A Survey of Data Augmentation for Audio Classification

Lucas Ferreira-Paiva [*] **Elizabeth Alfaro-Espinoza** [**]
Vinicius M. Almeida [***] **Leonardo B. Felix** [*]
Rodolpho V. A. Neves [*]

[*] *Núcleo Interdisciplinar de Análise de Sinais (NIAS), Universidade Federal de Viçosa, MG (e-mail: {lucas.f.paiva,leobonato, rodolpho.neves}@ufv.br)*
[**] *Programa de Pós-Graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, MG (e-mail: elizaespinoza@ufmg.br)*
[***] *Centro de Ciências Exatas e Tecnológicas - Engenharia de Computação, Centro Universitário de Viçosa, MG (e-mail: viniciusmartins@univicosa.com.br)*

Abstract: One of the most effective methods for reducing overfitting in deep learning models for audio classification is data augmentation. The range of techniques available, as well as a lack of understanding of the most efficient ones, can result in severe time and processing power costs. This survey covers numerous techniques, tools, and datasets for offline data augmentation to assist in the selection and implementation of data augmentation strategies to improve audio classification models in Environmental Sound Classification, Music Information Retrieval, and Automatic Speech Recognition. Finally, we present a short review of papers that apply data augmentation in Environmental Sound Classification which indicates that the use of spectrogram and audio augmentation has considerable potential for improving the performance of convolutional models, especially for small datasets with increases in accuracy of up to 30%. However, the accuracy gains achieved may be insufficient to justify the additional computer burden depending on the application. Furthermore, the usage of image data augmentation is unsuitable for audio data.

Keywords: Sound systems; Acoustic noise; Data processing; Artificial neural networks; Multimedia systems.

1. INTRODUCTION

Data Augmentation (DA) is defined as the creation of new data by adding deformations to increase the variety of the data so that these deformations do not change their semantic value. DA application for audio signals, including natural, and non-natural sounds, can be categorized accordingly to where the DA techniques are applied: the raw audio or to its spectrogram. Classification is the most important task in Environmental Sound Classification (ESC) and is highly noted in Music Informational Retrieval (MIR) and Automatic Speech Classification (ASR).

In ESC, the applications include urban noise recognition and mitigation (J.Cao et al., 2019; Bello et al., 2019), and identification of animals by their sounds (Pandeya and Lee, 2018a; Nanni et al., 2020). Genre (Aguiar et al., 2018), instruments (Wu et al., 2018), and emotion (Seo and Huh, 2019) classification are prominent areas in MIR, while in ASR, speech commands classification (Solovyev et al., 2020; Dominguez-Morales et al., 2018) is a traditional task.

[⋆] Authors thank to the Coordination for the Improvement of Higher Education Personnel (CAPES) - Financing code 001 for providing financial support for the development of this work.

Convolutional Neural Network (CNN) is the most widely model used in audio applications (Aguiar et al., 2018; Wu et al., 2018; Ephrat et al., 2018; Adavanne et al., 2017; Nanni et al., 2020; Salamon and Bello, 2017; Mushtaq and Su, 2020; Mushtaq et al., 2021).

However, when faced with small datasets, CNN's capacity for information retention becomes a flaw; the models memorize the training data and lose performance on new data (Shorten and Khoshgoftaar, 2019). To tackle the issue of overfitting, DA techniques can be used to improve the performance of the model (Salamon and Bello, 2017; Aguiar et al., 2018; Mushtaq and Su, 2020; Mushtaq et al., 2021).

These explored techniques can be implemented in a variety of programming languages using multiple deformations. The variety of options available can make the DA application in audio a challenging task, resulting in excessive use of time and computational power, and a decline in model performance, especially for newcomers.

In order to enhance the process of learning and applying data augmentation strategies, this survey aims: (i) to provide an overview of the most used strategies to current

