Mixed-Type Data Augmentations for Environmental Sound Classification

Tadas Turskis [1], Marius Teleiša [2], Rūta Buckiūnaitė [3] and Dalia Čalnerytė [4]

[1] *Kaunas University of Technology, Studentų g. 50, Kaunas, 51368, Lithuania*
[2] *Kaunas University of Technology, Studentų g. 50, Kaunas, 51368, Lithuania*
[3] *Kaunas University of Technology, Studentų g. 50, Kaunas, 51368, Lithuania*
[4] *Kaunas University of Technology, Studentų g. 50, Kaunas, 51368, Lithuania*

Abstract

The goal of environmental sound classification is to accurately identify and classify sounds in order to provide valuable insights about the environment. The classification task can be solved by training machine learning models, such as convolutional neural networks, on a dataset of labeled sound samples. Due to the small size of available datasets in this field, time-consuming and expensive labeling process, data augmentations have become a popular practice to artificially generate additional data. The purpose of this study is to analyze whether using Mixed-Type data augmentations improves the classification performance compared to results with no augmentations. Mixed-Type data augmentation methods were evaluated on ESC-50 and UrbanSound8K datasets for the pretrained ResNet-18 model with extracted mel-frequency cepstral coefficients as feature inputs. Results for both datasets show that data augmentations can improve model performance with certain mixup probabilities and coefficients but specific methods and parameters used may vary for each dataset and task.

Keywords

Environmental sound classification, mixed augmentation, Mel-Frequency cepstral coefficients.

1. Introduction

Sound classification is the process of identifying and labelling sounds based on their characteristics, such as pitch, duration, and timbre. It is a fundamental task in the field of audio signal processing and has numerous applications, including music information retrieval [1], speech recognition [2], and environmental monitoring [3]. Improving the efficiency and accuracy of sound classification process may enable more low-power devices to perform this task, help in mitigation of noise pollution, and increase the robustness of early-warning systems, such as bee hive health monitoring systems [4].

There are various approaches to sound classification, including traditional machine learning techniques [5] and more recent deep learning approaches [6]. Traditional methods typically involve extracting hand-crafted features from the audio signal and using them as input to a classifier, such as a Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor (KNN), Gaussian Mixture Modeling (GMM) and Hidden Markov Model (HMM) [7–9]. Due to their limited modeling capabilities, that lead to the lack of time and frequency invariance, deep neural network-based models have been proven to perform better in classifying environmental sounds than traditional methods [10]. Deep learning methods involve training a neural network to learn features directly from the raw audio data. A hybrid approach of combining both types of methods has also been explored [11].

Deep learning models for tasks in all audio domains are limited in size and complexity due to the small size of available datasets [12]. With the possible exception of speech recognition, environmental sound classification (ESC) suffers from the lack of universal database [13]. Nonetheless ESC tasks are a popular classification problem to solve therefore there have been several public datasets created. ESC-

Information Society and University Studies (IVUS 2023), May 12, 2023, Kaunas, Lithuania
EMAIL: tadas.turskis@ktu.edu; marius.teleisa@ktu.edu; ruta.buckiunaite@ktu.edu; dalia.calneryte@ktu.lt
ORCID: 0000-0003-4185-0397 (D.Čalnerytė)

50 [14] and UrbanSound8K (US8K) [15] are the leading datasets with the best results achieved for ESC tasks [16]. ESC-50 is a balanced dataset published in 2015 with 2000 recordings for 50 classes of various environmental sounds. Its subset of 10 classes ESC-10 with 400 recordings is often used as a smaller version. US8K is a dataset published in 2014 with 8732 sound recordings of diverse city sounds. Other occasionally used datasets are: CHIME-HOME [17] with 6138 recordings of house indoor sounds; AudioSet [18] with over 2 million recordings of very diverse 632 classes of sounds; FSD50K [19] unbalanced dataset with 51197 recordings of various indoor, outdoor and instrument sounds; SONYC-UST [20] with 18510 recordings of New York city sounds. It can be noted that besides the AudioSet dataset, most of these datasets are of limited size and, as mentioned before, are too small for deep learning models to be trained properly.

The state-of-the-art accuracy for the ESC-50 and UrbanSound8K (US8K) datasets was 97.15% and 96% respectfully[20, 21]. Both [21, 22] equipped deep learning models. Data augmentation techniques such as time scaling, time inversion, random crop or padding, and random noise were used in [21]. On the other hand, [22] opted to use various feature extraction methods like NGCC [23], MFCC [24], GFCC [25], LFCC and BFCC. An approach of using a simple CNN network without any data augmentations or signal pre-processing on US8K dataset has demonstrated 89% of mean accuracy which outperforms most recent solutions [6].

