Article

Forest Sound Classification Dataset: FSC22

Meelan Bandara[1], Roshinie Jayasundara[1], Isuru Ariyarathne[1], Dulani Meedeniya[1] and Charith Perera[2,*]

[1] Department of Computer Science & Engineering, University of Moratuwa, Moratuwa 10400, Sri Lanka.;
[2] School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K.;
* Correspondence: pereraC@cardiff.ac.uk;

Abstract: The study of environmental sound classification (ESC) has become popular over the years due to the intricate nature of environmental sounds and the evolution of deep learning (DL) techniques. Forest ESC is one use case of ESC, which has been widely experimented with recently to identify illegal activities inside a forest. However, at present, there is a limitation of public datasets specific to all the possible sounds in a forest environment. Most of the existing experiments have been done using generic environment sound datasets such as ESC-50, U8K, and FSD50K. Importantly, in DL-based sound classification, the lack of quality data can cause misguided information, and the predictions obtained remain questionable. Hence, there is a requirement for a well-defined benchmark forest environment sound dataset. This paper proposes FSC22, which fills the gap of a benchmark dataset for forest environmental sound classification. It includes 2025 sound clips under 27 acoustic classes, which contain possible sounds in a forest environment. We discuss the procedure of dataset preparation and validate it through different baseline sound classification models. Additionally, it provides an analysis of the new dataset compared to other available datasets. Therefore, this dataset can be used by researchers and developers who are working on forest observatory tasks.

Keywords: forest acoustic dataset; environment sound classification; machine learning; freesound; deep learning

1. Introduction

Environmental sound recognition is a widely used technique when identifying various sound events for surveillance or monitoring systems based on the acoustic environment. Several investigations have been carried out with different techniques in the context of a forest monitoring system, to protect forest reserves. For example, prior studies have experimented with different sound classification approaches for the recognition of various species and possible forest threats like illegal logging, poaching, and wildfire [1–5]. In such systems, environmental sounds are captured, processed using a modelling algorithm, and classified into different sound classes.

With the technical advancement, sound classification approaches evolved from Machine Learning (ML) models such as K-Nearest Neighbor (KNN) [3,6,7], XGBoost [8,9], Gaussian Mixture Modelling (GMM) [5,10], and Support Vector Machine (SVM) [6,11,12] to Deep Learning. Deep neural networks (DNNs) such as Convolutional Neural networks (CNN) and Recurrent Neural Networks (RNN) require a large amount of labelled data compared to ML for a promising result. Hence, when using DL-based approaches, a well-biased and rich dataset with relatively high data size is essential as the performance keeps increasing with a quality dataset.

Although several studies have been carried out in the forest acoustic monitoring context, still, a standard benchmark dataset specific to forest sounds is unavailable. Therefore most of the existing studies have utilized publicly available environmental sound datasets like ESC-50 [4,13–17], UrbanSound8K (U8k) [14,18–21], FSD50K [22,23] and SONYC-UST [24,25]. These datasets contain a large amount of audio data categorized into several groups covering a broad area of sound events. However, a limited number of classes can be used

Citation: Bandara, M.; Jayasundara, R.; Ariyarathne, I.; Meedeniya, D.; Perera, C. Title. *Sensors* **2023**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

for forest environment sound classification, and most data are irrelevant for such a domain. Since a significant number of resources need to be utilized to extract data from datasets and to annotate the data points according to a suitable taxonomy, the use of public datasets directly for the classification model is impotent.

Additionally, some studies have utilized datasets such as BIRDZ [26,27] and Xeno-canto Archive [28–30], which contain only bird sounds. Xeno-canto Archive is an open audio collection dedicated to sharing bird sounds, and BIRDZ is a control audio dataset originating from the Xeno-canto Archive, which contains a subset of 11 bird species. As it contains audio data specific to one class, several such datasets need to be used in the forest sound classification system. Moreover, several researchers have experimented with private datasets due to the unavailability of forest-specific sound datasets. For instance, in such studies, they have deployed sound sensors in a forest environment and recorded the sound events to create a dataset according to their requirements [6,31,32]. In contrast, some studies have created datasets using audio clips collected from online sound data repositories like free sound [3,5,11]. With a closer look at the literature, it can be identified that the forest acoustic monitoring domain suffers from certain shortcomings including the lack of a standard taxonomy and the unavailability of a public benchmark dataset. These limitations motivated us to introduce a new dataset for the domain. Accordingly, the novelty of this paper is to present a standard dataset for forest sound classification and to provide a comprehensive overview of the procedure for creating and validating the dataset. Addressing the current research gaps we introduce FSC22 [33], a novel benchmark dataset for the acoustic-based forest monitoring domain. It contains 5 seconds long 2025 audio clips originating from an online audio database FreeSound. All sound events are categorized into 6 major classes, which are further divided into 34 subclasses. For the initial phase of dataset composition, 27 subclasses were picked, and 75 audio samples were collected per class. Each audio clip was manually annotated and verified to ensure the quality of the dataset. The key contributions of this paper can be summarized as follows.

- Introduces a novel public benchmark dataset consisting of forest environmental sounds, which can be utilized for acoustic-based forest monitoring.
- Presents a comprehensive description of the methodology used for dataset creation, including data acquisition from FreeSound, filtering, and validation to normalization.
- Explains the baseline models used for the sound classification and the selection criteria for those models.
- Provides a detailed evaluation of the dataset using human classification, ML-based and DL-based classification.
- Presents a comprehensive discussion of the results obtained with the proposed FSC22 dataset and compares them with the publicly available datasets.

We have created the FSC22 dataset and made it freely available to support and motivate future researchers in this domain [33]. We expect that this dataset will help research communities to better understand forest acoustic surveillance and experiment with the domain. The rest of the paper is structured as follows. Section 2 explores the related datasets used in previous research. Section 3 provides an overview of the taxonomy of the proposed dataset. Section 4 introduces the FSC22 dataset, including the data collection methodology and its importance to the acoustic domain. Section 5 provides a comprehensive description of the baseline model-based dataset evaluation approach. Section 6 describes the experiments conducted on the dataset namely human classification and baseline model-based classification, with the results and observations. Finally, Section 7 concludes the paper.

2. Related Work

Seminal contributions have been made to the ESC context in recent years. Among those, several instances of research carried out for forest acoustic monitoring can be identified. Forest acoustic monitoring is crucial as it provides a firm basis of evidence to arrive at conclusions to conserve forest coverage and species. However, due to the unavailability of a comprehensive forest-specific sound dataset, most of the previous research on forest monitoring was done using a common environmental sound dataset or a private dataset according to the requirement. This section provides an overview of the publicly available environmentally sound datasets and other datasets utilized by previous researchers in this domain. However, to the best of our knowledge, there is no forest-specific sound dataset in the literature.

Among the available datasets, ESC-50 [34] is a frequently used environmental sound dataset for forest acoustic monitoring. For instance, Andreadis et al. [4], have utilized ESC-50 to detect illegal tree-cutting and identify animal species. ESC-50 is a dataset consisting of 2000 environmental audio clips under 50 classes of common sound events. It contains 5-second long 40 recording samples per class, extracted from FreeSound. Figure 1 shows a section of the ESC-50 dataset taxonomy emphasizing forest-specific sounds. Moreover, U8K [35] is another popular dataset used in many types of research on audio-based monitoring systems [18,36]. U8K is a subset of the main Urban Sound dataset, which contains 8732 labelled sound clips of urban sounds from 10 classes. The classes of this dataset are drawn from the urban sound taxonomy [37], and all the recordings are extracted from Freesound. Figure 2 includes a part of the U8K dataset taxonomy mostly relevant to the forest environment sound domain. FSD50K [38] is an open dataset of human-labelled sound events. It consists of over 51K audio clips totalling over 100h of audio manually labelled using 200 classes. The classes of this dataset are drawn from AudioSet Ontology [39]. All the above-mentioned 3 datasets were created using the audio extracted from the Freesound project. It is an audio-based public dataset that contains more than 500 000 audio clips.

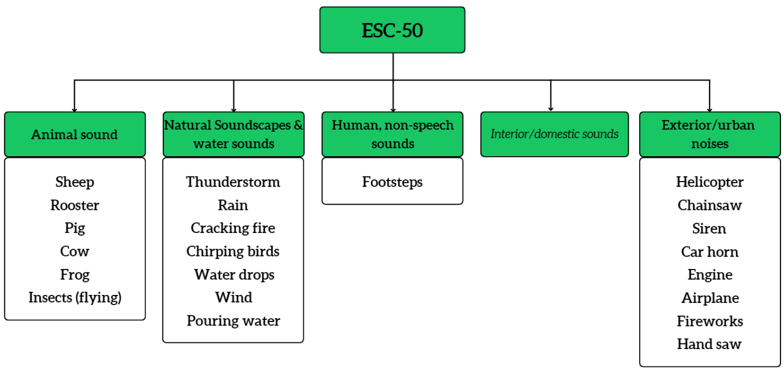

Figure 1. ESC-50 dataset

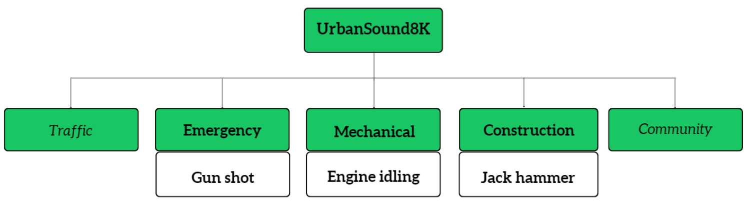

Figure 2. Urbansound8K dataset

Moreover, SONYC-UST [40] is another quality dataset, where data is grouped into 8 main classes and further divided into 23 fine-grained classes. This can be considered a more realistic dataset as it was created using the audio data acquired using the acoustic sensors deployed in New York City. Figure 3 shows a part of the SONYC-UST dataset taxonomy highlighting the audio classes specific to forest monitoring and surveillance. AudioSet [41] is another audio event dataset, including over 2M tracks from Youtube videos. Every

10-second video is annotated using over 500 sound classes derived from AudioSet ontology [39]. The main concern with AudioSet is it cannot be considered an open dataset due to the copyright issues and Terms of Services constraint from Youtube. However, as the clips are collected from Youtube, they may consist of clips with poor quality and can disappear after a certain time due to privacy issues or copyright claims. Table 1 presents a summary of the existing environmental sound datasets.

