Intelligent Systems with Applications 16 (2022) 200115

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Environmental Sound Classification: A descriptive review of the literature

Anam Bansal [*,1,a], Naresh Kumar Garg [2,b]

[a] Research Scholar, Computer Science and Engineering, GZS Campus College of Engineering and Technology, Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India

[b] Professor, Computer Science and Engineering, GZS Campus College of Engineering and Technology, Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India

ARTICLE INFO

Keywords:
Environmental Sound Classification
Feature extraction
Feature selection
Machine learning classifiers
Deep neural networks

ABSTRACT

Automatic environmental sound classification (ESC) is one of the upcoming areas of research as most of the traditional studies are focused on speech and music signals. Classifying environmental sounds such as glass breaking, helicopter, baby crying and many more can aid in surveillance systems as well as criminal investigations. In this paper, a vast range of literature in the field of ESC is elucidated from various facets like preprocessing, feature extraction, and classification techniques. Researchers have used various noise removal and signal enhancement techniques to preprocess the signals. This paper explicates multitude of datasets used in recent studies along with the year of publication and maximum accuracy achieved with the dataset. Deep Neural Networks surpass the traditional machine learning classifiers. The future challenges and prospective research in this field are proposed. Since no recent review on ESC has been published, this study will open up novel ways for certain business applications and security systems.

1. Introduction

Context Recognition using audio is one of the most prevalent research areas. Sounds, technically called acoustics help to recognize the environment(Fan et al., 2020), speech(Bhat et al., 2020), and music (Elbir and Aydin, 2020), Virtanen and Helén. Acoustics can be used to recognize events and scenes. Acoustic scene recognition (Plata, 2019) is identifying and classifying the scenes such as offices(Hossain and Muhammad, 2018), parks, hospitals, and buses. Acoustic event classification (Sharan and Moir, 2019) is recognizing temporary changes in ongoing acoustic scenes such as dog barking, gunshot, door knock, and engine sounds. ESC is acoustic event classification in which various activities in the surroundings are identified and classified so that certain applications can be activated.

Most of the past studies on acoustics have primarily focused on speech and music. In a study (Duan et al., 2014), the tagging techniques for speech, music, and environment sounds are surveyed. ESC is complex as compared to speech and music classification (Mushtaq et al., 2021). The main reason behind it is that environmental sounds are non-static and do not have a particular structure. Speech recognition models complex words by breaking them down into phonemes. Environmental sounds do not have a phonetic structure. Also, environmental sounds, unlike music do not have stationary aspects such as rhythm and melody (Chachada and Kuo, 2014). Environmental sounds have a low signal-to-noise ratio as the microphone or the source capturing sounds are not placed exactly near to the sound production. The environmental scene consists of numerous overlapping sounds which pose a problem for ESC (Chandrakala and Jayalakshmi, 2019). Though the video cameras can also be used for environmental scene recognition, cameras are not omnidirectional as microphones. The audios are less prone to errors as compared to video cameras (Crocco et al., 2016).

Research in ESR has increased tremendously focussing on different aspects of ESR. ESC involves the collection of data, preprocessing, feature extraction, and feature selection and classification of the data (Fig. 1). Customizing any of the stages in ESC and introducing new methods can help significantly improve the performance. There are few standard datasets that the researchers have mostly used in their studies. Preprocessing is essential to remove background noises and process the data in the form that it can be used for feature extraction. Researchers have worked with different kinds of features such as temporal, spectral,

* Corresponding author. Research Scholar, Computer Science and Engineering, GZS Campus College of Engineering and Technology, Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India. Tel.: 8437800233.
E-mail addresses: anambansal19@gmail.com (A. Bansal), naresh2834@rediffmail.com (N.K. Garg).
[1] Research Scholar
[2] Professor

https://doi.org/10.1016/j.iswa.2022.200115
Received 12 April 2022; Received in revised form 4 August 2022; Accepted 14 August 2022
Available online 20 August 2022

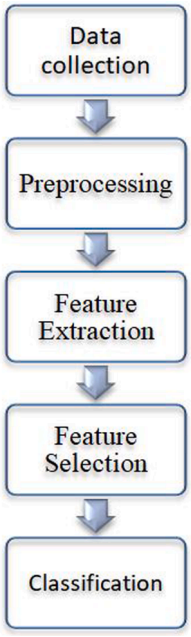

Fig. 1. Stages of Environmental Sound Classification.

and dynamic time wrapping features. The important features can be selected to reduce the number of features. Various feature selection approaches are used by researchers. Finally, the different researches show that machine learning classifiers such as Support Vector Machines (SVM), K- Nearest Neighbour (K-NN), Decision Trees, and Hidden Markov Models (HMM) are used for ESC. Novel neural networks are used extensively in ESR recently. Convolutional Neural Networks (CNN), Multilayer Perceptron(MLP), Deep Neural Networks, and Recurrent Neural Networks (RNN) open up new doors in ESC.

1.1. Motivation

Recognizing environmental sounds can aid in several applications. Smart homes (Vafeiadis et al., 2017) can be developed that can assist elders staying at home (Saraubon et al., 2018). The audio surveillance system works under the concept of environmental sound recognition (ESR)(Chandrakala and Jayalakshmi, 2019; Rabaoui et al., 2008). ESR also finds application in Robot Navigation (Aziz et al., 2019; Tsunoda et al., 2019; Yamakawa et al., 2011). ESR can be customized for detecting criminal activities and can be used in various other security systems. Wildlife monitoring like the classification of birds(Tuncer et al., 2021), animals(Kim et al., 2020), frogs(Brodie et al., 2020), and bats (Mac Aodha et al., 2018) is made possible using ESR. ESR can be incorporated to develop noise monitoring systems so that operational policies can be made(Mydlarz et al., 2017). Recently, spectral temporal analysis of hive sounds has helped in monitoring hive health(Soares et al., 2022).

1.2. Significant Contributions of Survey work

Cowling and Sitte are the first ones to survey the techniques for ESC (Cowling and Sitte, 2003). The most recent survey conducted for ESR was in 2014 by Chachada and Kuo (2014). To our knowledge, there is no other survey article related to ESC. The review articles in the field of ESC have not distinguished the techniques used in different phases of ESC. The study in this paper explains every phase in the process of ESC in detail. The datasets used, techniques used for preprocessing, different feature extraction techniques and audio features, and classification techniques used by researchers in the past are illustrated in different sections. The most recent study on ESR is using spectrogram along with data augmentation (Mushtaq et al., 2021).

In this review, the updated techniques and research done by researchers in the field of ESR and ESC are described. Section II describes the datasets available and data collection techniques used by the researchers in past studies. Section III describes the preprocessing techniques. Section IV and Section V focus on different types of features used by researchers and various feature selection techniques respectively. Section VI details the classifiers and experimental results in the past studies. Finally, Section VII analyzes different parameters used by main references in work. In the end, the future challenges along with the concluding remarks are presented.