In the domain of environmental sound, it has been noted that time-frequency representations are especially useful as learning features due to the non-stationary and dynamic nature of the sounds [26]. These representations can be grouped into two broad categories: time-domain methods and frequency-domain methods. Time-domain methods involve computing statistics such as the mean, standard deviation, skewness, and variance over different time windows of the signal. Other time-domain methods include calculation of zero-crossing rate, amplitude envelope and the root mean square energy. Frequency-domain methods include techniques such as the calculation of the Power Spectral Density (PSD) and the Mel-Frequency Cepstral Coefficients (MFCCs). For an increase in performance, it is advised to combine several feature extraction methods and types of methods [27].

Combination of suitable audio feature extraction, deep learning methods and data augmentation has been proven to help boost the classification performance [28]. Data augmentation is a widely used technique in various machine learning tasks, including environmental sound classification, to virtually enlarge the datasets [29]. Augmentations can be divided into two general categories: image and audio signal. Image augmentation methods include adding noise, sample pairing [30], cropping, adding filters (e.g. blur, sharpen) [31]. Audio signal augmentations include random cropping, frequency filtering, equalized mixture data augmentation [32], tone shifting.

A recent approach is to use various mixup methods [28] to provide higher prediction accuracy and robustness. Generally, the process of data augmentation is context and dataset dependent, which requires expert knowledge to select augmentation methods. Mixup augmentation technique is data-agnostic, and is performed by generating a random mixing coefficient, which is used to produce a new image and label as a convex combination of two selected images/labels.

In this paper the proposed augmentation technique is based on the mixed-example data augmentation methods, which combine multiple examples from the training set to create a new, augmented example. This technique aims to increase the diversity of the training data, which can lead to better generalization and improved model performance.

The rest of this paper is organized as follows. Section 2 presents the materials and methods. Section 3 describes the details about the chosen augmentation methods. The experimental results are shown in Section 4. The results and future prospects are discussed in Section 5. Finally, the conclusions are presented in the last Section 6.

2. Materials and methods

2.1. Datasets

The study focuses on two publicly available datasets for ESC, that is ESC-50 and UrbanSound8K (US8K). These datasets consist of audio recording of various indoor and outdoor environmental sounds. For example, the ESC-50 dataset consists of 2000 sound clips, such as animal, natural environment,

water, human produced sounds (not speech), household indoor sounds, city sounds. The UrbanSound8K dataset consists of 8732 short sounds of various city noises people usually complain about.

2.2. Classification Model

ResNet-18 [33] is an 18 layers deep convolutional neural network that has shown strong performance on a variety of tasks, including image classification and object detection. It is relatively lightweight and efficient, while still being able to capture complex patterns in the data. Additionally, ResNet18 has been pre-trained on a large dataset (ImageNet), which means that it has already learned to recognize a wide range of features that may be useful for environmental sound classification.

2.3. Data Pre-processing

To apply ResNet-18 model for classification, raw audio recordings were converted to image representation of sound as Mel-frequency cepstral coefficients (MFCCs). The scheme for feature extraction steps is demonstrated in Figure 1.

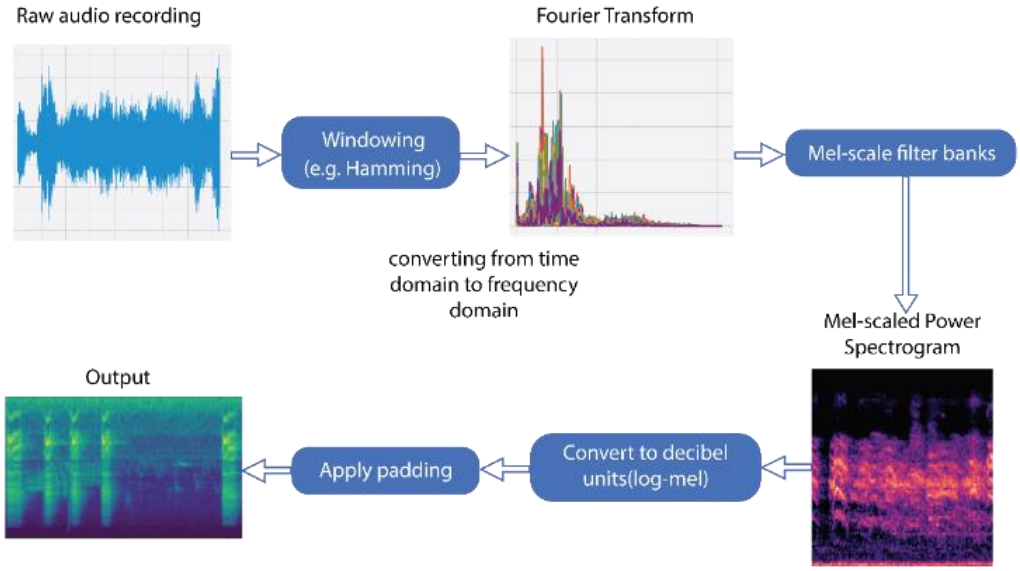

Figure 1: Feature extraction steps

3. Data Augmentation Methods

3.1. Gaussian Noise

Gaussian noise augmentation is based on modifying the original image by adding the random values generated using normal (Gaussian) distribution with a mean of 0 and a standard deviation equal to 3% of absolute minimal value in the matrix of the original spectrogram.