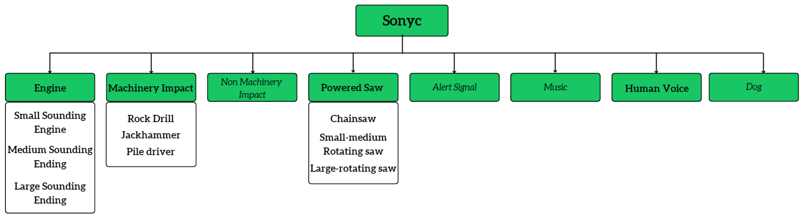

Figure 3. SONYC-UST-V2 dataset

Table 1. Summary of existing ESC datasets.

Dataset	Source	Total clips	Clip length	Classes
ESC-50 [34]	Free-sound	2000	5s	50
Urban-Sound8K [35]	Free-sound	8732	More than 4s	10
AudioSet [41]	Youtube	2M	10s	527
FSD50K [42]	Free-sound	51197	0.3s to 30s	200
SONYC- UST-v2 [40]	SONYC acoustic	18510	10s	23

Additionally, several other domain-specific dataset usages were reported in prior studies on environment sound observatory systems. For bird sound identification studies, xeno-canto-archive [43], which is a bird sound-sharing portal, was used to acquire the audio data essential for the experiment [28,30,44]. BIRDZ dataset, which is a real-world audio dataset made using the xeno-canto archive was also used in the related literature [45,46]. Similarly, the usage of the BirdCLEF dataset was identified in the prior studies, which consists of 62902 audio files and is publicly available on Kaggle [47]. As all these datasets are specific to a certain sound class, a combination of several such datasets is required when developing a complete forest monitoring system.

Many researchers have experimented with a private dataset they have created according to their requirements, due to the scarcity of forest-specific sound datasets. Such datasets were generated using the audio data acquired from online sound repositories or audio recorded by acoustic sensors or as a combination of both. Mporas et al. [3], have created a chainsaw sound dataset, including the background noises such as rain and wind, using the sounds acquired from freely available sound repositories. Ying et al. [11], have experimented with an animal sound recognition system, and the required animal sounds are acquired from Freesound. In contrast, Assoukpou et al. [6], combined the chainsaw sounds recorded from acoustic sensors deployed in three different forest areas and other sounds acquired from online websites to create a dataset to identify chainsaw sounds.

Accordingly, many environmental-sound classification studies have utilized the datasets mentioned above with different sound classification approaches. In most of the studies, CNN models were widely adopted as a firm basis for prominent audio classification models [20,36,48]. Besides, there are instances where ML algorithms were utilized for audio classification [49]. One of the key distinctions when choosing between DL and ML was the availability of well-labelled and high volumes of data. DL algorithms scale with the data while increasing the performance, whereas ML plateaus at a certain level of performance when adding more data. Table 2 shows an overview of DL and ML approaches deployed for sound classification using the ESC-50 and U8K datasets.

Table 2. Related studies on sound classification using ESC-50 and UrbanSound 8k Dataset.

Study	ML/DL	ESC-50		UrbanSound-8K	
		Model	**Accuracy**	**Model**	**Accuracy**
[20]	DL(CNN)	DenseNet	98.50%	DenseNet	97.10%
		AlexNet	88.10%	AlexNet	93%
		ResNet	96.80%	ResNet	99.20%
[36]	DL(CNN)	DenseNet	97.57%	DenseNet	99.20%
		ResNet	96.80%	ResNet	99.49%
[48]	DL(CNN)	DenseNet	92.80%	DenseNet	87.40%
[50]	ML			SVM	71%

3. FSC22 Taxonomy

Prominent research efforts carried out in the forest acoustic classification domain have been based on a subset of an already established public dataset like ESC50, U8K, or on small self-made datasets. Thus, the requirement for a well-defined dataset dedicated to forest acoustics can be identified. As the first step of creating a benchmark dataset, a standard taxonomy that can showcase and capture all the different acoustic scenarios present in forest ecosystems needs to be established.

In the parent level of the proposed taxonomy, all the acoustic scenarios are classified into six classes: mechanical sounds, animal sounds, environmental Sounds, vehicle Sounds, forest threat sounds, and human sounds. Further, each class is divided into subclasses that capture specific sounds which fall under the main category. For example, under the main class, mechanical sounds, four subclasses can be identified, namely axe, chainsaw, handsaw, and generator. This subdivision aims to introduce specific class labels to prevent the usage of generalized labels like tree cutting, animal roar, etc. Figure 4. presents the complete forest sound taxonomy developed to base the creation of the FSC22 dataset. Further, it showcases the complete subdivision of the main six classes into 34 sub-classes. We have selected only 27 subclasses for the FSC22 dataset ignoring 7 subclasses shown in blue colour, due to the unavailability of a sufficient number of sound clips in Freesound. Though all the left-out classes have more than 200 search results in the Freesound platform, most of the audio clips were artificially generated or included unnecessary noise making them unsuitable to be included in the FSC22 dataset.

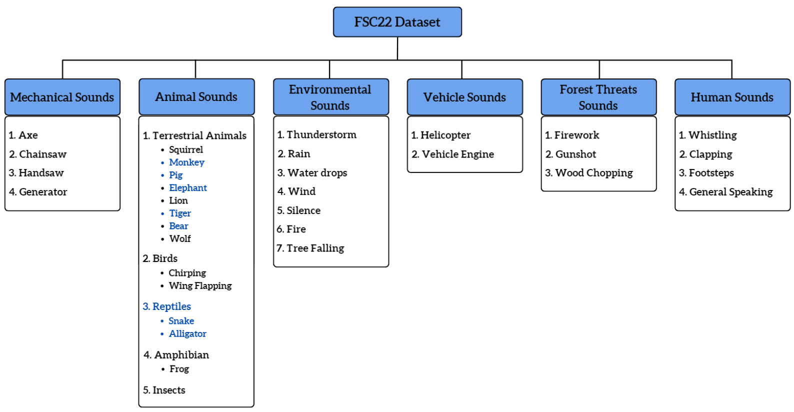

Figure 4. FSC22 Taxonomy

The proposed taxonomy is aimed at covering two main objectives. The first objective is to completely cover fundamental acoustic scenarios such as chainsaw sounds, tree felling, and wildfire, which are extensively used for research works. The second objective is to provide high-quality, normalized audio under unambiguous class labels. We have extensively analyzed related literature, which has utilized forest acoustics and has identified the most essential and frequent types of acoustic phenomenon that should be available in a benchmark dataset to fulfil the first objective, as explained in Section 4. It should be noted that the proposed taxonomy is not fixed and with time, more related acoustic classes under forest acoustics need to be added while refining the taxonomy to achieve saturation.

4. FSC22 Dataset

The proposed FSC22 dataset [33] in this paper is a public benchmark dataset containing 2025 audio samples normalized to 44100 Hz sample rate, 16-bit depth, and stereo channel configuration. All the audio samples are distributed between six major parent-level classes. Each audio is further divided into scenario-specific low-level classes, which capture the context of the considered audio sample as described in Section 3. The FSC22 dataset serves two major objectives, the first one being the requirement to provide sufficient audio samples for widely researched forest-related acoustic classes. The second objective is to present high-quality normalized audio samples under event-specific class labels. This section describes the procedure which was followed to develop the FSC22 dataset while ensuring the objectives. Figure 5 shows the overall procedure of creating the FSC22 dataset and each sub-process is described in this section.

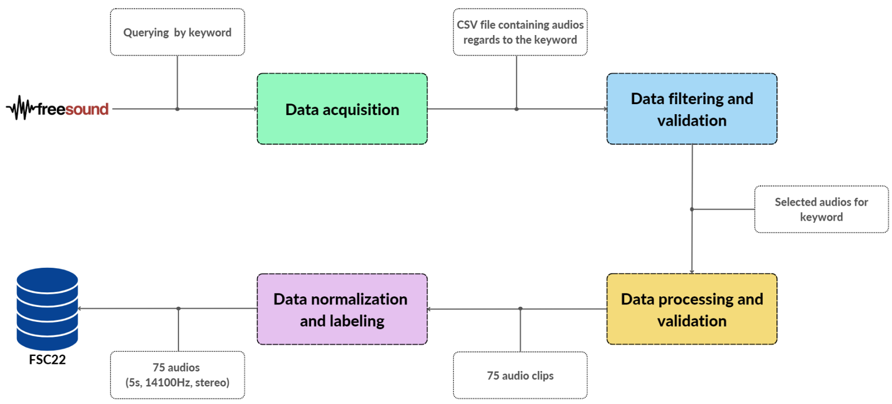

Figure 5. Overall Procedure

4.1. Dataset Preparation

4.1.1. Data acquisition

The development of major datasets governing the acoustic classification domain is mainly based on online audio collection portals such as YouTube, BBC Sound Effects Library, and FreeSound Org. The usage of such sources presents unique advantages and disadvantages. Therefore, it was initially required to select the source that FSC22 is based upon, to develop a high-quality benchmark dataset. Although both YouTube and BBC Sound Effects Library are rich when some acoustic labels are considered, they publicly present copyright issues when publishing the final dataset. FreeSound, available at https://freesound.org/, is a free, public, online platform where thousands of audio data are published, and it was identified that by basing the content of FSC22 on the FreeSound platform, we could easily navigate the publishing issue. Further, the API endpoints available in the FreeSound Platform allowed users to write python scripts to search for different audio scenarios and download the metadata and the corresponding audio files without manually searching and downloading the audio.

