

Review

# Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review

Olusola O. Abayomi-Alli <sup>1</sup>, Robertas Damaševičius <sup>1,\*</sup>, Atika Qazi <sup>2</sup>, Mariam Adedoyin-Olowe <sup>3</sup> and Sanjay Misra <sup>4</sup>

<sup>1</sup> Department of Software Engineering, Kaunas University of Technology, 44249 Kaunas, Lithuania

<sup>2</sup> Centre for Lifelong Learning, Universiti Brunei Darussalam, Gadong BE1410, Brunei

<sup>3</sup> School of Computing and Digital Technology, Birmingham City University, Birmingham B4 7XG, UK

<sup>4</sup> Department of Computer Science and Communication, Østfold University College, 1757 Halden, Norway

\* Correspondence: robertas.damasevicius@ktu.lt

**Abstract:** The aim of this systematic literature review (SLR) is to identify and critically evaluate current research advancements with respect to small data and the use of data augmentation methods to increase the amount of data available for deep learning classifiers for sound (including voice, speech, and related audio signals) classification. **Methodology:** This SLR was carried out based on the standard SLR guidelines based on PRISMA, and three bibliographic databases were examined, namely, Web of Science, SCOPUS, and IEEE Xplore. **Findings:** The initial search findings using the variety of keyword combinations in the last five years (2017–2021) resulted in a total of 131 papers. To select relevant articles that are within the scope of this study, we adopted some screening exclusion criteria and snowballing (forward and backward snowballing) which resulted in 56 selected articles. **Originality:** Shortcomings of previous research studies include the lack of sufficient data, weakly labelled data, unbalanced datasets, noisy datasets, poor representations of sound features, and the lack of effective augmentation approach affecting the overall performance of classifiers, which we discuss in this article. Following the analysis of identified articles, we overview the sound datasets, feature extraction methods, data augmentation techniques, and its applications in different areas in the sound classification research problem. Finally, we conclude with the summary of SLR, answers to research questions, and recommendations for the sound classification task.

**Keywords:** sound data; audio data; data augmentation; feature extraction; deep learning



**Citation:** Abayomi-Alli, O.O.; Damaševičius, R.; Qazi, A.; Adedoyin-Olowe, M.; Misra, S. Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review. *Electronics* **2022**, *11*, 3795. <https://doi.org/10.3390/electronics11223795>

Academic Editor: Amir Mosavi

Received: 12 September 2022

Accepted: 16 November 2022

Published: 18 November 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The continuous growth of the application of artificial intelligence (AI) methods in a diverse range of scientific fields has played a significant role in solving real-life problems, especially in classification tasks in various domains such as computer vision [1], natural language processing (NLP) [2], healthcare [3], industrial signal processing [4], etc. Interestingly, the success of these AI methods has also spread across other domains, including speech recognition and the music recommendation task [5]. The need for effective and automatic sound classification systems is on the rise, as its relevance cannot be underestimated in our everyday life. Automatic sound classification technologies are widely applied in surveillance systems [6], voice assistants [7], chatbots [8], smart safety devices [9], and in different real-world environments, such as engineering [10], industrial [11], domestic [12], urban [13], road [14], and natural [15].

Machine learning approaches such as random forest (RF), decision tree (DT), logistic regression (LR), multilayer perceptron (MLP), etc. have been applied for sound recognition systems [16]. In the last decade, the advancement of machine learning algorithms (including deep learning methods) has shown great capabilities in extracting high-level features that have helped to effectively learn complex level characteristics from raw input data, thus

improving performance of classification models [17]. Recently, the paradigm shift in improvement in these deep learning algorithms, such as finetuning hyperparameters such as enhancing dropout and regularization, momentum methods of gradient descent, etc. [18], has played an important role in the advancement of researchers' contributions in many areas such as computer vision, natural language processing (NLP), finance, and biomedical imaging [19–22].

The exceptional performance of deep learning, especially convolution neural network (CNN) for pattern recognition, has continued to show great impact in an effective modern classification task. Recently, the application of deep learning methods in different kinds of sound/audio classification tasks has shown great progress, especially in domains such as environmental sound detection [23], automatic speech recognition (ASR) [24], music/acoustic classification [25], medical diagnostics [26], etc. However, this approach to deep learning methods still struggles with poor performance due to the availability of insufficient data to solve audio/sound related issues, noisy audio signals [27], and industrial sounds [28]. Considering the vast application of deep learning methods, several researchers have been fascinated by applying different machine learning algorithms in sound classification [29–31]. However, audio signals have high dimensionality, indicating that more than one thousand floating point values are required to represent a short audio signal, raising the need for exploring dimensionality reduction and feature extraction methods.

Deep learning models in sound recognition systems can be seriously affected by environmental noise, which could possibly result in loss of detailed information [32]. Another important challenge in developing an efficient sound recognition system is accessing a large and well-annotated dataset. In addition, challenges with data scarcity in sound classification systems include privacy [33] and ethical and legal considerations [34].

The poor performance of deep learning models can be attributed to the following:

- Insufficient sound or audio data makes it extremely difficult to train deep neural networks, as efficient training and evaluation of audio/sound systems are only dependent on large training data [35].
- Traditional audio feature extraction methods lack strong abilities to effectively identifying better feature representations, thus affecting the performance of sound recognition [36].
- Robustness and generalization are the key challenges in building a high-performance sound recognition system, and some of the existing systems degrade due to scenario mismatch due to some factors such as reverberations, noise types, channels, etc. [37].
- Dependency on expert knowledge for reliable annotation of audio data [38].

Another problem that negatively affects research progress in sound classification research is imbalance of data [39]; this plays a major factor by deteriorating the performance of deep learning systems because most audio recordings are susceptible to environmental noise [40]. In addition, the creation of a sound recording dataset is extremely time-consuming and a resource constraint; thus, the need for data augmentation techniques is not negotiable, as this approach has become powerful in generating a synthesis dataset (images, sounds, text, etc.) and has significantly contributed to improving the performance of deep learning models.

Contrary to the popular claim that training large datasets is essential to achieve optimal results for deep architecture models [41], the advancement of data augmentation in sound classification tasks has shown its consistency in improving the performance of training models for small data [42]. The need for data augmentation cannot be overlooked, as previous research studies have shown in the application of neural network models in the sound/audio classification task [43,44] because it is a typical over-parameterized model and therefore requires larger datasets to mitigate overfitting and reduce sensitivity to background noise and information redundancy [45]. Furthermore, the application of neural network models is highly dependent on initializing and carefully adjusting hyperparameters during the training process to improve the classification model [46].

This article presents a systematic review of the comprehensive past, present, and future trends of data augmentation techniques in sound classification tasks. This study aims to present the different data enhancement methods applied in the literature to increase data generalization and detection rate using statistical analyses (quantitative and qualitative) measures to showcase research trends in combating insufficient/imbalance datasets. In this paper, a systematic review of the literature is presented based on a comprehensive methodology with the purpose of identifying current progress and progress of related studies in sound classification tasks with respect to sources, data repositories, feature extraction methods, data enhancement steps, and classification models is presented.

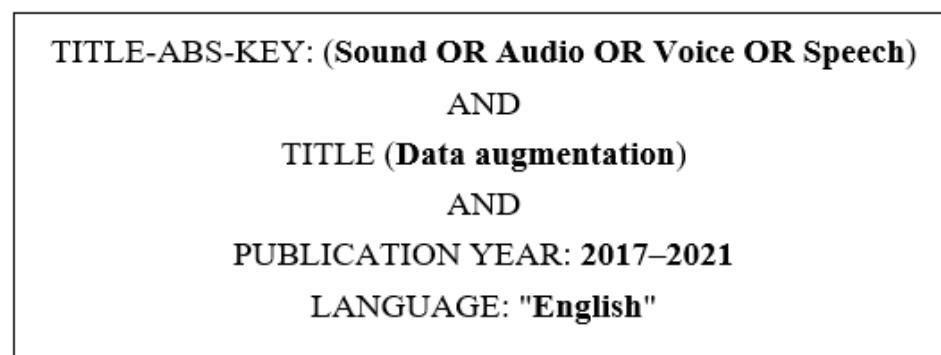
The following research questions (RQs) are raised to define the scope of the systematic review, and analytical results are presented to enhance future research studies on sound classification as follows:

1. Are existing papers based on the sound classification task applied to specific and established data sources for experimentation?
2. What are the data repository or dataset sources used?
3. Which feature extraction methods are used and which data are extracted?
4. What are the different data augmentation techniques applied in sound classification?
5. How can we measure the importance of data augmentation techniques in learning algorithms?
6. What is the future research recommendation for augmentation techniques?
7. What obstacles are identified in the application of data augmentation for sound classification?

The rest of this paper is organized as follows: Section 2 discusses in detail the overview of sound datasets, feature extraction methods, data augmentation techniques, and its applications in different areas in sound classification. The SLR methodology steps are presented in Section 3 with a description of search selection methods. The details of our results/findings and the analysis of selected studies with respect to the research questions raised are addressed in Section 4, and finally Section 5 concludes with a summary of the SLR and future recommendations in the sound classification task.

## 2. Methodology of Literature Search and Selection

We adopted the guidelines [47] that cover the systematic literature review in the formulation of research questions, the structure of the search study, and the data extraction criteria. In addition to this, we also adopted a simple and effective process in accordance with the methodological approach proposed in [48] as shown in Figure 1.



**Figure 1.** Search query and its parameters.

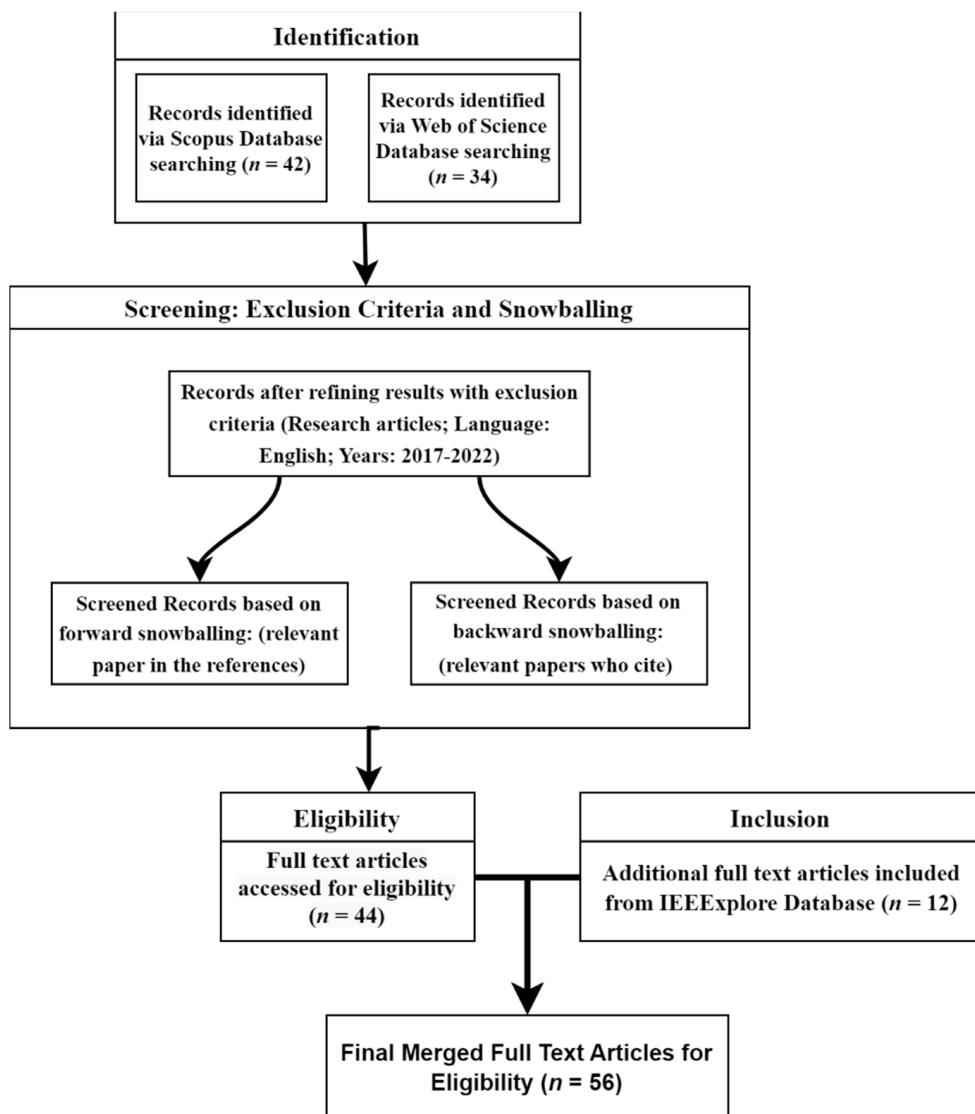
In this context, a research plan was initially outlined that involves the research objective and questions, and a variety of combinations/keywords metadata was constructed as represented in Figure 1. For consistency throughout the article and to avoid confusion and misinterpretation of words, this article decides on the basis that the term ‘sound’ encompasses audio, speech, and voice, which are all related to a category of sound. The search was carried out on two databases, the Web of Science and Scopus databases as summarized in Table 1. The article search using combinations of keywords carried out from January 2017 up to 2022 returned a total of 42 for Scopus and 34 for Web of Science, and these articles passed the title, abstract, and keyword selection, and supplemented with forward and backward snowballing [49]. Other databases were also selected, and a total of 12 relevant articles are included in this study. Therefore, the final set of relevant selected articles included only 55 full text articles which were compared to the SLR research focus. These selected groups of articles were read, reviewed, categorized, and analyzed to ensure openness and detailed reporting of the systematic literature review process [50] as depicted using the PRISMA workflow diagram in Figure 2.