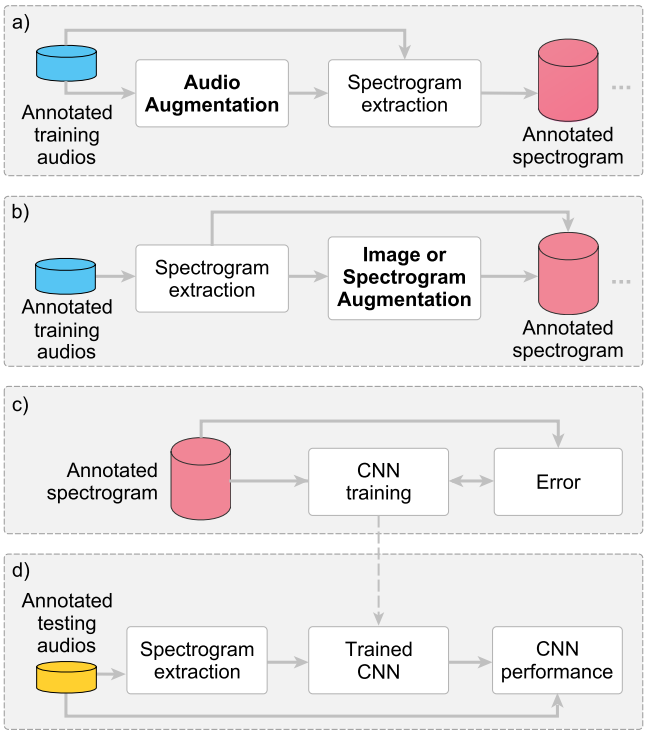

Fig. 1. Offline augmentation approach. a) Audio augmentation. b) Image/spectrogram augmentation. c) CNN training with added augmented data. d) Validating the model with new data without augmented data.

augment audio data research; (ii) to present the main techniques for each data augmentation tool and packages; (iii) to discuss open datasets for implementing and validating CNN models and data augmentation techniques; and (iv) to thoroughly examine the advantages and shortcomings of data augmentation for convolutional audio classification models using ESC research articles as a case of study.

2. OFFLINE DATA AUGMENTATION

Data Augmentation can be applied to audio samples directly or after the spectrogram has been extracted. Regardless of where the augmentation occurs, the samples generated receive the same annotation as the original sample. Fig. 1 depicts the usage of convolutional models with supervised training for audio classification (Nanni et al., 2020; Mushtaq et al., 2021). In this approach, the augmentation can be done to audio files and spectrograms are applied to both the original and augmented samples (Fig. 1a). Another possibility is to convert the original samples into spectrograms and then image or spectrograms augmentation techniques are used (Fig. 1b). Following this process, the spectrograms utilized to train the models are made up of both original and augmented data (Fig. 1c). In the test step (Fig. 1d), the trained model is applied to the new data without deformations.

3. DATA AUGMENTATION FOR AUDIO CLASSIFICATION

This section presents the deformation techniques, deployed in recent papers that used data augmentation for audio classification in ESC, MIR, and ASR. As the nomenclature of the procedures employed differed throughout the examined papers and tools, a standardization of the terms used was sought. In this section we organized the techniques into three major groups: Audio Data Augmentation (ADA), Image Data Augmentation (IDA), and Spectrogram Data Augmentation (SDA).

3.1 Audio Data Augmentation

Audio augmentation approaches introduce deformations directly to the audio raw, with the spectrogram generated from the previous ones (Salamon and Bello, 2017; Aguiar et al., 2018; Nanni et al., 2020; Mushtaq and Su, 2020; Mushtaq et al., 2021). An example of ADA technique is shown in Fig. 2.

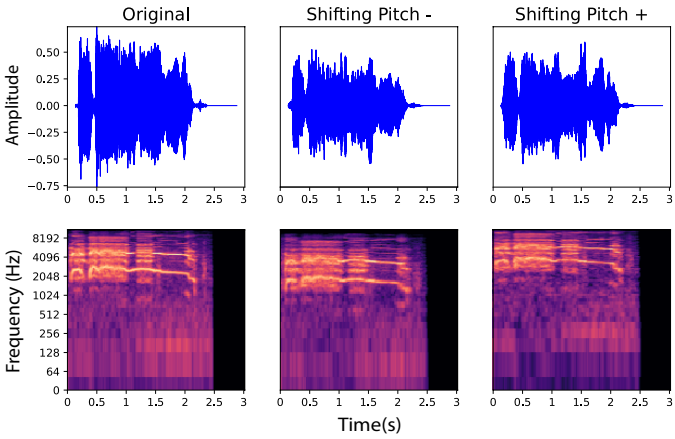

Fig. 2. The decrease (-) and increase (+) in pitch for a rooster crowing and its effects on the log-mel spectrogram.

Shifting Pitch (SP) In this technique, the pitch of each audio signal in the datasets is increased or decreased by a factor and the duration remains the same as shown in Fig. 2.

Time Stretching (TS) It slows down or speeds up an audio sample by a preset ratio without altering the pitch drastically. In Mushtaq and Su (2020), the authors used 1.2 and 0.7 to produce quicker and slower samples.

Volume Adjustment (VA) It is done by varying the loudness of the audio file.

