ICIC Express Letters
Volume **20**, Number **7**, July **2026**

ICIC International ©2026 ISSN 1881-803X
pp. **793–802**

COMPARATIVE ANALYSIS OF AUGMENTATION AND FEATURE EXTRACTION TECHNIQUES FOR ENVIRONMENTAL SOUND CLASSIFICATION

Theresia Herlina Rochadiani[1,*], Yulyani Arifin[1], Derwin Suhartono[2] and Widodo Budiharto[2]

[1]Computer Science Department, BINUS Graduate Program – Doctor of Computer Science
[2]Computer Science Department, School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
yulyaniarifin@binus.ac.id; { dsuhartono; wbudiharto }@binus.edu
*Corresponding author: theresia.rochadiani@binus.ac.id

Received August 2025; accepted October 2025

Abstract. *A limited labeled dataset presents a significant challenge in achieving accurate environmental sound classification models. This study addresses this by investigating the effectiveness of data augmentation techniques and feature extraction methods. Experiments were conducted using four public datasets: BDLib2, ESC-50, UrbanSound8K (US8K), and FSD50K, and evaluated five augmentation techniques, namely Additive Gaussian Noise (AGN), Pitch Shift (PS), Shifting (S), Time Masking (TM), and Time Stretching (TS). Three feature extraction methods, Mel spectrogram, log-mel spectrogram, and Mel-frequency cepstral coefficients (MFCC), were applied to convolutional neural network (CNN) models. The experimental results demonstrate that the AGN augmentation technique combined with Mel spectrogram feature extraction is remarkably effective. This combination achieved high accuracies: 99.72% for BDLib2, 99.60% for ESC-50, 96.04% for US8K, and 94.73% for FSD50K. Overall, the study concludes that selecting the appropriate augmentation technique and feature extraction method, aligned with the model architecture, is crucial for achieving optimal model performance.*

Keywords: Audio signal processing, Data augmentation, Convolutional neural network, Environmental sound classification, Feature extraction

1. **Introduction.** The audio signal processing and acoustic analysis sector has rapidly advanced due to machine learning and deep learning innovations, excelling in sound recognition, classification, and human-machine interactions. Model success hinges on high-quality, diverse training data. A major challenge is sourcing sufficiently varied and representative data, as many public audio datasets are limited and unevenly distributed, risking model overfitting and poor generalization to unseen data. Audio augmentation addresses these challenges by editing sound signals without altering labels, generating new data variations that represent real conditions better. This technique enhances feature richness and reduces overfitting, serving as a regularization method in deep learning. In self-supervised and contrastive learning, augmented data drives the formation of meaningful positive and negative pairs for better representations.

Many studies have explored and investigated the leverage of data augmentation, especially in environmental sound dataset. [1] enhanced urban environmental sound classification on edge devices via waveform and spectrogram augmentations, addressing noise, overlapping sounds, and class imbalance, achieving a 4.31 percentage point increase in F1-score. Another study found that advanced spectrogram augmentation methods, particularly Mixed Frequency Masking, significantly improve classification accuracy on the

DOI: 10.24507/icicel.20.07.793

Freesound Dataset compared to traditional raw audio methods [2]. [3] evaluated mixed-method data augmentation for environmental sound classification. Findings indicate random column shuffling yields optimal performance information. In addition, there is a study applying diffusion probabilistic models (DPMs) with a top-k discriminator for augmentation, enhancing model precision in US8K dataset [4].

Audio feature extraction transforms original audio signals into compact representations for environmental sound classification, enhancing machine learning performance. Recent research focuses on deep learning methods that automate feature development, surpassing traditional hand-crafted approaches. [5] proposed using a CNN with three convolutional layers, max-pooling, and k-nearest neighbors (KNN) classifiers, improving classification accuracy and training speed on US8K and DCASE-2017 ASC datasets. Another recent study demonstrated the transition from handcrafted features to automatically learned representations. It presents a novel environmental sound classification system using multi-channel deep convolutional neural networks (MC-DCNNs) for feature extraction, achieving high classification accuracy on ESC-10, ESC-50, and US8K without traditional features [6]. Furthermore, there is growing interest in feature fusion to exploit the strengths of each representation. [7] introduced eight hybrid feature combinations that integrate frequency-domain features and mid-level features, with MFCC-Log-Mel-CST reaching cutting-edge accuracy 93.4% on US8K and 85.6% on ESC-50, surpassing current models. Another study combined L2M and L3M attributes with traditional features for environmental sound classification, achieving 99.22% accuracy using DenseNet-161 and data augmentation techniques [8].