**Table 1.** Number of studies found per keyword combination.

Database	Search (Title, Abstract or Keyword)	No. of Results	Filtered by Exclusion Criteria	Forward Snowballing	Backward Snowballing	Final No.
Scopus	T-A-K: ((Sound OR Audio OR Voice OR Speech)) AND T: (Data Augmentation) AND (L-T: (SRCTYPE, “j”)) AND (L-T (PY: 2021) OR L-T (PY: 2020) OR L-T (PY: 2019) OR L-T (PY: 2018) OR L-T (PY: 2017)) AND (L-T (LANG, “English”))	42	32	4	2	38
Web of Science	T: ((Sound OR Audio OR Voice OR Speech)) AND T: (Data Augmentation) AND ((PY:2021) OR (PY:2020) OR (PY: 2019) OR (PY: 2018) OR (PY = 2017))	34	28	-	4	32
IEEE Xplore	T: ((Sound OR Audio OR Voice OR Speech)) AND T: (Data Augmentation) AND ((PY:2021) OR (PY:2020) OR (PY: 2019) OR (PY: 2018) OR (PY = 2017))	55	20	-	-	20

T = title, A = abstract, K = keyword, PY = publication year, LANG = language, L-T = limit-to.

Considering the research questions raised, the context of this study aims to identify the concepts of changes and main perspectives related to sound classification in diverse research domains using the following groups of keywords as presented in Figure 1 and a summary of studies found per keyword combinations as depicted in Table 1. Table 1 summarizes the keyword combinations and the number of papers identified using the keywords in the title, abstract, and keywords of the papers.

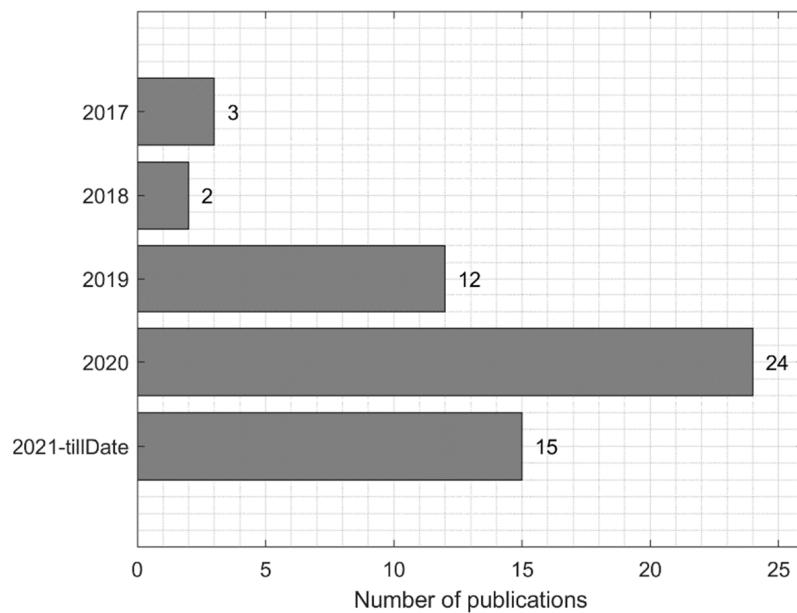


**Figure 2.** PRISMA workflow for article selection and retention in each step.

### 3. Results of SLR

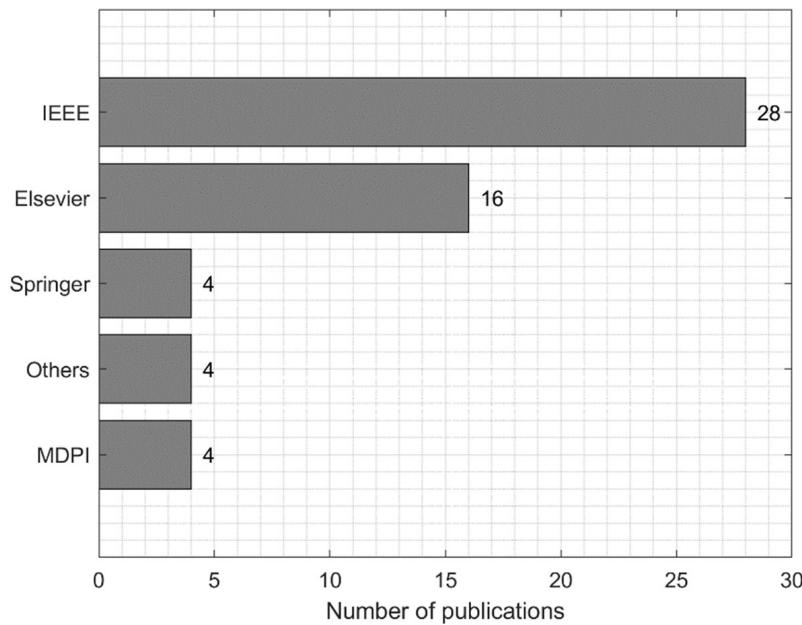
This section presents the results of the SLR including the quantitative and qualitative analyses of selected articles that meet the inclusion criteria. We describe the main statistics of the 56 primary studies [51–106] selected for this SLR.

We have included 56 primary studies in this SLR, and the highest percent of the studies were published in the year 2020, more than 26% of selected articles were published in 2021, 23% were published in 2019, and less than 10% of the selected article was published in 2018 and 2017 as depicted in Figure 3. The type of publication was classified as Workshop (which includes workshop and symposium papers), Conference, and Journal. In the publication type, 59% of the selected articles are journals, 34% were conference papers, and 7% of studies were workshop papers.



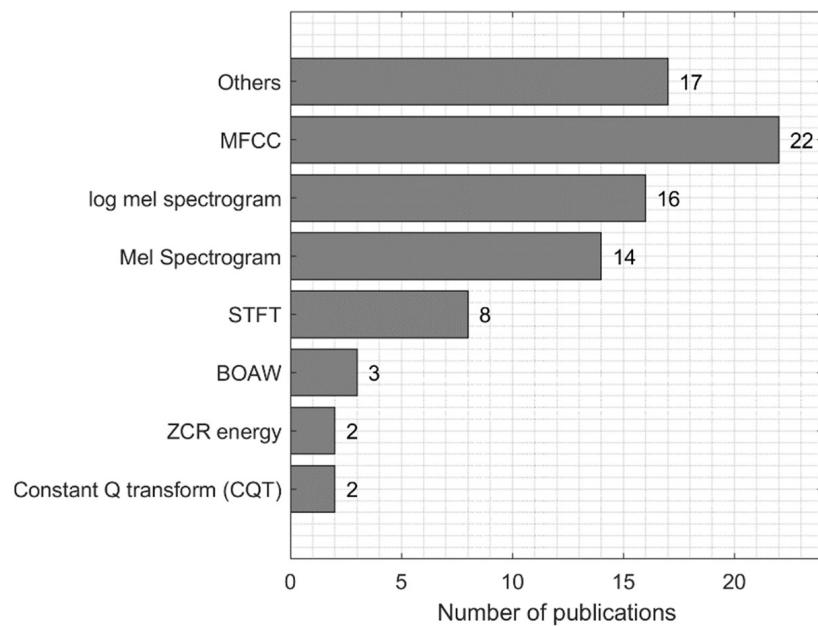
**Figure 3.** Number of publications year-wise.

As shown in Figure 4, the most popular databases for primary studies are IEEE and ScienceDirect with 28 (50%) and 16 (29%), respectively, and the remaining 21% of our selected articles are equally from MDPI, Springer, and other publication sources.



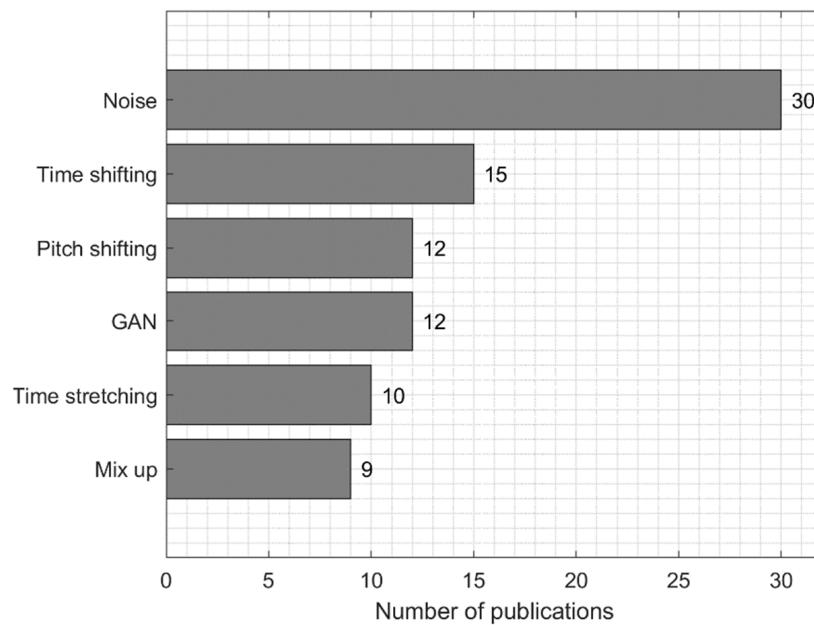
**Figure 4.** Number of publications per publisher.

The importance of feature extraction methods in the sound classification task cannot be overlooked in this SLR as we analyzed the different methods presented in the selected publications. From the selected studies, we realized that some feature extraction methods were used more by previous researchers than others and the results of our analysis showed that 25.6% of our selected publications used MFCC-based representation, while 18.6% of selected publications applied the log-mel spectrogram, 16.3% of selected publications used Mel spectrogram methods, 9.3% applied the STFT approach, and the remaining 30% of the publications applied other feature representation methods such as bag of words, CQT, and ZCR energy as depicted in Figure 5.



**Figure 5.** Feature extraction methods used for sound representation.

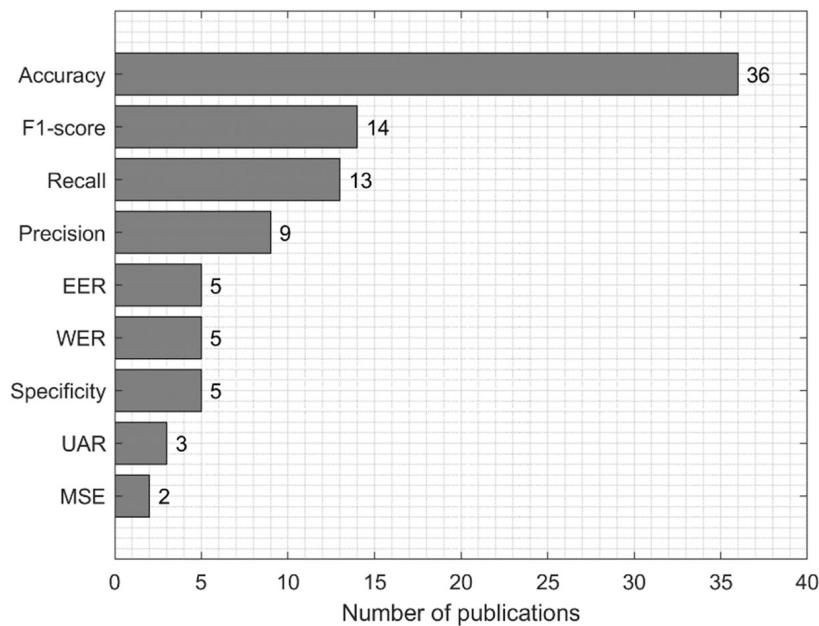
The results of data augmentation methods in Figure 6 show that the data augmentation based on addition of noise has the highest number of publications of 22 (39.2%), while the second highest is the time shift method with 15 (26.7%) of the selected publications. Next to that are the GAN based models and pitch shift with 12 (21.4%) each, followed by other methods such as time stretching, mix-up, and background noise with 10 (17.8%), 9 (16.1%), and 8 (13.6%), respectively. Methods such as speed modulation, masking, VTLP, trim silences, flipping, etc., have less than 10% application in selected publications.



**Figure 6.** Data augmentation approaches from analyzed articles.

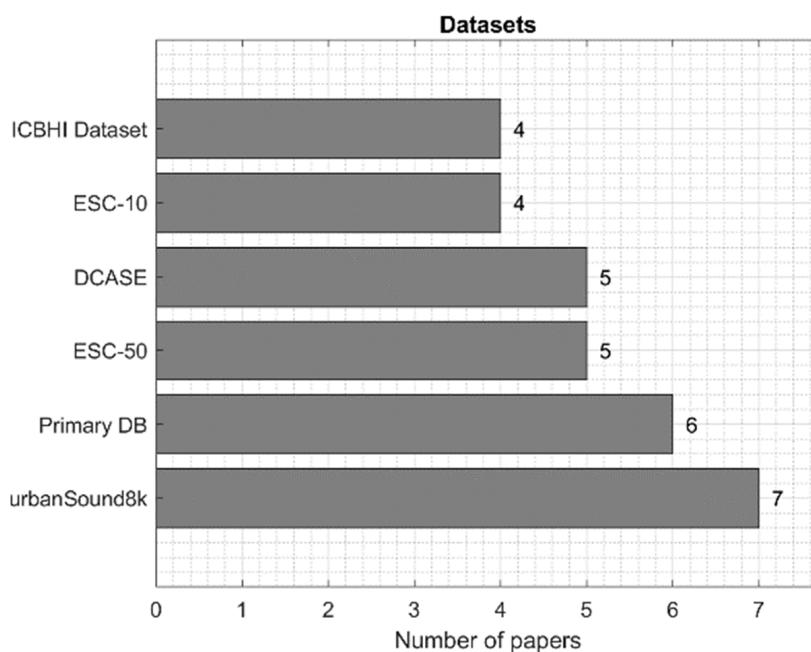
The results of performance evaluation methods in Figure 7 show that accuracy was the most used performance evaluation method used in 36 publications (64.3%), while the second highest is F1-score with 14 (25%) of analyzed publications. Next to that is recall with 13 (23.2%) publications, followed by precision with 9 (16.1%) publications. Other

measures such as equal error rate (EER), word error rate (WER), mean square error (MSE), and specificity were used less commonly.



