sciences* 11, 11 (2021), 4880.

[30] Thi Thuy An Dang and Thi Kieu Tran. 2016. Audio scene classification using gated recurrent neural network. In *Proceedings of the Conference on Information Technology and Its Applications (CITA'16)*. IEEE, Los Alamitos, CA, 48–51.

[31] Joy Krishan Das, Ghosh Arka, Pal Abhijit Kumar, Dutta Sumit, and Chakrabarty Amitabha. 2020. Urban sound classification using convolutional neural network and long short term memory based on multiple features. In *Proceedings of the 2020 4th International Conference on Intelligent Computing in Data Sciences (ICDS'20)*. IEEE, Los Alamitos, CA, 1–9.

[32] Aveen Dayal, Sreenivasa Reddy Yeduri, Balu Harshavardan Koduru, Rahul Kumar Jaiswal, J. Soumya, M. B. Srinivas, Om Jee Pandey, and Linga Reddy Cenkeramaddi. 2022. Lightweight deep convolutional neural network for background sound classification in speech signals. *Journal of the Acoustical Society of America* 151, 4 (2022), 2773–2786.

[33] Jonathan Dennis, Huy Dat Tran, and Haizhou Li. 2011. Spectrogram image feature for sound event classification in mismatched conditions. *IEEE Signal Processing Letters* 18, 2 (2011), 130–133.

[34] Itxasne Diez Gaspon, Ibon Saratxaga, and Karmele Lopez de Ipiña. 2019. Deep learning for natural sound classification. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Vol. 259. Institute of Noise Control Engineering, Madrid, Spain, 5683–5692.

[35] Golnoosh Elhami and Romann M. Weber. 2019. Audio feature extraction with convolutional neural autoencoders with application to voice conversion. In *Infoscience EPFL Scientific Publications*. EPFL Scientific Publications, Lausanne, Switzerland, 1–5. http://infoscience.epfl.ch/record/261268

[36] David Elliott, Evan Martino, Carlos E. Otero, Anthony Smith, Adrian M. Peter, Benjamin Luchterhand, Eric Lam, and Steven Leung. 2020. Cyber-physical analytics: Environmental sound classification at the edge. In *Proceedings of the 2020 IEEE 6th World Forum on Internet of Things (WF-IoT'20)*. IEEE, Los Alamitos, CA, 1–6.

[37] David Elliott, Carlos E. Otero, Steven Wyatt, and Evan Martino. 2021. Tiny transformers for environmental sound classification at the edge. *arXiv abs/2103.12157* (2021).

[38] Marcelo Fernandes, Weverton Cordeiro, and Mariana Recamonde-Mendoza. 2021. Detecting *Aedes aegypti* mosquitoes through audio classification with convolutional neural networks. *Computers in Biology & Medicine* 129 (2021), 104152.

[39] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2020. FSD50K: An Open Dataset of Human-Labeled Sound Events. Retrieved July 20, 2022 from https://doi.org/10.5281/zenodo.4060432

[40] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 30, 24 (2022), 829–852.

[41] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andrés Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. 2017. Freesound datasets: A platform for the creation of open audio datasets. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR'17)*. 486–93.

[42] Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, New York, NY, 411–412.

[43] Avi Gazneli, Gadi Zimerman, T. Ridnik, Gilad Sharir, and Asaf Noy. 2022. End-to-end audio strikes back: Boosting augmentations towards an efficient audio classification network. *arXiv abs/2204.11479* (2022).

[44] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'17)*. 776–780.

[45] Marius Vasile Ghiurcau, Corneliu Rusu, Radu Ciprian Bilcu, and Jaakko Astola. 2012. Audio based solutions for detecting intruders in wild areas. *Signal Processing* 92, 3 (2012), 829–840.

[46] N'tcho Assoukpou Jean Gnamélé, Yelakan Berenger Ouattara, Tokpa Arsene Kobea, Geneviève Baudoin, and Jean-Marc Laheurte. 2019. KNN and SVM classification for chainsaw identification in the forest areas. *International Journal of Advanced Computer Science and Applications* 10, 12 (2019), 531–536.

[47] Alex Graves. 2012. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*. Studies in Computational Intelligence, Vol. 385. Springer, 37–45.

[48] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2021. ESResNet: Environmental sound classification based on visual domain models. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR'21)*. IEEE, Los Alamitos, CA, 4933–4940.