Although there have been many studies investigating the effectiveness of different audio augmentation methods, there have been few studies that systematically compare the strengths of five different augmentation techniques across different datasets and feature extraction methods. The study addressed these shortcomings through a structured experimental approach to evaluate the effectiveness of each augmentation technique in improving model accuracy. In this study, we investigated five frequently used audio augmentation methods: AGN [2,9,10], PS [2,9,10], S [9], TM [2,10], and TS [2,9]. In the feature extraction stage, we use three well-known and efficient audio representations: MFCC [11,12], Mel spectrogram [12], and log-mel spectrogram [11,13]. These three types of features capture frequency and temporal information from audio signals that are crucial in machine learning modeling. In assessing the effectiveness of each augmentation technique, we used four diverse public datasets: BDLib2, ESC-50, US8K, and FSD50K. Contributions from this study include 1) a comprehensive evaluation of five audio augmentation techniques and their impact on the performance of classification models; 2) investigate the effectiveness of augmentation techniques influenced by the feature extraction techniques used; 3) provide practical recommendations on the best audio augmentation techniques that can be used in the development of limited data-based sound classification systems.

This paper is structured into several sections following this section. Section 2 outlines the research methods applied in this study. Section 3 contains the results and their analyses. Section 4 offers the study's conclusion and outlines future efforts in environmental sound classification.

2. **Methodology.**

This experimental study, in which the scenario is shown by Figure 1, utilizes a WAV audio dataset with five augmentation techniques. After pre-processing, feature extraction involves MFCC, Mel spectrogram, and log-mel spectrogram. A CNN classifier assesses performance through accuracy and F1-score metrics, with detailed descriptions provided in subsequent subsections.

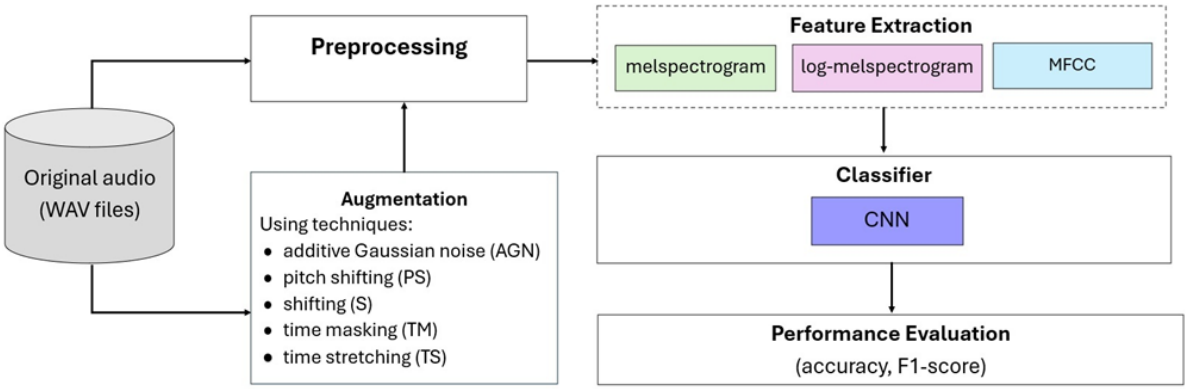


FIGURE 1. The experimental scenario to investigate the effects of augmentation techniques across different feature extraction methods and datasets

2.1. **Data collection.** Gather a large and varied collection of audio samples. The audio samples are picked up from 4 audio public datasets: BDLib2, ESC-50, US8K, and FSD50K.

- **BDLib2** [14,15] consists of 10 categories of environmental sounds, each with 18 balanced, uncompressed WAV files at 44.1 kHz. Categories include planes, alarms, applause, birds, dogs, motors, rain, rivers, waves, and lightning, representing diverse real-life conditions.
- **ESC-50** [16] includes 2000 labeled 5-second environmental sound recordings across 50 semantic classes, grouped into 5 categories.
- **US8K** [17] includes 8732 labeled urban sound clips across 10 categories like air conditioning, car horns, dog barks, gunshots, jackhammers, and street music in WAV format.
- **FSD50K** [18] is a dataset of 51197 labeled voice events from Freesound, over 100 hours of audio across 200 classes, with clips ranging from 0.3 to 30 seconds.

2.2. **Data augmentation.** To expand the training dataset, this study creates a variety of real and varied sample from the existing samples by applying each of these augmentation techniques: AGN, PS, S, TM, TS. The parameter settings applied in each augmentation technique are shown in Table 1.

