Journal of
Sensor and Actuator Networks

MDPI

Article

Impact Evaluation of Sound Dataset Augmentation and Synthetic Generation upon Classification Accuracy

Eleni Tsalera [1,*], Andreas Papadakis [2], Gerasimos Pagiatakis [2] and Maria Samarakou [1,*]

[1] Department of Informatics and Computer Engineering, School of Engineering, University of West Attica, 11521 Athens, Greece

[2] Department of Electrical and Electronic Engineering Educators, School of Pedagogical and Technological Education (ASPETE), 15122 Athens, Greece; apapadakis@aspete.gr (A.P.); pagiatakis@aspete.gr (G.P.)

* Correspondence: etsalera@uniwa.gr (E.T.); marsam@uniwa.gr (M.S.)

Abstract

We investigate the impact of dataset augmentation and synthetic generation techniques on the accuracy of supervised audio classification based on state-of-the-art neural networks used as classifiers. Dataset augmentation techniques are applied upon the raw sound and its transformed image format. Specifically, sound augmentation techniques are applied prior to spectral-based transformation and include time stretching, pitch shifting, noise addition, volume controlling, and time shifting. Image augmentation techniques are applied after the transformation of the sound into a scalogram, involving scaling, shearing, rotation, and translation. Synthetic sound generation is based on the AudioGen generative model, triggered through a series of customized prompts. Augmentation and synthetic generation are applied to three sound categories: (a) human sounds, (b) animal sounds, and (c) sounds of things, with each category containing ten sound classes with 20 samples retrieved from the ESC-50 dataset. Sound- and image-orientated neural network classifiers have been used to classify the augmented datasets and their synthetic additions. VGGish and YAMNet (sound classifiers) employ spectrograms, while ResNet50 and DarkNet53 (image classifiers) employ scalograms. The streamlined AI-based process of augmentation and synthetic generation, enhanced classifier fine-tuning and inference allowed for a consistent, multicriteria-comparison of the impact. Classification accuracy has increased for all augmentation and synthetic generation scenarios; however, the increase has not been uniform among the techniques, the sound types, and the percentage of the training set population increase. The average increase in classification accuracy ranged from 2.05% for ResNet50 to 9.05% for VGGish. Our findings reinforce the benefit of audio augmentation and synthetic generation, providing guidelines to avoid accuracy degradation due to overuse and distortion of key audio features.

Keywords: dataset enhancement; wavelet manipulation; scalogram augmentations; synthetic sound generation; classification accuracy

check for updates

Academic Editors: Manolo Dulva Hina and Amar Ramdane-Cherif

Received: 21 July 2025
Revised: 6 September 2025
Accepted: 7 September 2025
Published: 9 September 2025

Citation: Tsalera, E.; Papadakis, A.; Pagiatakis, G.; Samarakou, M. Impact Evaluation of Sound Dataset Augmentation and Synthetic Generation upon Classification Accuracy. *J. Sens. Actuator Netw.* **2025**, *14*, 91. https://doi.org/10.3390/jsan14050091

1. Introduction

Sound classification employs labeled sound datasets of generic or specific thematic orientation, including environmental, urban, animal, and machine-made sounds. Datasets are frequently characterized by asymmetrical distribution of audio samples per class type, differences in excerpt duration and quality, and noise. In parallel, the quick succession and/or overlapping of audio types create difficulties in audio labeling.

Improved sound classification accuracy strengthens the detection, decision, and feedback loop in AI-based systems. The correct identification of a sound, despite noise, rarity, or ambiguity, typically leads to the detection of an event. The event, based on its contextualized meaning, is interpreted and can support a decision. Such decision may trigger, automatically or through human-in-the-loop intervention, actuation and a corrective or responsive/feedback action. Subsequently, sound classification may take place again to verify the persistence, adaptation, or ceasing of the causal sound. By closing the loop, the contextual feedback can allow for improvement of the classification in terms of accuracy.

Examples of such systems come from multiple areas, including healthcare monitoring (regarding patient safety detection of distress sounds, coughing, or abnormal breathing which may support diagnosis or trigger prompt interventions), in predictive maintenance (where acoustic anomalies and irregular vibration patterns may precede engine faults and distinguish operational noise variations from fault conditions), and smart cities (where detecting sound events such as glass breaking and pedestrian cues can support context awareness, improve incident response times, and trigger assistive applications). The availability of audio samples per audio type may vary especially for infrequent or difficult to capture and/or reproduce sound types. The collection of high-quality, annotated recordings for such classes is resource-intensive and contextually constrained. Such objective difficulties create asymmetries in class population and representation. Overall, labeling imprecision, noisy environments, insufficient annotation data, and imbalanced classes creates a persistent challenge in sound classification affecting the performance of relevant algorithms [1]. Improved sound classification accuracy allows for lower false-positive triggers preventing unnecessary interventions or alarms as well as false-negative rates so that events of interest are not missed. In addition, in real time settings, reliable accuracy allows for event-based data recording and persistence, instead of continuous audio surveillance, which can be beneficial for technical groundings and data privacy concerns.

Sound augmentation techniques confront such bottlenecks allowing the expansion of the diversity and quantity of available training data and improving the model generalization particularly for underrepresented classes. Manipulation of existing audio recordings can be based on techniques applied on the waveform such as time stretching, pitch shifting and noise addition.

In parallel, recent evolutions in multimedia generative artificial intelligence (AI) may be applied for sound dataset enhancements through the generation of synthetic sound clips. Generative audio models include Generative Adversarial Networks (GANs) [2], autoregressive models [3], and transformer-based audio generation. These generate sounds from labeled latent variables or textual prompts.

The main contributions and innovations of this work include (a) the application of different types of sound augmentation and synthetic generation techniques to increase the population and variability of the training sets of a wide pool of potentially underrepresented sound classes and (b) the investigation of the potential impact upon the classification accuracy for these classes. Synthetic generation has employed the AudioGen auto-regressive deep learning model, generating audio samples from text descriptions. The intensity of applying the augmentation and generative techniques, quantified as the volume of the new (or modified) sound clips added to the training sets, has been configurable to three discrete values to evaluate the corresponding impact. Three sound datasets (categories) have been created, consisting of human, animal, and thing sounds, as subsets of the ESC-50 sound dataset. Including ten discrete classes per category (human, animals, and things) allows an adequate representation of different sound types, with some of them more difficult to capture or potentially under-represented under real conditions typically (e.g., teeth brushing, frog, can opening).