**Figure 7.** Performance evaluation measures from analyzed articles.

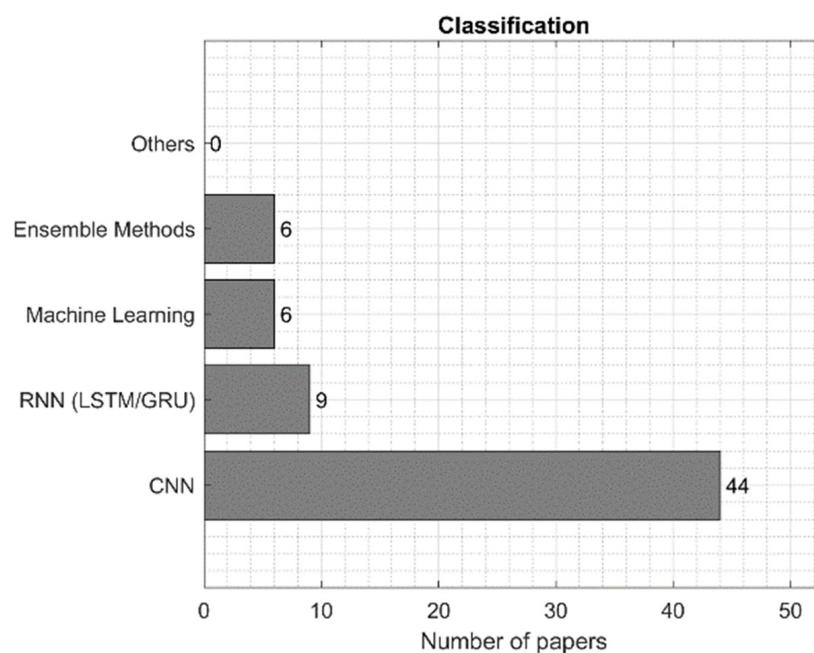
The dataset usage results in Figure 8 show that urbanSound8k was the most popular dataset used in seven publications (12.5%), while the second most popular is Primary DB with six (10.7%) of the analyzed publications. Next to that is ESC-50 and DCASE with five (8.9%) publications each, followed by ESC-10 and ICBHI datasets with four (7.1%) publications each.



**Figure 8.** Audio/sound datasets used in the analyzed articles.

The results of classification methods in Figure 9 show that convolutional neural networks (CNN) were the most used classification methods used in 44 publications (78.5%), while the second most used were various variants of recurrent neural networks (RNN)

with nine (16.1%) of analyzed publications. The ensemble learning and machine learning methods were each used in six (10.7%) publications.

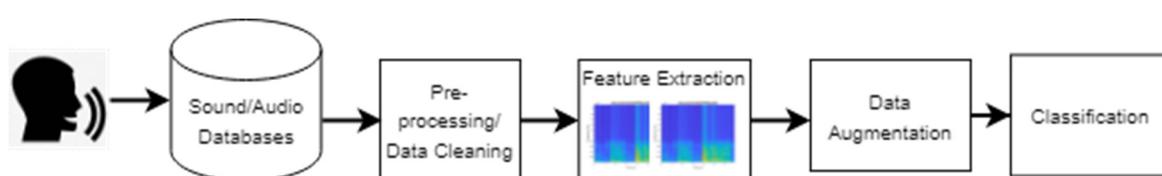


**Figure 9.** Classification methods used in the analyzed articles.

For more in-depth analysis, in the following sections we also use bubble plots, which allow visualizing of the relations between different aspects of analysis by different dimensions. It allows identifying the research gaps in the existing literature. The methods used in the analyzed studies are discussed in Section 4.

#### 4. Discussion of the Results of Review

We discuss in detail the overview of the different categories of sound classification modules as depicted in Figure 10. In the last decades, interesting findings and research methods have been introduced and implemented by researchers ranging from the creation of sound databases from environmental sounds, medical sounds, etc.



**Figure 10.** A simple sound classification architecture.

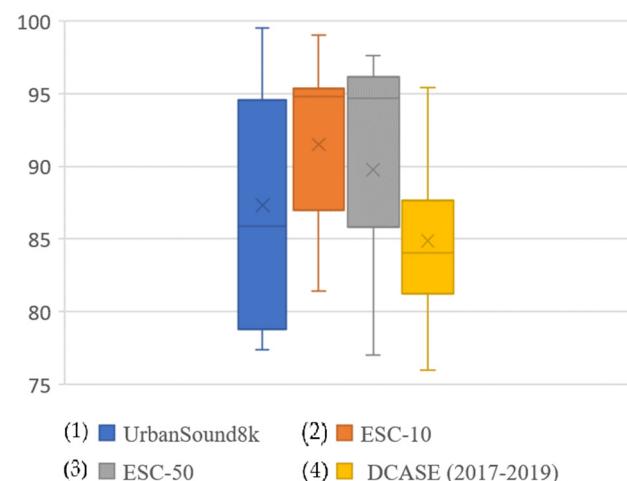
In addition, some studies have improved data cleaning methods since sound datasets are susceptible to noise and these play a major role in either improving or degrading the performance of learning models. An interesting overview of feature extraction methods was also presented, and the uniqueness of existing feature extraction techniques was also addressed. To further widen the scope of this SLR, we concentrated on the data augmentation methods applied to both audio/sound datasets and images generated from feature extractions. Finally, in this section, we analyze the various classification methods presented in previous studies and emphasize the pros and cons of the classifiers with respect to their overall performance.

#### 4.1. Sound Datasets

The real-life application of sound datasets ranges from the automatic speech recognition (ASR) system for developing smart systems (smart cities, smart healthcare, etc.) [107], acoustic scene recognition [69], music classification [108], speaker recognition [78], voice rehabilitation [109], voice disorder detection [110], speech emotion detection [111], cardiac auscultation [61], etc. Datasets presented in related studies include the following Audioset tagging [21], INTERSPEECH 2017 computational paralinguistics challenges for automatic snore sound recognition [112], animal audio datasets such as Birdz and CAT [77], and speaker verification datasets including PRISM [113], NIST SRE10, SRE08 [114], etc.

The lack of sufficient amounts of labeled data has become one of the major barriers to the advancement of sound classification. The major reason behind these can be outlined into the following: class imbalance, data privacy issues, time constraints involved in data collection, high dependency of expertise for effective annotation, etc. Another interesting factor to consider in existing sound datasets is the problem of a noisy environment within the dataset, especially when recognizing children's speech [63].

In recent years, the exceptional performance of deep learning methods in pattern recognition tasks has continued to have a great impact on modern sound classification tasks. With the advancement of deep learning, some state-of-the-art possibilities have emerged; however, this approach still struggles with poor performance (see Figure 11) due to the insufficient availability of data to solve audio/sound-related issues. On this note, the lack of sufficient sound data negatively impacts the performance of deep learning methods, especially CNN [100]. Based on studies, we were able to summarize some of the problems of existing sound dataset as follows: problem with weakly labeled data [96], noisy environment, insufficient data [115], and imbalance classes within existing sound datasets [116].



**Figure 11.** Summary of performance metrics (accuracy) based on most popularly used sound datasets: (1) Urbansound8k, (2) ESC-10, (3) ESC-50, and (4) DCASE datasets from 2017 to 2019.

The datasets used in analyzed studies are summarized in Table 2. Four types of sound datasets are distinguished: speech datasets (including emotion recognition from speech data), medical sound datasets including various sounds originating from the living body, natural sound datasets including various sounds originating from the natural world (animals), and environmental sound datasets combining sounds from the environment.

**Table 2.** Summary of sound datasets.

Datasets	Number of Categories	Number of Samples (Training/Test/Validation)	References
<b>Speech datasets</b>			
Acted Emotional Speech Dynamic Database	Five emotion classes (anger, disgust, fear, happiness, sadness)	600 phrases	Vryzas et al. [94]
AMI (meeting transcription)	n/a	100 h of meeting recordings	Qian et al. [85]
ASVspoof 2017 corpus	Two utterance classes (bona fide, spoofing)	3014/1710/13,306	Zhao et al. [103]
Dysarthric Speech Corpus in Tamil	n/a	22 dysarthric speakers, 262 sentences and 103 words	Celin et al. [53]
Baum-1a	13 emotional & mental states	1184 clips	Lalitha et al. [66]
Indian Institute of Technology Kharagpur Simulated Emotion Speech Corpus (IITKGP-SESC)	8 emotions	40 sentences	Lalitha et al. [66]
Indonesian Ethnic Speaker Recognition	70 classes	280 records	Nugroho et al. [79]
OC16-CE80 Mandarin-English mix lingual speech	50 speakers	80 h	Long et al. [70]
Punjabi Children speech corpus	n/a	1887 utterances (39 speakers)/and 485 (6 speakers)	Kadyan et al. [62]
Speech Command Dataset v1, v2	30 words	V1: 28,410 (22,236/3093/3081) V2: 46,258 (36,923/445/4890)	Pervaiz et al. [83]
Toronto Emotional Speech Set (TESS)	7 emotion classes	200 target words	Praseetha and Joby [84]
WSJCAM0 adults' speech corpus	92/20 speakers	16 h of records	Kathania et al. [63], Vecchiotti et al. [93]
PF-STAR children's speech corpus	122/60 speakers	9.4 h	Kathania et al. [63], Vecchiotti et al. [93]
EmotionDB 1999	7 emotion classes	3188 records	Garcia-Ceja et al. [58]
Surrey Audio Visual Expressed Emotion (SAVEE) database	7 emotions	4 subjects, 480 utterances	Lalitha et al. [66], Vryzas et al. [94]
VOCE Corpus Database	Stress levels	38 raw recordings (638 min of speech)	Shahnawazuddin et al. [89]
Universal Access research (UA-corpus)	Two health classes (palsy/no palsy)	19 speakers	Celin et al. [53]
<b>Medical sound datasets</b>			
Gastrointestinal Sound Dataset	6 kinds of body sounds	43,200 audio segments	Zheng et al. [106]
PhysioNet CinC Dataset	2 classes (normal/abnormal)	3240 audio files	Jeong et al. [61], Koike et al. [64]

**Table 2.** Cont.

Datasets	Number of Categories	Number of Samples (Training/Test/Validation)	References
<b>Medical sound datasets</b>			
PASCAL Heart Sound Challenge (HSC) A and B datasets	5 classes of heart sounds	A: 176 records B: 656 records	Jeong et al. [61]
Munich-Passau Snore Sound Corpus (MPSSC)	219 subjects	828 snore events	Zhang et al. [101]
ICBHI Challenge database	4 classes	6898 cycles (5.5 h)	Basu and Rana [51], Chanane and Bahoura [54], Zhao et al. [104], Rituerto-González et al. [87]
<b>Natural datasets</b>			
BIRDZ	12 classes	3101 samples	Nanni et al. [77]
CAT	10 sound classes	3000 samples	Nanni et al. [77]
NARW calls dataset	2 classes	24 h	Padovese et al. [82]
<b>Environmental sound datasets</b>			
Audioset dataset	632 audio event classes	2,084,320 sound clips	Padhy et al. [81]
DCASE dataset	11 sound classes	20 sound files	Wyatt et al. [97], Ykhlef et al. [100], Zhang et al. [102], Esmaeilpour et al. [57], Imoto [60]
Emotional Soundscapes database	na	1213 clips	Mertes et al. [74]
TAU Urban Acoustic Scenes	10 acoustic scenes	64 h of audio	Difallah et al. [56], Ma et al. [72]
Mivia Road Audio Events Dataset	2 classes (car crash/tire skidding)	400 records	Greco et al. [59]
Urbansound8K (US8K)	10 sound event classes	302 labeled sound recordings	Davis and Suresh [55], Esmaeilpour et al. [57], Lu et al. [71], Madhu and Kumaraswamy [73], Singh and Joshi [90], Mushtaq and Su [75], Mushtaq et al. [76], Salamon et al. [88]
ESC-10, ESC-50	50 classes	2000 (ESC-50)	Esmaeilpour et al. [57], Mushtaq and Su [75], Mushtaq et al. [76], Zhang et al. [102], Wyatt et al. [97]
Real Word Computing Partnership Sound Scene Database (RWCP-SSD)	105 kinds of environmental sounds	155,568 words	Ozer et al. [80]
Sound Events for Surveillance Applications (SESA)	4 sound classes	585 (480/105)	Greco et al. [59]
TUT acoustic scenes	15 acoustic scenes	312 segments (52 min)	Leng et al. [69], Yang et al. [98]
YBSS-200	10 sound classes	2000 (1600/400)	Singh and Joshi [90]

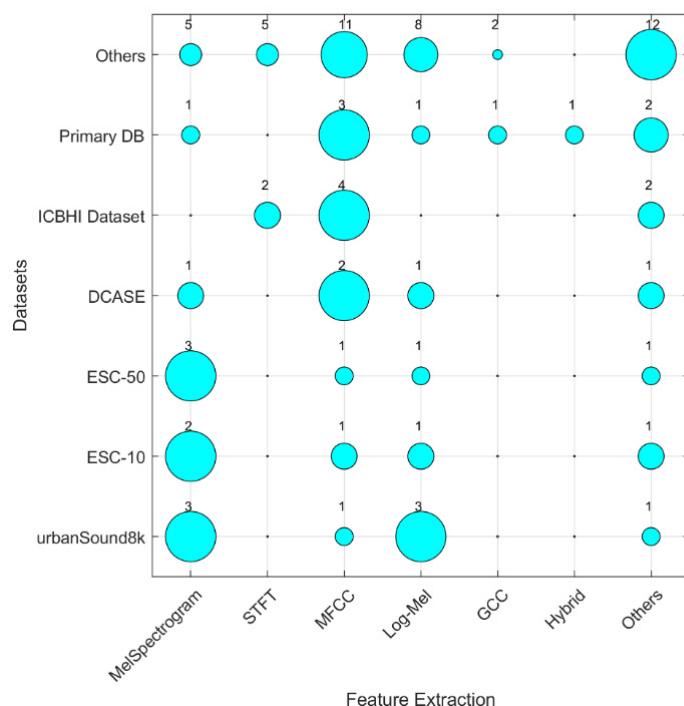
#### 4.2. Feature Extraction Methods in Sound Classification

The degree of how great or poor a model performs is also determined by the choice of features used. Therefore, it is completely important to consider the various state-of-the-art feature extraction methods used in previous studies. Some of the findings of previous work show that the use of handcrafted features suffers with the generic representation of audio signals [117].

Spectrogram characteristics have been widely used by previous researchers in different domains of sound classification, such as heartbeat sounds to detect heart diseases [64]. The features of mel-frequency cepstral coefficients (MFCC) have shown good achievement in representing sounds for the detection of respiratory diseases [51]. Ramesh et al. [86] presented combinations of different feature extraction methods for the detection of respiratory diseases based on lung sounds and the examples of feature extraction methods proposed are as follows: ZCR, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, and MFCC. Similar studies were carried out using four feature extraction methods to represent lung sound such as CQT, STFT, and mel-STFT, and finally the combination with empirical mode decomposition (EMD) [54]. The authors in [118] introduced an aggregated feature extraction scheme based on the combination of local and global acoustic features.