[49] Osman Günay, Kasım Taşdemir, B. Uğur Töreyin, and A. Enis Çetin. 2009. Video based wildfire detection at night. *Fire Safety Journal* 44, 6 (2009), 860–868.

[50] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN architectures for large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'17)*. IEEE, Los Alamitos, CA, 131–135.

[51] Jonas Höchst, Hicham Bellafkir, Patrick Lampe, Markus Vogelbacher, Markus Mühling, Daniel Schneider, Kim Lindner, Sascha Rösner, Dana G. Schabo, Nina Farwig, and Bernd Freisleben. 2022. Bird@Edge: Bird species recognition at the edge. In *Proceedings of the 10th International Conference on Networked Systems (NETYS'22)*. 69–86.

[52] Zilong Huang, Chen Liu, Hongbo Fei, Wei Li, Jinghu Yu, and Yi Cao. 2020. Urban sound classification based on 2-order dense convolutional network using dual features. *Applied Acoustics* 164 (2020), 107243.

[53] Muhammad Huzaifah. 2017. Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. *arXiv abs/1706.07156* (2017).

[54] Rakib Hyder, Shabnam Ghaffarzadegan, Zhe Feng, John Hansen, and Taufiq Hasan. 2017. Acoustic scene classification using a CNN-supervector system trained with auditory and spectrogram image features. In *Proceedings of Interspeech 2017*. 3073–3077.

[55] Christian Janiesch, Patrick Zschech, and Kai Heinrich. 2021. Machine learning and deep learning. *Electron Markets* 31, 31 (2021), 685–695.

[56] Gayan Kalhara, Vishan Jayasinghearachchi, Achala Dias, Vishwa Ratnayake, Chandimal Jayawardena, and Nuwan Kuruwitaarachchi. 2017. TreeSpirit: Illegal logging detection and alerting system using audio identification over an IoT network. In *Proceedings of the 2017 11th International Conference on Software, Knowledge, Information Management, and Applications (SKIMA'17)*. IEEE, Los Alamitos, CA, 1–7.

[57] Kalyanaswamy Banuroopa and Shanmuga Priyaa. 2022. MFCC based hybrid fingerprinting method for audio classification through LSTM. *International Journal of Nonlinear Analysis and Applications* 12 (2022), 2125–2136.