3.2. Mixup

Mixup constructs virtual training examples $\bar{X}$ from two examples X_i, X_j drawn at random from the training data, and mixup coefficient $\lambda \sim U(a;b)$ [34] as follows:

$$\bar{X} = \lambda X_i + (1 - \lambda) X_j, \quad (1)$$

Mixing is done between the data of the same class label. The scheme for mixup method is shown in Figure 2.

Figure 2: Example output mixup data augmentation method. The blue and orange squares represent the original spectrograms, the diagonal pattern represents the mixup part of the resulting image

The rest of the augmentation methods are derived from mixup with different part of the image mixed.

3.3. Vertical/horizontal mixup

This method is used to vertically/horizontally mixup the top fraction of spectrogram image X_i with the bottom fraction of image X_j. A cutpoint is generated by multiplying the width/height of the first image with mixing coefficient λ. Cutpoint is a pair of row r and column c indices of an image X. The resulting merged image is then created by mixing the top cutpoint rows/columns from both images and selecting the bottom cutpoint rows/columns from the first image. The scheme for horizontal/vertical mixup method is shown in Figure 3. The outlined part shows the mixed part of the image and the solid blue part is the original X_i image.

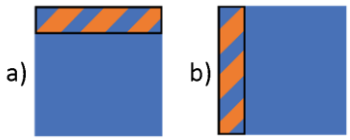

Figure 3: Example of a) horizontal, b) vertical mixup data augmentation method

3.4. Random 2x2

This method divides the images into 4 quadrants using two randomly generated height and width cutpoints and then for each quadrant either X_i section or mixup of X_i and X_j with mixing coefficient λ is used. A constraint p = 0.5 on 2x2 grid has been found to be helpful in preventing the image content from becoming too long, narrow or missing [30]. Example random 2x2 mixup method illustration is shown in Figure 4. The outlined part shows the part of the image that is mixed and the solid blue part is the original X_i image.

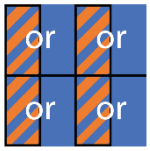

Figure 4: Example of random 2x2 mixup data augmentation method

3.5. Random column/row interval

This method picks a random interval of columns/rows and replaces that part of X_i spectrogram image with the mixed columns/rows from X_j. The start and end indices are generated randomly for the column/row interval to be mixed. Random interval method is somewhat similar to the previously mentioned vertical/horizontal mixup method with the difference being that the random interval does not start from the first column/row. The scheme for random column/row interval mixup method is shown in Figure 5. The outlined part shows the part of the image that is mixed and the solid blue part is the original X_i image.

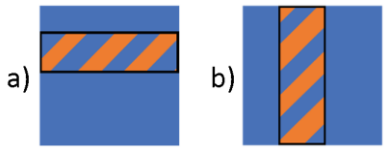

Figure 5: Example of random a) row, b) column interval mixup data augmentation method

3.6. Random columns/rows

This method involves randomly selecting rows/columns from X_j to be mixed up. Probability of choosing a row/column from X_j is determined by p, and for experimental testing we used p value of 0.5. The scheme for random column/rows mixup method is shown in Figure 6. The outlined part shows the part of the image that is mixed and the solid blue part is the original X_i image.

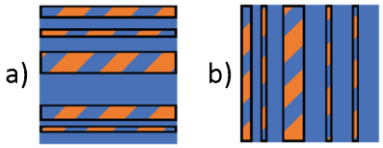

Figure 6: Example output random a) rows, b) columns mixup data augmentation method

4. Experimental Results

4.1. Experiment setup

The fixed duration of an audio sample for ESC-50 is 5 seconds, for UrbanSound8K – 4 seconds. For each audio file in the ESC-50 and US8K datasets a log-mel spectrogram is generated. Features are extracted from all recordings with Hamming window size of 512, hop length of 512, 128 Mel bands and sampling rate of 44.1 kHz. The resulting spectrograms are padded or their length is fixed. Bootstrap validation with 5 runs is performed for dataset using Stratified Shuffle Split with 0.25 test set size and static random seed of 42. All data is then standardized according to train set. Augmentation is performed online (when data sample is being provided to model for training).