- *Loudness (L).* It increases and decreases the volume of all samples at a random or fixed rate. For example, -10 and +10 dB were used in Aguiar et al. (2018).
- *Dynamic Range Compression (DRC).* It distorts samples by altering the loudness of the original sample using different noises. Salamon and Bello (2017) used music standards, film standards, speech, and radio.

Noise (N) It introduces noises into the original samples.

- *Background Noise (BN).* It consists of mixing the original audios with everyday noise. Salamon and Bello (2017) combined the original audio with noises from street workers, street traffic, street people, and parks.

- *Synthetic Noise (SN).* It creates new samples by combining audio and synthetic noise. White noise, for example, as seen in Mushtaq and Su (2020).

Silence Trimming (ST) It eliminates the silence present at the start and end of each sample.

Time Shifting (TiS) It shifts the audio to the left/right by a random factor.

SpeedUp (SU) The signal is re-sampled at a preset sampling rate and later returned at the original sampling rate, resulting in a speed change.

Wow Resampling (WR) The resampling frequency oscillates around the original sampling rate with a given frequency and amplitude, similar to SP, but with the intensity changing over time. The transformation is provided in (1) where x is the input signal. Nanni et al. (2020) have used the amplitude $a_m = 3$ and the fundamental frequency $f_m = 2$.

$$F(x) = x + a_m \frac{sin(2\pi f_m x)}{2\pi f_m} \tag{1}$$

Clipping (C) The audio sample is normalized so that a specific amount of points are saturated. The out-of-range samples are then clipped.

Harmonic Distortion (HD) The transformation $sin(x)$ is applied many times, resulting in a saturation effect.

Impulse Response (IR) An audio signal is convolved with a unitary response.

Filter (F) It applies several kinds of filtering to the input audio. These are some common filters: band-pass, band-stop, high-pass, high-shelf, low-pass, low-shelf, and peaking filter.

Random Mask (RM) The random frequencies or audio parts are masked.

MP3 Compression (MC) This function compresses the audio to reduce its quality by using an encoder.

Inversion (I) It can be performed on the y-axis, multiplying the audio by -1 or inverting the audio along the x-axis.

Peak Normalization (PN) The highest signal level in the song is set to 0 dBFS. The loudest level must be at [-1,1].

Tangent Distortion (TD) It adds distortion to guitars changing the timbre of the sound when applied hyperbolic tangent function.

3.2 Image Data Augmentation

In this section, it is presented some traditional deformations techniques used in computer vision applications as data augmentation, which has been used for audio classification (Nanni et al., 2020; Mushtaq and Su, 2020).

Flip (F) The rows and columns of pixels can be reversed, as presented in Fig. 2.

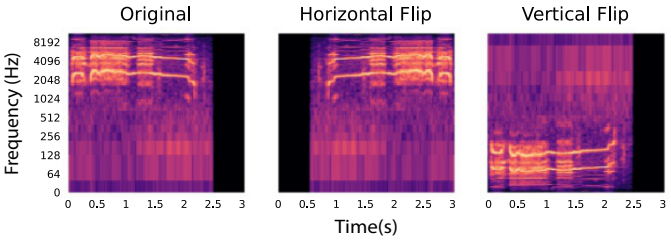

Fig. 3. Horizontal and vertical flip for the rooster crowing present in Fig. 2.

Zoom Range (ZR) It applies zooms randomly and augments new pixels around the image.

Shift (S) The pixels are moved in one direction that can either be vertically or horizontally, maintaining the size of the image.

Rotation Angle (RA) The image is randomly rotated clockwise in the range from 0° to 360°.

Brightness Range (BR) Both randomly darkening and brightening can augment the brightness of the image. This technique simulates the VA technique, which was presented previously.

Shear Range (SR) It causes distortions along an axis that simulate the visualization of an object from different perspectives.

The S and SR techniques must be applied with low intensity since these deformations can change the semantic value of the signal. On the other hand, F, ZR, and RA techniques completely distort the signal and cannot be considered data augmentation techniques in the context of audio signals.

3.3 Spectrogram Data Augmentation

This class of data augmentation is similar to image augmentation because it is applied to spectral images, however, they are selected especially for audio applications (Maguolo et al., 2021; Nanni et al., 2020). Fig. 4 shows an example of a random mask technique.

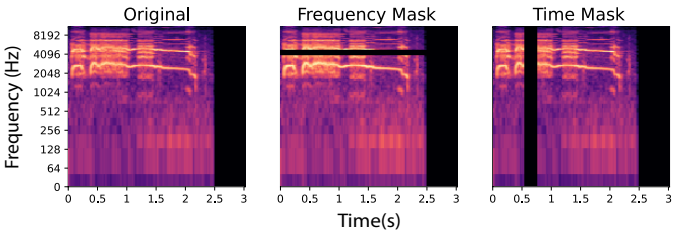

Fig. 4. Frequency and time mask for the rooster crowing present in Fig. 2.