2. Datasets

The first stage in ESC is data acquisition. In this section, both public and self-collected datasets used by researchers for ESC are discussed. Most of the researchers have used three publicly available datasets- ESC-10(Piczak, 2015b), ESC-50(Piczak, 2015b), and UrbanSound8k (Salamon et al., 2014). ESC-10 dataset comprises 400 recordings of 10 classes (dog bark, rain, sea waves, baby cry, clock tick, person sneeze, helicopter, chainsaw, rooster, fire crackling). Each class has 40 recordings and each recording is of 5 seconds. ESC-50 dataset comprises 2000 recordings from 50 different classes. 50 classes are placed in 5 major groups- animals, natural soundscapes and water sounds, human non-speech sounds, interior/domestic sounds, and exterior/urban noises. Each class has 40 recordings of 5 seconds each. ESC-10 is a subset of ESC-50. UrbanSound8k dataset consists of 8732 labeled sound recordings from 10 classes - air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, and street music. The duration of each recording is less than 4 seconds. These sounds from the free sound repository(Font et al., 2013) are filtered and labeled manually to generate the urbansound8k dataset for ESR. Another dataset for ESR is BDLib which is collected by authors in Bountourakis et al. (2015)from Sony Pictures Sounds Effect Series, BBC Complete Sound Effects Library, and a free sound repository (Font et al., 2013). BDLib dataset has 120 recordings from 10 categories. Each class has 10 audio files of ten seconds each. In Zhang et al. (2017), the CICESE dataset is used which contains 7 classes of indoor sound events of duration 14 minutes.

Few other researchers have used self-collected datasets containing the different numbers of recordings. Researchers need to contact the dataset owners for conducting further research. A dataset of 22 recordings is collected to detect human activity from the background environmental sounds using a wearable device(Zhan and Kuroda, 2014). In Uzkent et al. (2012), 258 non-speech recordings from seven different classes are used for surveillance. A dataset organized into 8 classes is collected from sources like internet recordings and BBC Sound Effects Library(Ntalampiras et al., 2010). The classes are train(82), motorcycle (79), thunder (60), wind (66), aircraft (110), crowd (60), car (81), and horns (194). A database of 1325 recordings belonging to 61 different classes is used in Gencoglu et al. (2014). This database was first introduced in Mesaros et al. (2010). In study (Rabaoui et al., 2008) and (Zhang et al., 2015), 1015(9 classes) and 4000 audio files are self-recorded respectively. A self-recorded dataset of 3500 samples from 15 different classes divided into five categories are used by researchers in Valero and Alías (2012a). Few other researchers used 1000 audio files from 10 classes(Muhammad et al., 2010) and 128 audio files from six classes(Han and Hwang, 2009). Table 1 lists the datasets used in past studies in the field of ESC.

The problem in ESR is that there is no consolidated dataset for the benchmark. As clear from table 1, most of the datasets are with limited samples due to which deep neural networks can not attain good accuracy. Several data augmentation techniques are used for increasing the number of audio samples artificially (Mushtaq and Su, 2020; Mushtaq et al., 2021). In past studies, time-stretching, linear interpolation, Pitch Shifting, Dynamic Range Compression (DRC), and Background noise

Table 1
Datasets used in literature.

Dataset	Dataset publication year	Data	Papers	Highest Accuracy Obtained
ESC-10	2015	400 recordings from 10 classes	(Ahmed et al., 2020; Boddapati et al., 2017; Khamparia et al., 2019; Li et al., 2018; Mushtaq and Su, 2020; Mushtaq et al., 2021; Piczak, 2015a; Tokozume et al., 2017)	99.04% (Mushtaq et al., 2021)
ESC-50	2015	2000 recordings from 50 classes	(Ahmed et al., 2020; Boddapati et al., 2017; Chi et al., 2019; Khamparia et al., 2019; Li et al., 2018; Mushtaq and Su, 2020; Mushtaq et al., 2021; Piczak, 2015a; Tokozume and Harada, 2017; Tokozume et al., 2017; Yao et al., 2019; Zhang et al., 2017)	97.57% (Mushtaq et al., 2021)
UrbanSound8k	2014	8732 recordings from 10 classes	(Ahmed et al., 2020; Boddapati et al., 2017; Chi et al., 2019; Dai et al., 2017; Demir et al., 2020; Li et al., 2018; Mendoza et al., 2018; Mushtaq and Su, 2020; Mushtaq et al., 2021; Piczak, 2015a; Salamon and Bello, 2017; Sang et al., 2018; Su et al., 2019; Tokozume et al., 2017; Zhang et al., 2017)	99.49%(Mushtaq et al., 2021)
BDLib	2015	120 recordings of 12 classes	(Bountourakis et al., 2015)	54.01%
CICESE	2017	14 minutes recordings from 7 classes of indoor events	(Zhang et al., 2017)	87.10%
Self Collected	2014	22 recordings from 22 personal and social activities	(Zhan and Kuroda, 2014)	96.90%
Self Collected	2012	258 recordings from 7 different classes(glass break, dog bark, gunshot, scream, engine, rain, restaurant)	(Uzkent et al., 2012)	88.70%
Self Collected from Internet Sources and BBC sounds effects library	2010	732 recordings from 8 classes	(Ntalampiras et al., 2010)	93.00%
Self Collected	2010	1325 recordings from 61 classes such as sneezing, dog barking, clapping, car door, beep, yelling	(Gencoglu et al., 2014; Mesaros et al., 2010)	60.30%(Gencoglu et al., 2014)
Self Collected	2008	1015 recordings from 9 classes	(Rabaoui et al., 2008)	96.89%
Self Collected	2015	4000 recordings from 50 classes	(Zhang et al., 2015)	97.53%
Self Collected	2012	3500 samples from 15 different classes	(Valero and Alías, 2012a)	91.00%
Self collected	2010	1000 audio files from 10 classes	(Muhammad et al., 2010)	96.00%
Self Collected	2009	258 audio files from 6 classes	(Han and Hwang, 2009)	86.09%

(BG) are used to generate the fake data to prevent overfitting(Salamon and Bello, 2017; Zhang et al., 2017). Using the data augmentation techniques, the accuracy of deep NNs is increased(Salamon and Bello, 2017).

3. Preprocessing

Preprocessing is required to remove the noise or enhance and smoothen the audio signal. The complexity while collecting the environmental sounds is that there can be certain noises that can be recorded. The accuracy of ESC is affected due to noise. So the audio signals need to be preprocessed so that they are ready to be used for feature extraction or classification. In this section, the preprocessing techniques employed by researchers in the field of ESC are discussed.

Silence is considered as noise and amplitude-based silence detection algorithm is used for preprocessing the sound signals(Ntalampiras et al., 2010). Dimensionality Reduction(Van Der Maaten et al., 2009) is the preprocessing technique to reduce the size of arbitrarily long spectrograms. The spectrograms are smoothened and denoised using this preprocessing technique(Zhang et al., 2015). Noise is removed by using certain signal enhancement techniques. In Wang et al. (2008), audio signals are enhanced using a perceptual filterbank and subspace-based method.

4. Feature Extraction

Features are the distinct characteristics of the sounds that are

extracted and fed to machine learning classifiers. It is one of the most important steps in ESC. In literature, different feature extraction techniques are used. Different features and feature extraction techniques used by researchers for the study of ESC are illustrated in detail in this section. The performance of ESR depends widely on the type of features extracted.

Various researchers investigated the audio features in detail(Alías et al., 2016; Mitrović et al., 2010; Sharma et al., 2020). The features for sound classification are basically characterised in four categories-cepstral features(Aziz et al., 2020; Bansal et al., 2018), temporal features (Yang and Krishnan, 2017), spectral features(Ma et al., 2018) and image-based features(Amiriparian et al., 2017)(Fig. 2).