Training was performed on ResNet-18 model with weights pre-trained on ImageNet and batch size of 16. Training was performed for 25 epochs with cases of augmentation probability of 0.3, 0.5, 1.0 and mixup coefficient generated uniformly in the intervals of (0.2; 0.3), (0.45; 0.55), (0.7; 0.8).

In total there were 76 distinct method configurations to test, that is, no augmentation, Gaussian noise with 3 augmentation probabilities, 3 augmentation probabilities with 3 mixup coefficients and 9 augmentations utilizing mixup.

Experimental testing was performed on a system provided by the Kaunas University of Technology, and this process took almost 64h (~5mins one augment method for 25 epochs * 5 runs * 76 augmentation methods * 2 datasets). Specifications of the system are 2x AMD EPYC 7452 32-Core Processor with 2x 256GB RAM and NVIDIA A100-PCIE-40GB GPU.

4.2. Results for ESC-50 dataset

There are many hyperparameters to consider, so the results were gradually filtered according to their performance. The first hyperparameter to consider is the augmentation probability. Mean results grouped by probability of augmentation for ESC-50 dataset can be seen in Table 1.

Table 1
Mean ESC-50 results grouped by augmentation probability and no augmentation

Probability	Mean accuracy	Mean loss	Min accuracy	Min loss	Max accuracy	Max loss	Q1 acc	Q1 loss	Q3 acc	Q3 loss
1	0.834	0.627	0.815	0.540	0.855	0.700	0.824	0.603	0.841	0.661
0.5	**0.846**	**0.580**	**0.827**	**0.483**	0.868	**0.648**	0.836	**0.560**	**0.854**	**0.616**
0.3	**0.845**	0.589	0.824	0.499	0.864	0.651	**0.837**	0.568	**0.854**	0.628
No augmentation	0.839	0.603	0.822	0.495	**0.870**	0.696	0.826	0.570	0.842	0.636

Looking at mean accuracy, results for 0.3 and 0.5 probabilities are slightly higher (0.7%) than no augmentation, which is expected as data augmentation should improve model performance when dataset is relatively small. Maximum accuracy of 0.87 was achieved during no augmentation, however almost all loss metrics indicate better performance for probabilities of 0.3 and 0.5. On the other hand, probability of 1.0 managed to degrade model performance compared to no augmentation in every metric that was recorded, so for further analysis results for probability of 1.0 was not considered.

Grouping results by mixup coefficient presents almost identical results for all values of mixup, as shown in Table 2, however upon closer inspection small trends can be seen. Mixup coefficient of 0.7-0.8 means that during mixup 70-80% of original image is used, and 20-30% of random image.

Table 2
Mean ESC-50 results grouped by mixup coefficient and no augmentation. Results exclude augmentation probability of 1.0.

Mixup coeff.	Mean accuracy	Mean loss	Min accuracy	Min loss	Max accuracy	Max loss	Q1 acc	Q1 loss	Q3 acc	Q3 loss
0.2-0.3	0.845	0.582	0.824	0.493	0.867	0.650	0.837	0.561	0.854	0.618
0.45-0.55	0.845	0.584	0.825	0.481	0.868	0.651	0.836	0.569	0.854	0.623
0.7-0.8	0.846	0.584	0.828	0.499	0.865	0.646	0.837	0.558	0.856	0.622
No augmentation	0.839	0.603	0.822	0.495	0.870	0.696	0.826	0.570	0.842	0.636

Most metrics show best results for mixup coefficient of 0.7-0.8, except for minimum loss and maximum accuracy, however the differences might be due to error with this sample size so hard conclusions cannot be drawn about the coefficient effect on model performance from this data alone.

Results show that mixup coefficient of 0.7-0.8 produces slightly better results ESC-50 dataset, which might suggest that the model prefers to have the main image “dominant” (image with higher mixup coefficient), and not the other way around.

Finally, mean results for each augmentation method can be seen in Table 3. Mixup augmentation performs best on almost all metrics except maximum loss and Q3, which suggests that the method produces less consistent. All proposed mixup methods performed better than the standard Gaussian Noise augmentation method. Interestingly, all column methods (random column interval, random column, horizontal mixup) performed better than their row counterparts.