Spectrogram Random Shifts (SRS) This technique randomly applies pitch shift and time shift simultaneously.

Spectrogram Sound Mix (SSM) SSM creates a new image by summing the two random spectrograms of the same class.

Vocal Tract Length Normalization (VTLN) It's an ASR technique that distorts the spectrum in the direction of a medium-level vocal treatment. In ASR, this technique is used to remove the variability that exists between the two vocal tracts length from each speaker (Jaitly and Hinton, 2013).

Equalized Mixture (EM) The weighted average of two randomly picked spectrograms with the same label (Szegedy et al., 2015).

Spectrogram Time Shift (STS) It is a change in time that consists of dividing the spectrum into two parts and later restoring them in a reverse order.

Spectrogram Random Mask (SRM) It consists of removing portions of spectrograms. The frequency and time mask are shown in Fig. 4.

Spectrogram Channel Shuffle (SCS) It shuffles the channels of the spectrogram. By using this function, bias may be mitigated.

4. AUGMENT DATA TOOLS

A set of data augmentation tools is presented in this section to help in their implementation. The functionalities available for each data audio augmentation library are listed in Table 1. In addition, all of the librosa, Keras, and audiomentations techniques were implemented in a GitHub-hosted Jupyter notebook [1].

Table 1. Audio and image augmentation techniques are available for each of the presented tools.

Tool	ADA	IDA/SDA
librosa (McFee et al., 2015b)	TS, SP, ST	
MUDA (McFee et al., 2015a)	TS, SP, BN, DRC	
SoX (Bagwell, 2010)	TS, SP, ST, L, SN, BN, ST, SU, F	
audiogmenter (Maguolo et al., 2021)	SP, L, DRC, SN, SU, WR, C, HD, IR, F	SRS, SSM, VTLN, EM, STS, SRM
Keras (Chollet, 2021)		F, ZR, S, RA, BR, SR
audiomentations (Jordal, 2021)	SP, TS, VA, L, BN, SN, ST, TiS, C, IR, F, RM, MC, I, PN, TD	SRM, SCS

4.1 librosa: Python Audio and Music Analysis

The librosa Python package [2] was designed to evaluate audio and music signals (McFee et al., 2015b). Its 0.8.1 version includes three techniques to perform offline data augmentation. It is widely used for the extraction of musical characteristics and data augmentation in audio (Mushtaq and Su, 2020; Mushtaq et al., 2021).

4.2 MUDA: A Software for Increasing Musical Data

The Musical Data Augmentation package [3] implements annotation-based musical data augmentation (McFee et al., 2015a). This software is based on JAMS (JSON annotated music specification) and it enables the creation of custom deformations as the tracking of data provenance. (Salamon and Bello, 2017). Its 0.4.1 version provides four DA techniques that can be customized to enlarge audio files.

4.3 Audiogmenter: A MATLAB Tool for Augmenting Audio Data

Audiogmenter [4] was the first MATLAB library designed specifically for audio data augmentation (Maguolo et al., 2021). It includes 23 data augmentation methods, such as image and audio augmentation. The library is free, but MATLAB is a paid software.

4.4 SoX: the Swiss Army Knife of Audio Manipulation

Sound eXchange (Bagwell, 2010) is a cross-platform command-line tool that can convert various audio file formats to others. It can apply various effects to audio files. In addition, pysox [5] is a Python wrapper for this tool (Bittner et al., 2016), which is widely used in combination with SoX. SoX was also used for DA by Aguiar et al. (2018) in the implementation of MUDA and audiogmenter packages. SoX 14.4.2 is the newest version.

4.5 Keras: An API for Deep Learning in Python

Keras (Chollet, 2021) is a Python-based machine learning package that can be used in TensorFlow and in other programming languages, such as R. It allows quick experimentation and results. The package includes image augmentation functionalities that can also be applied to spectrograms for audio problems (Mushtaq et al., 2021).

4.6 Audiomentations: A Python library for Audio Data Augmentation

Audiomentations [6] is a Python package (Jordal, 2021) that can be used in Tensorflow/Keras or Pytorch training pipelines. Its 0.20.0 version has 28 ADA and two SDA techniques.

5. DATASETS

In this section, we briefly introduce ESC, MIR, and ASR popular open datasets for audio classification tasks.