TABLE 1. Parameter settings of augmentation techniques applied in this experimental study

Augmentation techniques	Parameter
AGN	min_amplitudo=0.001, max_amplitudo=0.015, p=0.5
PS	min_semitones=−4, max_semitones=4, p=0.5
S	min_shift=0.5, max_shift=0.5, shift_unit='fraction', p=0.5
TM	min_band_part=0.1, max_band_part=0.15, fade=True, p=0.5
TS	min_rate=0.8, max_rate=1.25, leave_length_unchanged=True, p=0.5

AGN is a technique that adds random Gaussian noise to audio signals, simulating real-world distortion and noise. Gaussian noise follows a normal distribution and resembles natural sounds like thermal noise. In this study, the noise's amplitude will range from 0.001 to 0.015 of the original signals, with a 50% chance of application for each audio sample. While PS is the second audio augmentation technique used in this study that adjusts pitch within a range of −4 to +4 semitones without altering duration or tempo, allowing for natural-sounding pitch variations.

S augmentation shifts an audio signal in time, controlled by min_shift and max_shift parameters. With both set to 0.5 as a fraction in this experiment, the signal shifts uniformly by 50% of its duration. Positive shifts move right; negative shifts move left.

TM is an audio augmentation method that masks 10% to 15% of a signal's duration, determined by min_band_part and max_band_part parameters. The fade=True option ensures smooth transitions into and out of the mask to avoid artifacts, while p=0.5 indicates a 50% chance of applying the augmentation. While TS is an augmentation altering audio playback speed without pitch change, defined by min_rate and max_rate parameters. With min_rate=0.8 and max_rate=1.25, speed can range from 80% to 125%. The leave_length_unchanged=True keeps audio duration constant, while p=0.5 dictates augmentation application likelihood.

To evaluate the quality of the samples generated by the five augmentation techniques, we used two ways. The first is through manual review, which involves listening directly to the augmented audio samples. If the audio sample still conveys the same meaningful information as the original audio, for example, if the siren sound is still audible when augmented by adding noise, then the augmentation result is considered high-quality. The second method is to examine the model's performance when the augmented audio is used as training data and the original audio as validation data. If the model's performance improves, then the augmentation result is high-quality.

2.3. **Data pre-processing.** Convert raw audio data into a neat, ready-to-use, and standardized format to ensure uniformity across the dataset for model analysis and training. These stages include padding or cutting duration into 4 seconds, resampling by modifying the rate into 16 kHz, and normalizing by setting the amplitude of the audio signal into a standard range, generally between -1 and 1, to avoid clipping or too weak a signal.

2.4. **Feature extraction.** Feature extraction identifies key attributes of audio signals. This research explores three techniques: Mel spectrogram, log-mel spectrogram, and MFCC for sound classification, with pseudocode for generating these features depicted in Figure 2. Beginning with sampling, the audio signal $x(n)$ is segmented into a distinct window employing the Framing and Windowing Technique $W(n)$. According to [19], a sampling rate of 16 kHz strikes a favourable balance between the model's computational expense and the input sample's quality for environmental sounds. In this experiment, we sampled the input using windows at a rate of 16 kHz, with 40 mel bands, an FFT window length of 2048, and a hop length of 512, representing the number of samples

```
FUNCTION FEATURE_EXTRACTION (audio_signal x(n), sample_rate Fs, method):
  // Framing and Windowing
  DEFINE frame_length_samples = DURATION_FRAME * Fs
  DEFINE hop_length_samples = OVERLAP_FRAME * Fs
  DEFINE window_function W(n) = HANN_WINDOW_FUNCTION
  audio_frames = []
  FOR each segment in audio_signal with frame_length_samples and hop_length_samples:
    frame = segment * W(n)
    ADD frame to audio_frames

  //  Fast Fourier Transform (FFT)
  FFT_output Y(f) = []
  FOR each frame in audio_frames:
    spectrum = ABSOLUTE_VALUE(FFT(frame))
    ADD spectrum to FFT_output Y(f)

  //  Mel Filtering Bank
  DEFINE mel_filter_bank F_mel = GENERATE_MEL_FILTER_BANK(Fs, FFT_SIZE, NUM_MEL_BINS)
  melspectrogram_raw = []
  FOR each spectrum Y_i(f) in FFT_output Y(f):
    mel_spectrum = DOT_PRODUCT(mel_filter_bank F_mel, spectrum Y_i(f))
    ADD mel_spectrum to melspectrogram_raw
  IF method=='melspectrogram':
    RETURN melspectrogram_raw
  ELSE IF method=='log-melspectrogram':
    // Log Transformation
    log_melspectrogram_raw = []
    FOR each mel_spectrum in melspectrogram_raw:
      log_mel_spectrum = LOGARITHM(mel_spectrum + EPSILON)
      ADD log_mel_spectrum to log_melspectrogram_raw
    RETURN log_melspectrogram_raw

  ELSE IF method=='mfcc:
    // Log Transformation
    log_melspectrogram_raw = []
    FOR each mel_spectrum in melspectrogram_raw:
      log_mel_spectrum = LOGARITHM(mel_spectrum + EPSILON)
      ADD log_mel_spectrum to log_melspectrogram_raw
    //  Discrete Cosine Transform (DCT)
    MFCC_output = []
    FOR each log_mel_spectrum in log_melspectrogram_raw:
      mfcc_coeffs = DCT(log_mel_spectrum, NUM_MFCC_COEFFICIENTS)
      ADD mfcc_coeffs to MFCC_output
    RETURN MFCC_output
```