[58] Jaehun Kim, Kyoungin Noh, Jaeha Kim, and Joon-Hyuk Chang. 2018. Sound event detection based on beamformed convolutional neural network using multi-microphones. In *Proceedings of the 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC'18)*. IEEE, Los Alamitos, CA, 170–173.
[59] Tomoya Koike, Kun Qian, Qiuqiang Kong, Mark D. Plumbley, Björn W. Schuller, and Yoshiharu Yamamoto. 2020. Audio for audio is better? An investigation on transfer learning models for heart sound classification. In *Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, Los Alamitos, CA, 74–77.
[60] Stanislaw Komorowski Dariusz, Pietraszek. 2015. The use of continuous wavelet transform based on the fast Fourier transform in the analysis of multi-channel electrogastrography recordings. *Journal of Medical Systems* 40 (2015), 10.
[61] Qiuqiang Kong, Turab Iqbal, Yong Xu, Wenwu Wang, and Mark D. Plumbley. 2018. DCASE challenge surrey cross-task convolutional neural network baseline. In *Detection and Classification of Acoustic Scenes and Events*. DCASE.
[62] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* 60, 6 (2017), 84–90.
[63] Anurag Kumar and Vamsi Krishna Ithapu. 2020. A sequential self teaching approach for improving generalization in sound event recognition. In *Proceedings of the 37th International Conference on Machine Learning*. 5447–5457.
[64] Ianis Lallemand, Diemo Schwarz, and Thierry Artières. 2012. Content-based retrieval of environmental sounds by multiresolution analysis. In *Proceedings of Content-Based Retrieval of Environmental Sounds by Multiresolution Analysis (SMC'12)*.1.
[65] Mario Lasseck. 2018. Audio-based bird species identification with deep convolutional neural networks. In *Proceedings of the Working Notes of CLEF*. 2125.
[66] Iurii Lezhenin, Natalia Bogach, and Evgeny Pyshkin. 2019. Urban sound classification using long short-term memory neural network. In *Proceedings of the 14th Federated Conference on Computer Science and Information Systems*. IEEE, Los Alamitos, CA, 57–60.
[67] Juncheng Billy Li, Shuhui Qu, Po-Yao Huang, and Florian Metze. 2022. AudioTagging done right: 2nd comparison of deep learning methods for environmental sound classification. *arXiv abs/2203.13448* (2022).
[68] Ying Li and Zhibin Wu. 2015. Animal sound recognition based on double feature of spectrogram in real environment. In *Proceedings of the 2015 International Conference on Wireless Communications and Signal Processing (WCSP'15)*. IEEE, Los Alamitos, CA, 1–5.
[69] Aswathy Madhu and Suresh Kirthi Kumaraswamy. 2021. EnvGAN: Adversarial synthesis of environmental sounds for data augmentation. *arXiv abs/2104.07326* (2021).
[70] Alina-Elena Marcu, George Suciu, Elena Olteanu, Delia Miu, Alexandru Drosu, and Ioana Marcu. 2019. IoT system for forest monitoring. In *Proceedings of the 42nd International Conference on Telecommunications and Signal Processing (TSP'19)*. IEEE, Los Alamitos, CA, 629–632.
[71] Lucas Martin Wisniewski, Jean-Michel Bec, Guillaume Boguszewski, and Abdoulaye Gamatié. 2022. Hardware solutions for low-power smart edge computing. *Journal of Low Power Electronics and Applications* 12, 4 (2022), 61.
[72] Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt Mcvicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference*. 18–24.
[73] Dulani Meedeniya. 2023. *Deep Learning: A Beginners' Guide*. CRC Press, Boca Raton, FL. https://books.google.lk/books?id=PiirzwEACAAJ
[74] Aska Mehyadin, Adnan Mohsin Abdulazeez, Dathar Abas Hasan, and Jwan Saeed. 2021. Birds sound classification based on machine learning algorithms. *Asian Journal of Research in Computer Science* 9 (2021), 1–11.
[75] Massimo Merenda, Carlo Porcaro, and Demetrio Iero. 2020. Edge machine learning for AI-enabled IoT devices: A review. *Sensors* 20, 9 (2020), 2533.
[76] B. Mishachandar and S. Vairamuthu. 2021. Diverse ocean noise classification using deep learning. *Applied Acoustics* 181 (2021), 108141.
[77] Md. Mohaimenuzzaman, Christoph Bergmeir, and Bernd Meyer. 2022. Pruning vs XNOR-Net: A comprehensive study of deep learning for audio classification on edge-devices. *IEEE Access* 10 (2022), 6696–6707.
[78] Md. Mohaimenuzzaman, Christoph Bergmeir, Ian West, and Bernd Meyer. 2023. Environmental sound classification on the edge: A pipeline for deep acoustic networks on extremely resource-constrained devices. *Pattern Recognition* 133 (2023), 109025.
[79] Iosif Mporas, Isidoros Perikos, Vasilios Kelefouras, and Michael Paraskevas. 2020. Illegal logging detection based on acoustic surveillance of forest. *Applied Sciences* 10, 20 (2020), 1–12.
[80] Seongkyu Mun, Sangwook Park, David K. Han, and Hanseok Ko. 2017. Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyperplane. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE'17)*. 93–102.
[81] Andrés Muñoz. 2014. Machine learning and optimization. *Courant Institute of Mathematical Sciences* 2014 (2012), 1–2.
[82] Zohaib Mushtaq and Shun-Feng Su. 2020. Efficient classification of environmental sounds through multiple features aggregation and data enhancement techniques for spectrogram images. *Symmetry* 12, 11 (2020), 1822.