Table 3
Mean ESC-50 results grouped by augmentation methods. Results exclude augmentation probability of 1.0

Method	Mean accuracy	Mean loss	Min accuracy	Min loss	Max accuracy	Max loss	Q1 acc	Q1 loss	Q3 acc	Q3 loss
No augmentation	0.839	0.603	0.822	0.495	0.870	0.696	0.826	0.570	0.842	0.636
Mixup	0.852	0.565	0.829	0.460	0.875	0.640	0.845	0.544	0.858	0.603
Gaussian noise	0.841	0.602	0.820	0.502	0.863	0.657	0.832	0.595	0.847	0.640
Random 2x2 mixup	0.848	0.570	0.826	0.491	0.867	0.627	0.840	0.551	0.858	0.610
Vert. mixup	0.840	0.599	0.820	0.528	0.854	0.669	0.835	0.578	0.851	0.628
Horiz. mixup	0.849	0.573	0.829	0.466	0.872	0.638	0.838	0.550	0.860	0.621
Random row interval mixup	0.842	0.596	0.823	0.498	0.863	0.652	0.833	0.581	0.849	0.628
Random column interval mixup	0.846	0.583	0.826	0.493	0.867	0.645	0.837	0.563	0.853	0.623
Random rows mixup	0.842	0.599	0.823	0.505	0.864	0.665	0.831	0.576	0.853	0.635
Random cols mixup	0.847	0.582	0.828	0.485	0.871	0.654	0.836	0.559	0.854	0.622

4.3. Results for UrbanSound8K dataset

Mean results grouped by probability of augmentation for ESC-50 dataset can be seen in Table 4.

Table 4
Mean US8k results grouped by augmentation probability and no augmentation

Probability	Mean accuracy	Mean loss	Min accuracy	Min loss	Max accuracy	Max loss	Q1 acc	Q1 loss	Q3 acc	Q3 loss
1	0.958	0.143	0.954	0.130	0.962	0.155	0.957	0.138	0.960	0.149
0.5	0.967	0.118	0.963	0.105	0.970	0.132	0.965	0.112	0.969	0.122
0.3	0.967	0.118	0.963	0.105	0.970	0.133	0.965	0.112	0.968	0.124
No augmentation	0.968	0.125	0.964	0.109	0.972	0.146	0.967	0.110	0.969	0.143

Applying augmentation with a probability of 1.0 results in the worst loss and accuracy for the model. The loss value is decreased by 5.6% when using either 0.3 or 0.5 augmentation probability compared to no augmentation. The highest accuracy is identified for the case with no augmentation applied, although the difference is insignificant with only a 0.001 increase compared to the best performing result with augmentation used.

The results once again indicate that using no augmentation leads to the highest accuracy, although the difference of only 0.001% is not significant. In contrast, Table 5 shows that using mixup with coefficients of 0.2-0.3 produces the best results in all loss metrics, resulting in a decrease of 7.2% in

mean loss compared to no augmentation. This suggests that the model has almost identical accuracy, but with lower loss, leading to a more robust model.

Table 5
Mean US8k results grouped by mixup coefficient and no augmentation. Results exclude augmentation probability of 1.0.

Mixup	Mean accuracy	Mean loss	Min accuracy	Min loss	Max accuracy	Max loss	Q1 acc	Q1 loss	Q3 acc	Q3 loss
0.2-0.3	0.967	0.116	0.963	0.102	0.970	0.131	0.965	0.109	0.968	0.121
0.45-0.55	0.967	0.117	0.963	0.105	0.970	0.132	0.965	0.111	0.969	0.123
0.7-0.8	0.967	0.120	0.963	0.108	0.970	0.134	0.965	0.115	0.968	0.125
No augmentation	0.968	0.125	0.964	0.109	0.972	0.146	0.967	0.110	0.969	0.143

Looking at Table 6 random column augmentation performs best overall. Although, as seen in the results from Table 4 and Table 5, the accuracy while using augmentation compared to no augmentation, it is still only 0.001% higher, which is not significant. However, what is clearly seen in the previous tables and this table in the mean loss column, is that we always get lower loss compared to no augmentation. In this column, all used augmentation methods performed better or at least the same as no augmentation when evaluating mean loss, with the random cols mixup method being the best, giving a decrease of 14.4% in mean loss compared to no augmentation. This demonstrates that the models trained with augmentation are more confident with their predictions. In addition, from Table 6 we see that the difference between the min and max loss and accuracy of using random columns and using no augmentation is lower, respectively 0.029 compared to 0.037 and 0.007 compared to 0.008, which means that the trained models are more consistent and stable when applying augmentation.