FIGURE 2. Pseudocode of Mel spectrogram, log-mel spectrogram, and MFCC extractions

between consecutive frames. The Fast Fourier Transform for each frame is subsequently computed to convert the signal from the time domain into the frequency domain, producing $Y(f)$. Afterward, a 40 mel-band Fmel is produced. A spectrogram is generated for every window. The log-mel spectrogram was obtained by applying a log transformation following the generation of mel bands. The outcome of the log-mel transformation is subsequently utilized to generate a log-mel spectrogram. The mel-scale cepstral coefficients MFCC were derived by computing the discrete cosine transform (DCT) from the log-mel spectrogram, which is its formula shown by Equation (1).

Let $S(m, n)$ be the log-mel spectrogram value at the m-th Mel frequency bin and the n-th time frame. Then, the MFCC for the k-th coefficient at the n-th time frame, denoted as $C(k, n)$, is calculated using the discrete cosine transform (DCT) on the log-mel spectrogram:

$$C(k,n) = \sum_{m=0}^{M-1} S(m,n) \cdot \cos\left[\frac{\pi k}{M}\left(m + \frac{1}{2}\right)\right] \tag{1}$$

where $C(k, n)$ is the k-th MFCC coefficient for the n-th time frame. M is the total number of Mel bins (the number of filters in the Mel filtering bank). k is the calculated MFCC coefficient index, ranging from 0 to $K-1$, where K is the number of MFCC coefficients. $\cos\left[\frac{\pi k}{M}\left(m + \frac{1}{2}\right)\right]$ is a basis of the DCT.

2.5. **Model development.** This study utilized a CNN model as the classifier. The architecture of the model is illustrated in Figure 3. It comprises two 2D convolutional layers, each of which is followed by a 2D max pooling layer. The pooling layer reduces the dimensionality of the feature maps [20]. At the end of the architecture, there are two fully connected layers that receive input as a flattened matrix obtained from the output of the last max pooling layer. Each generated feature map in the convolutional layer and first fully connected layer was passed individually through the ReLU function. The last fully connected layer used the SoftMax function as the activation function.

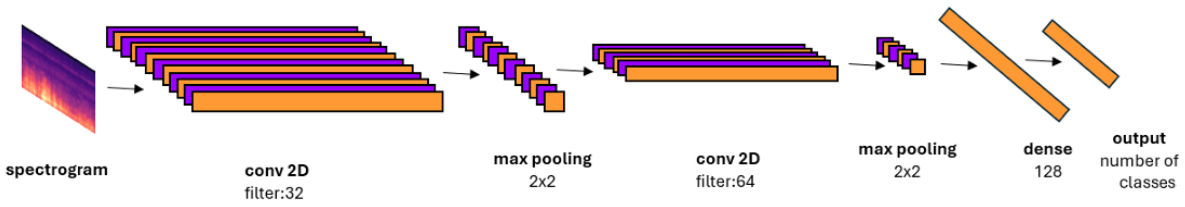


FIGURE 3. Architecture of CNN. It consists of two 2D convolutional layers followed by a 2D max pooling layer.

The initial dataset of N samples was split using stratified sampling, with 60% allocated for training and 40% for testing. After this division, an augmented dataset, containing an additional N sample, was added to the training set. Consequently, the final training set included 1.6N samples, while the testing set remained at 0.4N samples. This approach ensures the model benefits from augmented variety during training while preserving the integrity of evaluations on the untouched testing data.

2.6. **Model evaluation.** Model performance evaluation is a crucial step in validating the effectiveness of the proposed approach. In this study, model performance was assessed using accuracy and F1-score metrics. The F1-score was chosen as a more reliable metric for class imbalance in the dataset. To obtain robust and reliable results, each model was trained and tested five times. The final results of each test are presented as the average performance along with a standard deviation. This approach ensures that the findings are consistent and reliable.