A streamlined and uniform AI-based pipeline has been designed involving state-of-the-art neural networks so that the results are directly comparable for each applied technique. Specifically, augmentation techniques have been performed directly upon the sound (wav) file as well as to its frequency-based, image representations (scalograms), referred to as sound- and image-based augmentations, respectively. The sound classification is performed using two groups of state-of-the-art, convolutional neural networks (CNNs). The first group (VGGish and YAMNet) are CNNs focusing on sound classification using sound spectrograms, while the second group (ResNet50 and DarkNet53) are generic image classification CNNs, employed upon sound scalograms. The increase in the training set population has been logged arithmetically escalating (50%, 100%, and 200%) to investigate its influence upon the classification accuracy. These flows have resulted in the performance of a series of 252 scenarios.

While sound augmentation techniques have attracted research interest to enhance the availability of difficult-to-capture sounds and have been investigated, in this work we combine, for the first time to our knowledge, synthetic generation (through AudioGen, https://audiocraft.metademolab.com/audiogen.html, accessed on 30 May 2025) and different types of augmentation techniques for enhancing the training sets and comparing the impact upon the sound classification accuracy through an AI-based workflow. This approach complements the separate studies of augmentation techniques and synthetic generation, providing observations on its application for different sound types (human, animals, things) and with escalating intensity/volume.

As the main area of this study is the intersection of data augmentation, synthetic generation (via AudioGen), and sound classification, its findings can be applicable to practical use-cases involving sound classification, especially involving types of sounds that are difficult to capture, or relevant datasets are not readily available. The mechanism to improve the neural networks training makes the application of AI more robust and reliable in a series of thematic areas, indicatively referring to smart cities detecting rare sounds (such as glass breaking and human distress sounds), the industry supporting predictive maintenance (detecting fault sounds in engines and underrepresented hazardous sounds), healthcare, monitoring, and consumer technology (classifying non-speech sounds, and generating rare environmental sounds).

The structure of the paper is as follows: Section 2 describes current works in augmentation techniques, synthetic sound generation, and classification methods. Section 3 provides a detailed description of the methodology followed, as well as the materials used. Section 4 presents the results obtained from the experimental evaluation, while Section 5 critically discusses the results to identify interdependences and the thresholds of extending datasets through augmentation and synthetic generation. Section 6 includes the conclusions of the work.

2. State of the Art

2.1. Sound Augmentation Techniques

An augmentation method applied on both the input and the hidden representations of features in deep neural networks is proposed [4]. This approach yields an improvement in classification accuracy of approximately 4% in DCASE 2018 and 2019 Task 1 datasets. In [5], four time-domain augmentation techniques are applied on the UrbanSound8K dataset. The results demonstrate consistent improvements across accuracy, precision, recall, and F1-score. Multiple data augmentation protocols applied on both raw audio signals of bird and cat sounds and spectrogram representations are examined in [6]. The findings demonstrate that training five distinct convolutional neural networks (CNNs) on both original and augmented data significantly improves classifier performance. In [7], audio effects are used as augmentation techniques to enhance the training data. The results present a notable improvement in classification accuracy, indicating improved generalization to effect-processed samples. The authors in [8] introduce time-overlapping spectrogram window and pitch shift as augmentation techniques to address limited labeled data in Music Information Retrieval (MIR) tasks. These augmentations increase the training data volume threefold, improving model's generalization. A suite of augmentation techniques is applied on the UrbanSound8K dataset to [9] achieve improved classification accuracy. Additionally, the authors analyze the impact of each augmentation technique on individual sound classes, showing that specific augmentations are more effective on specific classes. In this study, time stretching, pitch shifting, noise injection, and volume control were employed as augmentation techniques to enhance temporal and spectral variability in the audio domain. Additionally, a series of geometric transformations were applied to the scalogram representations. The application of these augmentations led to an average of 4.86% increase in classification accuracy compared to the baseline models trained on non-augmented data. In [10], the authors analyze the performance of data augmentation techniques and conclude that data augmentation improves classification accuracy, Ref. [6] investigated data augmentation techniques on raw audio signals and their visual representation (spectrogram) improving the classification accuracy, Ref. [11] also applied augmentation techniques (incl. shift pitch, time stretch, and adding noise) for Sound Event Recognition (SER) improving the classification accuracy.

2.2. Synthetic Sound Generation

Challenges in collecting and annotating sound data involve data quality, scarcity, privacy, and distribution fairness among classes. The need for synthetic data, i.e., data generated by computer algorithms or simulation artificially annotated, is discussed in [12]. Synthetic audio generation expands dataset diversity without manual audio recording and labeling which can be challenging in noisy, uncontrolled environments. Synthetic data can be generated using statistical methods and deep-learning models [13]. The former is based on the estimation of statistical parameters of the existing real data, the design of the distribution model, and the subsequent creation of artificial data samples following the distribution model using random number generators or inverse transform sampling techniques. Deep learning-based methodologies compress the input data into a latent space, i.e., a more compact representation, used to reconstruct the original data. While initial efforts used statistical frameworks, deep learning models are increasingly used. These involve techniques including Large Language Models (LLMs), Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs).

Audio can be generated through physical modeling synthesis employing models to simulate the physical properties of the sound source (e.g., vocal cords, a machine). Sound generation has advanced mainly for categories related to human sounds. For example, neural speech synthesis models, such as WaveRNN [14] and LPCNet [15] create high-quality speech for text-to-speech, while the generation of synthetic datasets through sound representation and configuration is investigated in [16]. The WaveNet [17] generative model for raw audio has been used to generate arbitrary sounds in a controllable way. Jukebox [18] allows high-fidelity music synthesis and uses hierarchical Vector Quantized Variational Autoencoders (VQ-VAE) with separate up-sampling networks, in the music domain.

AudioGen [19] is a generative transformer-based mechanism for semantic-enabled, text-to-audio synthesis based on an autoregressive, decoder-only transformers trained to generate sequences of discrete audio tokens. The model architecture integrates a pretrained EnCodec neural audio codec to discretize waveform data into latent audio tokens. The multi-band Residual Vector Quantization (RVQ) scheme reduces waveform redundancy and produces compact representations. AudioGen autoregressively predicts the discrete codebook indices across time, conditioned on both past audio tokens and the prompt encoding derived from the text. The model has been trained on large-scale datasets of paired text and audio, using teacher-forcing and token-level, cross-entropy objectives.

2.3. Sound Classification Methods

Sound classification techniques have evolved from statistical (e.g., Gaussian Mixture Models and Hidden Markov Models), to machine learning (e.g., Support Vector Machines, Decision Trees, and Random Forests) and deep learning (such as Convolutional and other types of Neural Networks) [20].