Class-dependent temporal-spectral structures and long-term descriptive statistics features were extracted for sound events. Other authors applied the Discrete Gabor Transform (DGT) audio image representation [119], multiresolution feature [53], hybrid method based on mel frequency cepstral coefficient and the gammatone frequency cepstral coefficient [62], inverted MFCC and extended MFCC [66], bag of audio words (BoAW) [120], narrow band auto-correlation features (NB-ACF) [121].

Figure 12 shows a distribution of publications by feature extraction method and the dataset used. Although melSpectrogram and MFCC are commonly used, the figure allows us to identify the gaps for several datasets, such as melSpectrogram that was not used with the ICBHI dataset, while STFT was also rarely used.



**Figure 12.** Bubble plot of publications by feature extraction method and the dataset used. The size of the bubble is proportional to the number of published publications (given above the bubble) and bubble coordinates correspond to the feature extraction method and the dataset used.

Based on our findings, feature representation is crucial to improve the performance of learning algorithms in the sound classification task. Audio signals usually have high dimensionality. Therefore, there is a need to develop effective feature recognition methods through the applications of better feature representation techniques with the aim of enhancing sound recognition.

The feature extraction methods used in analyzed studies are summarized in Table 3.

**Table 3.** Summary of feature extraction methods.

Features/Feature Extraction Methods	Application	References
BoAW	Medical sound recognition	Zhang et al. [101]
cepstral mean, variance normalization (CMVN)	Speech recognition	Pervaiz et al. [83]
chromagram, spectral contrast, spectral centroid, spectral roll-off	Speech and breathing sound recognition	Tran and Tsai [92]
Constant Q transform (CQT)	Lung (respiratory) sound classification	Chanane and Bahoura [54], Zhao et al. [103]
Data De-noising Auto Encoder	Respiratory sound classification	Lella and Pja [68]
Empirical mode decomposition (EMD)	Lung (respiratory) sound classification	Chanane and Bahoura [54]
Filter-bank features	Speech recognition	Long et al. [70]
Gammatonegram	Sound event recognition	Greco et al. [59]
GCC-PHAT Pattern features	Speaker recognition	Wang, Yu et al. [96]
IFFT	Speaker recognition	Zheng et al. [105]
log-gammatone spectrogram	Environmental sound classification	Zhang et al. [102]
Log-Mel	Multiple applications	Leng et al. [69], Qian et al. [85], Salamon et al. [88], Sugiura et al. [91], Wang, Yang et al. [95], Diffallah et al. [56], Ma et al. [72], Wang, Yu et al. [96], Yang et al. [98], Yella and Rajan [99], Lu et al. [71], Rituerto-González et al. [87], Koszewski and Kostek [65], Mushtaq and Su [75], Singh and Joshi [90]
Mel filter bank energy	Speech emotion recognition	Praseetha and Joby [84]
Mel Frequency Cepstral Coefficient (MFCC)		Mushtaq and Su [75], Basu and Rana [51], Vecchiotti et al. [93], Zheng et al. [106], Davis and Suresh [55], Novotny et al. [78], Nugroho et al. [79], Padovese et al. [82], Pervaiz et al. [83], Shah Nawazuddin et al. [89], Ykhlef et al. [100], Imoto [60], Ramesh et al. [86], Tran and Tsai [92], Zhao et al. [104], Wang, Yang et al. [95], Koszewski and Kostek [65], Garcia-Ceja et al. [58]
Mel spectrogram	Multiple applications	Mushtaq and Su [75], Mushtaq et al. [76], Padhy et al. [81], Billah and Nishimura [52], Tran and Tsai [92], Vryzas et al. [94], Madhu and Kumaraswamy [73], Wyatt et al. [97]

**Table 3.** Cont.

Features/Feature Extraction Methods	Application	References
Mel-Frequency Cepstral Coefficient, Gammatone frequency cepstral coefficient (MF-GFCC)	Speech recognition	Kadyan et al. [62]
Mel-STFT	Lung sound classification	Chanane and Bahoura [54]
MFCC, inverted MFCC (IMFCC), extended MFCC, extended IMFCC, LPC, Mel, Bank filter bank-derived features	Speech recognition	Lalitha et al. [66]
Multi-resolution feature extraction	Speech recognition	Celin et al. [53]
Short Time Fourier Transform (STFT)	Medical sound recognition, speech recognition	Greco et al. [59], Jeong et al. [61], Kathania et al. [63]
SIFT	Sound classification	Ozer et al. [80]
Spectral features (Spectral Centroid, RMS, Spectral Bandwidth, Spectral Contrast, Spectral Flatness, Spectral Roll-off)	Speaker recognition	Mertes et al. [74], Zhao et al. [104]
Spectrogram	Heart sound recognition, sound quality evaluation, animal audio classification	Koike et al. [64], Lee and Lee [67], Nanni et al. [77]
Speed Up Robust Feature (SURF)	Environmental sound classification	Esmaeilpour et al. [57]
STFT	Lung sounds classification	Chanane and Bahoura [54], Zheng et al. [105]
Waveform based features	Music processing	Koszewski and Kostek [65]
Zero crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off	Emotion recognition, respiratory sound classification	Garcia-Ceja et al. [58], Ramesh et al. [86]

#### 4.3. Data Augmentation Methods in Sound Classification

The increasing numbers of articles based on the application of data augmentation techniques in sound classification studies show the importance of these techniques for effective classification of sound data in various research domains such as from medical disease detection to environmental sound classification.

With some of the challenges identified with existing sound datasets, especially the lack of sufficient sound datasets, and class imbalance have huge impact on performance of classifiers, therefore, the need to increase the data samples and balancing class distributions is essential for improving sound recognition systems. A general term used to increase the overall data samples by generating a synthetic dataset is data augmentation [122]. When adopted to the audio domain, data augmentation can be achieved by using some filters on an audio signal such as pitch shifting, removal of noise, compression, time stretching, etc. [88]. In addition to increasing generalization capabilities and representation of input training data, the augmentation of data also allows the system designed to improve data significance, regardless of the available data samples [59].

Considering the vast application of deep learning methods, several researchers have been fascinated by applying different machine learning algorithms to sound classification. Data augmentation (DA) is a very popular approach that is used to enhance the number of training data. A simple description of DA is the process of creating synthetic samples by transforming training data with the purpose of enhancing performance and robustness of learning classifiers. This is one of the most effective ways to solve the problem of overfitting, which is most prevalent when using deep learning models, and thereby improves generalization ability [105]. In audio or speech data, the means of corrupting clean training speech by adding noise has been said to improve the robustness of speech

recognition systems [123]. Furthermore, previous work has shown that the accuracy of a data-driven algorithm improves significantly when more data are used for the training model, reducing the chances of overfitting [93]. On a similar note, Diffallah et al. [56] described the main objective of data augmentation as disrupting training data by injecting variation of transformed synthetic data, thus increasing training data. The importance of data augmentation cannot be overemphasized, as it has played a major role in obtaining state-of-the-art results. Another study by the authors in [124] applied the window removal method to increase training samples and improve knee health classification.

The effect of wrong choices of data augmentation schemes is more likely to generate synthetic samples with poor generalization, which would result in poor performance of classifiers. Unfortunately, in the medical application domain, the issue of insufficient data in terms of images or sound has been a major gap, as existing data still suffer from class imbalance in available databases, a typical case being the ICBHI dataset [125]. Some of the existing sound datasets also suffer from background noise, which also affect overall performance of the learning models. In addition to traditional data augmentation methods, some interesting transformation methods were also proposed to generate synthetic training samples, such as random erasing, scaling, masking (frequency and time), standardization, and trimming [126]. Another study by [71] introduced a two-stage data enhancement framework for environmental sound classification that is categorized as brute-force, class-conditional, and metric-based augmentation. Interesting data augmentation methods such as Griffin Lim and the WORLD Vocode technique were proposed by [103].

A major challenge of environmental sound recognition task is the lack of a universal database; this is because the majority of the existing acoustic databases are specific to some applications tasks and are indirectly related to natural environmental sounds [55]. The application of data enhancement methods in environmental sound classification task helps create artificial input data from existing sound or audio samples that are altered in such a way that they differ from raw samples and maintain vital information for the task at hand [74]. Traditional data augmentation or transformation methods popularly used by previous authors in generating synthetic datasets for both sound and image features extracted from sound data are as follows: time shift, pitch shift, random noise, volume gain range, vocal tract length perturbation VTLP (54), etc. In this SLR, we realized that the use of random noise data augmentation applied in our selected publication include examples such as white noise [75,79], babble noise, MUSAN noises, static noise [78], factory noise, destroyer control room, factory floor, jet cockpit [80], volvo [63], shouting, brass [91], background noise [55,90], salt and pepper noise, etc. Other data augmentation techniques from Leng et al. [69] are based on the topic model—latent Dirichlet allocation (LDA) algorithm—for generating synthetic data, mixup approaches [81], etc.

More sophisticated data augmentation methods include the application of generative adversarial networks (GANs) [85,86] and another variants of GAN such as ACGAN [104], WCCGAN [57], GAN and VAE [95], and WaveGAN [74,99]. This SLR has shown some state-of-the-art methods used for data augmentation in different sound classification tasks, speech recognition, music retrieval, etc. The generation of artificial data samples is very challenging considering the complex sequential structure of audio/sound data [74].

The most notable contributions that GANs have been made to realistic image synthesis and the modeling of motion patterns in videos. Synthetic datasets can be improved using GANs such that the statistical distribution mimics that of a real-world dataset. Numerous methods investigate how to alter spectrogram images more effectively by using GAN models [85,86]. Additionally, GANs' successes are in modeling highly dimensional data, their capacity to manage missing data, and their ability to produce accurate results.

Recently, the one-dimensional RNN models, long short-term memory (LSTM) and gated recurrent unit (GRU), have recently been merged with CNN. The 3D convolutions are yet another method for analyzing picture temporal sequences. To gain temporal information in this instance, the third dimension is employed to stack numerous consecutive frames. A possible alternative is to extend GANs with a time series-specific model, such as 3D

convolutions or RNN. Encoder-decoder networks convert a high-dimensional input into a lower-dimensional vector in latent space, which is subsequently converted back into the original high-dimensional or structural input. Data augmentation techniques decode vector samples taken from the latent space to produce novel patterns. In one instance, artificial data was produced using an LSTM-based autoencoder (LSTM-AE) [95].

Despite the importance and improvement of data augmentation in increasing training samples and overall sound classification task, it still suffers from some drawbacks, as described in previous analytical studies, which show that the expansion of data volumes may limit the structural complexity of neural networks [106]. In addition, the dataset generated from data augmentation methods lacks a proper representation and thereby leads to a poorly learned model [127]. Other advanced methods such as the conventional GAN augmentation approach are based on general random outputs resulting in an uncontrolled expansion of training samples with little or no impact on the learning classifier [74]. Another issue with GAN is the high computational complexity involved in generating synthetic data. Therefore, based on some of the shortcomings of existing data augmentation techniques in sound recognition, it is extremely crucial for future research endeavors to consider improving synthetic data representations through the adoption of effective hybridized models for generating better training samples which can be better represented, and thus improving classification models for sound classification.

The data augmentation methods used in analyzed studies are outlined in Table 4.

**Table 4.** Summary of data augmentation methods.

Reference	Data Augmentation Methods	Applications
Basu and Rana [51]	random noise, time stretching	Respiratory sound classification
Billah and Nishimura [52]	mixup	Chewing and swallowing sound classification
Celin et al. [53]	multi resolution	Speech recognition
Chanane and Bahoura [54]	Time stretching, spectrogram flipping, Vocal tract length perturbation (VTLPI),	Lung sounds classification
Davis and Suresh [55]	Time Stretching, pitch shifting, Dynamic range compression (DRC), background noise, Linear prediction cepstral coefficients (LPCC)	Environmental sound classification
Diffallah et al. [56]	Mix up	Acoustic scene classification
Esmaeilpour et al. [57]	Weighted Cycle-Consistent Generative Adversarial Network (WCCGAN)	Environmental sound classification
Garcia-Ceja et al. [58]	Random oversampling	Emotion recognition
Greco et al. [59]	Adding noise attenuating or amplifying the energy	Sound event recognition
Imoto [60]	Mask, overwrite, random copy, swap	Acoustic scene classification
Jeong et al. [61]	random noise, salt, pepper noise, SpecAugmentation (random frequency masking, time masking)	Cardiac sound classification
Kadyan et al. [62]	Adding noise (factory, babble, white)	Speech recognition
Kathania et al. [63]	Adding noise (factory, babble, white, volvo)	Speech recognition
Koike et al. [64]	trimming, scaling frequency masking, time masking, isation, random erase	Heart sound classification
Koszewski and Kostek [65]	mixup approach (linear interpolation), scale augmentation	Music classification

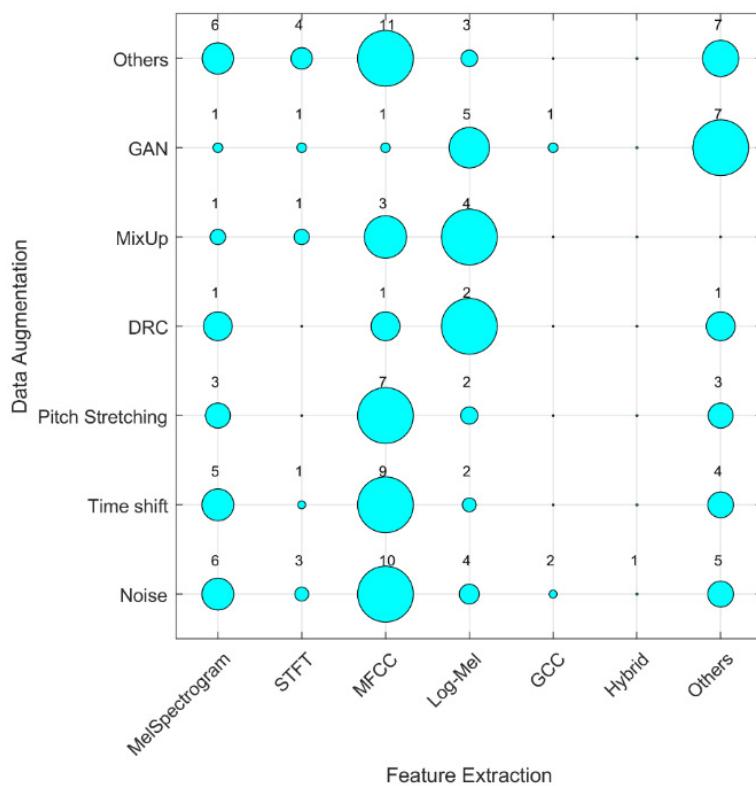
**Table 4.** Cont.