3. **Result and Discussion.** In this section, we describe the result of the experimental study. The results of the classification using the CNN model for each dataset are listed in Tables 2 and 3. From both tables, it can be seen that without augmentation of the three feature representations, MFCC, Mel spectrogram, and log-mel spectrogram have quite low performance and vary across datasets, especially on ESC-50 and FSD50K where the accuracy and F1-score are often below 40%. The implementation of the AGN technique consistently produces the highest jump for all combinations. On MFCC, the accuracy increases from around 69%-82% to around 92%-95% for BDLib2 and US8K, while on Mel spectrogram and log-mel spectrogram, it reaches over 94% except for FSD50K which is only around 78% with log-mel spectrogram feature extraction. The application of the PS technique also significantly improves the metrics but is slightly behind AGN with an average decrease of around 2-5 percentage points and a higher standard variation. The performance of the S technique shows high efficiency on MFCC but is less stable on Mel spectrogram and log-mel spectrogram, where the accuracy and F1-score decrease significantly. The temporal TM technique shows progress and consistency.

TABLE 2. The accuracy of classification for each dataset. This experimental study investigates the impact of each of 5 augmentation and 3 feature extraction techniques.

Representation	Augmentation	Accuracy (%) ± std			
		BDLib2	ESC-50	US8K	FSD50K
MFCC	–	69.44 ± 0.06	40.00 ± 0.03	82.12 ± 0.01	30.37 ± 0.01
	AGN	95.00 ± 0.01	72.17 ± 0.02	92.01 ± 0.01	34.68 ± 0.01
	PS	91.94 ± 0.01	69.75 ± 0.02	89.13 ± 0.00	35.85 ± 0.03
	S	96.11 ± 0.02	88.92 ± 0.06	94.54 ± 0.01	38.88 ± 0.02
	TM	97.50 ± 0.01	92.62 ± 0.01	94.24 ± 0.00	38.32 ± 0.02
	TS	$\mathbf{99.72 \pm 0.01}$	$\mathbf{93.87 \pm 0.01}$	$\mathbf{94.68 \pm 0.00}$	$\mathbf{36.84 \pm 0.03}$
mel	–	45.56 ± 0.05	18.60 ± 0.04	64.21 ± 0.04	24.93 ± 0.02
	AGN	$\mathbf{99.72 \pm 0.01}$	$\mathbf{99.60 \pm 0.00}$	$\mathbf{96.04 \pm 0.00}$	$\mathbf{94.73 \pm 0.01}$
	PS	89.72 ± 0.09	83.95 ± 0.04	90.68 ± 0.01	63.36 ± 0.03
	S	73.89 ± 0.05	36.47 ± 0.02	80.70 ± 0.02	43.88 ± 0.01
	TM	90.28 ± 0.03	95.97 ± 0.02	94.28 ± 0.01	90.64 ± 0.01
	TS	72.50 ± 0.07	47.60 ± 0.02	83.57 ± 0.03	46.09 ± 0.04
log-mel	–	49.17 ± 0.07	28.80 ± 0.04	69.75 ± 0.02	30.38 ± 0.02
	AGN	$\mathbf{98.89 \pm 0.01}$	$\mathbf{95.27 \pm 0.01}$	$\mathbf{95.90 \pm 0.00}$	78.80 ± 0.01
	PS	98.61 ± 0.02	94.83 ± 0.02	91.67 ± 0.01	66.97 ± 0.03
	S	87.78 ± 0.02	59.22 ± 0.07	88.09 ± 0.00	47.81 ± 0.05
	TM	95.83 ± 0.02	90.08 ± 0.02	94.25 ± 0.01	$\mathbf{84.67 \pm 0.04}$
	TS	89.72 ± 0.04	79.05 ± 0.04	90.07 ± 0.01	58.20 ± 0.02

The majority of augmentation methods significantly improve performance compared to no augmentation. In Figure 4, without augmentation, median accuracy was about 47%, and median F1-score was about 40%, highlighting performance instability and sensitivity to data variations. After applying AGN, median accuracy rose to approximately 95%, and median F1-score to 96%, with a much smaller IQR, indicating increased model consistency and noise robustness. The shorter AGN whiskers confirmed effective noise-based augmentation. The PS technique improved performance, achieving median accuracy and F1-score of 92%, but had longer whiskers and outliers below 50%, suggesting that extreme pitch shifts can hinder timbral pattern recognition. The TM method led to median accuracy and F1-score of about 96%, with a more focused distribution and fewer outliers, maintaining class content integrity despite duration and delay variations. In contrast, S and

TABLE 3. The F1-score of classification for each dataset. This experimental study investigates the impact of each of 5 augmentation and 3 feature extraction techniques.