[1] https://github.com/lucas-fpaiva/survey-audio-aug
[2] https://github.com/librosa/librosa
[3] https://github.com/bmcfee/muda
[4] https://github.com/LorisNanni/Audiogmenter
[5] https://github.com/rabitt/pysox
[6] https://github.com/iver56/audiomentations

5.1 Urbansound8k

Urbansound8k [7] or US8K (Salamon et al., 2014) is a subset of 4-second audio clips containing 8732 audio files. It is not uniformly distributed across its ten folds. The folds are air conditioner, dog bark, car horn, children playing, gunshot, engine idling, street music, siren, jackhammer, and drilling. It is derived from UrbanSound which was created by manually filtering and labeling the tracks from Freesound (Font et al., 2013).

5.2 ESC-10 and ESC-50

ESC-10 and ESC-50 are part of the ESC dataset [8] (Piczak, 2015) of urban environment 5-second audio recordings. Both datasets are extracted from Freesound (Font et al., 2013), and organized into 5 uniformly sized cross-validation folds. ESC-10 has 400 clip recordings with a total time duration of 33 minutes and a clip rate of 50 clips per fold. These folds include sounds like a dog barking, a crackling fire, baby cries, rain, sneezing, a rooster, sea waves, a helicopter, a chainsaw, and a clock ticking. ESC-50 is more complex than the other due to its 50 folds, which are divided into five major categories. Sounds of animal, non-speech human, urban or outdoor noises, indoor noises, and various natural soundscapes are included in these fold. It contains 2000 sound clips and runs for 168 minutes.

5.3 CatSound

The CatSound Classification dataset [9] (Pandeya and Lee, 2018b; Pandeya et al., 2018) contains over three hours of domestic cat 4-second audio recordings divided into ten folds, every each with over 300 samples. The folds are resting, warning, angry, defending, fighting, happy, hunting mind, mating, mother call, and paining.

5.4 Audio Set

Audio Set (Gemmeke et al., 2017) helps in the development of audio event recognition systems. It is an ensemble of 632 audio folds in a hierarchy over 6 levels of manually annotated ten-second audio files, containing 1,789,621 ten-second video segments from YouTube. Its seven main folds are human sounds, animal sounds, natural sounds, music, sounds of things, source-ambiguous sounds, and channel, environment, and background.

5.5 Speech Commands

The Speech Commands dataset (Warden, 2018) has 3.8 GB of 105,829 one-second long utterances of 35 short words recording by 2,618 members of the Artificial Intelligence Yourself community and stored in WAVE format file [10] . Their 35-word categories include actions, numbers, people names, animal names and objects.

[7] https://urbansounddataset.weebly.com/
[8] https://github.com/karolpiczak/paper-2015-esc-dataset
[9] https://zenodo.org/record/4724180
[10] http://download.tensorflow.org/data/speech_commands_v0.02.tar.gz

5.6 FMA

The Free Music Archive or FMA [11] is a large dataset suitable for evaluating several tasks in MIR (Defferrard et al., 2017). FMA consists of 917 GB and 343 days of audio from 106,574 tracks from 16,341 artists and 14,854 albums, arranged in 16 genres and 145 subgenres with a track, album and artist metadata.

5.7 Nsynth

NSynth [12] holds 306,043 four-second prerecorded notes of 1006 instruments, ranging over a standard MIDI piano with an average of 4.75 unique velocities per pitch (Engel et al., 2017). Nsynth metadata is subdivided into sources according to the instrument sound, family of the note instrument, and sonic qualities of the note. The eleven label families are bass, brass, flute, guitar, keyboard, mallet, organ, reed, string, synth lead, and vocal.

6. DATA AUGMENTATION IN ENVIRONMENTAL SOUND CLASSIFICATION

We chose the ESC area due to its wide application in intelligent systems and the availability of current works that allow us to exemplify the benefits and challenges of using data augmentation in audio data. Table 2 presents the references found in literature, the analyzed datasets, the used augmentation techniques and the tool used. All of the works presented used accuracy as a performance measure, where $accuracy = \frac{VP+VN}{N}$, with VP representing the true positives, VN the true negatives and N the number of samples.

Table 2. Environmental classification sounds works. *BIRD's dataset is unavailable.