Reference	Data Augmentation Methods	Applications
Lalitha et al. [66]	Synthetic Minority Oversampling Technique (SMOTE)	Emotion recognition
Lee and Lee [67]	Bayesian approach, grayscale	Sound quality evaluation
Lella and Pja [68]	Stretching Time, Shift Pitch, Compression of Range Dynamically, Background of Noise	Respiratory sound classification
Leng et al. [69]	Topic model-LDA (Latent Dirichlet Allocation)	Acoustic scene classification
Long et al. [70]	speed, volume, noise perturbation; SpecAugment	Speech recognition
Lu et al. [71]	Time stretch, pitch shift 1, pitch shift, dynamic range compression, background noise	Environmental sound classification
Ma et al. [72]	Mix-up, Image Data Generator, temporal corp	Acoustic scene classification
Madhu and Kumaraswamy [73]	GAN, time stretching, Pitch shifting, background noise (BG), Dynamic range compression (DRC)	Environmental sound classification
Mertes et al. [74]	WaveGAN	Soundscape classification
Mushtaq and Su [75]	Offline augmentation (pitch shifting, silence trimming, time stretch, adding white noise)	Environmental sound classification
Mushtaq et al. [76]	Augmentation 1: (Zoom, Width shift, Fill mode, Brightness, Rotation, Height shift, Shear, Horizontal flip). Augmentation 2: (pitch shift, time stretch, trim silences)	Environmental sound classification
Nanni et al. [77]	Audiogmenter, image augmentation (Reflection, Rotation, Translation); Signal augmentation (speed scaling, pitch shift, volume Gain range, random noise, Time shift); Spectrogram augmentation (Randomshift, SameClass Sum, VTLN, Equalized Mixture Data Augmentation, Timeshift, random Image Warp)	Animal audio classification
Novotny et al. [78]	Reverberation, MUSAN noises, music, Babble noise, static noise	Speaker recognition
Nugroho et al. [79]	Adding white noise, pitch shifting, time stretching	Speaker recognition
Ozer et al. [80]	Adding noise: Destroyer Control Room, Speech Babble, Factory Floor-1, Jet Cockpit-1	Sound event recognition
Padhy et al. [81]	Background white noise, Time shifting, Speed Tuning, Mixing white noise with stretching or shifting, Mixup	Audio classification
Padovese et al. [82]	SpecAugment (Time warping, Masking; time, Frequency masking), Mixup	Animal sound classification
Pervaiz et al. [83]	Noise (six types)	Speech recognition
Praseetha and Joby [84]	Time stretching, embedding noise	Speech emotion recognition
Qian et al. [85]	GAN under all noisy condition, additive noise, channel distortion, reverberation.	Speech recognition
Ramesh et al. [86]	GAN	Respiratory sound classification
Rituito-González et al. [87]	Time Domain, Time-Frequency Domain (Vocal tract length perturbation (VTLP), volume adjusting, noise addition, pitch adjusting, speed adjusting	Speaker identification

**Table 4.** Cont.

Reference	Data Augmentation Methods	Applications
Salamon et al. [88]	offline augmentation (pitch shifting, time shifting, dynamic range compression (DRC), background noise)	Environmental sound classification
Shahnawazuddin et al. [89]	oversampling (SMOTE), Pitch, Speed Modifications	Speech recognition
Singh and Joshi [90]	Background noise	Sound classification
Sugiura et al. [91]	Mixup, synthetic noise (shouting, brass)	Audio classification
Tran and Tsai [92]	Background noise addition, time-shifting, time-stretching, value augmentation, a combination	Speaker identification
Vecchiotti et al. [93]	Pitch modification, time shift	Speaker identification
Vryzas et al. [94]	(1) Pitch alterations with constant tempo; (2) Overlapping windows	Speech emotion recognition
Wang, Yang et al. [95]	Generative adversarial network (GAN) and variational autoencoder (VAE)	Speech recognition
Wang, Yu et al. [96]	Room Impulse Response generator	Speaker recognition
Wyatt et al. [97]	noise	Environment sound classification
Yang et al. [98]	label smoothing mixup (spatial-mixup) technique	Acoustic scene classification
Yella and Rajan [99]	waveGAN	Respiratory sound recognition
Ykhlef et al. [100]	not disclosed	Sound event detection
Zhang et al. [101]	semi-supervised conditional Generative Adversarial Networks (scGANs)	Snore sound classification
Zhang et al. [102]	time, frequency masking, mixup	Environmental sound classification
Zhao et al. [103]	Auxiliary classifier generative adversarial network (AC-GAN), shifting, stretching traditional method	Respiratory sound classification
Zhao et al. [104]	GriffinLim, WORLD Vocode	Speaker recognition
Zheng et al. [105]	spectrum interference-based data augmentation (random cropping, random label, soft label, amplitude interference, spectrum interference)	Radio signal classification
Zheng et al. [106]	data sampling, class balance sampling, audio transformation	Biomedical sound detection

Figure 13 shows a distribution of publications using feature extraction and data augmentation methods. The figure shows that pitch stretching, time shift, and other data augmentation methods were more commonly used with MFCC features, while GAN, Mix-up and dynamic range compression (DRC) augmentations are more commonly used with log-mel features. The MFCC-based features are most used and allow achieving of state-of-the-art results in the domain.



**Figure 13.** Bubble plot of publications by feature extraction and data augmentation methods. The size of a bubble is proportional to the number of published publications (given above the bubble) and bubble coordinates correspond to the feature extraction and data augmentation methods used.

#### 4.4. Classification Methods for Sound Classification

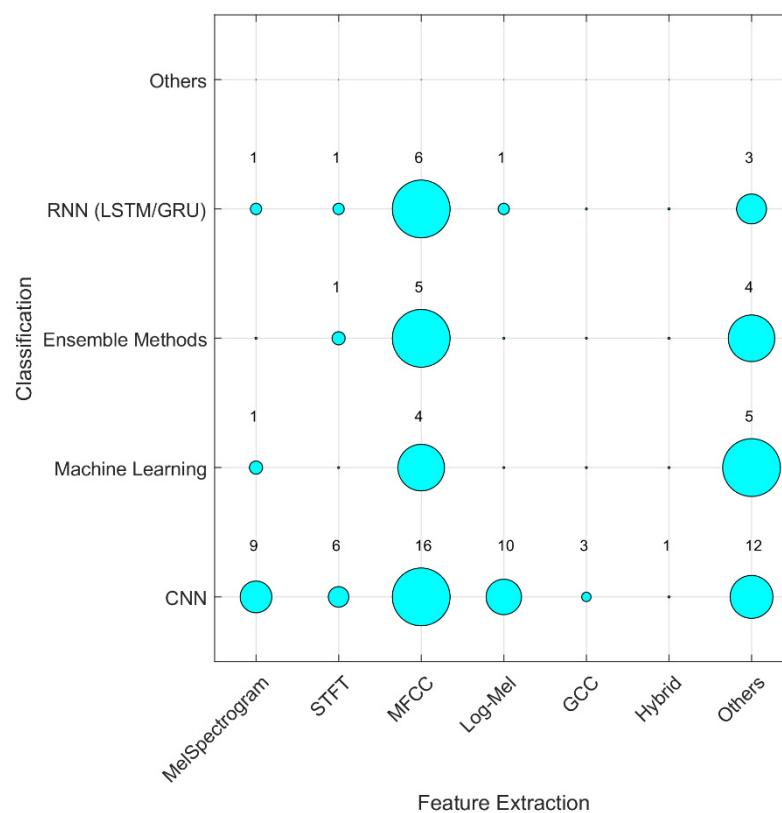
Classification is a common task in machine learning and pattern recognition. The application of deep learning methods, such as CNN models, often performs weakly in comparison to machine learning methods, such as random forest, Adaboost, etc., especially in small data [100]. For a better performance of the classifier, a larger amount of data is required to achieve a reliable estimate of the generalization error. In contrast, typical machine learning algorithms, such as ensemble classifiers, have been shown to adapt really well in learning features with improved generalization ability even in the case of a small and imbalanced dataset.

In recent years, different machine learning algorithms have been applied in the detection of sound events and in medical sound detection, and its achievement has also been of great significance. Interestingly, some single classifiers have shown to be very useful in automatic sound classification tasks such as support vector machine (SVM) [66,74,126], multilayer perceptron in person identification using speech and breath sounds [92], hidden Markov model (HMM) [53], logistic regression and linear discriminant analysis [118], etc. Other studies applied ensemble methods such as random forest [86,100], XgBoost [99], etc. Although, considering the complexity of sound and the need for the learning classifier to be extremely sensitive in order to identify different representations of sound features, traditional machine learning algorithms still suffer with the complex tasks involved in effective classification of sound data. Therefore, the choice of deep learning methods has been proven to be more efficient in a sound classification task. Deep learning, which is a subdivision of machine learning, differs from other branches of machine learning due to its ability to extract meaningful features from data through the application of a hierarchical structure and without human intervention [67]. Sound classification methods have shown a great transition from simple machine learning classifiers to advanced deep

learning classifiers, and CNNs were able to achieve significant and more accurate training results [65].

Previous research works have shown the diverse application of deep learning methods and neural network architectures for sound classification tasks, such as MLP and another four machine learning algorithms, was presented by [66], namely, VGG network [90], long short-term memory network [69], DCNN [82,83], TDNN [70,96], and GoogleNet [77]. Another study by Mushtaq et al. [76] further applied different pretrained deep learning models including DenseNet, ResNet [56,60], AlexNet, SqueezeNet, in environmental sound classification, and BiGRU Attention XGBoost [103]. Greco et al. [59] also implemented a CNN-based on the audio event recognition network (AReN). The hybrid method using the combination of SVM and GRU-RNN was presented by Zhang et al. [101], and another hybrid method by Celin et al. [53] applied the combination of the hidden Markov model of deep neural networks (DNN) for sound classification.

Figure 14 shows the distribution of the publications by feature extraction and classification methods. Most of the articles used MFCC for feature extraction and CNN for classification. MFCC was also frequently used with RNN, ensemble learning, and machine learning, however, other features have been rarely used with ensemble and machine learning methods.



**Figure 14.** Bubble plot of publications by feature extraction and classification methods. The size of a bubble is proportional to the number of published publications (given above the bubble) and bubble coordinates correspond to the feature extraction and classification methods used.

The classification methods used in analyzed studies are summarized in Table 5.

**Table 5.** Summary of classification methods.

Classification Method	References	Application
Adaboost	Ykhlef et al. [100]	Sound event detection
AlexNet	Diffallah et al. [56], Esmaeilpour et al. [57], Mushtaq et al. [76]	Environmental sound and acoustic scene classification
Audio Event Recognition Network (AReN)	Greco et al. [59]	Sound event recognition
Autoencoder DNN	Ma et al. [72], Novotny et al. [78]	Acoustic scene classification, speaker recognition
Bidirectional Encoder Representations from Transformers (BERT)	Wyatt et al. [97]	Environmental sound classification
Bidirectional Gated Recurrent Neural Networks (BiGRU)	Zhao et al. [104], Zheng et al. [106]	Sound event detection, speaker recognition
Convolutional Neural Network (CNN)	Chanane and Bahoura [54], Davis and Suresh [55], Jeong et al. [61], Koike et al. [64], Lee and Lee [67], Lella and Pja [68], Lu et al. [71], Pervaiz et al. [83], Salamon et al. [88], Sugiura et al. [91], Tran and Tsai [92], Vryzas et al. [94], Wang, Yu et al. [96], Yella and Rajan [99], Ykhlef et al. [100], Zhao et al. [103], Zheng et al. [106]	Various applications
Deep CNN (DCNN)	Madhu and Kumaraswamy [73], Mushtaq and Su [75], Zheng et al. [105]	Various applications
Deep neural network (DNN)	Kadyan et al. [62], Nugroho et al. [79], Padovese et al. [82], Pervaiz et al. [83], Shahnawazuddin et al. [89]	Various applications
DenseNet	Koszewski and Kostek [65], Mushtaq et al. [76]	Music classification, environmental sound classification
DNN-hidden Markov model (HMM)	Celin et al. [53], Kathania et al. [63]	Speech recognition
DNN trained with Restricted Boltzmann Machine	Ozer et al. [80]	Sound classification
Ensemble CNN	Riterto-González et al. [87]	Speaker identification
Gated Recurrent Unit (GRU)	Basu and Rana [51], Praseetha and Joby [84], Zhang et al. [101]	Respiratory sound classification, speech emotion recognition, snore sound classification
GoogLeNet	Esmaeilpour et al. [57], Nanni et al. [77]	Environmental sound classification
LSTM	Billah and Nishimura [52], Leng et al. [69], Long et al. [70], Pervaiz et al. [83], Vecchiotti et al. [93], Zheng et al. [106]	Multiple applications
Multi-channel CNN	Padhy et al. [81]	Audio sound recognition
Multilayer perceptron (MLP)	Lalitha et al. [66], Leng et al. [69], Tran and Tsai [92]	Speech emotion recognition, acoustic scene classification, medical sound classification

**Table 5.** Cont.

Classification Method	References	Application
Random forest (RF)	Garcia-Ceja et al. [58], Lalitha et al. [66], Ramesh et al. [86], Ykhlef et al. [100]	Speech emotion recognition, medical sound classification, sound event detection
Recurrent Neural Network (RNN)	Praseetha and Joby [84], Zhang et al. [102], Zheng et al. [106]	Speech emotion recognition, environmental sound classification, sound event detection
REPTree (RT)	Lalitha et al. [66]	Speech emotion recognition
ResNet	Diffallah et al. [56], Imoto [60], Mushtaq et al. [76], Wang, Yang et al. [95]	Acoustic scene classification, environmental sound classification
SqueezeNet	Mushtaq et al. [76]	Environmental sound classification
Support Vector Machine (SVM)	Lalitha et al. [66], Mertes et al. [74], Ramesh et al. [86], Tran and Tsai [92], Ykhlef et al. [100], Zhang et al. [101]	Speech emotion recognition, audio classification, sound event detection, biomedical sound classification
Time-Delay Neural Network (TDNN)	Long et al. [70], Wang, Yang et al. [95]	Speech recognition
VGG	Imoto [60], Mushtaq et al. [76], Nanni et al. [77], Singh and Joshi [90], Leng et al. [69]	Acoustic scene and environmental sound classification, animal sound classification
Xception	Yang et al. [98]	Acoustic scene classification

#### 4.5. What Are the Obstacles in Application of Data Augmentation in Sound Classification?

Table 6 shows the summary of information identified from the selected studies with the total number of corresponding studies. Most of the selected articles in this SLR highlighted that the increasing or high computational complexity of data augmentation methods with respect to training time is a serious obstacle. More importantly, another interesting obstacle is creation of noisy synthetic data from a noisy dataset (e.g., captured using low quality microphone) would possibly result in poor sound quality and thereby lead to poor performance of the machine learning classifier. Furthermore, obstacles such as high misclassification evidenced by high false positive rate and poor data generalization of existing data augmentation methods also play a crucial role in the sound classification task.