[83] Zohaib Mushtaq, Shun-Feng Su, and Quoc-Viet Tran. 2021. Spectral images based environmental sound classification using CNN with meaningful data augmentation. *Applied Acoustics* 172 (2021), 107581.
[84] Satoshi Nakamura, Kazuo Hiyane, Futoshi Asano, and Takashi Endo. 1999. Sound scene data collection in real acoustical environments. *Journal of the Acoustical Society of Japan (E)* 20, 3 (1999), 225–231.
[85] Loris Nanni, Gianluca Maguolo, Sheryl Brahnam, and Michelangelo Paci. 2021. An ensemble of convolutional neural networks for audio classification. *Applied Sciences* 11, 13 (2021), 5796.
[86] Loris Nanni, Gianluca Maguolo, and Michelangelo Paci. 2020. Data augmentation approaches for improving animal audio classification. *arXiv abs/1912.07756* (2020).
[87] Alireza Nasiri and Jianjun Hu. 2021. SoundCLR: Contrastive learning of representations for improved environmental sound classification. *arXiv abs/2103.01929* (2021).
[88] Masaki Okawa, Takuya Saito, Naoki Sawada, and Hiromitsu Nishizaki. 2019. Audio classification of bit-representation waveform. In *Proceedings of Interspeech 2019*. 2553–2557.
[89] Elena Olteanu, Victor Suciu, Svetlana Segarceanu, Ioana Petre, and Andrei Scheianu. 2018. Forest monitoring system through sound recognition. In *Proceedings of the International Conference on Communications (COMM'18)*. IEEE, Los Alamitos, CA, 75–80.
[90] Heshan Padmasiri, Jithmi Shashirangana, Dulani Meedeniya, Omer Rana, and Charith Perera. 2022. Automated license plate recognition for resource-constrained environments. *Sensors* 22, 4 (2022), 1434.
[91] Kamalesh Palanisamy, Dipika Singhania, and Angela Yao. 2020. Rethinking CNN models for audio classification. *arXiv abs/2007.11154* (2020).
[92] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2010), 1345–1359.
[93] Yagya Raj Pandeya, Dongwhoon Kim, and Joonwhoan Lee. 2018. Domestic cat sound classification using learned features from deep neural nets. *Applied Sciences* 8, 10 (2018), 1949.
[94] Ning Peng, Aibin Chen, Guoxiong Zhou, Wenjie Chen, Wenzhuo Zhang, Jing Liu, and Fubo Ding. 2020. Environment sound classification based on visual multi-feature fusion and GRU-AWS. *IEEE Access* 8 (2020), 191100–191114.
[95] Karol J. Piczak. 2015. Environmental sound classification with convolutional neural networks. In *Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP'15)*. IEEE, Los Alamitos, CA, 1–6.
[96] Karol J. Piczak. 2015. ESC-50: Dataset for Environmental Sound Classification. Retrieved July 20, 2022 from https://github.com/karolpiczak/ESC-50
[97] Karol J. Piczak. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*. ACM, New York, NY, 1015–1018.
[98] Anupriya Prasad and Pradeep Chawda. 2018. Power management factors and techniques for IoT design devices. In *Proceedings of the 2018 19th International Symposium on Quality Electronic Design (ISQED'18)*. IEEE, Los Alamitos, CA, 364–369.
[99] Dirga Chandra Prasetyo, Giva Andriana Mutiara, and Rini Handayani. 2018. Chainsaw sound and vibration detector system for illegal logging. In *Proceedings of the 2018 International Conference on Control, Electronics, Renewable Energy, and Communications (ICCEREC'18)*. IEEE, Los Alamitos, CA, 93–98.
[100] Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara Sainath. 2019. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing* 13, 2 (2019), 206–219.
[101] Mohammad Rastegari, Vincente Ordonez, Joseph Redmon, and Ali Farhadi. 2016. NOR-Net: ImageNet classification using binary convolutional neural networks. In *Proceedings of the 14th European Conference on Computer Vision (ECCV'16)*. 525–542.
[102] Imran Mohammed Safwat, Rahman Afia Fahmida, Sifat Tanvi, Kadir Hamim Hassan, Iqbal Junaid, and Mostakim Moin. 2021. An analysis of audio classification techniques using deep learning architectures. In *Proceedings of the 2021 6th International Conference on Inventive Computation Technologies (ICICT'21)*. IEEE, Los Alamitos, CA, 805–812.
[103] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*. ACM, New York, NY, 1041–1044.
[104] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. UrbanSound8k. Retrieved July 20, 2022 from https://urbansounddataset.weebly.com/urbansound8k.html
[105] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello. 2017. Scaper: A library for soundscape synthesis and augmentation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'17)*. IEEE, Los Alamitos, CA, 344–348.
[106] Sina Sanaei, Babak Majidi, and Ehsan Akhtarkavan. 2018. Deep multisensor dashboard for composition layer of Web of Things in the smart city. In *Proceedings of the 9th International Symposium on Telecommunications*. IEEE, Los Alamitos, CA, 211–215.

[107] Svetlana Segarceanu, Elena Olteanu, and George Suciu. 2020. Forest monitoring using forest sound identification. In *Proceedings of the 2020 43rd International Conference on Telecommunications and Signal Processing (TSP'20)*. IEEE, Los Alamitos, CA, 346–349.

[108] Svetlana Segarceanu, George Suciu, and Inge Gavat. 2021. Neural networks for automatic environmental sound recognition. In *Proceedings of the 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD'21)*. IEEE, Los Alamitos, CA, 7–12.