As the first step of data acquisition, we selected 27 classes from the FSC22 taxonomy to complete in the first phase of the FSC22 dataset. For each of the selected class labels, we queried for audio samples, which contain the considered label in the title or the description, using the API endpoint for text search. The querying process was completed through a python script. After all the matching audio samples were identified, their metadata was class-wise written to spreadsheet files to be fed to the filtering and validation stage. We selected 47832 audios and sent them for the filtering and validation step. Figure 6 showcases the number of audio samples identified via the data-acquiring step for each selected 27 classes

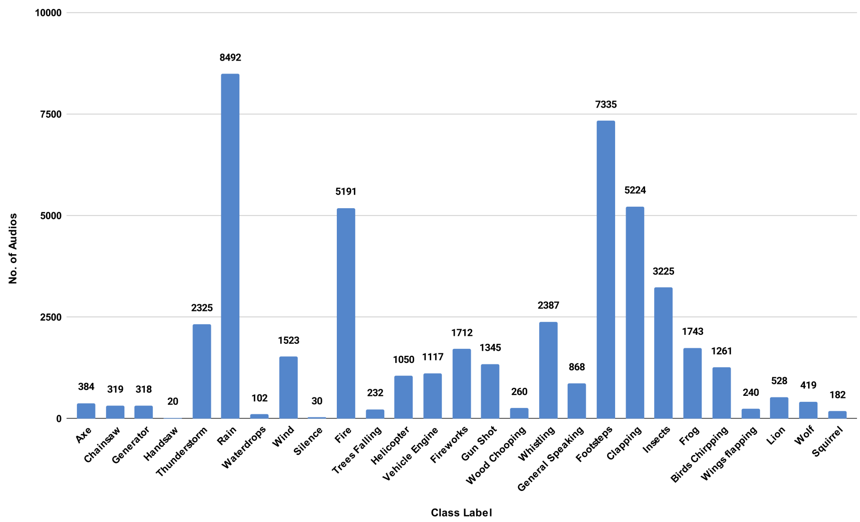

Figure 6. The number of audio per class

4.1.2. Data filtering and validation phase

After spreadsheet files were completed for all the selected classes, all the sheets were traversed to remove non-suitable query results which were present in the sheets due to the noise associated with the API endpoint. After the filtering of suitable audio was completed, each selected audio sample was manually checked by listening to them and downloaded for further processing to begin. All the unclear or unsuitable audios for further processing were removed to refine the dataset quality. Figure 7 shows the number of audio samples selected from each class to be further processed to complete the FSC22 dataset.

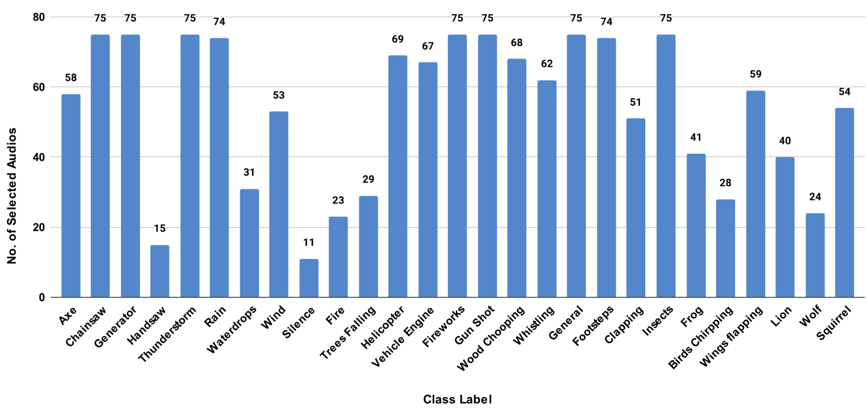

Figure 7. The number of selected audios per class

4.1.3. Data processing and Validation

In order to generate 75 audio clips for each audio class, downloaded audio was processed based on the duration of the original file. Audacity software was used for this procedure, which is an open-source application for audio editing and tagging. Downloaded audio files were uploaded to the Audacity application and trimmed to 5 seconds. Selected audios with longer duration were spliced into multiple recordings of 5 seconds. This step was necessary for some classes due to the lack of suitable audio samples to complete the considered audio sample limit per class. This process was repeated for all the sound classes, and 75 audio recordings were validated and finalized at the end.

4.1.4. Data normalization and labelling

After the filtering and validation process was completed, all 27 classes, which were selected for the first phase were finalized with 75 audio recordings. As the first step of normalization, the sampling frequency was set to 44100 Hz, the bit depth was set to 16, and the channel setting was configured to stereo for all the selected audio recordings, using the load function of Librosa. In the audio extraction step, from the original audios in the earlier phase, audios with nearly 5 seconds of duration were extracted. Hence as the second step of normalization, the duration of all the selected audio was set to 5 seconds by trimming excess parts or by padding with silence accordingly.

At the end of the normalization process, all the original audio samples were renamed accordingly. In this step, the source file name was mapped into the dataset file name in the format of UniqueClassIndex_UniqueAudioID.wav. The first part of the label indicates the class related to the audio sample and is followed by a unique audio ID. Proper labelling of the audio files will make it easier to navigate through the dataset. Once the audio files were labelled, the corresponding metadata was entered into the base metadata file to complete the development of the FSC22 dataset.

4.2. Content Description

FSC22 is a public benchmark dataset that can be utilized in research work governing forest acoustic monitoring and classification. The dataset is developed according to the taxonomy proposed in section 3. Out of the thirty-four subclasses listed in the taxonomy, 27 subclasses were completed for the first phase of the FSC22 dataset. Each subclass contains 75 selected audio samples, which have been manually checked for any inconsistencies. Overall, the dataset contains 2025 audio samples, each with a duration of 5 seconds, resulting in 2.81 hours of forest acoustics under the specified class labels. All the required information about the audio samples available in the dataset is listed under the metadata file located in the FSC22 master folder. The FSC22 master folder contains two subfolders, audio wise V1.0 which includes the 2025 audio samples, and the Metadata folder which holds the Metadata.csv file.

Readers of this study and the users of the FSC22 dataset should note that each audio sample was renamed according to the following convention to better support the usage of the new dataset.

- UniqueClassIndex_UniqueAudioID.wav eg: 1_10101.wav

Table 3 provides a snapshot of the Metadata.csv file for the convenience of the readers. As shown for each audio file, the Metadata file provides:

- Source File Name - ID of the original audio sample, used to extract the corresponding audio.
- Dataset File Name - ID of the audio, in the context of FSC22
- Class ID - Class Identification index (An integer from the range 1 to 27)
- Class Name - Class Name in which the audio is classified.

Table 3. Sample of meta-data of the FSC22 Dataset.

Source File Name	Dataset File Name	Class ID	Class Name
17548__A.wav	1_10101.wav	1	Fire
17548__B.wav	1_10102.wav	1	Fire
17548__C.wav	1_10103.wav	1	Fire

4.3. Importance of FSC22 to the Forest Acoustics Domain

Analyzing the research contributions made towards the forest acoustic domain, it becomes evident that a publicly available forest-specific sound dataset is unattainable. Due to the scarcity of a standard dataset for forest sounds, the research community has experimented with different approaches for data acquisition. Few can be identified as obtaining sound recordings by employing sound sensors, collecting sound clips available in online sound repositories and extracting the sounds from YouTube videos. Table 4 summarizes the sound acquisition approaches used in previous forest acoustic domain research for a better overview.

Table 4. Sound acquisition approaches in related studies.

Study	Domain	Source	Dataset acquiring approach
[3]	Illegal logging detection	Freely available online sound data repositories	Collected audio recordings of chainsaws and environment background noises (rain, wind, birds)
[11]	Animal sound recognition	Freesound	Collected bird sounds, mammal sounds, and insect sounds
[10]	Tree cutting detection	Sensor recordings from an urban environment	Collected 18 chainsaw sounds, 27 vehicle sounds, 20 forest-specific sounds, 28 background sound clips
[51]	Animal sound recognition	HU-ASA database	Collected 1418 animal sound clips from the archive
[44]	Bird species detection	Xeno-Canto	Collected 2104 sound clips for 5 bird species
[4]	Illegal Tree Cutting	ESC-50	Selected specific 7 classes related to forest environment (wind, chainsaw, rain, birds, etc.)
[6]	Chainsaw sound identification	Sensor recordings from a forest environment and online sound repositories	Collected 301 chainsaw sounds and 2964 other sounds (bird, insects, animals, etc.)
[5]	Chainsaw and vehicle sound detection	Sensor recordings from the forest and urban environments	Acquired 57 chainsaw recordings, 70 vehicle/engine sounds, 62 forest sounds, 28 general urban sounds
[31]	Illegal Logging Detection	Sensor recordings from a forest environment	Collected 100 chainsaw sounds

Findings in Table 4 confirmed that in most of the early studies, authors have prepared a separate dataset according to their requirements due to the unavailability of a proper forest acoustic dataset. However, data collection is a complex and time-consuming task which could be an overhead for research tasks. Hence, the requirement for a standard dataset arises. Addressing the problem of the unavailability of a standard dataset, this paper introduces FSC22, which includes forest-specific sounds under 27 classes. This dataset covers most of the general acoustic classes identified in a forest environment. The FSC22 dataset will be a great contribution to any further research performed under the forest acoustic domain.