Despite the shortfalls identified in the selected articles, we also identified that the application of data augmentation methods in sound classification research has shown significant progress in the last five years, between 2017 and 2022. Advancement of classification or recognition of a sound dataset with integration of data augmentation techniques has helped to improve the generalization ability as recorded by the authors in [62,69,72,92,98,103]. Second, the introduction of class-specific data augmentation techniques in imbalanced datasets has helped to overcome the problem of overfitting [67,86,92,104] and thus increasing prediction performance [58,61] and classification stability [59,65,66,69,91,106]. Some of the reports from selected articles implied that the implementation of augmentation techniques achieved better classification results [51,63,67,75,76,79,84,87,89,93,95] and reduction in misclassification or error rate [62].

**Table 6.** Identified obstacles reported in the selected studies.

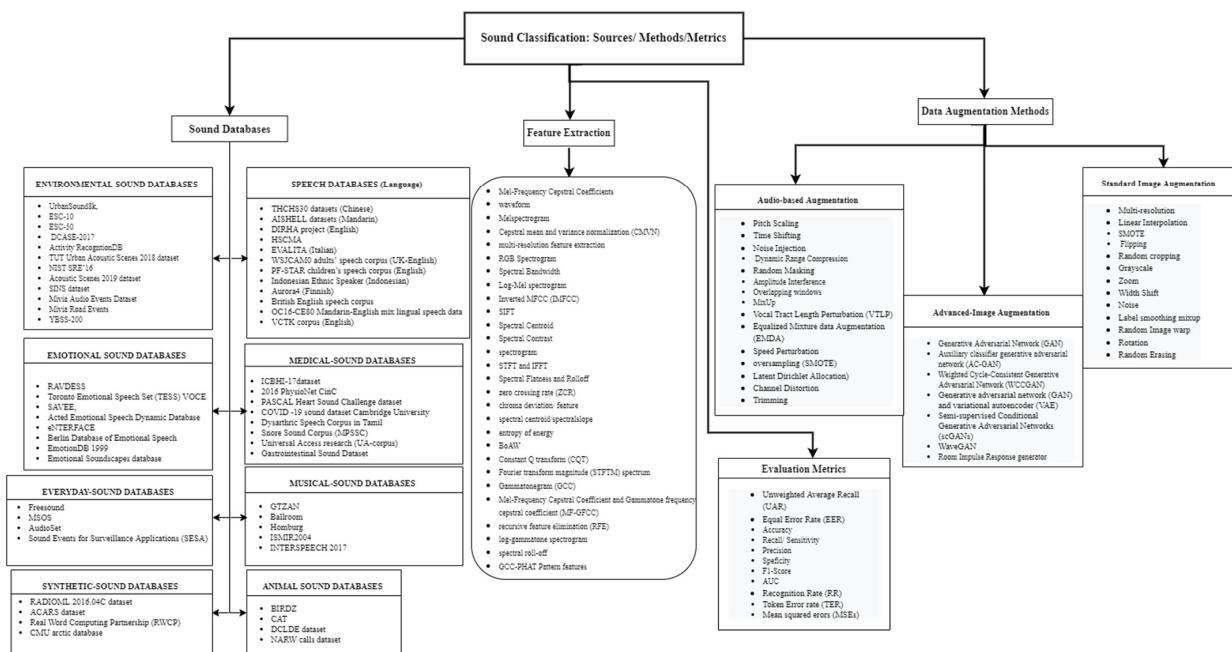
Obstacles	References	No of Papers
Limited amount of data volume	Garcia-Ceja et al. [58], Lee and Lee [67], Zhang et al. [101], Ykhlef et al. [100]	4
Lack generalization between data classes	Jeong et al. [61]	1

**Table 6.** Cont.

Obstacles	References	No of Papers
Noisy dataset/poor sound quality	Jeong et al. [61], Rituerto-González et al. [87], Lu et al. [71], Mertes et al. [74], Tran and Tsai [92], Wang et al. [96]	6
High computational complexity (Training time)	Lella and Pja [68], Zheng et al. [106], Vryzas et al. [94], Mushtaq and Su [75], Zhao et al. [103], Kadyan et al. [62], Mushtaq et al. [76], Padovese et al. [82], Pervaiz et al. [83], Sugiura et al. [91], Singh and Joshi [90], Wyatt et al. [97]	12
High Misclassification errors	Vecchiotti et al. [93], Kathania et al. [63], Lalitha et al. [66]	3
Over-smoothing effect	Esmaeilpour et al. [57]	1
Poor performance of classifier	Yang et al. [98], Vryzas et al. [94], Zhang et al. [102], Salamon et al. [88], Novotny et al. [78], Zhao et al. [103], Long et al. [70]	7
Degradation of synthetic data	Shahnawazuddin et al. [89], Chanane and Bahoura [54]	2
Class imbalance	Chanane and Bahoura [54]	1
Overfitting	Padhy et al. [81], Chanane and Bahoura [54]	2

#### 4.6. Summary of Results

We summarize the results of SLR as a taxonomy of methods used in sound classification (Figure 15), which is based on a summary of methods presented in Figure 16. The taxonomy includes the feature extraction and data augmentation methods as well as the datasets used in the research field of sound classification. The taxonomy is expected to be useful for the researchers working in the domain.



**Figure 15.** Taxonomy of methods using in sound classification.

**Figure 16.** A summary of methods analyzed in this SLR. Symbol ‘✓’ means that a corresponding method is used in the reference [51–106].

#### 4.7. Recommendations

Different augmentation methods were identified and most augmentation techniques applied had a strong impact on improving overall classification performance such as adding random noise, time shifting, time stretching and warping, pitch modification, mixup, scaling, and spectrogram transformations. These methods are recommended to be used by the researchers in the area to improve the performance of classification in case of small data availability. Furthermore, this study identified and reported the contributions and limitations of selected studies in the application of data augmentation for sound classification. All the highlighted issues arising with respect to applications of data augmentation techniques will aid future research endeavors and help future researchers in the development or implementation of more robust augmentation methods for better sound generalization and improving sound recognition systems for real-world systems.

Currently, various convolutional deep learning methods with memory such as RNN, LSTM, and their flavors such as time-delay neural network (TDNN) and bidirectional gated recurrent neural networks (BiGRU), provide best results in the field of sound classification. Deep learning models often require input data formatted as matrices and interpreted as images. In that case, the appropriate feature extraction methods commonly used are mel frequency cepstral coefficients and mel spectrograms, which allow converting one-dimensional audio sequences into images that can be used for training deep learning models. Furthermore, in addition to integrating an effective data augmentation technique, the need to identify the best representation of sound data is also a crucial factor for improving performance results in sound classification.

#### 4.8. Potential Threats to Validity

To ensure the complete findings of an SLR, the need to identify potential threats to validity is very important. Possible threats to validity could be categorized as internal, external, and conclusive validity. This SLR addressed the internal category and was constructed as a discussion among authors who agreed to formulate all research questions in this study and to identify the synergy between our formulated RQs and research objectives. Contrary to recent SLRs in other domains that focused only on journal papers, this SLR applies a thorough and careful literature search in high-quality journals and conferences. Based on this factor, the major literature search was in journal articles and conference papers considering quality of the proposed findings and the ability to provide sufficient information needed to address our defined research objectives.

The 56 selected articles in this research can represent the state-of-the-art in the field of application data augmentation techniques for sound classification. Additionally, we systematically applied automated and manual search strategies to identify existing literature. A snowballing method was adopted in the search for relevant articles with the aim of avoiding the possibility of missing any relevant article using our search query, and the search results were evaluated by thorough reading through the abstract for relevance.

For the validity of our conclusions, we reported all the details of our findings and analyzed the search results. Therefore, we were able to obtain reliable and meaningful results and that all potential threats to this study have been carefully addressed.

### 5. Conclusions

The aim of this review study is to identify the progress of data augmentation methods in sound classification tasks related to environmental and medical sounds detection, and secondly, to recognize the best methods with respect to classification and feature extraction model. To this end, we applied a quantitative and qualitative mix of systematic literature review. To the best of our knowledge, this study is one of the first to analyze the different perspectives of data augmentation methods; however, there have been some previous reviews based specifically on sound classification methods, but this SLR combines all the progress of data augmentation, classification, and feature extraction methods into a single coherent paper. Unlike the progress of data augmentation methods in computer vision,

through our systematic literature review, we identified that there is still a lagging in the application of advanced methods for generating synthetic sound data.

The literature search was based on three databases (Web of Science (WoS), Scopus, and IEEEXplore) with the intention of searching only for high-quality academic publications in journals and conferences. This systematic review of the literature shows that studies in sound classification belong to a wide variety of applications areas, environmental sound classification, music instrument recognition, mechanical sound recognition, medical disease detection, etc. In this study, to focus on recent trends and future projections of previous studies in sound classification, we narrowed down our search to studies within a five-year span (2017–2022) and 56 selected publications were identified after thorough selection criteria were applied. For better interpretation, these selected publications were analyzed based on quantitative and qualitative methods with the aim of addressing the research questions raised regarding adopted data augmentation techniques; measuring the efficiency of these techniques on learning algorithms, future recommendation of augmentation techniques to improve the efficiency of classifier in developing smarter sound classification systems that are applicable in real-life scenarios were made.

This study contributes to the state-of-the-art by thoroughly creating a new and comprehensive systematic literature review that structures data augmentation techniques for advancing the sound classification task. This study showcases the influence and the shortcomings of data augmentation methods by identifying the obstacles and areas of improvement that would benefit substantially in future research focusing on the need to improve data augmentation methods for small datasets and improving sound classification.

**Author Contributions:** Conceptualization, O.O.A.-A. and R.D.; methodology, R.D.; validation, O.O.A.-A., R.D., and S.M.; investigation, O.O.A.-A., R.D., A.Q., M.A.-O., and S.M.; resources, R.D.; writing—original draft preparation, O.O.A.-A. and R.D.; writing—review and editing, A.Q., M.A.-O., and S.M.; visualization, O.O.A.-A. and R.D.; supervision, R.D.; funding acquisition, R.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Patrício, D.I.; Rieder, R. Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. *Comput. Electron. Agric.* **2018**, *153*, 69–81. [[CrossRef](#)]
2. Le Glaz, A.; Haralambous, Y.; Kim-Dufor, D.H.; Lenca, P.; Billot, R.; Ryan, T.C.; Marsh, J.; De Vylder, J.; Walter, M.; Berrouiguet, S.; et al. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *J. Med. Internet Res.* **2021**, *23*, e15708. [[CrossRef](#)] [[PubMed](#)]
3. Rong, G.; Mendez, A.; Bou Assi, E.; Zhao, B.; Sawan, M. Artificial intelligence in healthcare: Review and prediction case studies. *Engineering* **2020**, *6*, 291–301. [[CrossRef](#)]
4. Liu, R.; Yang, B.; Zio, E.; Chen, X. Artificial intelligence for fault diagnosis of rotating machinery: A review. *Mech. Syst. Signal Process.* **2018**, *108*, 33–47. [[CrossRef](#)]
5. Zinemanas, P.; Rocamora, M.; Miron, M.; Font, F.; Serra, X. An Interpretable Deep Learning Model for Automatic Sound Classification. *Electronics* **2021**, *10*, 850. [[CrossRef](#)]
6. Crocco, M.; Cristani, M.; Trucco, A.; Murino, V. Audio surveillance: A systematic review. *ACM Comput. Surv.* **2016**, *48*, 1–46. [[CrossRef](#)]
7. Azimi, M.; Roedig, U. Room Identification with Personal Voice Assistants (Extended Abstract). In *Computer Security, Lecture Notes in Computer Science, Proceedings of the ESORICS 2021 International Workshops, Online, 4–8 October 2021*; Katsikas, S., Zheng, Y., Yuan, X., Yi, X., Eds.; Springer: Cham, Switzerland, 2022; Volume 13106, p. 13106. [[CrossRef](#)]
8. Kapočiutė-Dzikienė, J. A Domain-Specific Generative Chatbot Trained from Little Data. *Appl. Sci.* **2020**, *10*, 2221. [[CrossRef](#)]
9. Shah, S.K.; Tariq, Z.; Lee, Y. Audio IoT Analytics for Home Automation Safety. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 5181–5186. [[CrossRef](#)]
10. Gholizadeh, S.; Lemana, Z.; Baharudinb, B.T.H.T. A review of the application of acoustic emission technique in engineering. *Struct. Eng. Mech.* **2015**, *54*, 1075–1095. [[CrossRef](#)]