4.1. Cepstral Features

Mel Frequency Cepstral Coefficients(MFCC) have been extensively used in audio classification in fields of music(Logan et al., 2000), speech (Palo et al., 2018) and environment(Sharma et al., 2019; Zhang et al., 2015). MFCCs are computed by first calculating the Fourier transform of the audio signal, mapping the powers to the mel scale, computing the log of the powers, and applying discrete cosine transformation on the mel log scales. The amplitude of these spectrums is called MFCC. Audio signals are cepstrally represented using MFCCs. MFCCs are used extensively in the ESC(Chu et al., 2006; Gencoglu et al., 2014; Ntalampiras et al., 2010; Sigtia et al., 2016). In Wang et al. (2008), variation of MFCC called Independent Component Analysis(ICA) transformed MFCCS are used and they provide sustainable gain in performance. Researchers claim that MFCCs are incapable if the audio signals are noisy(Ahmed et al., 2020) and MFCCs can not reflect the non-stationary properties of environmental sounds (Uzkent et al., 2012). Code Excited Linear Prediction(CELP) based features outperforms MFCCs for ESC(Tsau et al., 2011). The combination of CELP-based features and MFCCs helps in attaining 95.1% accuracy.

4.2. Temporal Features

Temporal features are also called time-domain features. These are extracted directly from sounds. Zero-Crossing Rate(ZCR), autocorrelation, Linear Predictive Coding, Energy Entropy(EE), Short Time Energy (STE), and Root Mean Square(RMS) are a few of the features that fall under the temporal domain. ZCR is the frequency of sign changes of the signal. ZCR, energy range are applied in the ESR domain as described in a study (Chu et al., 2006). Narrowband autocorrelation features (NB-ACF) can attain higher accuracy as compared to MFCCs and Discrete Wavelet Coefficients(Valero and Alías, 2012a). Linear Predictive Coding (LPC) is a linear representation of the audio signal, it fails to take into account non-linear aspects of the audio signal (Ahmed et al., 2020)

4.3. Image-Based Features:

A Spectrogram is the time-frequency representation of an audio sample. The image-based features have proved to be effective for ESC as all the neural network models applied in image classification tasks can work on audios. In Zhang et al. (2015), spectrogram and Cross Recurrence Plot(CRP) are used. Spectrogram represents the audio signal visually at various frequencies. CRP visualizes the times at which states in two dynamical systems occur simultaneously(Boddapati et al., 2017)

Log Mel Spectrogram features(LMS) have performed well as compared to MFCC. LMS features are computed by simply calculating the fourier transform of the audio signal, then log of these frequencies is calculated and mapped to mel scale to generate the spectrograms. LMS is generated for each audio clip(Ahmed et al., 2020; Khamparia et al., 2019; Mu et al., 2021). LMS can be concatenated with Log Gammatone spectrogram to achieve good accuracy of 83.80% as compared to the case when only LMS is used(81.00%)(Chi et al., 2019). LMS are fused with raw waveform input features and considerable improvement in accuracy is achieved for ESC(Li et al., 2018). Static delta log mel features and static log mel features are given as input to the convolutional Neural Network(CNN) and accuracy is improved (Tokozume and Harada, 2017) as compared to state of art static delta log mel CNN(Piczak, 2015a).

4.4. Spectral Features

Spectral Features are derived from temporal features by exposing temporal features to some transformations. Different transformations that can be applied are Discrete Chirplet Transform, Discrete Curvelet Transform, Discrete Hilbert Transform along with Fast Fourier Transforms. In work (Han and Hwang, 2009), an accuracy of 86.09% is achieved using the transformations. Certain Spectral features such as spectral contrast (Gencoglu et al., 2014), spectral centroid, spectral bandwidth, spectral asymmetry, spectral flatness(Chu et al., 2006), Spectral dynamic Features(Karbasi et al., 2011), MPEG-7 feature set (Ntalampiras et al., 2010) are used widely in ESC. Using three MPEG-7 audio features- spectrum spread, spectrum centroid, and spectral flatness, the accuracy of 85.10% is attained(Wang et al., 2006). Selected MPEG-7 features along with MFCCs perform well(Muhammad et al., 2010). The MPEG-7 audio features are first prioritized using fisher's discriminant ratio and then the PCA is applied to the top 30 MPEG-7 features to get 13 features. These 13 features are combined with MFCCs for ESC.

4.5. Other domain features

Certain other features such as perceptual domain features-chroma, and tonnetz are also used in ESC(Gencoglu et al., 2014). Three different features- MFCCs, Log Mel Energies, Mel Energies are used for experiments out of which Mel Energies has proved to be most effective for ESC with NN (Gencoglu et al., 2014). Few other researchers have

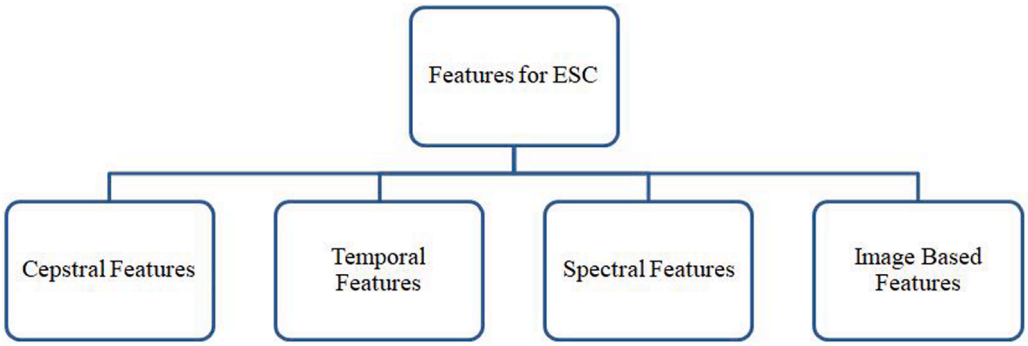

Fig. 2. Features used in Environmental Sound Classification.

used different features for sounds like haar-like sound features inspired by haar-like filtering in the case of 2-d face detection(Zhan and Kuroda, 2014),constant-Q transform (CQT) features(Mendoza et al., 2018), pitch range(PR) based feature set(Uzkent et al., 2012), Gammatone wavelet features(Valero and Alías, 2012b). Label smoothing(LS) method and additive margin softmax loss(AM-softmax) are combined and a deep feature is extracted to get the accuracy of 81.90% for a VGG-style deep neural network(Yao et al., 2019).

Feature extraction is not included as a step in ESC using the deep neural networks as features are extracted implicitly. Although the deep neural networks act as a black box and can classify the sounds directly without extracting the features but extracting features reduces the number of parameters to be trained(Khamparia et al., 2019)

5. Feature Selection

There are several features that can be extracted from audio samples but not all the features are informative for every application. Different applications require different sets of features. The optimal subset of features needs to be selected so that the computational complexity can be reduced(Liu et al., 2010). The calculation of higher dimensional features leads to an increase in computational time. Features that do not contribute to the classification or the features which are correlated can be discarded. Researchers have experimented with different feature sets for ESC. This section discusses noticeable feature selection techniques and feature combinations employed by researchers in the past for ESC.

Principal Component Analysis can be used for feature selection (Rabaoui et al., 2008) Bountourakis et. al had experimented with three feature sets. The combination of MFCC, LPCC, SFM, SCF, ZCR, Spectral Centroid, Spectral Spread, Spectral Roll-off, Spectral Skewness, Spectral Sharpness, and Spectral Smoothness give the highest classification accuracy with the classifiers k-NN, SVM, and ANN(Bountourakis et al., 2015). In a study (Su et al., 2019), two feature sets are created from five auditory features-: log-mel spectrogram (LM), MFCC, chroma, spectral contrast, and tonnetz(CST). Log-mel spectrogram and CST features are combined (LMC feature set) and MFCC and CST features are combined (MC feature set). Both the feature sets helped in attaining the accuracy of 95.20% and 95.30% respectively. In Rabaoui et al. (2008), feature vector sets are chosen rather than feature sets.