[109] Romain Serizel, Nicolas Turpault, Ankit Shah, and Justin Salamon. 2020. Sound event detection in synthetic domestic environments. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'20)*. IEEE, Los Alamitos, CA, 86–90.

[110] Sulis Setiowati, Zulfanahri, Eka Legya Franita, and Igi Ardiyanto. 2017. A review of optimization method in face recognition: Comparison deep learning and non-deep learning methods. In *Proceedings of the 9th International Conference on Information Technology and Electrical Engineering (ICITEE'17)*. IEEE, Los Alamitos, CA, 1–6.

[111] Sayed Khushal Shah, Zeenat Tariq, and Yugyung Lee. 2019. IoT based urban noise monitoring in deep learning using historical reports. In *Proceedings of the IEEE International Conference on Big Data*. IEEE, Los Alamitos, CA, 4179–4184.

[112] Roneel V. Sharan and Tom J. Moir. 2015. Noise robust audio surveillance using reduced spectrogram image feature and one-against-all SVM. *Neurocomputing* 158 (2015), 90–99.

[113] Garima Sharma, Kartikeyan Umapathy, and Sridhar Krishnan. 2020. Trends in audio signal feature extraction methods. *Applied Acoustics* 158 (2020), 107020.

[114] Jithmi Shashirangana, Heshan Padmasiri, Dulani Meedeniya, Charith Perera, Soumya R. Nayak, Janmenjoy Nayak, Shanmuganthan Vimal, and Seifidine Kadry. 2021. License plate recognition using neural architecture search for edge devices. *International Journal of Intelligent Systems* 36, 7 (2021), 1–38.

[115] Sungho Shin, Jongwon Kim, Yeonguk Yu, Seongju Lee, and Kyoobin Lee. 2021. Self-supervised transfer learning from natural images for sound classification. *Applied Sciences* 11, 7 (2021), 3043.

[116] Siddharth Sigtia, Adam M. Stark, Sacha Krstulovic, and Mark D. Plumbley. 2016. Automatic environmental sound recognition: Performance versus computational cost. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 11 (2016), 2096–2107.

[117] Rajesh Singh, Anita Gehlot, Shaik Vaseem Akram, Amit Kumar Thakur, Dharam Buddhi, and Prabin Kumar Das. 2021. Forest 4.0: Digitalization of forest using the Internet of Things (IoT). *Journal of King Saud University—Computer and Information Sciences* 34, 8 (2021), 5587–5601.

[118] Marina Sokolova, Nathalie Japkowicz, Stan Szpakowicz, Abdul Sattar, and Byeong-Ho Kang. 2006. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In *AI: Advances in Artificial Intelligence*. Springer, Berlin, Germany, 1015–1021.

[119] MathWorks Inc. 2019. Continuous Wavelet Transform and Scale-Based Analysis. Retrieved July 20, 2022 from https://www.mathworks.com/help/wavelet/gs/continuous-wavelet-transform-and-scale-based-analysis.html

[120] Po-Jung Ting, Shanq-Jang Ruan, and Lieber Po-Hung Li. 2021. Environmental noise classification with inception-dense blocks for hearing aids. *Sensors* 21, 16 (2021), 5406.

[121] Achyut Mani Tripathi and Aakansha Mishra. 2021. Self-supervised learning for Environmental Sound Classification. *Applied Acoustics* 182 (2021), 108183.

[122] Nicolas Turpault, Romain Serizel, Scott Wisdom, Hakan Erdogan, John R. Hershey, Eduardo Fonseca, Prem Seetharaman, and Justin Salamon. 2021. Sound event detection and separation: A benchmark on DESED synthetic soundscapes. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, Los Alamitos, CA, 840–844.

[123] Jurgen Vandendriessche, Nick Wouters, Bruno da Silva, Mimoun Lamrini, Mohamed Yassin Chkouri, and Abdellah Touhafi. 2021. Environmental sound recognition on embedded systems: From FPGAs to TPUs. *Electronics* 10, 21 (2021), 2622.

[124] Mário P. Véstias, Rui Policarpo Duarte, José T. de Sousa, and Horácio C. Neto. 2020. Moving deep learning to the edge. *Algorithms* 13, 5 (2020), 125.

[125] H. L. Wang, D. Z. Song, Z. L. Li, X. Q. He, S. R. Lan, and H. F. Guo. 2020. Acoustic emission characteristics of coal failure using automatic speech recognition methodology analysis. *International Journal of Rock Mechanics and Mining Sciences* 136 (2020), 104472.