Representation	Augmentation	F1-score (%) ± std			
		BDLib2	ESC-50	US8K	FSD50K
MFCC	–	68.88 ± 0.05	38.21 ± 0.03	82.24 ± 0.00	26.62 ± 0.01
	AGN	94.92 ± 0.01	72.33 ± 0.02	92.20 ± 0.01	31.39 ± 0.01
	PS	91.76 ± 0.01	69.44 ± 0.02	89.31 ± 0.00	32.38 ± 0.03
	S	96.01 ± 0.03	88.83 ± 0.06	94.69 ± 0.01	**36.02 ± 0.03**
	TM	97.51 ± 0.01	92.61 ± 0.01	94.37 ± 0.00	35.48 ± 0.03
	TS	**99.72 ± 0.01**	**93.82 ± 0.01**	**94.83 ± 0.00**	33.81 ± 0.03
mel	–	42.53 ± 0.06	16.89 ± 0.04	64.52 ± 0.04	21.67 ± 0.03
	AGN	**99.72 ± 0.01**	**99.60 ± 0.00**	**96.21 ± 0.00**	**94.93 ± 0.01**
	PS	89.28 ± 0.10	84.17 ± 0.04	90.79 ± 0.01	63.46 ± 0.03
	S	72.34 ± 0.05	37.84 ± 0.03	80.99 ± 0.02	42.96 ± 0.02
	TM	90.31 ± 0.03	96.07 ± 0.02	94.43 ± 0.01	90.73 ± 0.01
	TS	72.13 ± 0.07	48.10 ± 0.02	83.70 ± 0.03	45.49 ± 0.04
log-mel	–	45.69 ± 0.07	27.22 ± 0.04	69.92 ± 0.02	27.85 ± 0.02
	AGN	**98.88 ± 0.01**	**95.26 ± 0.01**	**96.06 ± 0.00**	78.81 ± 0.01
	PS	98.58 ± 0.02	94.79 ± 0.02	91.77 ± 0.01	66.82 ± 0.03
	S	87.52 ± 0.02	59.01 ± 0.08	88.21 ± 0.01	46.60 ± 0.06
	TM	95.89 ± 0.01	90.10 ± 0.02	94.39 ± 0.01	**84.70 ± 0.04**
	TS	89.72 ± 0.04	78.98 ± 0.04	90.21 ± 0.01	57.97 ± 0.02

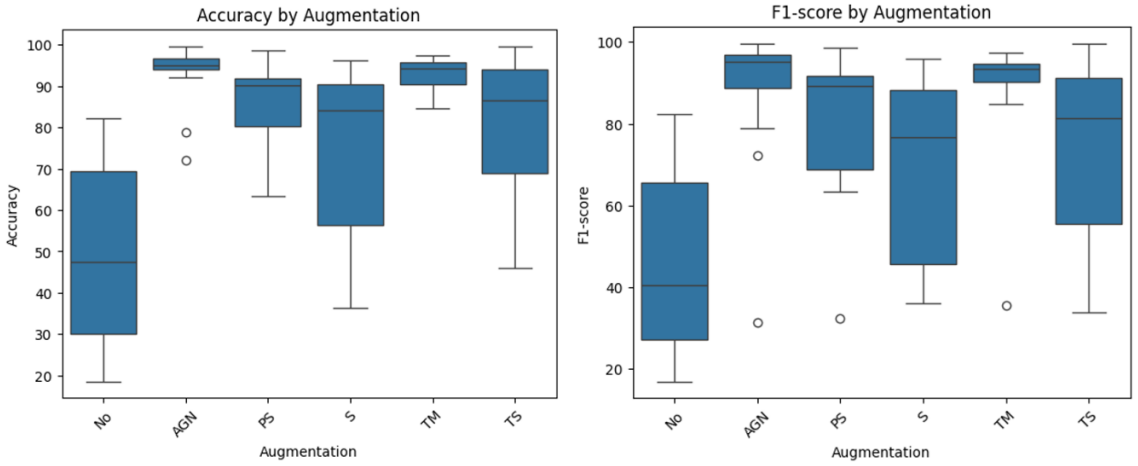


FIGURE 4. Boxplot of accuracy and F1-score for each augmentation (No: No Augmentation, AGN: Additive Gaussian Noise, PS: Pitch Shift, S: Shift, TM: Time Mask, TS: Time Stretch) applied for four datasets (BDLib2, ESC-50, US8K, FSD50K)

TS techniques showed high variability and long whiskers, as changes in duration significantly distorted the temporal spectrum, negatively affecting pattern detection. Variations in recording sources further influenced performance based on the nature of the audio.

Overall, this visualization clearly shows that AGN, TM, and PS emerged as the most effective and consistent augmentation methods in improving accuracy as well as F1-score. This highlights the importance of these techniques in building models that are resilient to data variation and improving generalization capabilities, as well as reducing the risk of model failure in the field.

This study also investigates the effects of various audio augmentation methods on model performance using MFCC, Mel spectrogram, and log-mel spectrogram feature representations. As shown in Figure 5, for MFCC, augmentations such as TM, TS, and S consistently achieved high accuracy, about 95-98%, with minimal variability, exceeding the performance of the unaugment baseline, 69-70%. This indicates that MFCC is highly advantageous due to temporal adjustments.