11. Henriquez, P.; Alonso, J.B.; Ferrer, M.A.; Travieso, C.M. Review of automatic fault diagnosis systems using audio and vibration signals. *IEEE Trans. Syst. Man Cybern. Syst.* **2014**, *44*, 642–652. [[CrossRef](#)]
12. Lozano, H.; Hernández, I.; Picón, A.; Camarena, J.; Navas, E. Audio Classification Techniques in Home Environments for Elderly/Dependant People. In *Computers Helping People with Special Needs. Lecture Notes in Computer Science, Proceedings of the 12th International Conference on Computers Helping People, Vienna, Austria, 14–16 July 2010*; Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A., Eds.; Springer: Cham, Switzerland, 2010; Volume 6179, p. 6179. [[CrossRef](#)]
13. Bear, H.L.; Heittola, T.; Mesaros, A.; Benetos, E.; Virtanen, T. City Classification from Multiple Real-World Sound Scenes. In Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019; pp. 11–15. [[CrossRef](#)]
14. Callai, S.C.; Sangiorgi, C. A review on acoustic and skid resistance solutions for road pavements. *Infrastructures* **2021**, *6*, 41. [[CrossRef](#)]
15. Blumstein, D.T.; Mennill, D.J.; Clemins, P.; Girod, L.; Yao, K.; Patricelli, G.; Kirschel, A.N.G. Acoustic monitoring in terrestrial environments using microphone arrays: Applications, technological considerations and prospectus. *J. Appl. Ecol.* **2011**, *48*, 758–767. [[CrossRef](#)]
16. Bountourakis, V.; Vrysis, L.; Papanikolaou, G. Machine learning algorithms for environmental sound recognition: Towards soundscape semantics. In Proceedings of the ACM International Conference Proceeding Series, Guangzhou, China, 7–9 October 2015. [[CrossRef](#)]
17. Najafabadi, M.M.; Villanustre, F.; Khoshgoftaar, T.M.; Naemmi, S.; Wald, R.; Muhammed, E. Deep learning applications and challenges in big data analytics. *J. Big Data* **2015**, *2*, 1. [[CrossRef](#)]
18. Marin, I.; Kuzmanic Skelin, A.; Grujic, T. Empirical Evaluation of the Effect of Optimization and Regularization Techniques on the Generalization Performance of Deep Convolutional Neural Network. *Appl. Sci.* **2020**, *10*, 7817. [[CrossRef](#)]
19. Bergstra, J.; Yamins, D.; Cox, D.D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In Proceedings of the 30th International Conference on Machine Learning Part 1 (ICML), Baltimore, MD, USA, 17–23 July 2013; pp. 115–123.
20. Khalid, R.; Javaid, N. A survey on hyperparameters optimization algorithms of forecasting models in smart grid. *Sustain. Cities Soc.* **2020**, *61*, 2275. [[CrossRef](#)]
21. Kalliola, J.; Kapočiūte-Dzikiene, J.; Damaševičius, R. Neural network hyperparameter optimization for prediction of real estate prices in helsinki. *PeerJ Comput. Sci.* **2021**, *7*, 1–25. [[CrossRef](#)] [[PubMed](#)]
22. Połap, D.; Woźniak, M.; Hołubowski, W.; Damaševičius, R. A heuristic approach to the hyperparameters in training spiking neural networks using spike-timing-dependent plasticity. *Neural Comput. Appl.* **2021**, *34*, 13187–13200. [[CrossRef](#)]
23. Saeed, N.; Nyberg, R.G.; Alam, M.; Dougherty, M.; Jooma, D.; Rebreyend, P. Classification of the Acoustics of Loose Gravel. *Sensors* **2021**, *21*, 4944. [[CrossRef](#)]
24. Castro Martinez, A.M.; Spille, C.; Roßbach, J.; Kollmeier, B.; Meyer, B.T. Prediction of speech intelligibility with DNN-based performance measures. *Comput. Speech Lang.* **2022**, *74*, 1329. [[CrossRef](#)]
25. Han, D.; Kong, Y.; Han, J.; Wang, G. A survey of music emotion recognition. *Front. Comput. Sci.* **2022**, *16*, 166335. [[CrossRef](#)]
26. Alonso Hernández, J.B.; Barragán Pulido, M.L.; Gil Bordón, J.M.; Ferrer Ballester, M.Á.; Travieso González, C.M. Speech evaluation of patients with alzheimer’s disease using an automatic interviewer. *Expert Syst. Appl.* **2022**, *192*, 6386. [[CrossRef](#)]
27. Tagawa, Y.; Maskeliūnas, R.; Damaševičius, R. Acoustic Anomaly Detection of Mechanical Failures in Noisy Real-Life Factory Environments. *Electronics* **2021**, *10*, 2329. [[CrossRef](#)]
28. Qurthobi, A.; Maskeliūnas, R.; Damaševičius, R. Detection of Mechanical Failures in Industrial Machines Using Overlapping Acoustic Anomalies: A Systematic Literature Review. *Sensors* **2022**, *22*, 3888. [[CrossRef](#)] [[PubMed](#)]
29. Domingos, L.C.F.; Santos, P.E.; Skelton, P.S.M.; Brinkworth, R.S.A.; Sammut, K. A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance. *Sensors* **2022**, *22*, 2181. [[CrossRef](#)]
30. Ji, C.; Mudiyanselage, T.B.; Gao, Y.; Pan, Y. A review of infant cry analysis and classification. *Eurasip. J. Audio Speech Music. Process.* **2021**, *2021*, 1975. [[CrossRef](#)]
31. Qian, K.; Janott, C.; Schmitt, M.; Zhang, Z.; Heiser, C.; Hemmert, W.; Schuller, B.W. Can machine learning assist locating the excitation of snore sound? A review. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 1233–1246. [[CrossRef](#)]
32. Meyer, J.; Dentel, L.; Meunier, F. Speech Recognition in Natural Background Noise. *PLoS ONE* **2013**, *8*, e79279. [[CrossRef](#)]
33. Bahle, G.; Fortes Rey, V.; Bian, S.; Bello, H.; Lukowicz, P. Using Privacy Respecting Sound Analysis to Improve Bluetooth Based Proximity Detection for COVID-19 Exposure Tracing and Social Distancing. *Sensors* **2021**, *21*, 5604. [[CrossRef](#)]
34. Holzapfel, A.; Sturm, B.L.; Coeckelbergh, M. Ethical Dimensions of Music Information Retrieval Technology. *Trans. Int. Soc. Music. Inf. Retriev.* **2018**, *1*, 44–55. [[CrossRef](#)]
35. Tsalera, E.; Papadakis, A.; Samarakou, M. Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning. *J. Sens. Actuator Netw.* **2021**, *10*, 72. [[CrossRef](#)]
36. Alías, F.; Socoró, J.C.; Sevillano, X. A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds. *Appl. Sci.* **2016**, *6*, 143. [[CrossRef](#)]
37. Wang, J.; Lee, Y.; Lin, C.; Siahaan, E.; Yang, C. Robust environmental sound recognition with fast noise suppression for home automation. *IEEE Trans. Autom. Sci. Eng.* **2015**, *12*, 1235–1242. [[CrossRef](#)]

38. Steinfath, E.; Palacios-Muñoz, A.; Rottschäfer, J.R.; Yuezak, D.; Clemens, J. Fast and accurate annotation of acoustic signals with deep neural networks. *eLife* **2021**, *10*, e68837. [CrossRef] [PubMed]
39. Sun, Y.; Wong, A.K.C.; Kamel, M.S. Classification of Imbalanced Data: A Review. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 687–719. [CrossRef]
40. Dong, X.; Yin, B.; Cong, Y.; Du, Z.; Huang, X. Environment sound event classification with a two-stream convolutional neural network. *IEEE Access* **2020**, *8*, 125714–125721. [CrossRef]
41. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [CrossRef] [PubMed]
42. Zhao, Y.X.; Li, Y.; Wu, N. Data augmentation and its application in distributed acoustic sensing data denoising. *Geophys. J. Int.* **2021**, *228*, 119–133. [CrossRef]
43. Abeßer, J. A review of deep learning based methods for acoustic scene classification. *Appl. Sci.* **2020**, *10*, 20. [CrossRef]
44. Bahmei, B.; Birmingham, E.; Arzanpour, S. CNN-RNN and data augmentation using deep convolutional generative adversarial network for environmental sound classification. *IEEE Signal Process. Lett.* **2022**, *29*, 682–686. [CrossRef]
45. Horwath, J.P.; Zakharov, D.N.; Mégret, R.; Stach, E.A. Understanding important features of deep learning models for segmentation of high-resolution transmission electron microscopy images. *NPJ Comput. Mater.* **2006**, *108*, 2–9. [CrossRef]
46. Feurer, M.; Hutter, F. Hyperparameter Optimization. In *Automated Machine Learning. The Springer Series on Challenges in Machine Learning*; Hutter, F., Kotthoff, L., Vanschoren, J., Eds.; Springer: Cham, Switzerland, 2019. [CrossRef]
47. Van Wee, B.; Banister, D. How to write a literature review paper? *Transp. Rev.* **2016**, *36*, 278–288.
48. Kitchenham, B.; Brereton, O.P.; Budgen, D.; Turner, M.; Bailey, J.; Linkman, S. Systematic literature reviews in software engineering—A systematic literature review. *Inf. Softw. Technol.* **2009**, *51*, 7–15. [CrossRef]
49. Badampudi, D.; Wohlin, C.; Petersen, K. Experiences from using snowballing and database searches in systematic literature studies. In Proceedings of the ACM International Conference Proceeding Series, Edinburgh, UK, 27–29 April 2015. [CrossRef]
50. Liberati, A.; Altman, D.G.; Tetzlaff, J.; Mulrow, C.; Gøtzsche, P.C.; Ioannidis, J.P.; Moher, D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *J. Clin. Epidemiol.* **2009**, *62*, e1–e34. [CrossRef] [PubMed]
51. Basu, V.; Rana, S. Respiratory diseases recognition through respiratory sound with the help of deep neural network. In Proceedings of the 2020 4th International Conference on Computational Intelligence and Networks (CINE), Online, 2–5 September 2020; pp. 1–6.
52. Billah, M.M.; Nishimura, M. A data augmentation-based technique to classify chewing and swallowing using LSTM. In Proceedings of the 2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech), Kyoto, Japan, 10–12 March 2020; pp. 84–85.
53. Celin, M.T.A.; Nagarajan, T.; Vijayalakshmi, P. Data augmentation using virtual microphone array synthesis and multi-resolution feature extraction for isolated word dysarthric speech recognition. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 346–354. [CrossRef]
54. Chanane, H.; Bahoura, M. Convolutional Neural Network-based Model for Lung Sounds Classification. In Proceedings of the 2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS), East Lansing, MI, USA, 7–10 August 2021; pp. 555–558.
55. Davis, N.; Suresh, K. Environmental sound classification using deep convolutional neural networks and data augmentation. In Proceedings of the 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS), Thiruvananthapuram, India, 6–8 December 2018; pp. 41–45.
56. Diffallah, Z.; Ykhlef, H.; Bouarfa, H.; Ykhlef, F. Impact of Mixup Hyperparameter Tuning on Deep Learning-based Systems for Acoustic Scene Classification. In Proceedings of the 2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI), Tebessa, Algeria, 21–22 September 2021; pp. 1–6.
57. Esmaeilpour, M.; Cardinal, P.; Koerich, A.L. Unsupervised feature learning for environmental sound classification using Weighted Cycle-Consistent Generative Adversarial Network. *Appl. Soft Comput.* **2019**, *86*, 105912. [CrossRef]
58. Garcia-Ceja, E.; Riegler, M.; Kvernberg, A.K.; Torresen, J. User-adaptive models for activity and emotion recognition using deep transfer learning and data augmentation. *User Model. User-Adapt. Interact.* **2020**, *30*, 365–393. [CrossRef]
59. Greco, A.; Petkov, N.; Saggese, A.; Vento, M. Aren: A deep learning approach for sound event recognition using a brain inspired representation. *IEEE Trans. Inf. Forens. Secur.* **2020**, *15*, 3610–3624. [CrossRef]
60. Imoto, K. Acoustic Scene Classification Using Multichannel Observation with Partially Missing Channels. In Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23–27 August 2021; pp. 875–879.
61. Jeong, Y.; Kim, J.; Kim, D.; Kim, J.; Lee, K. Methods for improving deep learning-based cardiac auscultation accuracy: Data augmentation and data generalization. *Appl. Sci.* **2021**, *11*, 4544. [CrossRef]
62. Kadyan, V.; Bawa, P.; Hasija, T. In domain training data augmentation on noise robust punjabi children speech recognition. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *13*, 03468. [CrossRef]
63. Kathania, H.K.; Kadiri, S.R.; Alku, P.; Kurimo, M. Using data augmentation and time-scale modification to improve ASR of children’s speech in noisy environments. *Appl. Sci.* **2021**, *11*, 8420. [CrossRef]