Classification techniques are applied upon features extracted from raw sound. Such features can be time-dependent (e.g., amplitude, zero-crossing rate), frequency-dependent (spectrum), or related to perception (e.g., Mel-Frequency Cepstral Coefficients, MFCCs) [21,22]. Deep learning classification algorithms are typically employed upon scalograms and spectrograms, learning spatial patterns in audio data. These are visual representations of the frequency content of the audio over time, and capture frequency-time relationships based on Short-Term Fourier Transforms (STFTs) or wavelet transforms [23]. STFTs convert the sound signal into a time-frequency representation, which is then converted through the Mel filter bank into a Mel spectrogram. The embeddings capture the spectral content (e.g., frequency patterns, timbre), the temporal patterns (e.g., rhythm and attack/decay) and perceptual qualities (e.g., harmonicity and roughness). According to [24], scalograms may outperform in non-stationary signals due to their multiresolution and spectrograms perform better in stationary due to their linear resolution.

Recurrent Neural Networks (RNNs) identify patterns in sequential data, for sound segmentation and speech recognition. AudioBERT uses self-attention mechanisms [25], while hybrid CNN-Transformers combine convolutional and transformer architectures for long sequences. Attention-based Long Short-Term Memory (LSTM) focuses on relevant parts of the input sequence. Specialized, large neural networks such as VGGish and YAMNet, based on the VGG and the MobileNet-v1 architectures, respectively, are trained in sound datasets. Similarly, convolutional neural networks use image-based sound representation (scalograms and/or spectrograms) [26].

In parallel, attention-based models, such as Audio Spectrogram Transformers (ASTs) capture global context in audio data and long-range dependencies. While neural networks perform local feature extraction via convolution, transformers perform global context modeling through self-attention [27]. DenseNet utilizes dense connections designed for image classification, benefiting from transfer learning to improve audio classification performance [28]. The transformer-based models (e.g., Wav2Vec2, AST) operate on raw waveform or patches of spectrograms and model temporal dependencies using attention mechanisms.

For the classification we employ well-established CNN models and use both main visual representations of sound, spectrograms, and scalograms. VGGish and YAMNet, which are specifically designed for audio classification tasks, use log-mel spectrograms. ResNet50 [29] and DarkNet53 [30] image classifiers work on scalograms. ResNet50 achieves the golden mean combination in classification accuracy, robustness to noise, and computational time [31] while DarkNet53 is comparable to ResNet50 in terms of architecture features, number of layers, and shortcut connections.

3. Materials and Methods

3.1. Selection of Sound Categories and Dataset Configuration

Sound categories have been selected considering by considering the diversity and wide representation of sound types, so that the classification task is challenging (non-trivial). The sound clips have been retrieved from the ESC-50 dataset [32], allowing for consistent labeling, a balanced distribution, and similar audio quality characteristics. ESC-50 is a widely used dataset for environmental sound classification and it has intrinsic characteristics, which make it reliable and appropriate for training and testing sound classification algorithms. ESC-50 offers non-overlapping sound categories with reliable labeling. Equally important, it provides a sufficiently large and heterogeneous set of sound classes, including types that are not easily captured in practice. At the same time, the homogeneous approach followed for its creation creates specific traits and even limitations, when considering more realistic settings. Specifically, sound recordings have been based on clean acoustic conditions, with limited background noise and without overlapping sources providing a single sound type per clip. The recording conditions also do not generalize across microphones, devices, or environments.

The ESC-50 dataset has been split into three sub-datasets: (a) human body operation and voluntary or involuntary movements, (b) animal sounds, and (c) sounds generated in an indoor environment by 'things' (tools and machines). This categorization is aligned with the hierarchical AudioSet ontology [33] as it includes three out of the seven main categories, i.e., human sounds, animals, and sounds of things.

The sound classes are depicted in Table 1. ESC-50 includes 40 samples per class. From these samples we use half of them, 20 samples per class, to resemble the realistic conditions of underrepresented classes.

Table 1. Three distinct datasets based on ESC-50.

Human Sounds	Animals	Sounds of Things
Crying baby	Dog	Pouring water
Sneezing	Rooster	Toilet flush
Clapping	Pig	Door knock
Breathing	Cow	Mouse click
Coughing	Frog	Keyboard typing
Footsteps	Cat	Door, wood cracks
Laughing	Hen	Can opening
Brushing teeth	Insects (flying)	Vacuum cleaner
Snoring	Crow	Clock alarm
Drinking/sipping	Chirping birds	Clock tick

ESC-50 sound waveforms (wav) are transformed into visual representations spectrograms and scalograms so that they can be used for augmentation (scalograms) and classification (both).

In Figure 1, each column includes three sound clips of an indicative sound class, i.e., laughing from human sounds, dog barking from animals, and keyboard typing from sounds of things. The figure indicates the distinct patterns observed across classes as well as the variations present within sounds of the same class. For example, in the case of laughing, harmonic energy bands cover human voice frequencies, while the dog barking is characterized by distinct intermitted patterns. The dashed lines correspond to percussive sounds in the case of keyboard sounds.

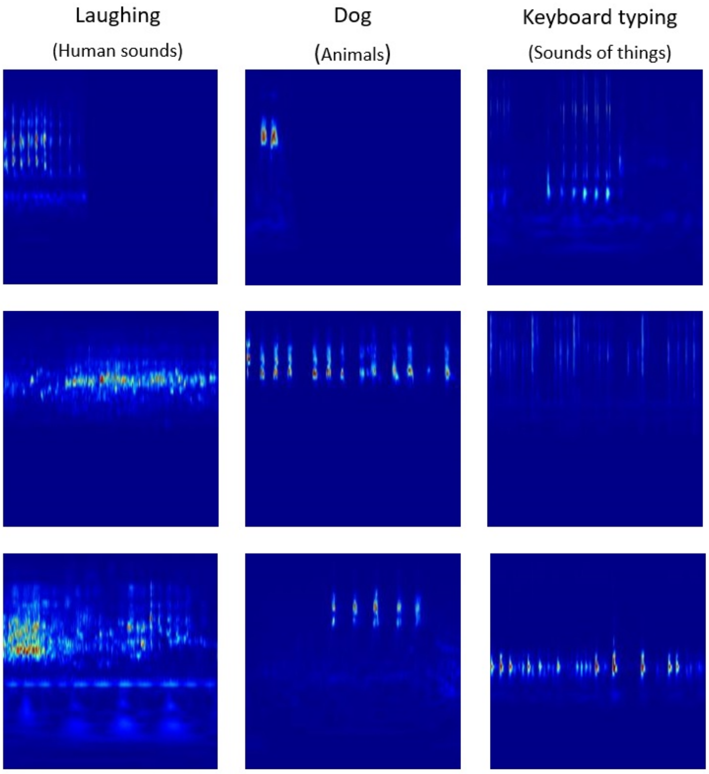

Figure 1. Scalograms of laughing (human), dog (animal) and keyboard typing (sounds of things).