5. Methods and Technical Implementation

For ESC, both ML and DL have been extensively used in related literature. Therefore, we provide classification experiments covering both architectures. An Extreme Gradient Boosting (XGBoost) based experiment is provided for the ML approach, while a CNN-based experiment is provided for the DL approach. These models were used as the baseline models.

5.1. Feature Engineering

Feature engineering is a principal requirement for a successful ML pipeline. Studies focusing on the audio classification domain properly emphasize the requirement of advanced feature engineering techniques like the usage of spectrograms to represent audios in the time and frequency domains [4,6,10,17,52], and the audio augmentation techniques to prevent overfitting of the prediction algorithm [13,14,46,53,54], to obtain state-of-the-art classification performances. This section provides an overview of the feature engineering techniques followed in the proposed experiments as shown in Figure 8.

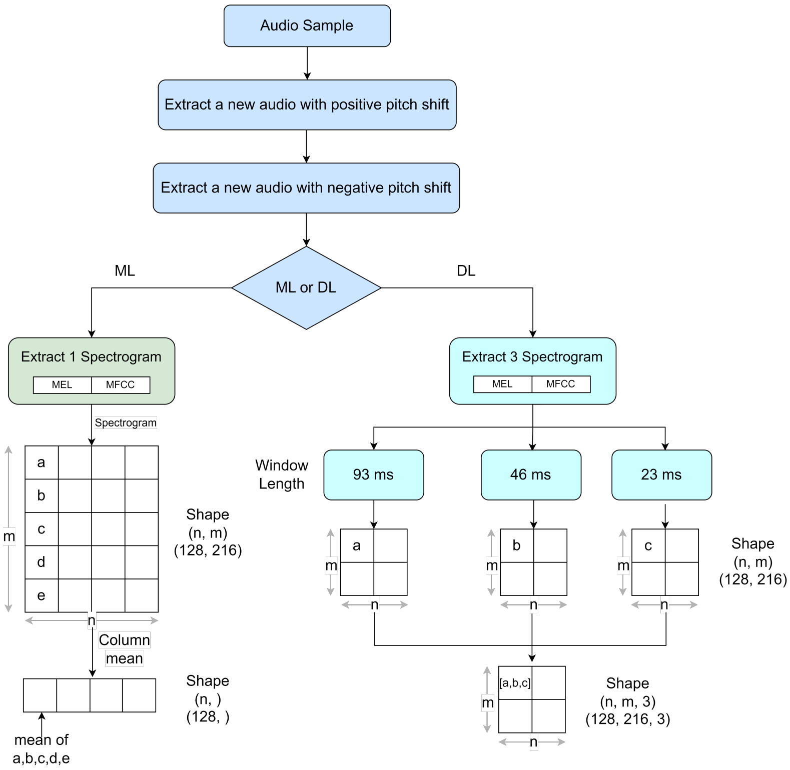

Figure 8. Feature preparation methodology

5.1.1. Considered Datasets

As described in subsection 4.3 quality audio data is scarce under the forest acoustics domain, thus a benchmark dataset that could be used to compare the quality of the proposed FSC22 dataset cannot be identified in the related literature. ESC50 dataset, which is a benchmark dataset used under the ESC domain is therefore used to compare the performance of the FSC22 dataset. For the study 2000 audio recordings, each of 5-second duration distributed into 50 unique classes from the ESC50 dataset, and 2025 audio recordings each of 5-second duration distributed into 27 unique classes from the FSC22 dataset were subjected.

5.1.2. Data Augmentation Technique

Data augmentation is an important step in the feature engineering phase to artificially expand the available data samples for training and testing ML and DL algorithms.

Especially when it comes to DL approaches, models suffer from overfitting when the amount of training data available is considerably less [55]. For the proposed experiments, Positive pitch shifting and negative pitch shifting, where the pitch of audio recordings is increased and decreased by two steps respectively, are utilised [56]. The pitch shifting was implemented with the pitch_shift function provided by Librosa.effects library for python.

As a result of a single audio sample, two new augmented audios were created increasing the amount of data available. In summary, due to the augmentation with pitch shift, the number of audio samples from ESC50 was increased to 6000, while the FSC22 dataset increased to 6025 audio samples. For both datasets, 80% of the audio samples were used for training the model, while 20% were used for validating the performance of the trained model, by following the Pareto Principle as in most of the general cases, 80% of effects come from 20% of causes.

5.1.3. Feature Extraction

Under the audio classification domain, the general practice is using spectrograms, representing an audio signal in both time and frequency domains, as the feature extraction mechanism. The Mel Spectrogram (MEL) [20,57] and the Mel Frequency Cepstral Coefficients (MFCC) [3,10], which can be identified as the two most utilized spectrograms, are used to extract the features for this study. In order to extract the spectrograms from the raw audio data, the Mel spectrogram and MFCC are provided by the librosa.feature library was used. Using both functions, each audio file gets sampled into overlapping frames, and for each frame model coefficient or Mel frequency, cepstral coefficients are calculated. Thus, calculated coefficients are returned as a 2-dimensional array of shapes (number of coefficients x number of samples). As a further improvement, for the Mel spectrograms obtained, all the coefficients were converted to the decibel scale from the power scale.

As shown in Figure 8, ML based classifications generally utilize 1-dimensional features. Therefore, it is required to reduce the dimensionality of the created spectrograms, before they were used with the XGBoost model. This was achieved by aggregating the 1-dimensional feature vectors extracted for each overlapping frame into a single vector by taking their mean value. For DL based classification, an image-like representation of the features according to the RGB mode is required. Hence for each audio sample, three spectrograms were created by changing the length of the window used for framing. Created spectrograms were of windowing length of 93 milliseconds, 46 milliseconds, and 23 milliseconds and this was achieved by keeping the sample rate parameter at 22050 Hz and the n_fft parameter in 2048, 1024, and 512, respectively.

5.2. Machine Learning based Classification

Related literature that explores the automated classification of acoustic phenomena that is abundant in forest ecosystems has utilized different ML algorithms to carry out the classification task. Among such efforts, ML algorithms like KNN, SVM, and Random forests can commonly be identified. Due to the superiority of the Extreme Gradient Boosting (XGBoost) algorithm against such traditional ML algorithms, this study explores the usability of XGBoost to properly classify forest acoustics.

XGboost is capable of handling non-linear relationships in the features. Handling non-linear relationships are important in sound classification as there are many non-linear relationships between the sound features and the class labels. Moreover, it has the ability of XGBoost to learn from the errors made by previous trees. Additionally, XGBoost use L1 and L2 regularization which is important to reduce overfittings.

The XGBoost library available for python was used to conduct the tests and the model parameters were used to fine-tune the performance of the implemented model. As the final set of parameters, num_class was set to 27, the multiclass classification error rate was used as the eval_metric, subsample, colsample_bytree and min_child_weight was set to 1, max_depth of 6, learning_rate of 0.3 and 100 n_estimators were used. Further, to improve the memory efficiency and the training speed of the XGBoost model, both the training and validation datasets were converted to the internal data structure (DMatrix) used by the model which is optimized for both memory efficiency and training speed. Then the configured model was trained with 80% of the considered dataset, and the evaluation was completed with the remaining 20% of the data using the trained XGBoost model.

5.3. CNN-based classification

Although it can be identified that a substantial number of studies have used ML-based algorithms to classify unstructured data like audio and images, DL based models can outperform the traditional ML models with considerable margins, due to their ability to extract features from raw data [58]. For the study, a Convolutional Neural Network [14,59] based model consisting of 9 layers has been utilized, based on the work of the authors of [36].

Similarly, as in the ML-based approach, 80% of the data were used to train and fine-tune the CNN model, while the remaining 20% was used for the validation procedure. The model was configured to run for 50 epochs; however, an early stopping callback function was used to stop the model from overfitting to the training data. Implementation of the model was completed using the Keras library provided by TensorFlow [60]. Figure 9 presents the architecture of the model accompanied by the parameters used to implement the model using the Keras library.

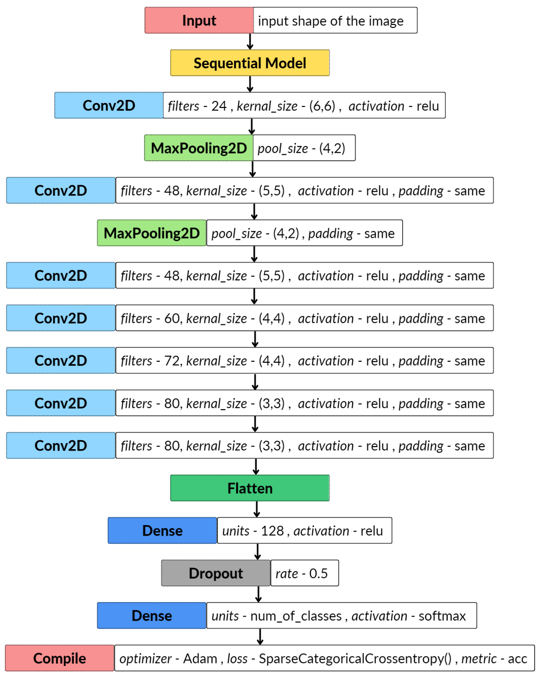

Figure 9. The CNN based architecture of the model

6. Dataset Evaluation

In order to analyze the performance and characteristics of the FSC22 dataset, three major classification experiments were performed. As the first phase, a human classification experiment was conducted to identify a baseline classification accuracy for the FSC22 dataset. An ML and DL based classification of the FSC22 dataset was conducted as the second phase, to generate comparable performance scores respective to related studies. Finally, the same ML and DL models were tested with the ESC50 dataset to present the general performance of the developed models. This section describes each experiment and the results obtained.