6. Classification

After the feature extraction and selection, the audio samples are classified into different categories. There exists a myriad of approaches for classification. Fig. 3 depicts the main classifiers used in ESC in literature. In this section, various machine learning and deep neural network classifiers used for ESC are reviewed.

6.1. Traditional machine Learning Classifiers

Researchers have tried to compare several machine learning algorithms for urban sound classification (Bountourakis et al., 2015; Jekic and Pester, 2018; da Silva et al., 2019).

6.1.1. Support Vector Machine(SVM)

SVM is the most popular supervised machine learning classifier used for sound applications. There are certain different types of SVM depending on the kernels used- binary, linear, polynomial, RBF, and Gaussian kernels. SVMs can also be categorized as multiclass SVM and One-class SVM. SVM has helped in attaining high accuracy in various past studies on ESC(Chu et al., 2006; Theodorou et al., 2015; Zhang et al., 2017). In a study (Uzkent et al., 2012), experiments are done with RBF and Gaussian kernels. In SVM, the parameter called C is varied which is the tradeoff between minimizing the model complexity and error in training. A classification rate of 87.30% is achieved with c=2 (Bountourakis et al., 2015). Multiclass SVM is used in Wang et al. (2008)

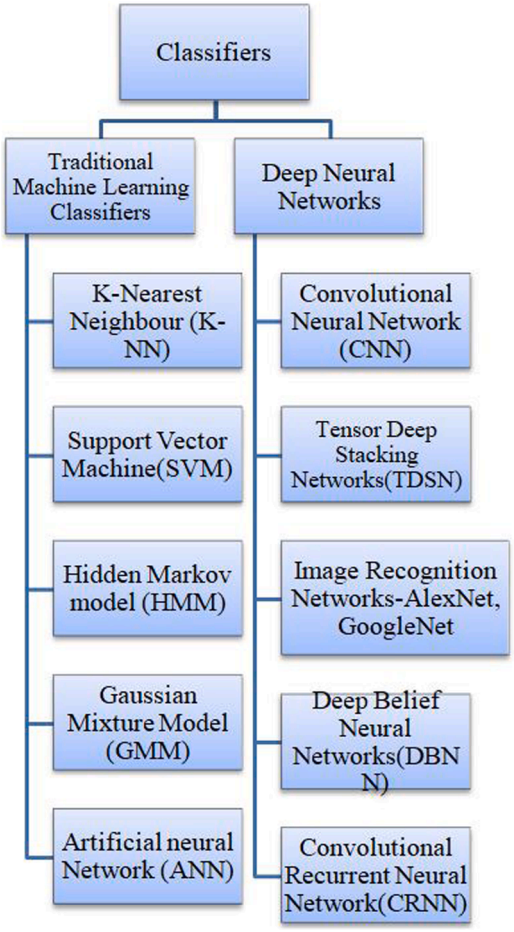

Fig. 3. Classification algorithms used in Environmental Sound Classification.

and considerable accuracy is obtained(91.10%). One class SVM is used for ESC in Rabaoui et al. (2008).

SVM is combined with KNN to outperform HMM for ESC using three low-level audio descriptors of the MPEG-7 feature set(Wang et al., 2006)

6.1.2. K-Nearest Neighbour Classifier(K-NN)

K-NN is mostly used for pattern recognition. Here k is the number of nearest neighbors. The new audio sample is assigned to the class to which most of the nearest neighbors belong. K-NN is widely used in ESC (Chu et al., 2006). The researchers tried to vary the value of k. The value of k as 8 gave the highest classification rate as 87.52% (Bountourakis et al., 2015)

6.1.3. Artificial Neural Network(ANN)

ANN is a classifier that works as a biological neuron. It consists of a set of neurons. Initially, the input layer is fed with random weights and inputs. The output is compared with the desired output. If both are different, the weights are adjusted. In ANN, the learning rate(LR) is varied. LR adjusts the bias and weight changes so that the algorithm learns adequately. A recognition rate of 87.30% is obtained by setting the value of LR=0.5(Bountourakis et al., 2015).

6.1.4. Hidden Markov Model(HMM)

HMM has been successfully used for ESR(Ntalampiras et al., 2010). In the study (Zhan and Kuroda, 2014) researchers have claimed that HMM provides good classification accuracy(96.90%) and consumes less power as compared to other algorithms. Gaussian Mixture based HMM provides less accuracy for ESC as compared to neural networks(54.80%) (Gencoglu et al., 2014; Su et al., 2011). In a study(Zhang et al., 2015),

researchers state that MFCCs based HMM models for ESC fail in case of noisy audio samples.

6.1.5. Gaussian Mixture Model(GMM)

GMM is a parametric classifier. GMM is an approach in which the model consists of several gaussian components. GMM has proved to be effective for ESC (Barchiesi et al., 2015; Chu et al., 2006; Muhammad et al., 2010; Ntalampiras et al., 2010).

6.2. Deep Neural Network Based Models

Neural Networks(NN) have performed well for ESC. Two-layer NN outperformed the GMM-based HMM as stated in Gencoglu et al. (2014).

6.2.1. Convolutional Neural Network(CNN)

CNN is deep learning neural network. The different proposed CNNs have surpassed the classification accuracy in ESR(Chi et al., 2019). The first use of CNN was done by PiczackCNN (Piczak, 2015a; Zhang et al., 2015) in which it was demonstrated that CNN outperforms the MFCCs-based machine learning model. Different hyperparameters such as padding, size of the max-pooling layer, and stride length are changed to find the best combination and get good accuracy. 92.90% accuracy is obtained on UrbanSound8k by using CNN(Ahmed et al., 2020). In research(Mendoza et al., 2018), sequential, parallel, and end-to-end CNN are used, and parallel CNN gives the highest accuracy(83.79%). Two-Stream CNN with Decision-Level Fusion(TSCNN-DS) model is proposed for ESR(Su et al., 2019). The two feature sets are given as inputs to two CNNs and then the output of softmax layers of both CNNs are combined using Dempster Shafer(DS) Evidence theory and an accuracy of 97.20% is attained. DS Evidence Theory has also improved the accuracy when end-to-end learning CNN and LMS-based CNN are combined(Li et al., 2018). Two-layer CNN gives an accuracy of 77.00% and 49.00% on ESC-10 and ESC-50 datasets respectively(Khamparia et al., 2019). End-to-end CNN(64.00%)is combined with static log mel CNN or static delta log mel CNN to achieve high accuracy(69.30% and 71.00% respectively). The authors in Abdoli et al. (2019) used one-dimensional end-to-end CNN which learns directly from audio representation and provides an accuracy of 89.00%.The authors claimed that this architecture uses less number of parameters as compared to the 2-D representations and 2-D CNN. In another study conducted in 2021 (Ragab et al., 2021), one-dimensional end-to-end CNN is employed along with Bayesian optimization and ensemble learning. The model learns the features directly from audio representation instead of hand-crafted features. In study (Zhang et al., 2017), activation functions (ReLu, PReLu,SoftPlus,LeakyReLu,ELU are varied to determine the best activation function for ESC.Leaky Relu gives the highest classification accuracy using dilated filters as the receptive field of convolution layers will store more contextual information. Very deep CNN with up to 34 weight layers increases the accuracy to 71.80% which is 15.56% more than CNN with 2 weight layers(Dai et al., 2017). Considerable improvement in accuracy(79.00%) is observed when deep CNN works on augmented data(Salamon and Bello, 2017). Authors in Fang et al. (2022), proposed Resource Adaptive CNN (RACNN) which can reduce the hardware requirements of traditional CNN and increase the speed and accuracy.