64. Koike, T.; Qian, K.; Schuller, B.W.; Yamamoto, Y. Transferring cross-corpus knowledge: An investigation on data augmentation for heart sound classification. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), Guadalajara, Mexico, 26 July 2021; pp. 1976–1979.
65. Koszewski, D.; Kostek, B. Musical instrument tagging using data augmentation and effective noisy data processing. *AES J. Audio Eng. Soc.* **2020**, *68*, 57–65. [CrossRef]
66. Lalitha, S.; Gupta, D.; Zakariah, M.; Alotaibi, Y.A. Investigation of multilingual and mixed-lingual emotion recognition using enhanced cues with data augmentation. *Appl. Acoust.* **2020**, *170*, 107519. [CrossRef]
67. Lee, H.; Lee, J. Neural network prediction of sound quality via domain knowledge-based data augmentation and bayesian approach with small data sets. *Mech. Syst. Signal Process.* **2021**, *157*, 107713. [CrossRef]
68. Lella, K.K.; Pja, A. Automatic COVID-19 disease diagnosis using 1D convolutional neural network and augmentation with human respiratory sound based on parameters: Cough, breath, and voice. *AIMS Public Health* **2021**, *8*, 240–264. [CrossRef] [PubMed]
69. Leng, Y.; Zhao, W.; Lin, C.; Sun, C.; Wang, R.; Yuan, Q.; Li, D. LDA-based data augmentation algorithm for acoustic scene classification. *Knowl.-Based Syst.* **2020**, *195*, 105600. [CrossRef]
70. Long, Y.; Li, Y.; Zhang, Q.; Wei, S.; Ye, H.; Yang, J. Acoustic data augmentation for mandarin-english code-switching speech recognition. *Appl. Acoust.* **2020**, *161*, 107175. [CrossRef]
71. Lu, R.; Duan, Z.; Zhang, C. Metric learning based data augmentation for environmental sound classification. In Proceedings of the 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 15–18 October 2017; pp. 1–5.
72. Ma, X.; Shao, Y.; Ma, Y.; Zhang, W.Q. Deep Semantic Encoder-Decoder Network for Acoustic Scene Classification with Multiple Devices. In Proceedings of the 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, 7–10 December 2020; pp. 365–370.
73. Madhu, A.; Kumaraswamy, S. Data augmentation using generative adversarial network for environmental sound classification. In Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO), Coruna, Spain, 2–6 September 2019; pp. 1–5.
74. Mertes, S.; Baird, A.; Schiller, B.W.; André, E. An evolutionary-based generative approach for audio data augmentation. In Proceedings of the 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Online, 21–24 September 2020; pp. 1–6.
75. Mushtaq, Z.; Su, S. Environmental sound classification using a regularized deep convolutional neural network with data augmentation. *Appl. Acoust.* **2020**, *167*, 107389. [CrossRef]
76. Mushtaq, Z.; Su, S.; Tran, Q. Spectral images based environmental sound classification using CNN with meaningful data augmentation. *Appl. Acoust.* **2021**, *172*, 107581. [CrossRef]
77. Nanni, L.; Maguolo, G.; Paci, M. Data augmentation approaches for improving animal audio classification. *Ecol. Inform.* **2020**, *57*, 101084. [CrossRef]
78. Novotný, O.; Plchot, O.; Glembek, O.; Černocký, J.; Burget, L. Analysis of DNN speech signal enhancement for robust speaker recognition. *Comput. Speech Lang.* **2019**, *58*, 403–421. [CrossRef]
79. Nugroho, K.; Noersasongko, E.; Purwanto; Muljono; Setiadi, D.R.I.M. Enhanced indonesian ethnic speaker recognition using data augmentation deep neural network. *J. King Saud Univ. Comput. Inf. Sci.* **2021**, *34*, 4375–4384. [CrossRef]
80. Ozer, I.; Ozer, Z.; Findik, O. Lanczos kernel based spectrogram image features for sound classification. *Procedia Comput. Sci.* **2017**, *111*, 137–144. [CrossRef]
81. Padhy, S.; Tiwari, J.; Rathore, S.; Kumar, N. Emergency signal classification for the hearing impaired using multi-channel convolutional neural network architecture. In Proceedings of the 2019 IEEE Conference on Information and Communication Technology, Surabaya, Indonesia, 18 July 2019; pp. 1–6.
82. Padovese, B.; Frazao, F.; Kirsebom, O.S.; Matwin, S. Data augmentation for the classification of north atlantic right whales upcalls. *J. Acoust. Soc. Am.* **2021**, *149*, 2520–2530. [CrossRef] [PubMed]
83. Pervaiz, A.; Hussain, F.; Israr, H.; Tahir, M.A.; Raja, F.R.; Baloch, N.K.; Zikria, Y.B. Incorporating noise robustness in speech command recognition by noise augmentation of training data. *Sensors* **2020**, *20*, 2326. [CrossRef]
84. Praseetha, V.M.; Joby, P.P. Speech emotion recognition using data augmentation. *Int. J. Speech Technol.* **2021**, *9883*. [CrossRef]
85. Qian, Y.; Hu, H.; Tan, T. Data augmentation using generative adversarial networks for robust speech recognition. *Speech Commun.* **2019**, *114*, 1–9. [CrossRef]
86. Ramesh, V.; Vatanparvar, K.; Nemati, E.; Nathan, V.; Rahman, M.M.; Kuang, J. CoughGAN: Generating synthetic coughs that improve respiratory disease classification. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), Online, 20–24 July 2020; pp. 5682–5688.
87. Rituerto-González, E.; Mínguez-Sánchez, A.; Gallardo-Antolín, A.; Peláez-Moreno, C. Data augmentation for speaker identification under stress conditions to combat gender-based violence. *Appl. Sci.* **2019**, *9*, 2298. [CrossRef]
88. Salamon, J.; Bello, J.P. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process. Lett.* **2017**, *24*, 279–283. [CrossRef]
89. Shah Nawazuddin, S.; Adiga, N.; Kathania, H.K.; Sai, B.T. Creating speaker independent ASR system through prosody modification based data augmentation. *Pattern Recognit. Lett.* **2020**, *131*, 213–218. [CrossRef]

90. Singh, J.; Joshi, R. Background sound classification in speech audio segments. In Proceedings of the 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timisoara, Romania, 10–12 October 2019; pp. 1–6.
91. Sugiura, T.; Kobayashi, A.; Utsuro, T.; Nishizaki, H. Audio Synthesis-based Data Augmentation Considering Audio Event Class. In Proceedings of the 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), Online, 12–15 October 2021.
92. Tran, V.T.; Tsai, W.H. Stethoscope-Sensed Speech and Breath-Sounds for Person Identification with Sparse Training Data. *IEEE Sens. J.* **2019**, *20*, 848–859. [[CrossRef](#)]
93. Vecchiotti, P.; Pepe, G.; Principi, E.; Squartini, S. Detection of activity and position of speakers by using deep neural networks and acoustic data augmentation. *Expert Syst. Appl.* **2019**, *134*, 53–65. [[CrossRef](#)]
94. Vryzas, N.; Kotsakis, R.; Liatsou, A.; Dimoulas, C.; Kalliris, G. Speech emotion recognition for performance interaction. *AES: J. Audio Eng. Soc.* **2018**, *66*, 457–467. [[CrossRef](#)]
95. Wang, E.K.; Yu, J.; Chen, C.; Kumari, S.; Rodrigues, J.J.P.C. Data augmentation for internet of things dialog system. *Mob. Netw. Appl.* **2020**, *27*, 1–14. [[CrossRef](#)]
96. Wang, S.; Yang, Y.; Wu, Z.; Qian, Y.; Yu, K. Data augmentation using deep generative models for embedding based speaker recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2598–2609. [[CrossRef](#)]
97. Wyatt, S.; Elliott, D.; Aravamudan, A.; Otero, C.E.; Otero, L.D.; Anagnostopoulos, G.C.; Lam, E. Environmental sound classification with tiny transformers in noisy edge environments. In Proceedings of the 2021 IEEE 7th World Forum on Internet of Things (WF-IoT), Online, 26 July 2021; pp. 309–314.
98. Yang, L.; Tao, L.; Chen, X.; Gu, X. Multi-scale semantic feature fusion and data augmentation for acoustic scene classification. *Appl. Acoust.* **2020**, *163*, 107238. [[CrossRef](#)]
99. Yella, N.; Rajan, B. Data Augmentation using GAN for Sound based COVID 19 Diagnosis. In Proceedings of the 2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Cracow, Poland, 22–25 September 2021; Volume 2, pp. 606–609.
100. Ykhlef, H.; Ykhlef, F.; Chiboub, S. Experimental Design and Analysis of Sound Event Detection Systems: Case Studies. In Proceedings of the 2019 6th International Conference on Image and Signal Processing and their Applications (ISPA), Mostaganem, Algeria, 24–25 November 2019; pp. 1–6.
101. Zhang, Z.; Han, J.; Qian, K.; Janott, C.; Guo, Y.; Schuller, B. Snore-GANs: Improving automatic snore sound classification with synthesized data. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 300–310. [[CrossRef](#)] [[PubMed](#)]
102. Zhang, Z.; Xu, S.; Zhang, S.; Qiao, T.; Cao, S. Learning attentive representations for environmental sound classification. *IEEE Access* **2019**, *7*, 130327–130339. [[CrossRef](#)]
103. Zhao, X.; Shao, Y.; Mai, J.; Yin, A.; Xu, S. Respiratory Sound Classification Based on BiGRU-Attention Network with XGBoost. In Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Republic of Korea, 16–19 December 2020; pp. 915–920.
104. Zhao, Y.; Tognoni, R.; Seeram, V. Replay anti-spoofing countermeasure based on data augmentation with post selection. *Comput. Speech Lang.* **2020**, *64*, 1115. [[CrossRef](#)]
105. Zheng, Q.; Zhao, P.; Li, Y.; Wang, H.; Yang, Y. Spectrum interference-based two-level data augmentation method in deep learning for automatic modulation classification. *Neural Comput. Appl.* **2021**, *33*, 7723–7745. [[CrossRef](#)]
106. Zheng, X.; Zhang, C.; Chen, P.; Zhao, K.; Jiang, H.; Jiang, Z.; Jia, W. A CRNN System for Sound Event Detection Based on Gastrointestinal Sound Dataset Collected by Wearable Auscultation Devices. *IEEE Access* **2020**, *8*, 157892–157905. [[CrossRef](#)]
107. Ismail, A.; Abdlerazek, S.; El-Henawy, I.M. Development of Smart Healthcare System Based on Speech Recognition Using Support Vector Machine and Dynamic Time Warping. *Sustainability* **2020**, *12*, 2403. [[CrossRef](#)]
108. Takahashi, N.; Gygli, M.; Van Gool, L. AENet: Learning deep audio features for video analysis. *IEEE Trans. Multimed.* **2018**, *20*, 513–524. [[CrossRef](#)]
109. Meltzner, G.S.; Heaton, J.T.; Deng, Y.; De Luca, G.; Roy, S.H.; Kline, J.C. Silent speech recognition as an alternative communication device for persons with laryngectomy. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2386–2398. [[CrossRef](#)] [[PubMed](#)]
110. Borsky, M.; Mehta, D.D.; Van Stan, J.H.; Gudnason, J. Modal and nonmodal voice quality classification using acoustic and electroglottographic features. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2281–2291. [[CrossRef](#)] [[PubMed](#)]
111. Lech, M.; Stolar, M.; Best, C.; Bolia, R. Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding. *Front. Comput. Sci.* **2020**, *2*, 14. [[CrossRef](#)]
112. Schuller, B.; Steidl, S.; Batliner, A.; Bergelson, E.; Krajewski, J.; Janott, C.; Zafeiriou, S. The interspeech 2017 computational paralinguistics challenge: Addressee, cold snoring. In *Computational Paralinguistics Challenge (ComParE)*; Interspeech: London, UK, 2017; pp. 3442–3446.
113. Ferrer, L.; Bratt, H.; Burget, L.; Cernocky, H.; Glembek, O.; Graciarena, M.; Lawson, A.; Lei, Y.; Matejka, P.; Pichot, O.; et al. Promoting robustness for speaker modeling in the community: The PRISM evaluation set. In Proceedings of the NIST Speaker Recognition Analysis Workshop (SRE11), Atlanta, GA, USA, 1–3 March 2011; pp. 1–7.
114. Sun, H.; Ma, B. The NIST SRE summed channel speaker recognition system. In *Interspeech 2014*; ISCA: New York, NY, USA, 2014. [[CrossRef](#)]
115. Xie, H.; Virtanen, T. Zero-Shot Audio Classification Via Semantic Embeddings. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1233–1242. [[CrossRef](#)]

116. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 1925. [[CrossRef](#)]
117. Papakostas, M.; Spyrou, E.; Giannakopoulos, T.; Siantikos, G.; Sgouropoulos, D.; Mylonas, P.; Makedon, F. Deep Visual Attributes vs. Hand-Crafted Audio Features on Multidomain Speech Emotion Recognition. *Computation* **2017**, *5*, 26. [[CrossRef](#)]
118. Ye, J.; Kobayashi, T.; Murakawa, M. Urban sound event classification based on local and global features aggregation. *Appl. Acoust.* **2017**, *117*, 246–256. [[CrossRef](#)]
119. Lachambre, H.; Ricaud, B.; Stempfel, G.; Torrésani, B.; Wiesmeyr, C.; Onchis-Moaca, D. Optimal Window and Lattice in Gabor Transform. Application to Audio Analysis. In Proceedings of the International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, Timisoara, Romania, 21–24 September 2015; Volume 17, pp. 109–112. [[CrossRef](#)]
120. Schmitt, M.; Janott, C.; Pandit, V.; Qian, K.; Heiser, C.; Hemmert, W.; Schuller, B. A Bag-of-Audio-Words Approach for Snore Sounds’ Excitation Localisation. In Proceedings of the 12th ITG Symposium on Speech Communication, Paderborn, Germany, 5–7 October 2016; pp. 1–5.
121. Valero, X.; Alías, F. Narrow-band autocorrelation function features for the automatic recognition of acoustic environments. *J. Acoust. Soc. Am.* **2013**, *134*, 880–890. [[CrossRef](#)]
122. Iwana, B.K.; Uchida, S. An empirical survey of data augmentation for time series classification with neural networks. *PLoS ONE* **2021**, *16*, e0254841. [[CrossRef](#)]
123. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. Audio augmentation for speech recognition. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
124. Ozmen, G.C.; Gazi, A.H.; Gharehbaghi, S.; Richardson, K.L.; Safaei, M.; Whittingslow, D.C.; Inan, O.T. An Interpretable Experimental Data Augmentation Method to Improve Knee Health Classification Using Joint Acoustic Emissions. *Ann. Biomed. Eng.* **2021**, *49*, 2399–2411. [[CrossRef](#)] [[PubMed](#)]
125. Rocha, B.M.; Filos, D.; Mendes, L.; Serbes, G.; Ulukaya, S.; Kahya, Y.P.; Jakovljevic, N.; Turukalo, T.L.; Vogiatzis, I.M.; Perantoni, E.; et al. An open access database for the evaluation of respiratory sound classification algorithms. *Physiol. Meas.* **2019**, *40*, 035001.
126. Wei, S.; Zou, S.; Liao, F.; Lang, W. A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification. *J. Phys. Conf. Ser.* **2020**, *1453*, 012085. [[CrossRef](#)]
127. Araújo, T.; Aresta, G.; Mendoca, L.; Penas, S.; Maia, C.; Carneiro, A.; Campilho, A.; Mendonça, A.M. Data Augmentation for Improving Proliferative Diabetic Retinopathy Detection in Eye Fundus Images. *IEEE Access* **2020**, *8*, 182462–182474. [[CrossRef](#)]