Ref.	Dataset	Techniques	Tool
(Salamon and Bello, 2017)	US8K	ADA, NoAug	MUDA
(Nanni et al., 2020)	BIRD*, CAT	ADA, SDA, IDA, NoAug	Audiogmenter
(Mushtaq and Su, 2020)	ESC-10, ESC-50, US8K	ADA, NoAug	librosa
(Mushtaq et al., 2021)	ESC-10, ESC-50, US8K	ADA, IDA	librosa

Salamon and Bello (2017) presented a CNN architecture for ambient sound classification that includes three convolutional layers interleaved with two maxpooling operations and two dense layers. Four audio augmentation techniques were used (TS, SP, DRC, and BN), yielding to five additional training datasets. Seven experiments were carried out for each model to evaluate the effects of each approach individually, collectively, and without data augmentation. The CNN model trained using all techniques together performed the best with an accuracy of 79% while the model without data augmentation reached 73%.

[11] https://github.com/mdeff/fma
[12] https://magenta.tensorflow.org/datasets/nsynth

To classify bird and cat sounds, Nanni et al. (2020) used pre-trained convolutional networks and ensemble techniques with a combination of five CNNs called "fusion". GoogleNet (Szegedy et al., 2015) and VGGnet (Simonyan and Zisserman, 2015) were the pre-trained networks utilized. The research examined the use of audio and image augmentation methods. For both datasets, fusion produced the best results. Table 3 shows the average accuracy of the fusion models for each data augmentation approach.

Table 3. Fusion model performance for each dataset and data augmentation approach in Nanni et al. (2020).

Approach	Techniques	CAT	BIRD
NoAug	-	87.36	95.81
ADA1	TS, SP, L, SN, TiS	89.22	96.16
ADA2	WR, C, SN, SU, HD	89.05	**96.56**
IDA	F, ZR, RA, S	82.71	92.89
SDA	SRS, SSM, VTLM, STS, SRM	**91.73**	94.30

Mushtaq and Su (2020) used SP, TS, and SN to improve CNNs models to classify ambient sounds. Mel, MFCC, and Log-Mel were tested as spectral extraction techniques. Two CNNs with five layers were proposed, one with and one without maxpooling. The best accuracy for all datasets was obtained by combining the model without maxpooling, the Log-Mel spectrogram, and ADA.

The accuracy of the best model for each dataset, the absolute accuracy gains and the cost of using audio augmentation over time are shown in Table 4. The costs were analyzed by (2), where T_{Aug} is the training duration in seconds with data augmentation and T_{NoAug} is the time used to train the model without using data augmentation.

$$Cost = \frac{T_{Aug}}{T_{NoAug}} \tag{2}$$

Table 4. The best model's performance (Accuracy, Gains and Costs) for each dataset in Mushtaq and Su (2020).

	ESC-10	ESC-50	Us8k
No Aug (%)	81.25	57.00	94.14
Aug (%)	94.94	89.28	95.37
Gains (%)	13.69	**32.28**	1.23
Costs	5.31	5.45	5.94

Mushtaq et al. (2021) has tested new ways to solve the problems presented in Mushtaq and Su (2020). Log-Mel was used to extract spectrograms and CNNs with seven (CNN-7) and nine layers were applied. Also, pre-trained models with millions of images using FastAi (`https://docs.fast.ai/vision.learner.html`) to get the weights of ResNet (He et al., 2016), DenseNet (Huang et al., 2017), SqueezeNet (Iandola et al., 2016), AlexNet (Krizhevsky et al., 2012), and VGG (Simonyan and Zisserman, 2015). The research explores the usage of image augmentation (ZR, S, BR, RA, sR and F) and audio augmentation (SP, Ts and TS) for all convolutional models studied.

CNN-7 has the best accuracy among the proposed CNNs, and ResNet-152 was the best of the pre-trained models. The performance of CNN-7 and ResNet-152 for the three datasets and absolute gains in accuracy between audio and image augmentation are shown in Table 5.

Table 5. Performance of the best models in Mushtaq et al. (2021) for each dataset.

	ESC-10	ESC-50	US8k
TAA CNN-7 (%)	77.86	40.46	69.13
NAA CNN-7 (%)	93.5	96.1	95.05
Gains (%)	15.64	**55.64**	25.92
TAA ResNet-152 (%)	95.23	87.49	98.29
NAA ResNet-152 (%)	**99.04**	97.30	**99.05**
Gains (%)	3.81	**9.81**	1.21

6.1 Spectrograms Methods

All convolutional models presented received the sound spectrogram, which describes the variation in the intensity of the spectral components over time as an input. The techniques used were Linear Spectrogram, Mel Spectrogram, Frequency Cepstral Coefficient, Log-Mel Spectrogram and Discrete Gabor Transform. Log-Mel presented a better performance than both Mel Spectrogram and Frequency Cepstral Coefficient with gains of up to 18% of accuracy (Mushtaq and Su, 2020).