6.1. *Human Classification Result of FSC22*

Hearing and identifying sound through the human auditory system depend on a series of complex steps. Scientists have discovered that a form of auditory learning occurs in daily life to help us identify and memorize sound patterns. Hence, when a certain sound pattern differs from a small factor like noisy background, humans find it difficult to recognize the exact sound type. Humans' identification of sound types comes with a high level of uncertainty, which may differ from machine classification. In order to identify this difference in human decisions, a human classification experiment was carried out for the created dataset. For this experiment, 25 participants in age groups 20-30 were selected. The survey includes audio-based questions where the participants were instructed to select the correct label after listening to the sound clips [61]. For the creation of the survey Free Online Survey Software and Tools | The QuestionPro® platform [62] was used. The questionnaire contains two randomly selected audio clips from each class and altogether 54 questions were included for the 27 classes. For each question, 4 choices of labels were given.

After the completion of the survey, an overall accuracy of 91% was observed for the selected audio samples. These survey responses were used to calculate the class-level recognition accuracies. It was identified that the human candidates achieved a maximum classification accuracy of 98% for the classes, Wolf, General Speaking, and Rain, while the two classes, Squirrel and Fire, achieved the lowest accuracies, showing the hardness to identify such sounds by the human auditory system. Figure 10 shows the human classification accuracies obtained for all the classes of the FSC22 dataset.

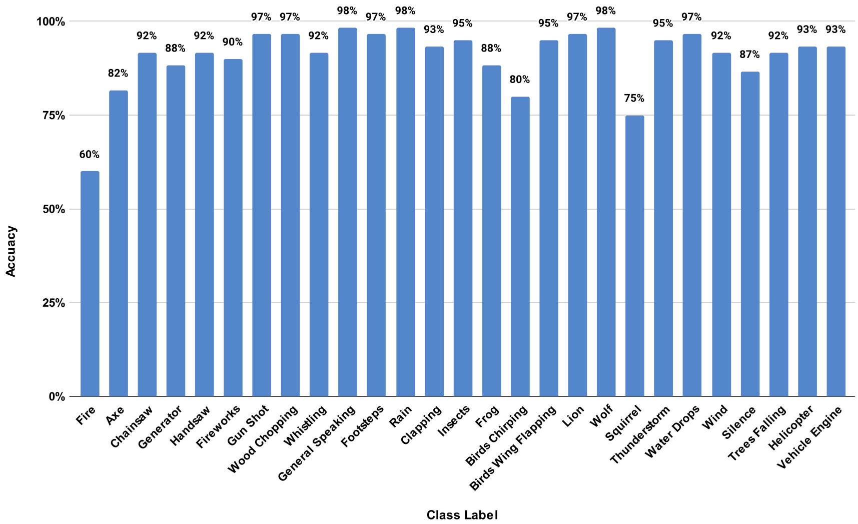

Figure 10. Class accuracies obtained in human classification

6.2. *Baseline Model-based classification results of FSC22*

As the second approach for dataset evaluation, a baseline classification analysis was performed using XGBoost and CNN based models. Section 5 provides a detailed overview of the baseline model selection and classification procedure. With the desired target accuracy results obtained through human classification in subsection 6.1, the next goal is to investigate the level of performance that can be achieved on a baseline model classification.

The baseline XGBoost and CNN based models were evaluated on the FSC22 dataset using the evaluation metrics accuracy, F1-score, precision, and recall. Accuracy is the most intuitive performance measure, and it provides the ratio of the correctly predicted samples to the total samples. While precision provides the ratio of correctly predicted positive samples to the total predicted positive samples and recall gives the ratio of correctly predicted positive samples to all samples in the actual class. The F1-Score is the weighted average of Precision and Recall. The metrics module of the Scikit-learn (Sklearn) library was used to calculate all the metrics, for the precision, recall, and F1-score, averaging was done using the unweighted mean as all the classes were balanced for both datasets. Table 5 and Table 6 provide the summary of results obtained by evaluating the FSC22 dataset against the baseline models XGBoost and CNN-based, respectively.

6.2.1. Results of ML-based Classification

As shown in Table 5, the FSC22 dataset had an average classification accuracy ranging from 48.14% to 62.17% for the selected XGBoost ML model. The highest classification accuracy of 62.71% was reported for the model with the MFCC feature extraction mechanism. In order to better analyze the results, the confusion matrix of the highest accuracy reported approach is displayed in Figure 11. A confusion matrix visualizes and summarizes the performance of a classification algorithm. According to the matrix, it can be identified that the Silence and Bird chirping classes obtained the highest-class level accuracy of 99.58% and 98.84%, respectively. Moreover, the Axe class and Generator class have shown the lowest accuracies among the 27 classes.

Table 5. Results of ML based classification of the FSC22 dataset.

Feature Representation	Augmentation	Accuracy	F1 - Score	Precision	Recall
MFCC	Applied	62.71%	0.62	0.63	0.62
MFCC	Not Applied	55.06%	0.54	0.55	0.55
Mel Spectrogram	Applied	56.04%	0.56	0.57	0.56
Mel Spectrogram	Not Applied	48.14%	0.47	0.48	0.48

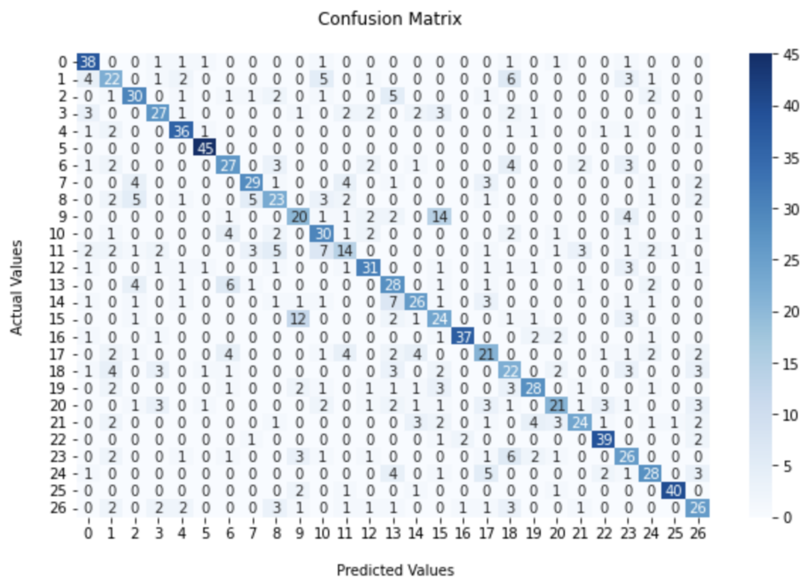

Figure 11. Confusion Matrix for Xgboost based Classification with MFCC for augmented data

6.2.2. Results of CNN-based Classification

When compared with the ML-based classification approach, CNN based classification has shown a significant performance with the FSC22 dataset. As reported in Table 6, the dataset had an average classification accuracy ranging from 53.08% to 92.59% for the CNN model. Out of the four classification accuracies, 92.59% is shown as the highest which is obtained for the CNN model with the MEL feature extraction mechanism. The confusion matrix given in Figure 12 for the approach which has the highest overall accuracy can be used to evaluate the class-level accuracy of the dataset. According to the matrix, it is apparent that almost all the classes have a very high accuracy level, while Generator and Rain classes obtained the lowest among them.

Table 6. Results of CNN based classification of the FSC22 dataset.

Feature Representation	Augmentation	Accuracy	F1 - Score	Precision	Recall
MFCC	Applied	89.30%	0.893	0.898	0.893
MFCC	Not Applied	53,82%	0.533	0.552	0.538
Mel Spectrogram	Applied	92.59%	0.925	0.929	0.925
Mel Spectrogram	Not Applied	53.08%	0.52	0.53	0.53

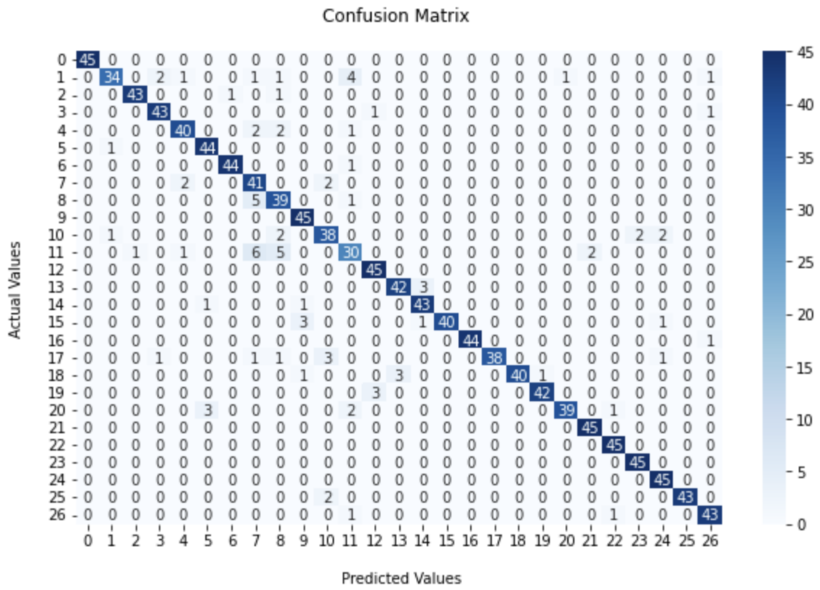

Figure 12. Confusion Matrix for CNN-based classification with Mel Spectrogram for the augmented data

6.3. *Model evaluation results of the ESC-50 dataset*

All the trials conducted with the two feature extraction approaches, for the ML and CNN-based classification of the FSC22 dataset were tested with the ESC50 dataset. All the conducted experiments were evaluated based on the metrics presented in Section 6.2. Table 7 showcases the results obtained with the ML approach, while Table 8 presents the CNN-based classification results. It can be identified that the trial which used data augmentation and the MFCC feature extraction obtained the highest accuracy of 53.25% for the ML-based approach. Moreover, the CNN-based approach which used Mel Spectrogram-based feature extraction supported with data augmentation generated the highest classification accuracy of 92.16%.