Models trained only on ESC-50 (and other clean, well-organized datasets) may be less effective in uncontrolled, noisy, real-life environments acoustic settings. Data augmentation techniques and synthetic audio generation during training are used to bridge the gap between controlled datasets and real-world audio complexity. These methods help strengthen the model by simulating various real-world situations and expanding the dataset's range.

3.2. Augmentation Techniques

Augmentation creates variations in available audio to increase the diversity of training data and improve model generalization. Augmentation techniques can be applied on the original sound format ('sound augmentations') or its visual transformation, scalograms, and/or spectrograms ('image augmentations').

Sound augmentation techniques include time stretching, pitch shifting, and noise addition. Time-Scale Modification (TSM), such as time stretching, increases or decreases the temporal speed of an audio signal while preserving pitch and timbre [34]. Pitch shift modifies the spectrum of the signal modifying the perceived pitch of the audio, performed either with digital signal processing and/or neural network techniques [35]. Noise addition depends on the noise spectrum and patterns. Time stretch and shift pitch simulate changes in speech rate, accent, and frequency shifts from different microphones. Noise addition mimics real-world sounds such as traffic wind and device noise. Volume control resembles the differences in gain application across devices and user distances. Time shift addresses problems such as undesirable variation in the exact timing of digital audio signals and latency in streaming applications. Time stretch, pitch shift, noise addition, control volume, and time shift audio augmentation techniques have been applied separately and in combination.

Noise addition simulates unpredictable environmental interference, such as background chatter, noise, wind, or machinery. Volume controlling reflects amplitude variability due to different recording devices, or microphones distance. Pitch shifting is for frequency variability in real-world acoustic events, like variations in speaker voices and animal vocalizations. Time shifting mimics delays in real-world acoustic capture due to asynchronous recordings. Time stretching captures temporal fluctuations in sound events, e.g., footsteps, alarms occurring at different speeds. The above is summarized in Table 2.

Table 2. Audio-specific augmentation techniques and their real-world relevance.

Augmentation Technique	Analogous Real-World Condition
Add Noise	Unexpected environmental interference
Control Volume	Differences in recording sensitivity or microphone distance
Pitch Shift	Variability in vocalizations or mechanical tones
Time Shift	Temporal misalignments in uncontrolled environments
Time Stretch	Temporal variability in sound events

Image augmentation techniques include geometric transformations, color space augmentations, kernel filters, mixing images, random erasing, and feature space augmentation [36]. Geometric transformations include scaling, reflection, rotation, shearing, and translation [37]. Scaling stretches or compresses the scalogram along the time or frequency axis, enhancing or decreasing the focus on specific regions of the scalogram. Rotation changes the orientation of the time-frequency plane by a specific angle. Reflection reverses the scalogram on the horizontal (time) or vertical (frequency) axis. Shearing tilts the scalogram along the time or frequency axis, distorting the shape of the features. Translation shifts the scalogram horizontally or vertically, moving the features to a different location in the time-frequency plane.

For image-specific augmentation techniques, reflection helps the model to generalize across spectral inversions of mirrored recording artifacts. Rotation resembles orientation differences in spectrogram visualization. Scale enhances robustness to recording-level variations due to different devices. Shear introduces variability in scalograms' structure by introducing recording or playback devices. Translation captures offset events. These are summarized in Table 3.

Table 3. Image-specific augmentation techniques and their real-world relevance.

Augmentation Technique	Analogous Real-World Condition
Reflection	Mirrored or inverted acoustic structures
Rotation	Changes in angular positioning of time-frequency features
Scale	Amplitude or frequency scaling differences due to different devices
Shear	Distortions from recording processes
Translation	Temporal or frequency shifts in the scalograms

Although the image augmentations do not directly correspond to physical processes/actions performed on the sound, they have been included considering that the classifier is trained on image features (e.g., edges, textures, intensity gradients) and the geometric transformations may contribute to the prevention of model bias and overfitting in case of image quality degradation, noise addition, and other changes such as position or orientation alteration and zooming. Rotation, translation, shear, and scaling of the scalograms make the model less sensitive to frequency or time shifts and data processing failures. Such cases may take place when, for example, the transformation of raw audio into a scalogram takes place locally (at the edge) and the images (scalograms) are subsequently transferred to the classifier (through a network).

This approach is aligned with relevant research including [38] which demonstrates the enhancement of data diversity through augmentation, and [39] where image augmentation contributes to overfitting prevention.

Figure 2 indicates the result of image-based enhancements in the scalogram of a 'baby crying' sound clip. The scalogram in its original form is depicted on the left, while the result of each image-based transformation is presented on the right (with the corresponding labeling). The reflection is applied on both the x- and the y- axes. The range [−45 45] specifies the rotation angle in degrees applied to the image. The image is rotated around its center. Scaling was conducted using factors from the interval [0.8 1.2], effectively multiplying each pixel coordinates by the corresponding factor. The range [−30 30] in shear refers to the angle in degrees which is applied on the image. This technique, contrary to rotation, slants the shape of the image along an axis (horizontal or vertical, or both). The range [−50 50] indicates the horizontal and/or vertical distance the image is translated. For each repetition of the augmentation, the value of each parameter is selected randomly from a continuous uniform distribution within the corresponding interval.