6.2 Data Augmentation Techniques Comparison

If available in a group, ADA and SDA approaches outperform IDA techniques overall (Mushtaq et al., 2021; Nanni et al., 2020). The ESC-50 dataset had the greatest difference regarding performance between audio and image augmentation. Both models, ResNet-152 and CNN-7 increased 9.81% and 55%, respectively (Mushtaq et al., 2021). Furthermore, when the individual improvements are measured, the SP approach has made the most relevant contributions, whereas the BN was in charge of the least ones (Salamon and Bello, 2017).

6.3 Data Augmentation versus No Augmentation

When comparing performance with and without data augmentation, DA approaches had a greater performance in most of the found papers. The results of Mushtaq and Su (2020) presented in Table 4, demonstrate the advantages of ADA for the three datasets, especially for ESC-50, which has more classes than ESC-10 and US8K, with an absolute increase of 32%. The authors state that the low gain for the dataset US8k is due to the large number of training samples in the original dataset that has enough data diversity to avoid overfitting.

Salamon and Bello (2017) have trained a CNN model using all techniques together that resulted in an absolute gain of 6% above the model trained without data augmentation. The most favored classes were idle engine and jackhammer, whereas DRC and BN impair the air conditioning sound class predictions. For even better outcomes, the authors recommend using the conditional ADA approach.

As shown in Table 3, the best performances for both datasets were obtained using augmentation techniques, with the highest gain for the CAT dataset using spectrogram augmentation (Nanni et al., 2020). However, the utilization of image augmentation deteriorated the performance of the models. The authors justified that the use of techniques such as reflection, when applied to spectrograms, drastically changes the sound and its semantic

value. This result suggests that the use of traditional IDA techniques is not suitable for application in spectrograms.

6.4 Trade Off between Cost and Performance

The offline data augmentation is a costly approach because for each transformation in an original sample, a new one is created and this reflects in training time. The use of six ADA strategies increased training time by 5 to 6 times (Mushtaq and Su, 2020). For ESC-10 and ESC-50, this cost is compensated with gains of 13.69% and 32.28% respectively. However, for US8k dataset, the gain was less than 2%. Therefore, it is necessary to consider the size of the dataset studied and the accuracy required for the respective task to define whether or not to use any data augmentation technique.

6.5 Transfer Learning and Ensemble Methods

The use of very deep pre-trained CNN models may provide more accuracy than convolutional networks with few layers (Mushtaq et al., 2021). When transfer learning and audio augmentation were used, the ResNet-152 accuracy outperformed all other networks, achieving 99% for ESC-10 and US8k using ADA (Mushtaq et al., 2021). In addition, the use of pre-trained ensemble models obtained higher accuracy than a unique model (Nanni et al., 2020).

7. CONCLUSIONS

Throughout this paper, several offline DA methods for audio and image were presented to improve the performance of CNNs in audio classification. In addition, we provide a list of audio DA tools and open datasets for ESC, MIR, and ASR tasks. In the ESC analyzed works, audio and spectrogram augmentation were more effective than without augmentation approach, with gains in accuracy reaching 50%. In addition, the Shifting Pitch technique was the one that individually provided the greatest increase regarding accuracy.

Shortcomings were found in the choice of deformations to perform data augmentation, since traditional techniques applied in computer vision tasks were used. In this way, these setbacks changed the semantic value of the audio signals when applied to the spectrograms, resulting in a worse of the models' performance. On the other hand, the effect of DA techniques may vary according to the sound type, standing out the SP as the most efficient audio augmentation technique. Furthermore, DA is especially important for small and complex datasets, with improvements of over 30%.

The choice of augmentation technique depends on each specific application. This paper has shown some shortcuts to be followed in the absence of computational power necessary to test each technique within a reasonable time budget. Future works will focus on including more papers, as well as different methods such as online data augmentation and deep learning applied to data augmentation.

REFERENCES

Adavanne, S., Drossos, K., Çakir, E., and Virtanen, T. (2017). Stacked convolutional and recurrent neural networks for bird audio detection. In *2017 25th European Signal Processing Conference*, 1729–1733. doi:10.23919/EUSIPCO.2017.8081505.

Aguiar, R.L., Y.M.G. Costa, M.G., and Silla, C. (2018). Exploring data augmentation to improve music genre classification with convnets. *Proceedings of the International Joint Conference on Neural Networks*, 1–8.

Bagwell, C. (2010). Sox - sound exchange — homepage. URL `http://sox.sourceforge.net/`.

Bello, J., Silva, C., Nov, O., Dubois, R., Arora, A., J.Salamon, Mydlarz, C., and Doraiswamy, H. (2019). Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution. *Communications of the ACM*, 62, 68–77.