Table 6
Mean US8k results grouped by augmentation methods. Results exclude augmentation probability of 1.0

Method	Mean accuracy	Mean loss	Min accuracy	Min loss	Max accuracy	Max loss	Q1 acc	Q1 loss	Q3 acc	Q3 loss
No augmentation	0.968	0.125	0.964	0.109	0.972	0.146	0.967	0.110	0.969	0.143
Mixup	0.966	0.117	0.963	0.104	0.969	0.131	0.964	0.112	0.967	0.122
Gaussian noise	0.968	0.121	0.961	0.106	0.973	0.143	0.967	0.113	0.972	0.123
Random 2x2 mixup	0.968	0.112	0.965	0.101	0.972	0.126	0.967	0.105	0.969	0.120
Vert. Mixup	0.965	0.122	0.961	0.111	0.969	0.135	0.963	0.117	0.968	0.127
Horiz. Mixup	0.967	0.119	0.962	0.103	0.972	0.138	0.965	0.111	0.969	0.124
Random row interval mixup	0.965	0.124	0.962	0.111	0.968	0.137	0.964	0.121	0.966	0.126
Random column interval mixup	0.969	0.115	0.965	0.102	0.972	0.129	0.967	0.108	0.970	0.121
Random rows mixup	0.964	0.125	0.960	0.115	0.967	0.137	0.962	0.118	0.965	0.130
Random cols mixup	0.969	0.107	0.965	0.094	0.972	0.123	0.967	0.100	0.971	0.114

5. Discussion

Column mixup methods performed better than row, which seems to suggest that having complete frequency data is more important than full temporal data in the ESC problem.

For ESC testing mixup coefficient of 0.2-0.3 and 0.7-0.8 were chosen, and one might argue that such values produce the same set of images. This may be true for an infinite set, however, image set for a training epoch is finite, and 0.7-0.8 mixup coefficient guarantees that the augmented set will always have one of each sample as the "dominant" image (higher mixup coefficient), whereas with 0.2-0.3 the reverse is true.

When comparing ESC-50 with US8K datasets, we see more improvements in former, and a probable reason could be that ESC-50 has only 40 examples per class, while US8K has up to 1000 examples per class. For other research it could be useful and interesting to apply these augmentations for dataset with very low number of samples, there we could see bigger improvements. For future improvements in accuracy, a combination of various feature extraction methods could be used with our proposed methods.

6. Conclussion

Results from the ESC-50 and UrbanSound8K datasets show that data augmentation can improve model performance, particularly when using probabilities of 0.3 or 0.5. Mixup augmentation with a coefficient of 0.7-0.8 was found to produce the best results for the ESC-50 dataset. The random column augmentation demonstrated the highest accuracy for the UrbanSound8K dataset. It is important to note that the results for the UrbanSound8K dataset showed lower and less significant improvements compared to the ESC-50 dataset. It is also worth noting that applying augmentation with probability of 100% resulted in worst results in loss and accuracy in both datasets. Overall, these results indicate that data augmentation can be a useful tool for improving model performance, but the specific methods and parameters used may vary depending on the dataset and task at hand.

7. References

[1] YANG, Gao. Research on Music Content Recognition and Recommendation Technology Based on Deep Learning. *Security and Communication Networks*. 14 March 2022. Vol. 2022. DOI 10.1155/2022/7696840.

[2] DOMINGUEZ-MORALES, Juan P., LIU, Qian, JAMES, Robert, GUTIERREZ-GALAN, Daniel, JIMENEZ-FERNANDEZ, Angel, DAVIDSON, Simon and FURBER, Steve. Deep Spiking Neural Network model for time-variant signals classification: a real-time speech recognition approach. In : *2018 International Joint Conference on Neural Networks (IJCNN)*. July 2018. p. 1–8. DOI 10.1109/IJCNN.2018.8489381.

[3] GHANNAM, Ryan B. and TECHTMANN, Stephen M. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Computational and Structural Biotechnology Journal*. Online. 2021. Vol. 19, p. 1092–1107. DOI 10.1016/j.csbj.2021.01.028.

[4] SOARES, Bianca Sousa, LUZ, Jederson Sousa, DE MACÊDO, Valderlândia Francisca, SILVA, Romuere Rodrigues Veloso e, DE ARAÚJO, Flávio Henrique Duarte and MAGALHÃES, Deborah Maria Vieira. MFCC-based descriptor for bee queen presence detection. *Expert Systems with Applications*. Online. 1 September 2022. Vol. 201, p. 117104. DOI 10.1016/j.eswa.2022.117104.

[5] EKPEZU, Akon O., KATSRIKU, Ferdinand, YAOKUMAH, Winfred and WIAFE, Isaac. The Use of Machine Learning Algorithms in the Classification of Sound: A Systematic Review. *https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJSSMET.298667*. Online. 1 January 1AD.

[6] ABDOLI, Sajjad, CARDINAL, Patrick and LAMEIRAS KOERICH, Alessandro. End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Systems with Applications*. Online. December 2019. Vol. 136, p. 252–263. DOI 10.1016/j.eswa.2019.06.040.