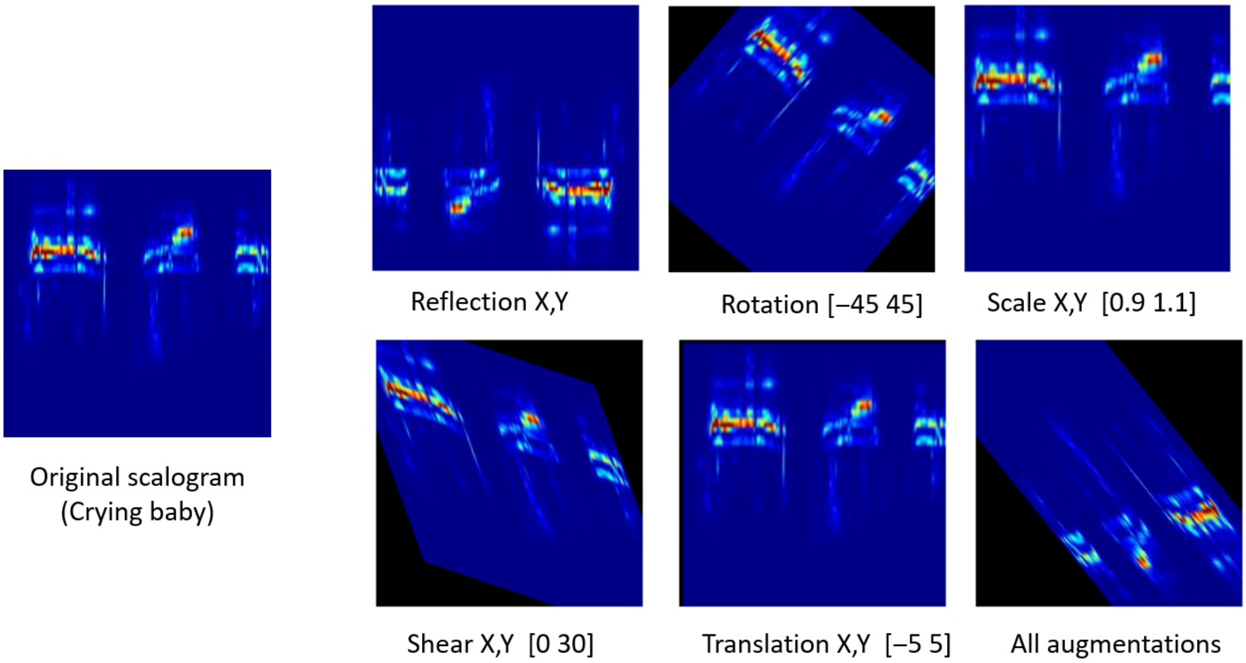

Figure 2. The effect of each image augmentation technique and the application of all techniques simultaneously.

3.3. Synthetic Audio Generation

The generation of synthetic sounds has been based on the AudioGen auto-regressive deep learning model. The medium version of Audiogen has been employed (1.5 billion parameters) operating at 16 kHz and using the EnCodec tokenizer with four codebooks. A python script has been prepared to load the AudioGen model from Audiocraft library and feed the customized prompts for each of the sound categories, also providing relevant parameters (clip duration).

Prompts were created for each of the three categories (human and animal sounds and sounds of things), included in JSON (JavaScript Object Notation files). Specifically, diverse prompts have been included to maximize acoustic variability based on context. For each sound class, 24 prompts generate content—specific sounds (e.g., "a crackling fire," "children playing in a park"). Human-in-the-loop has been involved in quality control, verifying the generated sounds. Indicative prompt examples are shown in Table 4.

Table 4. Examples of audio generation prompts with and without background context.

Prompts (Including Background)	Prompts (Without Background)
Dog	Dog
"A puppy whining softly in its crate."	"A dog growling protectively."
"Excited barking during a game of fetch."	"A dog panting after exercise."
"A dog growling protectively at night."	"A small dog barking at a cat."
Baby crying	Baby crying
"A newborn baby softly crying in a hospital room."	"A baby crying loudly."
"A baby crying loudly in a small apartment."	"A baby crying intermittently."
"A baby crying while being rocked in a cradle."	"A baby crying continuously."
Water pouring	Water pouring
"Pouring water into a glass in a quiet kitchen.",	"Pouring water into a kettle."
"Water being poured from a kettle into a teacup."	"Pouring water into a jug."
"Pouring water into a bathtub with echoes in the bathroom."	"Pouring water into a sink."

All audio clips have A fixed duration of 5 s, 16 kHz sampling frequency, and are exported in wav format. Each of the synthetically generated sounds has been inspected (heard) by the human operator for verification. The generated and the original training files enhanced the training set.

AudioGen has produced acoustically plausible, labeled samples to expand training corpora. In the pipeline synthetic samples have been manually validated, involving human-in-the-loop (HITL). The scope has been to confirm the semantic match between the sound output and its label. In terms of fidelity, no sample of the produced set has been excluded, while no unacceptable artifacts (unnatural timbre, looping) have been detected in the synthetic sound clips. Differentiations have been observed in the clarity of the produced sounds mainly due to the existence and intensity of background sounds. Prompt engineering allowed sound clarity, diversity (through the characterization of the sound itself such as complaining, raring, intense, long-drawn-out sounds), and regulation of the intensity of background sounds. Synthetic file generation expands the dataset's diversity, introducing plausible sound events, while maintaining perceptual realism.

3.4. Methodology

Classification accuracy has been evaluated starting from the baseline synthesis (i.e., 20 clips per class) of the three datasets (human, animal and things sounds) and re-evaluated after enhancing the training sets with additional samples applying augmentation or synthetic generation techniques. Spectrograms and scalograms of the original files were generated. Datasets were split into 60% for training, 20% for validation, and 20% for testing. The augmentation and synthetic sound generation techniques were applied uniformly across classes, to avoid biasing the classifier towards synthetic-heavy classes and skewing decision boundaries. Mixing real and synthetic sounds allowed the increase in training set population by 50%, 100%, and 200% (i.e., with 6, 12, and 24 additional audio clips).

Time-frequency analysis was performed using log-mel spectrograms and scalograms. Log-mel spectrograms were computed via MATLAB's 2025a vggishPreprocess with audio resampled to 16 kHz, a 25 ms Hann window, 10 ms hop length, 512-point FFT, and 64 mel bins. Scalograms were generated using the Continuous Wavelet Transform (CWT) with the Morse wavelet. These settings were chosen to provide an optimal balance between time and frequency resolution.

The sound augmentation techniques were applied upon the sound (wav) and spectrograms (image) files using MATLAB 2025a version. Each sound clip coming from the ESC-50 dataset had a duration of 5 s and sampling rate of 16 kHz. The image-based transformations applied upon the scalograms, image of resolution 224 × 224 pixels, 96 dpi were applied. The augmented and synthetic data were used exclusively in the training process.

For the classification, sound-oriented convolutional neural networks (CNNs) and generic image classifiers (CNNs) were employed. Two networks per category were chosen for robustness, matched in parameter count and computational capacity. VGGish (audio specific version of VGG) and YAMNet (MobileV1-based audio model), both pretrained on AudioSet, were selected as they can support classification tasks upon arbitrary sounds. ResNet50 was selected as an image classifier due to its balanced approach between classification accuracy and computational cost [31], as well as DarkNet53 which has a comparable number of layers.

Quantitative architecture and computational characteristics of the selected CNNs are given in Table 5.

Table 5. Computational requirements of the four selected CNN architectures. Depth refers to the number of layers; parameters are expressed in millions; memory footprint corresponds to model size in MB; and computational complexity is measured in giga–floating point operations (GFLOPs).