Table 7. Results of ML based classification of ESC50 dataset.

Feature Representation	Augmentation	Accuracy	F1 - Score	Precision	Recall
MFCC	Applied	53.25%	0.525	0.529	0.532
MFCC	Not Applied	43.50%	0.431	0.455	0.435
Mel Spectrogram	Applied	48.18%	0.478	0.493	0.481
Mel Spectrogram	Not Applied	31.75%	0.309	0.325	0.317

Table 8. Results of CNN based classification of ESC50 dataset .

Feature Representation	Augmentation	Accuracy	F1 - Score	Precision	Recall
MFCC	Applied	85.41%	0.855	0.866	0.854
MFCC	Not Applied	42.25%	0.408	0.443	0.422
Mel Spectrogram	Applied	92.16%	0.921	0.925	0.921
Mel Spectrogram	Not Applied	44.75%	0.429	0.448	0.447

7. Discussion

7.1. Lessons Learned

We have conducted eight performance comparisons over the FSC22 dataset, as shown in Table 5 and Table 6. These experiments are listed as follows.

- E1: Accuracy of XGBoost model with MFCC and data augmentation
- E2: Accuracy of XGBoost model with MFCC and no data augmentation
- E3: Accuracy of XGBoost model with Mel Spectrogram and data augmentation
- E4: Accuracy of XGBoost model with Mel Spectrogram and no data augmentation)
- E5: Accuracy of CNN model with MFCC and data augmentation
- E6: Accuracy of CNN model with MFCC and no data augmentation
- E7: Accuracy of CNN model with Mel Spectrogram and data augmentation
- E8: Accuracy of CNN model with Mel Spectrogram and no data augmentation

The same experiments were conducted over the ESC50 dataset to further support the observations as shown in Table 7 and Table 8. This section provides a discussion of the observations made after the experiments were completed.

7.1.1. ML vs DL for environmental sound classification

ML and DL techniques have been extensively used in related literature for environmental sound classification. To establish a performance comparison between ML and DL architectures over the FSC22 and ESC50 datasets, eight comparisons were done based on the above defined performance measures. For FSC22, the CNN model outperformed the XGBoost model by a significant margin for all the comparisons, E1 vs E5, E2 vs E6, E3 vs E7 and E4 vs E8. For the ESC50 dataset, CNN based approach outperformed the XGBoost approach in comparisons E1 vs E5, E3 vs E7 and E4 vs E8. The XGBoost outperformed the CNN model when MFCC was used for feature extraction of the non-augmented dataset (E2 vs E6). Careful evaluation of results published under related literature provides similar evidence, to identify that DL algorithms perform better when it comes to complex classification tasks such as audio data tagging. It can be identified that this is due to reasons like the ability of DL algorithms to extract inherent features from the raw data avoiding selective invariance [55], the ability of DL algorithms to learn from large volumes of data [36], and less requirement of feature engineering before the training of the model. Although DL presents high accuracies compared to ML, they need high resources for the training to complete and the resulting models are complex and suffer from low interpretability and explainability [63]. Thus, for proper real-world deployment of a DL-based sound classification system, further research is required to understand and improve the underlying dynamics.

7.1.2. Importance of Data Augmentation techniques

A major requirement to develop proper artificial intelligence models is the availability of large volumes of quality data. When the forest sound classification domain is considered, the availability of well-defined, quality public data is limited. Although the proposed FSC22 dataset provides 2025 audio recordings providing 2.81 hours of record time, the data volume is not sufficient to properly train a CNN, RNN, or ML model to achieve state-of-the-art results. Data augmentation techniques can be successfully used to expand the available data points and to present the significance. As observable by the results of Table 7 and Table 8, the performance of the XGBoost and CNN-based models show a significant improvement in accuracy, when augmentation techniques were employed, compared to the performances obtained without augmentation. The CNN model used with the FSC22 dataset shows accuracy degradations of 40% and 43% for the MFCC-based and Mel Spectrogram-based approaches, respectively when augmentations were not applied. Similarly the XGBoost model shows decrements of 12% and 14% for the two feature extraction approaches MFCC based and Mel Spectrogram, respectively. Accuracy reductions can be identified for the tests conducted with the ESC50 dataset as well. This empirical evidence showcases the importance of using data augmentation techniques when training artificial intelligence algorithms. Although we have successfully implemented baseline data augmentation techniques to increase model performance, further research is required to understand novel techniques that can solve data insufficiency issues while preventing models from overfitting.

7.1.3. Feature Representation Methodology

In the domain of audio classification, extracting feature embeddings that can accurately represent the audio signal is of utmost importance. For the ML and DL models implemented in this study, Mel Spectrograms and MFCC spectrograms were employed as discussed in subsection 5.1.3. With the experiments conducted for both FSC22 and ESC50 datasets using the ML-based approach, it can be identified that the usage of MFCC-based feature extraction outperforms the tests conducted with Mel Spectrograms as the feature representation. However, for the DL-based approach, Mel Spectrogram-based feature extraction provided the highest accuracies, except for the test conducted without augmentation for the FSC22 dataset. Hence, in the context of this study, a clear separation cannot be drawn between the two spectrogram methods, for the task of representing audio signals.

7.2. Comparison with the Existing Sound Datasets

Due to the unavailability of a publicly available benchmark dataset to be used for forest acoustic classification tasks, researchers have utilized different techniques to fulfil their data requirements as explained in subsection 4.3. Table 9 provides a comparison between the results of the existing studies and the highest-performing approach proposed in this paper. Accordingly, it can be seen that this study has utilized the highest number of audio recordings distributed in 27 unique forest acoustic classes while achieving state-of-the-art classification accuracies for the forest sound classification-based studies. However, when the model performances are compared to the state-of-the-art performances achieved for the broader ESC domain, it can be identified that the results published in this paper require further refinement. Therefore as future directions, applying transfer learning using the ImageNet dataset [36,64,65], exploring different data augmentation techniques [32,46,66], and feature representation methodologies [67,68] are suggested by the authors.

Table 9. Comparison of existing datasets

Paper	Model	Amount of data	Types of data used	Feature	Metric	Result
[3]	SVM	The total duration of around 5 min	Chainsaw sounds with background noise	MFCC	accuracy	91.07%
[11]	Random forest	40	Bird sounds, Mammal sounds, Insect sounds from Freesound	Double Features	Average accuracy rates in different environments(Rain, Wind, Traffic, Average)	86.28%
[51]	Cyclic HMM	1418	Animal sounds from HU-ASA database	MFCC	accuracy	64%
[4]	Configuration based on a CNN	280	Chainsaw sounds, Chirping birds, Crackling fire, Crickets, Handsaw, Rain, and Wind extracted from ESC50	MFCC	accuracy	85.37%
[6]	SVM with Log Kernel	3265	Chainsaw Sounds	MFCC	TPR	53.16%
[5]	Feed Forward Network	217	Chainsaw sounds, Vehicle/Engine Sounds, Forest sounds, Urban sounds	Fourier power spectrum coefficients	accuracy	79.50%
[31]	CNN	100	chainsaw	Fourier Spectrogram	accuracy	96%
This Study	CNN	2025	27 Unique classes	Mel Spectrogram	accuracy	92.59%

7.3. Future research directions

This study introduces the FSC22 dataset and proposes a baseline architecture for the classification of forest acoustics. As presented in section 7.2, the developed CNN based classification model outperforms existing forest acoustics classifier systems. Authors identify following directions for the reference of researchers working in the forest acoustics domain.

7.3.1. Practical deployment of forest acoustic classification systems

Forest acoustic classification systems can provide valuable information to protect forest reserves from natural and artificial phenomena. A practical implementation will require the classification model to be deployed in a resource constrained edge device, which will be challenging. The best performing CNN model proposed in this study contains 4.6 million parameters and to be deployed in an edge device, complexity needs to be reduced substantially. Techniques like pruning, XNOR-NET and bottleneck layers can be effectively used to reduce the model complexities, but will reduce the model performance by a significant amount [41]. Hence future work is required to identify methodologies to generate reduced complexity models for FSC while preserving the classification accuracy on a reasonable scale.

7.3.2. Explainability and interpretability of FSC models

Explainability and the interpretability of machine learning models is an emerging domain which presents interesting effects to the way that ML models are utilised. Explainability refers to the ability of a learning model to provide human-understandable explanations for its predictions. Interpretability refers to the ability to understand the internal workings of a model and how it arrives at its predictions. Forest sound classification systems with the potential of being deployed in the forest ecosystems to help authorities, can greatly benefit from a transparent classification model. Amount of studies covering the explainability and the interpretability of ESC or FSC models is scarce. Thus authors recommend future researchers working in this domain to contribute to develop more explainable and interpretable audio classification models.