[7] ANWAR, Muhammad Zohaib, KALEEM, Zeeshan and JAMALIPOUR, Abbas. Machine Learning Inspired Sound-Based Amateur Drone Detection for Public Safety Applications. *IEEE Transactions on Vehicular Technology*. Online. March 2019. Vol. 68, no. 3, p. 2526–2534. DOI 10.1109/TVT.2019.2893615.

[8] GIANNOULIS, Dimitrios, BENETOS, Emmanouil, STOWELL, Dan, ROSSIGNOL, Mathias, LAGRANGE, Mathieu and PLUMBLEY, Mark D. Detection and classification of acoustic scenes and events: An IEEE AASP challenge. In : *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Online. New Paltz, NY, USA : IEEE, October 2013. p. 1–4. ISBN 978-1-4799-0972-8. DOI 10.1109/WASPAA.2013.6701819.

[9] ZHANG, Haomin, MCLOUGHLIN, Ian and SONG, Yan. Robust sound event recognition using convolutional neural networks. In : *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Online. South Brisbane, Queensland, Australia : IEEE, April 2015. p. 559–563. ISBN 978-1-4673-6997-8. DOI 10.1109/ICASSP.2015.7178031.

[10] AL-HATTAB, Yousef Abd, ZAKI, Hasan Firdaus and SHAFIE, Amir Akramin. Rethinking environmental sound classification using convolutional neural networks: optimized parameter tuning of single feature extraction. *Neural Computing and Applications*. Online. November 2021. Vol. 33, no. 21, p. 14495–14506. DOI 10.1007/s00521-021-06091-7.

[11] ULLO, Silvia Liberata, KHARE, Smith K., BAJAJ, Varun and SINHA, G. R. Hybrid Computerized Method for Environmental Sound Classification. *IEEE Access*. 2020. Vol. 8, p. 124055–124065. DOI 10.1109/ACCESS.2020.3006082.

[12] PURWINS, Hendrik, LI, Bo, VIRTANEN, Tuomas, SCHLUTER, Jan, CHANG, Shuo-Yiin and SAINATH, Tara. Deep Learning for Audio Signal Processing. *IEEE Journal of Selected Topics in Signal Processing*. Online. May 2019. Vol. 13, no. 2, p. 206–219. DOI 10.1109/JSTSP.2019.2908700.

[13] DAVIS, Nithya and SURESH, K. Environmental Sound Classification Using Deep Convolutional Neural Networks and Data Augmentation. In : *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. Online. Thiruvananthapuram, India : IEEE, December 2018. p. 41–45. ISBN 978-1-5386-7336-2. DOI 10.1109/RAICS.2018.8635051.

[14] PICZAK, Karol J. ESC: Dataset for Environmental Sound Classification. In : *Proceedings of the 23rd ACM international conference on Multimedia*. Online. Brisbane Australia : ACM, 13 October 2015. p. 1015–1018. ISBN 978-1-4503-3459-4. DOI 10.1145/2733373.2806390.

[15] SALAMON, Justin, JACOBY, Christopher and BELLO, Juan Pablo. A Dataset and Taxonomy for Urban Sound Research. In : *Proceedings of the 22nd ACM international conference on Multimedia*. Online. Orlando Florida USA : ACM, 3 November 2014. p. 1041–1044. ISBN 978-1-4503-3063-3. DOI 10.1145/2647868.2655045.

[16] BANSAL, Anam and GARG, Naresh Kumar. Environmental Sound Classification: A descriptive review of the literature. *Intelligent Systems with Applications*. Online. 1 November 2022. Vol. 16, p. 200115. DOI 10.1016/j.iswa.2022.200115.

[17] FOSTER, Peter, SIGTIA, Siddharth, KRSTULOVIC, Sacha, BARKER, Jon and PLUMBLEY, Mark D. Chime-home: A dataset for sound source recognition in a domestic environment. In : *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Online. New Paltz, NY, USA : IEEE, October 2015. p. 1–5. ISBN 978-1-4799-7450-4. DOI 10.1109/WASPAA.2015.7336899.

[18] GEMMEKE, Jort F., ELLIS, Daniel P. W., FREEDMAN, Dylan, JANSEN, Aren, LAWRENCE, Wade, MOORE, R. Channing, PLAKAL, Manoj and RITTER, Marvin. Audio Set: An ontology and human-labeled dataset for audio events. In : *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Online. New Orleans, LA : IEEE, March 2017. p. 776–780. ISBN 978-1-5090-4117-6. DOI 10.1109/ICASSP.2017.7952261.