6.2.2. Tensor Deep Stacking Network(TDSN)

TDSN is similar to a deep stacking network(DSN) but it has parallel hidden layers in each module as compared to sequential hidden layers in the case of DSN. TDSN achieves an accuracy of 56.00% on the ESC-10 dataset(Khamparia et al., 2019)

6.2.3. Convolutional Recurrent Neural Network(CRNN)

CNN is combined with a recurrent neural network(RNN) for ESC (Sang et al., 2018). The features are extracted using CNN and temporal aggregation of extracted features is done through RNN. CRNN has proved effective for ESC. In the study (Bahmei et al., 2022), features are extracted using a deep Convolutional Generative Adversial Network, and further classification is done using CRNN.

6.2.4. Image Recognition Networks

Very deep CNN originally developed for image classification can be used to ESC. In a study (Boddapati et al., 2017), AlexNet and GoogLeNet are used on the image-based features for ESC and considerable accuracy is attained. VGG-style deep neural network is used (Yao et al., 2019).

6.2.5. Deep Belief Neural Networks(DBNN):

DBNN became popular as traditional DNN had problems like slow learning and requiring a lot of training data. DBNN surpasses the GMM-based HMM and NN with two or five layers for ESC(Gencoglu et al., 2014).

Certain other classifiers such as Self Organizing Maps (Sitte and Willets, 2007), self-supervised learning based deep classifier(Tripathi and Mishra, 2021), and Bayesian Belief Networks(Tsau et al., 2011) are studied by researchers for ESR.

Table 2 describes the literature review in ESC.

There are many classifiers used in literature for ESC. It becomes challenging to choose the appropriate classifier. There is a tradeoff between performance and computational cost in classification like feature extraction. There is no study that compares the performance of all the classifiers used in past studies. In a study (Sigtia et al., 2016), three classification algorithms- Deep Learning Neural networks, SVM, and GMMs are compared in terms of performance and computational costs for ESR. Deep Learning Neural networks provide considerable accuracy but a high computational cost is required. The SVM provides a tradeoff between accuracy and computational cost. GMMs provide acceptable accuracy at a low computational cost.

To choose the classifier, the following parameters can be considered:

Computational Complexity

A good classifier should have less computational complexity. Computational complexity stands for the time required and power consumed by a classifier. The time required for a classifier to produce the result should be less. The classifier should consume less power.

Recognition Accuracy

The classifier should have high accuracy. It should be able to classify the feature vectors accurately.

Robustness to Noise

A good classifier is robust to noise. It should ignore the variations caused by amplitude or bandwidth scaling in an audio signal.

7. Parameter Analysis of main References

Applying many algorithms to the task before selecting the algorithm to use is impractical. Comparing the machine learning algorithms used in infant research, we analyze them from the following aspects. Readers can choose the appropriate algorithm accordingly for their datasets and tasks.

Certain parameters are found by analyzing the past literature. An algorithm can be chosen considering the following aspects according to the datasets and tasks.

1. Zhang et al. (2021): In this paper, a sampling rate of 44.1 kHz and momentum of 0.8 is used. A batch size of 64 segments and 300 epochs with learning rate of 0.01 are used. The learning rate is decreased by dividing it by 10 for every 100 epochs. The following parameters are considered in this paper:
   (a) *Scaling Function:* The experiments are performed with two scaling functions-softmax and a sigmoid. It is found that sigmoid gives better accuracy as softmax function focusses on the frames with larger weight values.
   (b) *Application of attention on different layers of CRNN:* The attention is applied to different layers of CRNN from layer l_2 to l_{10}. It is

Table 2
Literature review in ESC.

Author,year	Technique/ Methodology	Validation Criteria/Measure	Dataset	Merit/Demerit
Zhang et al. (2021)	CRNN	Frame level attention mechanism	ESC-50 and ESC-10	**Merit:**Semantically relevant frames are focussed. **Demerit:** Robustness to noise of proposed model is not studied.
Chu et al. (2006)	SVM	Accuracy using leave-one-out cross-validation	Self-Collected	**Merit:** Focused on global characterization of the environment. **Demerit:** It has not focussed on localization and the effect of various sound sources on recognition. It is not scalable and robust to new environment
Uzkent et al. (2012)	SVM	Accuracy using gaussian kernel and new feature sets	Self-Collected	**Merit:** New feature sets are proposed and kernels of SVM are studied. **Demerit:** It does not work for speech environmental sounds.
Bountourakis et al. (2015)	SVM, K-NN, ANN,Logistic Regression, Naive Bayes	Precision, Recall, F-measure	BDLib	**Merit:**Time required to build the model is also compared. **Demerit:**Work focussed only on discrete sound events.
Ntalampiras et al. (2010)	New MPEG-7 feature sets and HMM and GMM	Accuracy using confusion matrices	Self-Collected	**Merit:**New Post processing algorithm is proposed. **Demerit:**It does not separate overlapping signals.
Zhan and Kuroda (2014)	1-d Haar like features with HMM	Recognition Accuracy and calculation cost	Self-collected	**Merit:**It outperforms other classifiers. **Demerit:** The number of recordings are very less(22).
Gencoglu et al. (2014)	Deep Neural Network and combination of GMM and HMM	Recognition accuracy	Self-collected	**Merit:**Effect of pretraining and effect of change of features is analyzed. **Demerit:** The effect of training parameters and network topology on performance is not examined.
Su et al. (2011)	Local discriminant bases (LDB) technique is used to identify the discriminatory time-frequency subspace for environmental sounds, HMM	Classification accuracy	Self-collected	**Merit:**Combination of LDB and MFCC gives high accuracy. **Demerit:**The database is not made standard.
Barchiesi et al. (2015)	MFCCs and HMM	Maximum Likelihood Criterion	Self-Collected	**Merit:**Testing on different datasets is not done. **Demerit:** Hierarchial classification is not considered.
Piczak (2015a)	CNN	Accuracy using confusion matrices	ESC-10, ESC-50, UrbanSound8k	**Merit:**First usage of CNN for ESC. **Demerit:** It doesnot explore ensembles of CNN.
Chi et al. (2019)	Two spectograms are concatenated-Log-Mel spectrogram and the Log-Gammatone spectrogram and CNN is used for classification	Classification accuracy with adam optimizer and 64 batch size	ESC-50 and UrbanSound8k dataset	**Merit:**Higher classification accuracy is attained. **Demerit:**More useful features can be extracted.
Zhang et al. (2015)	Spectograms are fed to CNN	Effect of noised spectrogram image feature(SIF) time span and frequency resolution on performance is examined	80 sound files are extracted randomly from Real Word Computing Partnership (RWCP) Sound Scene Database in Real Acoustic Environments(Nakamura et al., 1999)	**Merit:** Works best in noise corrupted conditions and first work in this case. **Demerit:**It only works for image based features.
Ahmed et al. (2020)	Log Mel Spectrogram are fed to CNN	Accuracy computation changing the padding and optimizer	ESC-50, ESC-10 and UrbanSound8k	**Merit:**Detailed comparison of the accuracy on benchmark datasets is presented. **Demerit:**It is not effective for sound samples with varying frequency and signal to noise ratio. Computational cost is high.
Mendoza et al. (2018)	Constant-Q transform(CQT) features are fed to parallel and sequential CNN	Classification accuracy for sequential, parallel and end to end CNN	UrbanSound8k	**Merit:**System is flexible and scalable in terms of number of nodes. **Demerit:**Due to continuous usage of sensor nodes, this model is not energy efficient.
Salamon and Bello (2017)	Data augmentation and CNN	Classification accuracy	UrbanSound8k dataset	**Merit:**Better performance as compared to CNN without augmented data. **Demerit:** Class conditional data augmentation is not applied.
Khamparia et al. (2019)	Spectrogram images of environmental sounds are fed to CNN and TDSNN	Classification accuracy	ESC-10 and ESC-50	**Merit:**The system is promising for critical areas. **Demerit:**The system works on compressed images rather than high definition images.
Sang et al. (2018)	Raw waveform are fed to CRNN	Classification accuracy	UrbanSound8k	**Merit:**The system time-series waveforms as input for audio classification and provide good accuracy. **Demerit:**It works only on raw waveforms.
Boddapati et al. (2017)	Image representations of environmental sounds(CRP, MFCCs, Spectrogram) are fed to AlexNet and GoogLeNet	Classification accuracy on test set	ESC-10, ESC-50, UrbanSoun8k	**Merit:**Same technology can be used for both object and sound recognition and classification. **Demerit:**The system is limited to image based features only.
Demir et al. (2020)	Features extracted using end-to-end CNN are fed to random subspaces KNN ensembles classifier	Classification accuracy	DCASE-2017 ASC and UrbanSound8K	**Merit:**The proposed CNN is flexible and not much deep. Sizes and numbers of layers can be freely changed and training time is less. **Demerit:** The obtained accuracy is less as compared to other benchmark results.