Apart from the above two major research directions, it can be identified that state of the art ESC models have comparatively high performance measures with respective to identified FSC models including the CNN based model proposed in this study. State of the art ESC models have utilised techniques like transfer learning from the Imagenet dataset, multiple aggregated feature representations, multiple data augmentation strategies to achieve very high performance measures. Therefore the authors recommend future researchers under FSC to explore such techniques and utilise them to improve the performance measures of current FSC models to a comparable scale.

8. Conclusion

Environment sound classification (ESC) using artificial intelligence is a prominent research area in audio recognition. Under ESC, forest sound classification (FSC), which focuses on identifying artificial and natural phenomena observable in forest ecosystems, receives a high research interest. Recognition of forest sounds generates highly valuable use cases when scenarios like illegal logging, poaching, and wildfires are considered. FSC suffers from the unavailability of a standard sound taxonomy and the unavailability of a sufficiently large public benchmark dataset. With the intention of resolving both issues, this study presents the FSC22 Taxonomy and the first version of the FSC22 dataset. The first version of the FSC22 dataset consists of 2025 human-annotated, 5-second-long audio recordings equally distributed into 27 unique classes. The authors intend to expand the first version of the FSC22 dataset in the future, capturing more acoustic classes according to the FSC22 taxonomy. Further, the study presents CNN-based and XGBoost-based classification experiments using the FSC22 dataset. CNN-based approach achieved a maximum classification accuracy of 92.59%, while the XGBoost model achieved a maximum accuracy of 62.71%. A survey conducted with 25 human candidates to identify different sounds from the classes listed in the FSC22 dataset was also conducted to establish a baseline accuracy score. Finally, the authors believe that the proposed FSC22 taxonomy, the created FSC22 V1.0 dataset, experiments conducted, and the discussions provided through this study will support future research work governing the FSC domain.

Author Contributions:

Conceptualization, D.M. and C.P.; methodology, M.B., R.J. and I.A.; software, M.B., R.J. and I.A.; validation, M.B., R.J. and I.A.; formal analysis, M.B., R.J. and I.A.; investigation, M.B., R.J. and I.A.; data curation, M.B., R.J. and I.A.; writing—original draft preparation, M.B., R.J. and I.A.; writing—review and editing, D.M., C.P; visualization, R.J. and I.A.; supervision, D.M. and C.P; project administration, D.M. and C.P; All authors have read and agreed to the published version of the manuscript.

Funding: Charith Perera's work is partially funded by grant number EP/W524219/1

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: "FSC22 Dataset", IEEE Dataport, doi: https://dx.doi.org/10.21227/40ds-0z76

Acknowledgments: Authors convey their gratitude to Mr. Savindu Rajapaksha, Mr. Charitha Dombawala and Mr. Sandaruth Siriwardana for the provided technical support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, C.; Zhan, H.; Hao, Z.; Gao, X. Classification of Complicated Urban Forest Acoustic Scenes with Deep Learning Models. *Forests* **2023**, *14*. https://doi.org/10.3390/f14020206.
2. Anđelić, B.; Radonjić, M.; Djukanović, S. Sound-based logging detection using deep learning. In Proceedings of the 2022 30th Telecommunications Forum (TELFOR), 2022, pp. 1–4. https://doi.org/10.1109/TELFOR56187.2022.9983766.
3. Mporas, I.; Perikos, I.; Kelefouras, V.; Paraskevas, M. Illegal Logging Detection Based on Acoustic Surveillance of Forest. *Applied Sciences* **2020**, *10*. https://doi.org/10.3390/app10207379.
4. Andreadis, A.; Giambene, G.; Zambon, R. Monitoring Illegal Tree Cutting through Ultra-Low-Power Smart IoT Devices. *Sensors* **2021**, *21*. https://doi.org/10.3390/s21227593.
5. Segarceanu, S.; Olteanu, E.; Suciu, G. Forest Monitoring Using Forest Sound Identification. In Proceedings of the 2020 43rd International Conference on Telecommunications and Signal Processing (TSP), 2020, pp. 346–349. https://doi.org/10.1109/TSP49548.2020.9163433.
6. Gnamélé, N.A.J.; Ouattara, y.B.; Kobea, T.A.; Baudoin, G.; Laheurte, J.M. KNN and SVM Classification for Chainsaw Identification in the Forest Areas. *International journal of advanced computer science and applications (IJACSA)* **2019**, *10*. https://doi.org/10.14569/IJACSA.2019.0101270.
7. Oo, M.M. Comparative Study of MFCC Feature with Different Machine Learning Techniques in Acoustic Scene Classification. *International Journal of Research and Engineering* **2018**, *5*, 439–444. https://doi.org/10.21276/ijre.2018.5.7.1.
8. Jin, W.; Wang, X.; Zhan, Y. Environmental Sound Classification Algorithm Based on Region Joint Signal Analysis Feature and Boosting Ensemble Learning. *Electronics* **2022**, *11*. https://doi.org/10.3390/electronics11223743.
9. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; et al. Xgboost: extreme gradient boosting. *R package version 0.4-2* **2015**, *1*, 1–4.
10. Olteanu, E.; Suciu, V.; Segarceanu, S.; Petre, I.; Scheianu, A. Forest Monitoring System Through Sound Recognition. In Proceedings of the 2018 International Conference on Communications (COMM), 2018, pp. 75–80. https://doi.org/10.1109/ICComm.2018.8484773.
11. Li, Y.; Wu, Z. Animal sound recognition based on double feature of spectrogram in real environment. In Proceedings of the 2015 International Conference on Wireless Communications & Signal Processing (WCSP), 2015, pp. 1–5. https://doi.org/10.1109/WCSP.2015.7341003.
12. Bansal, A.; Garg, N.K. Environmental Sound Classification: A descriptive review of the literature. *Intelligent Systems with Applications* **2022**, *16*, 200115. https://doi.org/https://doi.org/10.1016/j.iswa.2022.200115.
13. Cerutti, G.; Prasad, R.; Brutti, A.; Farella, E. Compact Recurrent Neural Networks for Acoustic Event Detection on Low-Energy Low-Complexity Platforms. *IEEE Journal of Selected Topics in Signal Processing* **2020**, *14*, 654–664. https://doi.org/10.1109/JSTSP.2020.2969775.
14. Piczak, K.J. Environmental sound classification with convolutional neural networks. In Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), 2015, pp. 1–6. https://doi.org/10.1109/MLSP.2015.7324337.
15. Elliott, D.; Otero, C.E.; Wyatt, S.; Martino, E. Tiny Transformers for Environmental Sound Classification at the Edge. *ArXiv* **2021**, *abs/2103.12157*. https://doi.org/10.48550/ARXIV.2103.12157.
16. Mohaimenuzzaman, M.; Bergmeir, C.; West, I.T.; Meyer, B. Environmental Sound Classification on the Edge: Deep Acoustic Networks for Extremely Resource-Constrained Devices. *ArXiv* **2021**, *abs/2103.03483*. https://doi.org/10.48550/ARXIV.2103.03483.
17. Huzaifah, M. Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks. *ArXiv* **2017**, *abs/1706.07156*. https://doi.org/10.48550/ARXIV.1706.07156.
18. Elliott, D.; Martino, E.; Otero, C.E.; Smith, A.; Peter, A.M.; Luchterhand, B.; Lam, E.; Leung, S. Cyber-Physical Analytics: Environmental Sound Classification at the Edge. In Proceedings of the 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), 2020, pp. 1–6. https://doi.org/10.1109/WF-IoT48130.2020.9221148.
19. Shah, S.K.; Tariq, Z.; Lee, Y. IoT based Urban Noise Monitoring in Deep Learning using Historical Reports. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 4179–4184. https://doi.org/10.1109/BigData47090.2019.9006176.