CNN	Depth	Millions of Parameters	Memory Size (MB)	FLOPs (GFLOPs)
VGGish	10	62	237	15.5
YAMNet	28	3.2	12	0.57
ResNet50	50	25.6	98	4.1
DarkNet53	53	41.6	159	7.3

VGGish and YAMNet, the sound-oriented models, use the log-mel spectrograms of the initial and augmented sounds for fine-tuning. The scalograms are used to fine-tune the image classifiers, i.e., the ResNet50 and DarkNet53. The fine tuning (transfer learning) initializes the CNNs with pretrained weights. Typically, the fully connected, softmax and classification layers are randomly initialized. As a result, each training run may lead to slightly different classification accuracy. To ensure statistical reliability each experiment was repeated 10 times and the results presented refer to the mean classification accuracy. Testing involved only original (not augmented or synthetically generated) sound clips.

The workflow is depicted in Figure 3. Parallel paths in the flow indicate the possibility of different practical scenarios. For example, the usage of the wav file or the visual representation (spectrogram or scalogram), the employment of image- or sound-based CNNs, the classification of the sound at the edge or their local transformation into image, transfer to cloud and subsequent classification.

The approach led to 252 scenarios, corresponding to the combination of 7 augmentation methods, 3 augmentation levels (50%, 100%, and 200%), 4 CNN architectures, and 3 datasets:

$$\begin{aligned}&(7\text{ augmentation methods}) \times (3\text{ augmentation levels}) \times (4\text{ CNNs}) \times (3\text{ datasets})\\&= 252\text{ scenarios}\end{aligned}$$

The results, in terms of classification accuracy, are presented in Section 4.

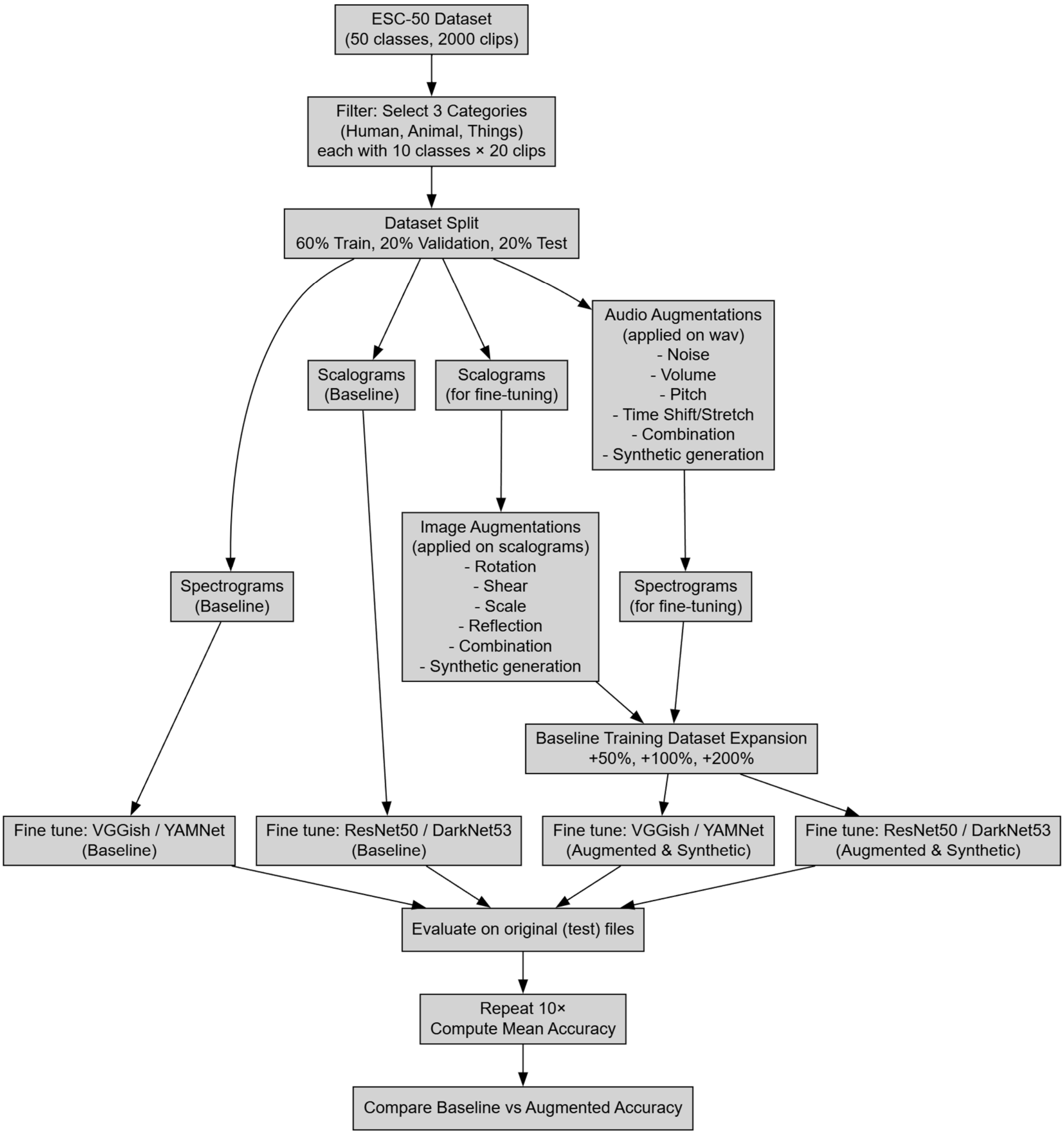

Figure 3. Workflow for evaluating the impact of each augmentation scenario on classification accuracy.

4. Results

4.1. Sound Augmentation Techniques

The training set was expanded by 6, 12, and 24 additional files (population increase of 50%, 100%, and 200%, respectively) generated with audio augmentation techniques. The baseline refers to the classification accuracy achieved with the original audio files (“No augmentation”). Table A1 presents the classification accuracy for VGGish and YAMNet models per augmentation technique and generative synthesis.