found that applying attention to layer l_{10} yields the highest accuracy.

(c) *Effect of data augmentation:* The authors have tried CRNN alone, CRNN with augmentation, CRNN with attention, and CRNN with attention and augmentation. It is found that CRNN with attention an d augmentation gives the best results.

2. Demir et al. (2020):

In this paper, Hamming window size of 1024, overlapping size of 256, and 3000 FFT parameters are used. The spectrogram images are initally 875 x 656 x 3 and are resized to 100 x 100 x 3 to feed the CNN model. CNN model has three convolutional layers, three max-pooling layers, and three fully connected (FC) layers. Thr following parameters are considered while performing the experiments:

(a) *Sizes of the FC layers:* The size of FC layers is varied. The size of FC1 is varied from 100 to 650 and the of FC2 is varied from 50 to 600 by incrementing by 50 on each experiment. It is found that FC1 of 500 and FC2 of 450 give the best results for both datasets.

(b) *Effect of input size:* Different input sizes: 20 x 20, 50x 50, 100 x 100 and 200 x 200 are taken. The input size of 100 x 100 gives the highest accuracy for both datasets.

(c) *Cross-Validation:* Two types of cross-validation- 5-fold and 10-fold are used for the UrbanSound8K dataset and 10-fold cross-validation outperforms 5-fold cross-validation.

3. Ahmed et al. (2020): Log-mel spectrogram features of size 128 x 128 are fed to the CNN. CNN consists of four convolutional layers, four max-pooling layers, two fully connected layers, and ReLu activation function are used for the experiments. 5-fold cross-validation is used for ESC-10 and ESC-50 and 10-fold cross-validation is used for UrbanSound8K. The following facets are considered:

(a) *Types of Padding:* Two types of padding- same and valid are considered. The same padding type gives the highest accuracy on all three datasets.

(b) *Optimizers:* Two types of optimizers - Adam and Rectified Adam (RAdam) are used for experiments and Adam outperforms RAdam for all the datasets.

8. Conclusion and Future work

In this paper, an in-depth survey of the work done in the field of ESC is demonstrated. The detailed study of datasets, features, and classifiers can help the researchers who work in this area. The lack of collaborated dataset for ESC hinders the research. Various researchers have worked with the combination of features as a single feature fails to provide accuracy. Recent studies show that deep neural networks have excelled. The highest accuracy is attained using a Convolutional neural network till now.

Finally, we would like to point out a few future research and development directions. In the future, vast datasets, more robust features, new feature combinations, and novel neural network architectures can be explored for environmental sound classification.

1. **Database expansion:**There is no larger benchmark dataset except UrbanSound8K which is publicly available. The existing dataset can be expanded by including more audio recordings and different classes of audio.
2. **Robust features and Feature combinations:**After critically analyzing the literature, it can be concluded that combinations of various features give better performance as compared to individual features. So, various other combinations of features can be explored in the future. Further, more robust features can be investigated. Probably, new features give better results.
3. **Novel neural network architectures:**Novel neural network architectures which are not been used till now can be investigated. Considerable effort is needed in combining the neural networks for ESC as the combination of the networks has given better results in similar areas.
4. **Meta Analysis of the past literature:** In this paper, only comprehensive review is done. In future, meta analysis and simulation of different research papers can be done.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.iswa.2022.200115

References

Abdoli, S., Cardinal, P., & Koerich, A. L. (2019). End-to-end environmental sound classification using a 1d convolutional neural network. *Expert Systems with Applications, 136*, 252–263.

Ahmed, M., Robin, T. I., Shafin, A. A., et al. (2020). Automatic environmental sound recognition (aesr) using convolutional neural network. *International Journal of Modern Education & Computer Science, 12*(5).

Alías, F., Socoró, J. C., & Sevillano, X. (2016). A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences, 6*(5), 143.

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., Baird, A., & Schuller, B. (2017). Snore sound classification using image-based deep spectrum features.

Aziz, S., Awais, M., Akram, T., Khan, U., Alhussein, M., & Aurangzeb, K. (2019). Automatic scene recognition through acoustic classification for behavioral robotics. *Electronics, 8*(5), 483.

Aziz, S., Khan, M. U., Alhaisoni, M., Akram, T., & Altaf, M. (2020). Phonocardiogram signal processing for automatic diagnosis of congenital heart disorders through fusion of temporal and cepstral features. *Sensors, 20*(13), 3790.

Bahmei, B., Birmingham, E., & Arzanpour, S. (2022). Cnn-rnn and data augmentation using deep convolutional generative adversarial network for environmental sound classification. *IEEE Signal Processing Letters, 29*, 682–686.

Bansal, A., Aggarwal, N., Vij, D., & Sharma, A. (2018). An off the shelf cnn features based approach for vehicle classification using acoustics. *International conference on ismac in computational vision and bio-engineering* (pp. 1163–1170). Springer.

Barchiesi, D., Giannoulis, D., Stowell, D., & Plumbley, M. D. (2015). Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Magazine, 32*(3), 16–34.

Bhat, G. S., Shankar, N., & Panahi, I. M. (2020). Automated machine learning based speech classification for hearing aid applications and its real-time implementation on smartphone. *2020 42nd annual international conference of the ieee engineering in medicine & biology society (embc)* (pp. 956–959). IEEE.

Boddapati, V., Petef, A., Rasmusson, J., & Lundberg, L. (2017). Classifying environmental sounds using image recognition networks. *Procedia computer science, 112*, 2048–2056.

Bountourakis, V., Vrysis, L., & Papanikolaou, G. (2015). Machine learning algorithms for environmental sound recognition: Towards soundscape semantics. *Proceedings of the audio mostly 2015 on interaction with sound* (pp. 1–7).