Bittner, R., Humphrey, E., and Bello, J. (2016). pysox: Leveraging the audio signal processing power of sox in python. In *Proceedings of the International Society for Music Information Retrieval Conference Late Breaking and Demo Papers*.

Chollet, F. (2021). Keras: the python deep learning api. URL `https://keras.io/`.

Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. (2017). Fma: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference*.

Dominguez-Morales, J.P., Liu, Q., James, R., Gutierrez-Galan, D., Jimenez-Fernandez, A., Davidson, S., and Furber, S. (2018). Deep spiking neural network model for time-variant signals classification: a real-time speech recognition approach. In *2018 International Joint Conference on Neural Networks*, 1–8.

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., and Simonyan, K. (2017). Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, 1068–1077.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W., and Rubinstein, M. (2018). Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37.

Font, F., Roma, G., and Serra, X. (2013). Freesound technical demo. In *Proceedings of the 21st ACM International Conference on Multimedia*, 411–412. doi:10.1145/2502081.2502245.

Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 776–780.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 770–778. doi:10.1109/CVPR.2016.90.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K.Q. (2017). Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2261–2269. doi:10.1109/CVPR.2017.243.

Iandola, F., Han, S., Moskewicz, M., Ashraf, K., Dally, W., and Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model

size. *arXiv preprint arXiv:1602.07360*.

Jaitly, N. and Hinton, G. (2013). Vocal tract length perturbation (vtlp) improves speech recognition. In *Proceedings of the 30th International Conference on Machine Learning*, volume 90, 42–51.

J.Cao, M.Cao, J.Wang, C.Yin, D.Wang, and Vidal, P. (2019). Urban noise recognition with convolutional neural network. *Multimedia Tools and Applications*, 78, 29021–29041.

Jordal, I. (2021). Audiomentations. doi:10.5281/zenodo.5470338.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25.

Maguolo, G., Paci, M., Nanni, L., and Bonan, L. (2021). Audiogmenter: a matlab toolbox for audio data augmentation. *Applied Computing and Informatics*.

McFee, B., Humphrey, E.J., and Bello, J.P. (2015a). A software framework for musical data augmentation. In *Proceedings of the 16th International Society for Music Information Retrieval Conference*, 248–254.

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., and Nieto, O. (2015b). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, 18–24. doi: 10.25080/Majora-7b98e3ed-003.

Mushtaq, Z., Su, S., and Tran, Q. (2021). Spectral images based environmental sound classification using cnn with meaningful data augmentation. *Applied Acoustics*, 172, 107581.

Mushtaq, Z. and Su, S.F. (2020). Environmental sound classification using a regularized deep convolutional neural network with data augmentation. *Applied Acoustics*, 167, 107389.

Nanni, L., Maguolo, G., and Paci, M. (2020). Data augmentation approaches for improving animal audio classification. *Ecological Informatics*, 57, 101084.

Pandeya, Y.R. and Lee, J. (2018a). Domestic cat sound classification using transfer learning. *The International Journal of Fuzzy Logic and Intelligent Systems*, 18, 154–160.

Pandeya, Y., Kim, D., and Lee, J. (2018). Domestic cat sound classification using learned features from deep neural nets. *Applied Sciences*, 8.

Pandeya, Y. and Lee, J. (2018b). Domestic cat sound classification using transfer learning. *The International Journal of Fuzzy Logic and Intelligent Systems*, 18, 154–160.

Piczak, K.J. (2015). Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, 1015–1018. doi: 10.1145/2733373.2806390.

Salamon, J. and Bello, J. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24, 279–283.

Salamon, J., Jacoby, C., and Bello, J.P. (2014). A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, 1041–1044. doi:10.1145/2647868.2655045.

Seo, Y. and Huh, J. (2019). Automatic emotion-based music classification for supporting intelligent iot applications. *Electronics*, 8, 164.

Shorten, C. and Khoshgoftaar, T.M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations*.

Solovyev, R.A., Vakhrushev, M., Radionov, A., Romanova, I.I., Amerikanov, A.A., Aliev, V., and Shvets, A.A. (2020). Deep learning approaches for understanding simple speech commands. In *IEEE 40th International Conference on Electronics and Nanotechnology*, 688–693.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1–9. doi: 10.1109/CVPR.2015.7298594.

Warden, P. (2018). Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*.

Wu, C.W., Dittmar, C., Southall, C., Vogl, R., Widmer, G., Hockman, J., Muller, M., and Lerch, A. (2018). A review of automatic drum transcription. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26, 1457–1483.