[126] Zhi Wang, Wentao Zha, Jin Chai, Yilin Liu, and Zhuoling Xiao. 2021. Lightweight implementation of FPGA-based environmental sound recognition system. In *Proceedings of the International Conference on UK-China Emerging Technologies (UCET'21)*. IEEE, Los Alamitos, CA, 59–66.

[127] Shengyun Wei, Shun Zou, Feifan Liao, and Weiman Lang. 2020. A comparison on data augmentation methods based on deep learning for audio classification. *Journal of Physics: Conference Series* 1453, 1 (2020), 012085.

[128] Felix Weninger and Björn Schuller. 2011. Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)*. IEEE, Los Alamitos, CA, 337–340.

[129] Zhizheng Wu and Simon King. 2016. Investigating gated recurrent networks for speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'16)*. IEEE, Los Alamitos, CA, 5140–5144.

[130] Steven Wyatt, David Elliott, Akshay Aravamudan, Carlos E. Otero, Luis D. Otero, Georgios C. Anagnostopoulos, Anthony O. Smith, Adrian M. Peter, Wesley Jones, Steven Leung, and Eric Lam. 2021. Environmental sound classification with tiny transformers in noisy edge environments. In *Proceedings of the IEEE 7th World Forum on Internet of Things (WF-IoT'21)*. IEEE, Los Alamitos, CA, 309–314.

[131] Nina Sofia Wyniawskyj, Milena Napiorkowska, David Petit, Pritimoy Podder, and Paula Marti. 2019. Forest monitoring in Guatemala using satellite imagery and deep learning. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS'19)*. IEEE, Los Alamitos, CA, 6598–6601.

[132] Jie Xie, Kai Hu, Mingying Zhu, Jinghu Yu, and Qibing Zhu. 2019. Investigation of different CNN-based models for improved bird sound classification. *IEEE Access* 7 (2019), 175353–175361.

[133] Lidong Yang, Jiangtao Hu, and Zhuangzhuang Zhang. 2019. Audio scene classification based on gated recurrent unit. In *Proceedings of the IEEE International Conference on Signal, Information, and Data Processing*. IEEE, Los Alamitos, CA, 1–5.

[134] Jiaxing Ye, Takumi Kobayashi, Nobuyuki Toyama, Hiroshi Tsuda, and Masahiro Murakawa. 2018. Acoustic scene classification using efficient summary statistics and multiple spectro-temporal descriptor fusion. *Applied Sciences* 8, 8 (2018), 1363.

[135] Yuzhong Wu and Tan Lee. 2018. Reducing model complexity for DNN based large-scale audio classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'18)*. IEEE, Los Alamitos, CA, 331–335.

[136] Shuo Zhang, Demin Gao, Haifeng Lin, and Quan Sun. 2019. Wildfire detection using sound spectrum analysis based on the Internet of Things. *Sensors* 19, 23 (2019), 5093.

[137] Sai-Hua Zhang, Zhao Zhao, Zhi-Yong Xu, Kristen Bellisario, and Bryan C. Pijanowski. 2018. Automatic bird vocalization identification based on fusion of spectral pattern and texture features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'18)*. IEEE, Los Alamitos, CA, 271–275.

[138] Zhao Zhao, Sai-Hua Zhang, Zhi-Yong Xu, Kristen Bellisario, Nian-Hua Dai, Hichem Omrani, and Bryan C. Pijanowski. 2017. Automated bird acoustic event detection and robust species classification. *Ecological Informatics* 39 (2017), 99–108.

[139] Pablo Zinemanas, Martín Rocamora, Marius Miron, Frederic Font, and Xavier Serra. 2021. An interpretable deep learning model for automatic sound classification. *Electronics* 10, 7 (2021), 850.

[140] Imran Zualkernan, Jacky Judas, Taslim Mahbub, Azadan Bhagwagar, and Priyanka Chand. 2021. An AIoT system for bat species classification. In *Proceedings of the IEEE International Conference on Internet of Things and Intelligence System (IoTaIS'21)*. IEEE, Los Alamitos, CA, 155–160.

[141] Incze Ágnes, Henrietta-Bernadett Jancsó, Szilágyi Zoltán, Farkas Attila, and Sulyok Csaba. 2018. Bird sound recognition using a convolutional neural network. In *Proceedings of the IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY'18)*. IEEE, Los Alamitos, CA, 295–300.

Received 2 August 2022; revised 31 May 2023; accepted 21 August 2023