Brodie, S., Allen-Ankins, S., Towsey, M., Roe, P., & Schwarzkopf, L. (2020). Automated species identification of frog choruses in environmental recordings using acoustic indices. *Ecological Indicators, 119*, 106852.

Chachada, S., & Kuo, C.-C. J. (2014). Environmental sound recognition: A survey. *APSIPA Transactions on Signal and Information Processing, 3*.

Chandrakala, S., & Jayalakshmi, S. (2019). Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies. *ACM Computing Surveys (CSUR), 52*(3), 1–34.

Chi, Z., Li, Y., & Chen, C. (2019). Deep convolutional neural network combined with concatenated spectrogram for environmental sound classification. *2019 ieee 7th international conference on computer science and network technology (iccsnt)* (pp. 251–254). IEEE.

Chu, S., Narayanan, S., Kuo, C.-C. J., & Mataric, M. J. (2006). Where am i? scene recognition for mobile robots using audio features. *2006 ieee international conference on multimedia and expo* (pp. 885–888). IEEE.

Cowling, M., & Sitte, R. (2003). Comparison of techniques for environmental sound recognition. *Pattern recognition letters, 24*(15), 2895–2907.

Crocco, M., Cristani, M., Trucco, A., & Murino, V. (2016). Audio surveillance: A systematic review. *ACM Comput. Surv., 48*(4). https://doi.org/10.1145/2871183

Dai, W., Dai, C., Qu, S., Li, J., & Das, S. (2017). Very deep convolutional neural networks for raw waveforms. *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 421–425). IEEE.

Demir, F., Abdullah, D. A., & Sengur, A. (2020). A new deep cnn model for environmental sound classification. *IEEE Access, 8*, 66529–66537.

Duan, S., Zhang, J., Roe, P., & Towsey, M. (2014). A survey of tagging techniques for music, speech and environmental sound. *Artificial Intelligence Review, 42*(4), 637–661.