20. Mushtaq, Z.; Su, S.F. Efficient Classification of Environmental Sounds through Multiple Features Aggregation and Data Enhancement Techniques for Spectrogram Images. *Symmetry* **2020**, *12*. https://doi.org/10.3390/sym12111822.
21. K., B.; D., S.P. MFCC based hybrid fingerprinting method for audio classification through LSTM. *International Journal of Nonlinear Analysis and Applications* **2022**, *12*, 2125–2136. https://doi.org/10.22075/ijnaa.2022.6049.
22. Mkrtchian, G.; Furletov, Y. Classification of Environmental Sounds Using Neural Networks. In Proceedings of the 2022 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO), 2022, pp. 1–4. https://doi.org/10.1109/SYNCHROINFO55067.2022.9840922.
23. Gong, Y.; Khurana, S.; Rouditchenko, A.; Glass, J. CMKD: CNN/Transformer-Based Cross-Model Knowledge Distillation for Audio Classification **2022**. https://doi.org/10.48550/ARXIV.2203.06760.
24. Cartwright, M.; Cramer, J.; Salamon, J.; Bello, J.P. Tricycle: Audio Representation Learning from Sensor Network Data Using Self-Supervision. In Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2019, pp. 278–282. https://doi.org/10.1109/WASPAA.2019.8937265.
25. Adapa, S. Urban Sound Tagging using Convolutional Neural Networks. In Proceedings of the Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), 2019, pp. 5–9. https://doi.org/10.33682/8axe-9243.
26. Nanni, L.; Maguolo, G.; Paci, M. Data augmentation approaches for improving animal audio classification. *ArXiv* **2020**, *abs/1912.07756*. https://doi.org/10.48550/ARXIV.1912.07756.
27. Chang, C.Y.; Chang, Y.P. Application of abnormal sound recognition system for indoor environment. In Proceedings of the 2013 9th International Conference on Information, Communications & Signal Processing, 2013, pp. 1–5. https://doi.org/10.1109/ICICS.2013.6782772.
28. Zhao, Z.; hua Zhang, S.; yong Xu, Z.; Bellisario, K.; hua Dai, N.; Omrani, H.; Pijanowski, B.C. Automated bird acoustic event detection and robust species classification. *Ecological Informatics* **2017**, *39*, 99–108. https://doi.org/10.1016/j.ecoinf.2017.04.003.
29. Zhang, S.h.; Zhao, Z.; Xu, Z.y.; Bellisario, K.; Pijanowski, B.C. Automatic Bird Vocalization Identification Based on Fusion of Spectral Pattern and Texture Features. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 271–275. https://doi.org/10.1109/ICASSP.2018.8462156.
30. Ágnes, I.; Jancsó, H.B.; Zoltán, S.; Attila, F.; Csaba, S. Bird Sound Recognition Using a Convolutional Neural Network. In Proceedings of the 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY), 2018, pp. 295–300. https://doi.org/10.1109/SISY.2018.8524677.
31. Kalhara, P.G.; Jayasinghearachchi, V.D.; Dias, A.H.A.T.; Ratnayake, V.C.; Jayawardena, C.; Kuruwitaarachchi, N. 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), 2017, pp. 1–7. https://doi.org/10.1109/SKIMA.2017.8294127.
32. Ryan, P.; Takafuji, S.; Yang, C.; Wilson, N.; McBride, C. Using Self-Supervised Learning of Birdsong for Downstream Industrial Audio Classification. In Proceedings of the ICML 2020 Workshop on Self-supervision in Audio and Speech, 2020.
33. Bandara, M.; Jayasundara, R.; Ariyarathne, I.; Meedeniya, D.; Perera, C. FSC22 Dataset. Accessed: Sep 27, 2022, https://doi.org/10.21227/40ds-0z76.
34. Piczak, K.J. ESC-50: Dataset for Environmental Sound Classification. https://github.com/karolpiczak/ESC-50. Accessed: July 20, 2022.
35. Salamon, J.; Jacoby, C.; Bello, J.P. URBAN SOUND DATASETS. https://urbansounddataset.weebly.com/urbansound8k.html. Accessed: July 20, 2022.
36. Mushtaq, Z.; Su, S.F.; Tran, Q.V. Spectral images based environmental sound classification using CNN with meaningful data augmentation. *Applied Acoustics* **2021**, *172*, 107581. https://doi.org/10.1016/j.apacoust.2020.107581.
37. Salamon, J.; Jacoby, C.; Bello, J.P. A Dataset and Taxonomy for Urban Sound Research. In Proceedings of the Proceedings of the 22nd ACM International Conference on Multimedia, 2014, p. 1041–1044. https://doi.org/10.1145/2647868.2655045.
38. Fonseca, E.; Favory, X.; Pons, J.; Font, F.; Serra, X. FSD50K: An Open Dataset of Human-Labeled Sound Events. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* **2022**, *30*, 829–852. https://doi.org/10.1109/TASLP.2021.3133208.

39. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. 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), 2017, pp. 776–780. https://doi.org/10.1109/ICASSP.2017.7952261.
40. Cartwright, M.; Cramer, J.; Mendez, A.E.M.; Wang, Y.; Wu, H.H.; Lostanlen, V.; Fuentes, M.; Dove, G.; Mydlarz, C.; Salamon, J.; et al. SONYC urban sound tagging (SONYC-UST): A Multilabel dataset from an Urban Acoustic Sensor Network. https://doi.org/10.5281/zenodo.3966543. Accessed: July 20, 2022.
41. AudioSet. https://research.google.com/audioset/download.html. Accessed: July 20, 2022.
42. Fonseca, E.; Favory, X.; Pons, J.; Font, F.; Serra, X. FSD50K. https://doi.org/10.5281/zenodo.4060432. Accessed: July 20, 2022.
43. xeno-canto archive. https://xeno-canto.org/. Accessed: Jan 06, 2023.
44. Chalmers, C.; Fergus, P.; Wich, S.; Longmore, S.N. Modelling Animal Biodiversity Using Acoustic Monitoring and Deep Learning. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), 2021, pp. 1–7. https://doi.org/10.1109/IJCNN52387.2021.9534195.
45. Elhami, G.; Weber, R.M. Audio Feature Extraction with Convolutional Neural Autoencoders with Application to Voice Conversion. In Proceedings of the Infoscience EPFL scientific publications, 2019.
46. Nanni, L.; Maguolo, G.; Brahnam, S.; Paci, M. An Ensemble of Convolutional Neural Networks for Audio Classification. *Applied Sciences* **2021**, *11*. https://doi.org/10.3390/app11135796.
47. Lasseck, M. Audio-based Bird Species Identification with Deep Convolutional Neural Networks **2018**. p. 2125.
48. Palanisamy, K.; Singhania, D.; Yao, A. Rethinking CNN Models for Audio Classification. *ArXiv* **2020**, *abs/2007.11154*. https://doi.org/10.48550/ARXIV.2007.11154.
49. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. https://doi.org/10.1126/science.aaa8415.
50. Huang, Z.; Liu, C.; Fei, H.; Li, W.; Yu, J.; Cao, Y. Urban sound classification based on 2-order dense convolutional network using dual features. *Applied Acoustics* **2020**, *164*, 107243. https://doi.org/10.1016/j.apacoust.2020.107243.
51. Weninger, F.; Schuller, B. 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), 2011, pp. 337–340. https://doi.org/10.1109/ICASSP.2011.5946409.
52. Segarceanu, S.; Suciu, G.; Gavat, I. Neural Networks for Automatic Environmental Sound Recognition. In Proceedings of the 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2021, pp. 7–12. https://doi.org/10.1109/SpeD53181.2021.9587378.
53. Ting, P.J.; Ruan, S.J.; Li, L.P.H. Environmental Noise Classification with Inception-Dense Blocks for Hearing Aids. *Sensors* **2021**, *21*. https://doi.org/10.3390/s21165406.
54. Madhu, A.; Kumaraswamy, S.K. EnvGAN: Adversarial Synthesis of Environmental Sounds for Data Augmentation. *ArXiv* **2021**, *abs/2104.07326*. https://doi.org/10.48550/ARXIV.2104.07326.
55. Chauhan, N.K.; Singh, K. A Review on Conventional Machine Learning vs Deep Learning. In Proceedings of the 2018 International Conference on Computing, Power and Communication Technologies (GUCON), 2018, pp. 347–352. https://doi.org/10.1109/GUCON.2018.8675097.
56. Wei, S.; Zou, S.; Liao, F.; weimin lang. A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification. *Journal of Physics: Conference Series* **2020**, *1453*, 012085. https://doi.org/10.1088/1742-6596/1453/1/012085.
57. Li, J.B.; Qu, S.; Huang, P.Y.; Metze, F. AudioTagging Done Right: 2nd comparison of deep learning methods for environmental sound classification. *ArXiv* **2022**, *abs/2203.13448*. https://doi.org/10.48550/ARXIV.2203.13448.
58. Setiowati, S.; Zulfanahri.; Franita, E.L.; Ardiyanto, I. 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), 2017, pp. 1–6. https://doi.org/10.1109/ICITEED.2017.8250484.
59. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems* **2022**, *33*, 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827.
60. Tensorflow. https://www.tensorflow.org/. Accessed: Jan 06, 2023.

61. Forest Sound Data Survey. https://questionpro.com/t/AWEf8ZuUxQ. Accessed: Oct 05, 2022.
62. QuestionPro. https://www.questionpro.com/. Accessed: Jan 06, 2023.
63. Zinemanas, P.; Rocamora, M.; Miron, M.; Font, F.; Serra, X. An Interpretable Deep Learning Model for Automatic Sound Classification. *Electronics* **2021**, *10*. https://doi.org/10.3390/electronics10070850.
64. Nasiri, A.; Hu, J. SoundCLR: Contrastive Learning of Representations For Improved Environmental Sound Classification. *ArXiv* **2021**, *abs/2103.01929*. https://doi.org/10.48550/ARXIV.2103.01929.
65. Bhat, K.M.; Bhandari, M.; Oh, C.; Kim, S.; Yoo, J. Transfer Learning Based Automatic Model Creation Tool For Resource Constraint Devices. *ArXiv* **2020**, *abs/2012.10056*. https://doi.org/10.48550/ARXIV.2012.10056.
66. Tripathi, A.M.; Mishra, A. Self-supervised learning for Environmental Sound Classification. *Applied Acoustics* **2021**, *182*, 108183. https://doi.org/10.1016/j.apacoust.2021.108183.
67. Peng, N.; Chen, A.; Zhou, G.; Chen, W.; Zhang, W.; Liu, J.; Ding, F. Environment Sound Classification Based on Visual Multi-Feature Fusion and GRU-AWS. *IEEE Access* **2020**, *8*, 191100–191114. https://doi.org/10.1109/ACCESS.2020.3032226.
68. Das, J.K.; Arka, G.; Kumar, P.A.; Sumit, D.; Amitabha, C. Urban Sound Classification Using Convolutional Neural Network and Long Short Term Memory Based on Multiple Features. In Proceedings of the 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS), 2020, pp. 1–9. https://doi.org/10.1109/ICDS50568.2020.9268723.