The baseline results indicate that VGGish outperforms YAMNet, especially for sounds of things and human sounds. Across all augmentation types, performance improvements were generally observed, particularly for VGGish, with diminishing returns or even accuracy degradation at higher augmentation levels. Synthetic sound generation consistently led to the largest classification accuracy, especially for VGGish where accuracy reached 90.00%. Sound augmentation methods such as adding noise, controlling volume, and pitch shifting contributed to improvements as well. Similarly, time-based augmentations (i.e., time shift and time stretch) demonstrated gains up to 50 or 100% expansion of the training set. For human noise sounds, addition was the most effective technique (where VGGish reached 86.50%), for animal sounds pitch shift allowed VGGish to reach 84.50% and for sounds of things noise addition again set VGGish accuracy to 89.50%. The combination of all augmentations led to the highest VGGish performance, demonstrating the cumulative effect of augmentation diversity. A consistent observation is that the accuracy increase is maximized with an extension of the training set up to 50 or 100%, while further extension (i.e., to 200%) causes accuracy decrease or degradation.

4.2. Scalogram Augmentation Techniques

The classification accuracy achieved by ResNet50 and DarkNet53 models after applying geometric transformations including reflection, rotation, scaling (horizontally and/or vertically), shearing and translation, and synthetic generation to extend the training set, as depicted in Table A2.

The baseline results indicate strong performance from ResNet50 (84.0% for human sounds, 71.0% for animals, 86.0% for sounds of things), outperforming DarkNet53 in most cases. DarkNet53 performed best on sounds of things (82.5%). In terms of augmentation techniques, moderate (i.e., up to 50% or 100%) expansion of the training set improved classification accuracy for all sound categories. Accuracy improvements have been similar for image augmentation and synthetic generation methods. Reflection offered modest gains across both models. Rotation leads to consistent improvements, especially for ResNet50, supporting models to generalize better to varied temporal or spectral orientations. Scaling also positively impacted classification. Shearing distortions notably improved performance on animal sounds, with ResNet50 reaching 77.5% and DarkNet53 peaking at 84.5%. Spatial image shifting (time-frequency displacement) benefited both models, especially DarkNet53. The combination of the augmentations achieved high results. Synthetic scalograms derived from AudioGen yielded high performance across all categories, especially for ResNet50, achieving 90.0% (things) and 87.5% (human). The observation made on sound augmentation results is also valid for image augmentation, i.e., accuracy is increased with an extension of the training dataset up to 50% or 100%, while further extension leads to accuracy drop.

YAMNet is the most lightweight (approximately 3.2 M of parameters, and 12 MB memory required) making it suitable for real-time applications (Table 3); however, its classification accuracy is lower compared to VGGish (approximately 62 M of parameters, and 237 MB memory required), as shown in Table 3. ResNet50 and DarkNet53 offer a balanced trade-off between accuracy and computational cost (approximately 25–41 M or parameters, and 98–159 MB memory). The experimental results presented in Tables A1 and A2 indicate that the performance (in terms of classification accuracy) of the selected CNNs is consistent with their complexity. The highest classification accuracy was demonstrated by VGGish, which is also more strongly affected (in a positive way) by the augmentation techniques. ResNet50 follows achieving high classification baseline with lower increase than VGGish. DarkNet53 and YAMNet achieve lower classification accuracy.

5. Discussion

5.1. Sound Augmentation

According to the results presented in Section 4, the baseline ranges from 62.50% (YAMNet for human sounds) to 76.50% (VGGish for sounds of things). Classification accuracy achieved with the VGGish model are superior to those of the YAMNet network performing better on sounds of things, followed by human sounds and animal sounds.

The highest increase in classification accuracy is observed in human sounds applying add noise and time shift augmentation techniques. For animal sounds this is attained using shift pitch and time stretch, whereas for sounds of things (object-oriented) with add noise and time stretch. The effectiveness of each augmentation differs per category. For example, pitch shift performs poorly for human sounds but proves more effective for animal sounds. Add noise enhances classification for human and object-related sounds but has limited impact on animal sound classification. Shift pitch is particularly effective for animal sounds possibly due to inherent physiological differences in their vocal expressions. Noise addition demonstrates the most beneficial impact on classification accuracy on average. Other effective techniques include control volume and time shift. Applying simultaneously all augmentation techniques yields accuracy superior to the individual techniques.

The classification accuracy achieved using synthetic audio files surpasses all results obtained with augmentation techniques. The maximum classification accuracy reached with each technique, and all techniques applied in parallel are depicted in Figure 4.