Elbir, A., & Aydin, N. (2020). Music genre classification and music recommendation by using deep learning. *Electronics Letters, 56*(12), 627–629.
Fan, X., Sun, T., Chen, W., & Fan, Q. (2020). Deep neural network based environment sound classification and its implementation on hearing aid app. *Measurement, 159*, 107790.
Fang, Z., Yin, B., Du, Z., & Huang, X. (2022). Fast environmental sound classification based on resource adaptive convolutional neural network. *Scientific Reports, 12*(1), 1–18.
Font, F., Roma, G., & Serra, X. (2013). Freesound technical demo. *Proceedings of the 21st acm international conference on multimedia* (pp. 411–412).
Gencoglu, O., Virtanen, T., & Huttunen, H. (2014). Recognition of acoustic events using deep neural networks. *2014 22nd european signal processing conference (eusipco)* (pp. 506–510). IEEE.
Han, B.-j., & Hwang, E. (2009). Environmental sound classification based on feature collaboration. *2009 ieee international conference on multimedia and expo* (pp. 542–545). IEEE.
Hossain, M. S., & Muhammad, G. (2018). Environment classification for urban big data using deep learning. *IEEE Communications Magazine, 56*(11), 44–50.
Jekic, N., & Pester, A. (2018). Environmental sound recognition with classical machine learning algorithms. *International conference on remote engineering and virtual instrumentation* (pp. 14–21). Springer.
Karbasi, M., Ahadi, S. M., & Bahmanian, M. (2011). Environmental sound classification using spectral dynamic features. *2011 8th international conference on information, communications & signal processing* (pp. 1–5). IEEE.
Khamparia, A., Gupta, D., Nguyen, N. G., Khanna, A., Pandey, B., & Tiwari, P. (2019). Sound classification using convolutional neural network and tensor deep stacking network. *IEEE Access, 7*, 7717–7727.
Kim, C.-I., Cho, Y., Jung, S., Rew, J., & Hwang, E. (2020). Animal sounds classification scheme based on multi-feature network with mixed datasets. *KSII Transactions on Internet and Information Systems (TIIS), 14*(8), 3384–3398.
Li, S., Yao, Y., Hu, J., Liu, G., Yao, X., & Hu, J. (2018). An ensemble stacked convolutional neural network model for environmental event sound recognition. *Applied Sciences, 8*(7), 1152.
Liu, H., Motoda, H., Setiono, R., & Zhao, Z. (2010). Feature selection: An ever evolving frontier in data mining. *Feature selection in data mining* (pp. 4–13). PMLR.
Logan, B., et al. (2000). Mel frequency cepstral coefficients for music modeling, *vol. 270. Ismir* (pp. 1–11). Citeseer.
Ma, N., Gonzalez, J. A., & Brown, G. J. (2018). Robust binaural localization of a target sound source by combining spectral source models and deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26*(11), 2122–2131.
Mac Aodha, O., Gibb, R., Barlow, K. E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G. R., Newson, S. E., et al. (2018). Bat detective?deep learning tools for bat acoustic signal detection. *PLoS computational biology, 14*(3), e1005995.
Mendoza, J. M., Tan, V., Fuentes, V., Perez, G., & Tiglao, N. M. (2018). Audio event detection using wireless sensor networks based on deep learning. *International wireless internet conference* (pp. 105–115). Springer.
Mesaros, A., Heittola, T., Eronen, A., & Virtanen, T. (2010). Acoustic event detection in real life recordings. *2010 18th european signal processing conference* (pp. 1267–1271). IEEE.
Mitrović, D., Zeppelzauer, M., & Breiteneder, C. (2010). Features for content-based audio retrieval, *vol. 78. Advances in computers* (pp. 71–150). Elsevier.
Mu, W., Yin, B., Huang, X., Xu, J., & Du, Z. (2021). Environmental sound classification using temporal-frequency attention based convolutional neural network. *Scientific Reports, 11*(1), 1–14.
Muhammad, G., Alotaibi, Y. A., Alsulaiman, M., & Huda, M. N. (2010). Environment recognition using selected mpeg-7 audio features and mel-frequency cepstral coefficients. *2010 fifth international conference on digital telecommunications* (pp. 11–16). IEEE.
Mushtaq, Z., & Su, S.-F. (2020). Environmental sound classification using a regularized deep convolutional neural network with data augmentation. *Applied Acoustics, 167*, 107389.
Mushtaq, Z., Su, S.-F., & Tran, Q.-V. (2021). Spectral images based environmental sound classification using cnn with meaningful data augmentation. *Applied Acoustics, 172*, 107581.
Mydlarz, C., Salamon, J., & Bello, J. P. (2017). The implementation of low-cost urban acoustic monitoring devices. *Applied Acoustics, 117*, 207–218.
Nakamura, S., Hiyane, K., Asano, F., Yamada, T., & Endo, T. (1999). Data collection in real acoustical environments for sound scene understanding and hands-free speech recognition,.
Ntalampiras, S., Potamitis, I., & Fakotakis, N. (2010). Automatic recognition of urban environmental sounds events,.
Palo, H. K., Chandra, M., & Mohanty, M. N. (2018). Recognition of human speech emotion using variants of mel-frequency cepstral coefficients. *Advances in systems, control and automation* (pp. 491–498). Springer.
Piczak, K. J. (2015a). Environmental sound classification with convolutional neural networks. *2015 ieee 25th international workshop on machine learning for signal processing (mlsp)* (pp. 1–6). IEEE.
Piczak, K. J. (2015b). Esc: Dataset for environmental sound classification. *Proceedings of the 23rd acm international conference on multimedia* (pp. 1015–1018).
Plata, M. (2019). Deep neural networks with supported clusters preclassification procedure for acoustic scene recognition. *Tech. Rep., DCASE2019 Challenge*.
Rabaoui, A., Davy, M., Rossignol, S., & Ellouze, N. (2008). Using one-class svms and wavelets for audio surveillance. *IEEE Transactions on information forensics and security, 3*(4), 763–775.
Ragab, M. G., Abdulkadir, S. J., Aziz, N., Alhussian, H., Bala, A., & Alqushaibi, A. (2021). An ensemble one dimensional convolutional neural network with bayesian optimization for environmental sound classification. *Applied Sciences, 11*(10), 4660.
Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters, 24*(3), 279–283.
Salamon, J., Jacoby, C., & Bello, J. P. (2014). A dataset and taxonomy for urban sound research. *Proceedings of the 22nd acm international conference on multimedia* (pp. 1041–1044).
Sang, J., Park, S., & Lee, J. (2018). Convolutional recurrent neural networks for urban sound classification using raw waveforms. *2018 26th european signal processing conference (eusipco)* (pp. 2444–2448). IEEE.
Saraubon, K., Anurugsa, K., & Kongsakpaibul, A. (2018). A smart system for elderly care using iot and mobile technologies, . In *ICSEB '18Proceedings of the 2018 2nd international conference on software and e-business* (p. 59?63). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3301761.3301769
Sharan, R. V., & Moir, T. J. (2019). Acoustic event recognition using cochleagram image and convolutional neural networks. *Applied Acoustics, 148*, 62–66.
Sharma, G., Umapathy, K., & Krishnan, S. (2020). Trends in audio signal feature extraction methods. *Applied Acoustics, 158*, 107020.
Sharma, J., Granmo, O.-C., & Goodwin, M. (2019). Environment sound classification using multiple feature channels and attention based deep convolutional neural network. *arXiv preprint arXiv:1908.11219*.
Sigtia, S., Stark, A. M., Krstulović, S., & Plumbley, M. D. (2016). Automatic environmental sound recognition: Performance versus computational cost. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(11), 2096–2107.
da Silva, B., W Happi, A., Braeken, A., & Touhafi, A. (2019). Evaluation of classical machine learning techniques towards urban sound recognition on embedded systems. *Applied Sciences, 9*(18), 3885.
Sitte, R., & Willets, L. (2007). Non-speech environmental sound identification for surveillance using self-organizing-maps. *Proceedings of the fourth conference on iasted international conference: Signal processing, pattern recognition, and applications* (pp. 281–286).
Soares, B. S., Luz, J. S., de Macêdo, V. F., e Silva, R. R. V., de Araújo, F. H. D., & Magalhães, D. M. V. (2022). Mfcc-based descriptor for bee queen presence detection. *Expert Systems with Applications, 201*, 117104.
Su, F., Yang, L., Lu, T., & Wang, G. (2011). Environmental sound classification for scene recognition using local discriminant bases and hmm. *Proceedings of the 19th acm international conference on multimedia* (pp. 1389–1392).
Su, Y., Zhang, K., Wang, J., & Madani, K. (2019). Environment sound classification using a two-stream cnn based on decision-level fusion. *Sensors, 19*(7), 1733.
Theodorou, T., Mporas, I., & Fakotakis, N. (2015). Automatic sound recognition of urban environment events. *International conference on speech and computer* (pp. 129–136). Springer.
Tokozume, Y., & Harada, T. (2017). Learning environmental sounds with end-to-end convolutional neural network. *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 2721–2725). IEEE.
Tokozume, Y., Ushiku, Y., & Harada, T. (2017). Learning from between-class examples for deep sound recognition,.
Tripathi, A. M., & Mishra, A. (2021). Self-supervised learning for environmental sound classification. *Applied Acoustics, 182*, 108183.
Tsau, E., Kim, S.-H., & Kuo, C.-C. J. (2011). Environmental sound recognition with celp-based features. *Isscs 2011-international symposium on signals, circuits and systems* (pp. 1–4). IEEE.
Tsunoda, Y., Sueoka, Y., & Osuka, K. (2019). Experimental analysis of acoustic field control-based robot navigation. *Journal of Robotics and Mechatronics, 31*(1), 110–117.
Tuncer, T., Akbal, E., & Dogan, S. (2021). Multileveled ternary pattern and iterative relieff based bird sound classification. *Applied Acoustics, 176*, 107866.
Uzkent, B., Barkana, B. D., & Cevikalp, H. (2012). Non-speech environmental sound classification using svms with a new set of features. *International Journal of Innovative Computing, Information and Control, 8*(5), 3511–3524.
Vafeiadis, A., Votis, K., Giakoumis, D., Tzovaras, D., Chen, L., & Hamzaoui, R. (2017). Audio-based event recognition system for smart homes. *2017 ieee smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, internet of people and smart city innovation (smartworld/scalcom/uic/atc/cbdcom/iop/sci)* (pp. 1–8). IEEE.
Valero, X., & Alías, F. (2012a). Classification of audio scenes using narrow-band autocorrelation features. *2012 proceedings of the 20th european signal processing conference (eusipco)*. IEEE.
Valero, X., & Alías, F. (2012b). Gammatone wavelet features for sound classification in surveillance applications. *2012 proceedings of the 20th european signal processing conference (eusipco)* (pp. 1658–1662). IEEE.
Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res, 10*(66-71), 13.
Virtanen, T., & Helén, M.. Probabilistic model based similarity measures for audio query-by-example.
Wang, J.-C., Lee, H.-P., Wang, J.-F., & Lin, C.-B. (2008). Robust environmental sound recognition for home automation. *IEEE transactions on automation science and engineering, 5*(1), 25–31.
Wang, J.-C., Wang, J.-F., He, K. W., & Hsu, C.-S. (2006). Environmental sound classification using hybrid svm/knn classifier and mpeg-7 audio low-level descriptor. *The 2006 ieee international joint conference on neural network proceedings* (pp. 1731–1735). IEEE.

Yamakawa, N., Takahashi, T., Kitahara, T., Ogata, T., & Okuno, H. (2011). *Environmental Sound Recognition for Robot Audition using Matching-pursuit, K.G. Mehrotra et al. (Eds), IEA/AIE-2011, Part II, pp.1–10*.

Yang, W., & Krishnan, S. (2017). Combining temporal features by local binary pattern for acoustic scene classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(6), 1315–1321.

Yao, K., Yang, J., Zhang, X., Zheng, C., & Zeng, X. (2019). Robust deep feature extraction method for acoustic scene classification. *2019 ieee 19th international conference on communication technology (icct)* (pp. 198–202). IEEE.

Zhan, Y., & Kuroda, T. (2014). Wearable sensor-based human activity recognition from environmental background sounds. *Journal of Ambient Intelligence and Humanized Computing, 5*(1), 77–89.

Zhang, H., McLoughlin, I., & Song, Y. (2015). Robust sound event recognition using convolutional neural networks. *2015 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 559–563). IEEE.

Zhang, X., Zou, Y., & Shi, W. (2017). Dilated convolution neural network with leakyrelu for environmental sound classification. *2017 22nd international conference on digital signal processing (dsp)* (pp. 1–5). IEEE.

Zhang, Z., Xu, S., Zhang, S., Qiao, T., & Cao, S. (2021). Attention based convolutional recurrent neural network for environmental sound classification. *Neurocomputing, 453*, 896–903.