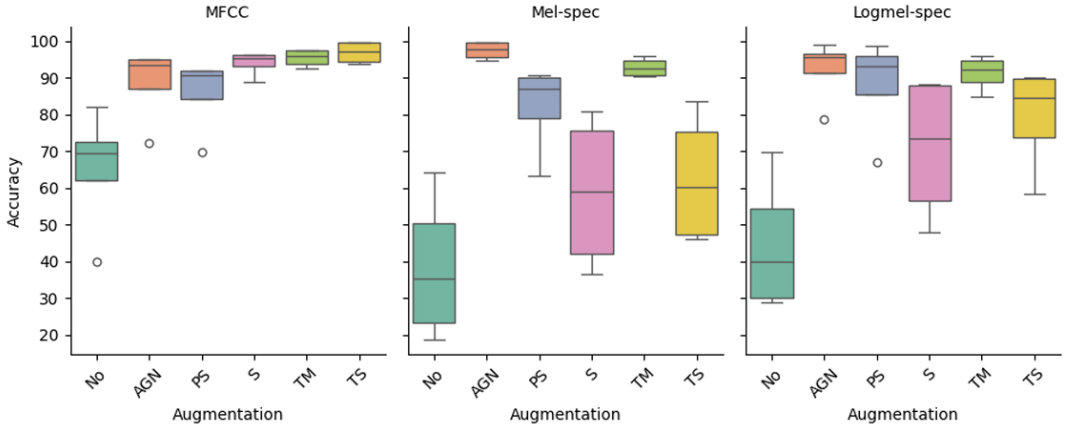


FIGURE 5. Boxplot of accuracy of each augmentation for each feature representation applied for four datasets (BDLib2, ESC-50, US8K, FSD50K)

In contrast, Mel spectrogram extraction starts from a lower median accuracy of around 35% without augmentation and exhibits more extreme variation than MFCC. AGN and PS techniques nearly close the gap, increasing the median to the 88-97% range, but the long whisker in PS indicates some trials still end up lower. Shifting exhibits the most fluctuating distribution, followed by TS, indicating that these two augmentations are less suitable for Mel spectrogram. Meanwhile, TM returns to a median of around 92% with a maintained range, indicating that temporal manipulation is more suitable for Mel spectrogram representation. Log-mel spectrogram exhibits a similar pattern to Mel spectrogram, with a low baseline and significant improvement when AGN, PS, and TM are applied, with the median reaching above 90%.

Overall, MFCC outperforms the baseline and maintains the highest accuracy when augmentation is applied. Mel spectrogram and log-mel spectrogram nearly match the performance of MFCC with appropriate augmentation, so the choice of augmentation technique significantly impacts the model's effectiveness.

4. **Conclusions and Future Works.** This study enhances environmental sound classification by examining data augmentation and feature extraction methods. Results indicate that data augmentation boosts model accuracy, with additive Gaussian noise and Mel spectrogram yielding impressive results 99.72% for BDLib2, 99.60% for ESC-50, 96.04% for US8K, and 94.73% for FSD50K. The study emphasizes the importance of appropriate augmentation and feature extraction techniques for optimal performance.

Future work can focus on optimizing augmentation techniques and feature extraction for self-supervised learning in environmental sound classification. Additionally, investigating the impact of various augmentation methods, including combinations, on model performance can enhance classification effectiveness.

Availability of Data and Materials. The datasets supporting the conclusions of this article are BDLib2 dataset (available in https://m3c.web.auth.gr/research/datasets/bdlib/), ESC-50 dataset (available in https://doi.org/10.7910/DVN/YDEPUT), US8K dataset (available in https://urbansounddataset.weebly.com/urbansound8k.html) and FSD50K dataset (available in https://doi.org/10.5281/zenodo.4060432).

Authors' Contributions. Theresia Herlina Rochadiani carried out the experiment of the research and wrote the manuscript. Yulyani Arifin reviewed the manuscript and provided critical feedback. Derwin Suhartono analyzed the experimental results. Widodo Budiharto provided directions regarding the methodology and supervised the implementation of the study.

REFERENCES

[1] D. M. D. Ferrão, *Data Augmentation for Urban Environmental Sound Classification on Edge Devices*, Master Thesis, Universidade de Lisboa, 2023.

[2] S. Wei, S. Zou, F. Liao and W. Lang, A comparison on data augmentation methods based on deep learning for audio classification, *Journal of Physics: Conference Series*, vol.1453, no.1, 012085, DOI: 10.1088/1742-6596/1453/1/012085, 2020.