[19] FONSECA, Eduardo, FAVORY, Xavier, PONS, Jordi, FONT, Frederic and SERRA, Xavier. *FSD50K: An Open Dataset of Human-Labeled Sound Events*. Online. 23 April 2022. arXiv. arXiv:2010.00475. arXiv:2010.00475 [cs, eess, stat]

[20] CARTWRIGHT, Mark, CRAMER, Jason, MENDEZ, Ana, WANG, Yu, WU, Ho-Hsiang, LOSTANLEN, Vincent, FUENTES, Magdalena, DOVE, Graham, MYDLARZ, Charlie, SALAMON, Justin, NOV, Oded and BELLO, Juan. *SONYC-UST-V2: An Urban Sound Tagging Dataset with Spatiotemporal Context*. . 2020.

[21] GUZHOV, Andrey, RAUE, Federico, HEES, Jörn and DENGEL, Andreas. *AudioCLIP: Extending CLIP to Image, Text and Audio*. . 2021.

[22] CHEN, Yunhao, ZHU, Yunjie, YAN, Zihui and CHEN, Lifang. *Effective Audio Classification Network Based on Paired Inverse Pyramid Structure and Dense MLP Block*. Online. 5 November 2022. arXiv. arXiv:2211.02940. arXiv:2211.02940 [cs, eess]
[23] ZHANG, Si, TONG, Hanghang, XU, Jiejun and MACIEJEWSKI, Ross. Graph convolutional networks: a comprehensive review. *Computational Social Networks*. Online. 10 November 2019. Vol. 6, no. 1, p. 11. DOI 10.1186/s40649-019-0069-y.
[24] ZHENG, Fang, ZHANG, Guoliang and SONG, Zhanjiang. Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*. Online. November 2001. Vol. 16, no. 6, p. 582–589. DOI 10.1007/BF02943243.
[25] XU, He, LIN, Lin, SUN, Xiaoying and JIN, Huanmei. A New Algorithm for Auditory Feature Extraction. In : *2012 International Conference on Communication Systems and Network Technologies*. Online. Rajkot, Gujarat, India : IEEE, May 2012. p. 229–232. ISBN 978-1-4673-1538-8. DOI 10.1109/CSNT.2012.57.
[26] MD SHAHRIN, Muhammad Huzaifah. Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks. . 21 June 2017.
[27] SU, Yu, ZHANG, Ke, WANG, Jingyu, ZHOU, Daming and MADANI, Kurosh. Performance analysis of multiple aggregated acoustic features for environment sound classification. *Applied Acoustics*. Online. January 2020. Vol. 158, p. 107050. DOI 10.1016/j.apacoust.2019.107050.
[28] XU, Kele, FENG, Dawei, MI, Haibo, ZHU, Boqing, WANG, Dezhi, ZHANG, Lilun, CAI, Hengxing and LIU, Shuwen. Mixup-Based Acoustic Scene Classification Using Multi-channel Convolutional Neural Network. In : HONG, Richang, CHENG, Wen-Huang, YAMASAKI, Toshihiko, WANG, Meng and NGO, Chong-Wah (eds.), *Advances in Multimedia Information Processing – PCM 2018*. Online. Cham : Springer International Publishing, 2018. p. 14–23. Lecture Notes in Computer Science. ISBN 978-3-030-00763-8.
[29] INOUE, Hiroshi. *Data Augmentation by Pairing Samples for Images Classification*. Online. 11 April 2018. arXiv. arXiv:1801.02929. arXiv:1801.02929 [cs, stat]
[30] SUMMERS, Cecilia and DINNEEN, Michael J. Improved Mixed-Example Data Augmentation. In : *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Online. Waikoloa Village, HI, USA : IEEE, January 2019. p. 1262–1270. ISBN 978-1-72811-975-5. DOI 10.1109/WACV.2019.00139.
[31] KANG, Guoliang, DONG, Xuanyi, ZHENG, Liang and YANG, Yi. PatchShuffle Regularization. . 22 July 2017.
[32] NANNI, Loris, MAGUOLO, Gianluca and PACI, Michelangelo. Data augmentation approaches for improving animal audio classification. *Ecological Informatics*. Online. May 2020. Vol. 57, p. 101084. DOI 10.1016/j.ecoinf.2020.101084.
[33] HE, Kaiming, ZHANG, Xiangyu, REN, Shaoqing and SUN, Jian. *Deep Residual Learning for Image Recognition*. Online. 10 December 2015. arXiv. arXiv:1512.03385. arXiv:1512.03385 [cs]
[34] ZHANG, Hongyi, CISSE, Moustapha, DAUPHIN, Yann N. and LOPEZ-PAZ, David. *mixup: Beyond Empirical Risk Minimization*. Online. 27 April 2018. arXiv. arXiv:1710.09412. arXiv:1710.09412 [cs, stat]