[3] T. Turskis, M. Teleiša, R. Buckiūnaitė and D. Čalnerytė, Mixed-type data augmentations for environmental sound classification, *CEUR Workshop Proceedings*, vol.3575, pp.184-194, 2023.

[4] Y. Chen, Z. Yan, Y. Zhu, Z. Ren, J. Shen and Y. Huang, Data augmentation for environmental sound classification using diffusion probabilistic model with top-k selection discriminator, *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol.14087, pp.283-295, DOI: 10.1007/978-981-99-4742-3_23, 2023.

[5] F. Demir, D. A. Abdullah and A. Sengur, A new deep CNN model for environmental sound classification, *IEEE Access*, vol.8, pp.66529-66537, DOI: 10.1109/ACCESS.2020.2984903, 2020.

[6] D. Chong, Y. Zou and W. Wang, Multi-channel convolutional neural networks with multi-level feature fusion for environmental sound classification, *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol.11296, pp.157-168, DOI: 10.1007/978-3-030-05716-9_13, 2019.

[7] Y. Su, K. Zhang, J. Wang, D. Zhou and K. Madani, Performance analysis of multiple aggregated acoustic features for environment sound classification, *Applied Acoustics*, vol.158, 107050, DOI: 10.1016/j.apacoust.2019.107050, 2020.

[8] Z. Mushtaq and S-F. Su, Efficient classification of environmental sounds through multiple features aggregation and data enhancement techniques for spectrogram images, *Symmetry*, vol.12, no.11, 1822, DOI: 10.3390/sym12111822, 2020.

[9] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe and S. Misra, Data augmentation and deep learning methods in sound classification: A systematic review, *Electronics*, vol.11, no.22, 3795, DOI: 10.3390/electronics11223795, 2022.

[10] X. Chen, M. Wang, R. Kan and H. Qiu, Improved patch-mix transformer and contrastive learning method for sound classification in noisy environments, *Applied Sciences*, vol.14, no.21, 9711, DOI: 10.3390/app14219711, 2024.

[11] A. Meghanani, C. S. Anoop and A. G. Ramakrishnan, An exploration of log-mel spectrogram and MFCC features for alzheimer's dementia recognition from spontaneous speech, *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp.670-677, DOI: 10.1109/SLT48900.2021.9383491, 2021.

[12] T. Al-Maashani, I. Mendonca and M. Aritsugi, Age classification based on voice using mel-spectrogram and MFCC, *2023 24th International Conference on Digital Signal Processing (DSP)*, pp.1-5, DOI: 10.1109/DSP58604.2023.10167887, 2023.

[13] H. S. L. Eduard and Y. B. Park, Audio deep fake detection using long exposure log-mel spectrogram, *Journal of Platform Technology*, vol.12, no.6, pp.3-16, https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE12101340, 2024.

[14] V. Bountourakis, L. Vrysis and G. Papanikolaou, Machine learning algorithms for environmental sound recognition, *Proc. of the Audio Mostly 2015 on Interaction With Sound*, pp.1-7, DOI: 10.1145/2814895.2814905, 2015.

[15] V. Bountourakis, L. Vrysis, K. Konstantoudakis and N. Vryzas, An enhanced temporal feature integration method for environmental sound recognition, *Acoustics*, vol.1, no.2, pp.410-422, DOI: 10.3390/acoustics1020023, 2019.

[16] K. J. Piczak, ESC: Dataset for environmental sound classification, *Proc. of the 23rd ACM International Conference on Multimedia*, pp.1015-1018, DOI: 10.1145/2733373.2806390, 2015.

[17] J. Salamon, C. Jacoby and J. P. Bello, A dataset and taxonomy for urban sound research, *Proc. of the 22nd ACM International Conference on Multimedia*, pp.1041-1044, DOI: 10.1145/2647868.2655045, 2014.

[18] E. Fonseca, X. Favory, J. Pons, F. Font and X. Serra, FSD50K: An open dataset of human-labeled sound events, *IEEE/ACM Trans. Audio, Speech, Language Processing*, vol.30, pp.829-852, DOI: 10.1109/TASLP.2021.3133208, 2022.

[19] S. Abdoli, P. Cardinal and A. L. Koerich, End-to-end environmental sound classification using a 1D convolutional neural network, *Expert Systems with Application*, vol.136, DOI: 10.1016/j.eswa.2019.06.040, 2019.

[20] K. Nugroho, E. Noersasongko, Purwanto, Muljono and D. R. I. M. Setiadi, Enhanced indonesian ethnic speaker recognition using data augmentation deep neural network, *Journal of King Saud University – Computer Information Science*, vol.34, no.7, pp.4375-4384, DOI: 10.1016/j.jksuci.2021.04.002, 2022.