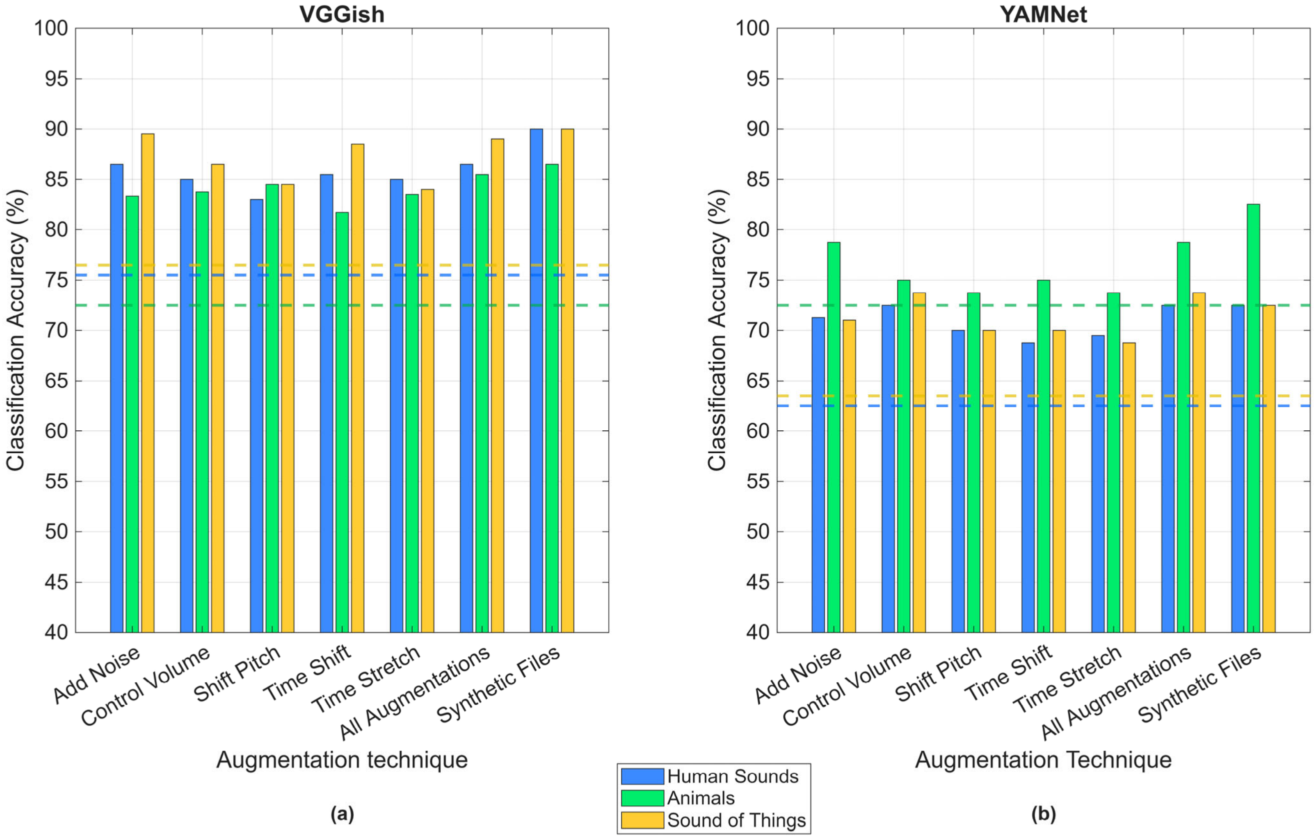

Figure 4. Maximum classification accuracy (%) per augmentation technique and synthetic generation using (**a**) the VGGish and (**b**) the YAMNet networks. Horizontal dashed lines indicate baseline accuracy without augmentation.

Classification accuracy improves, in comparison with the baseline, when the training set is increased by 50%, and even more when it is increased by 100%; however, a 200% increase does not lead to further improvement, and in some cases, results in performance degradation (i.e., in the control volume technique) for the VGGish network. In the case of the YAMNet architecture, the highest classification performance is achieved with an enrichment of the training set by 50%, while furhter enhancement leads to lower classification accuracy.

Figure 5 depicts the classification accuracy when all augmentation techniques are applied simultaneously with an augmentation of the datasets by 50, 100, and 200%, for both VGGish and YAMNet architectures.

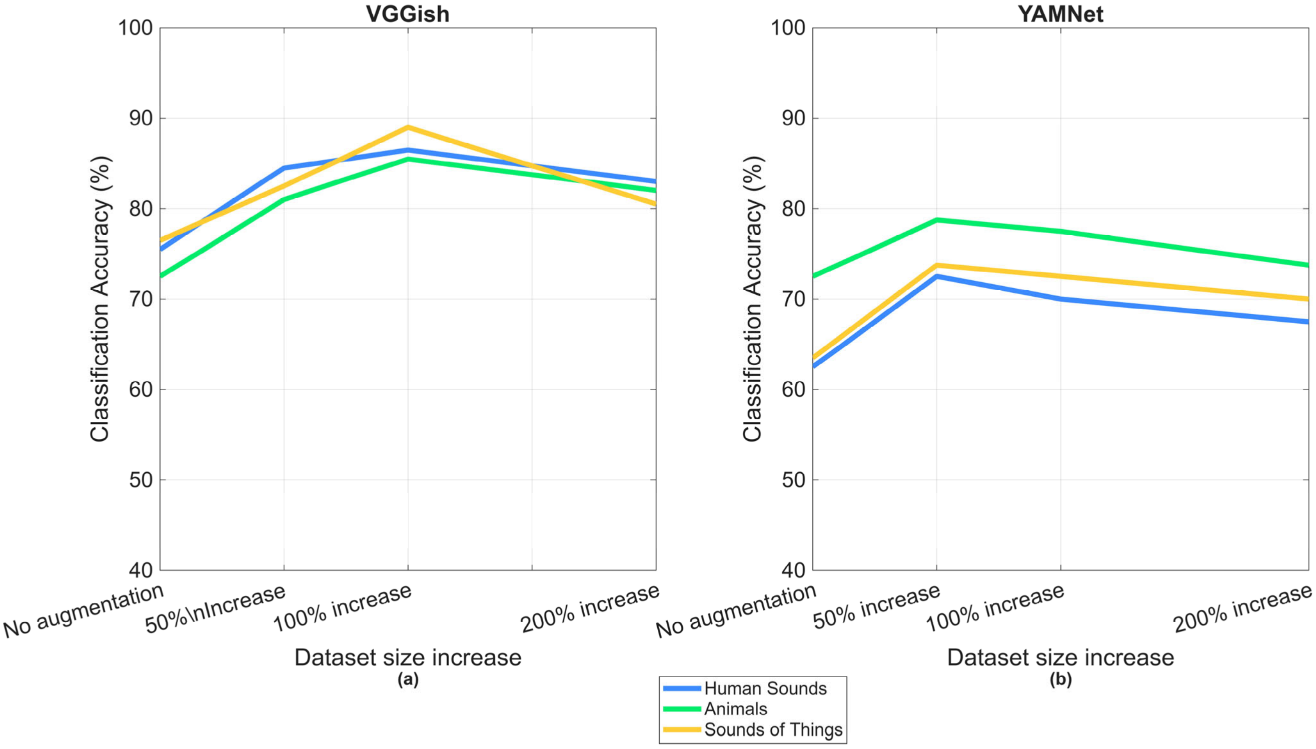

Figure 5. The classification accuracy (%) achieved when the training set is increased by 50%, 100%, and 200% applying all augmentation techniques simultaneously for (**a**) the VGGish and (**b**) for the YAMNet architectures.

While there is a positive correlation between sound dataset extension with synthetic data and improvements in classification accuracy, this correlation is not linear, and it is positive up to a threshold. In any attempt to explain this observation, we consider that variation in synthetic sounds is modest with mild changes in the prompt, i.e., similar prompts (i.e., having small semantic differences) are expected to correspond to vicinal semantic embeddings. These lead to similar output token sequences and acoustically similar outputs, which do not substantially enhance the dataset. Dataset population with synthetic audio beyond a threshold may make the model memorize synthetic patterns (not present in real data) and lead to overfitting causing reduced classification accuracy.

To further investigate this, the differences between an original sound file (of the 'baby crying' class) and the augmented and synthetic versions are quantified through the SNR (Signal to Noise Ratio), MSE (Mean Squared Error), spectral centroid, and MFCC (Mel-frequency cepstral coefficients) as depicted in Figure 6.

The Signal-to-Noise Ratio (SNR) metric indicates that time shift causes the greatest divergence from the original waveform, while add noise results in smaller deviation. In contrast, control volume, shift pitch, and time stretch produce minimal distortion. The Mean Square Error (MSE) metric which directly measures the numerical waveform differences indicates the same observations as the SNR. The spectral centroid, a metric related to the brightness of the sound, is affected by adding noise, while pitch and volume modifications preserve the spectral distribution. The MFCC distance, a metric of perceptual similarity, indicates perceptual dissimilarity in both the synthetically generated file and the file modified with add noise [40].

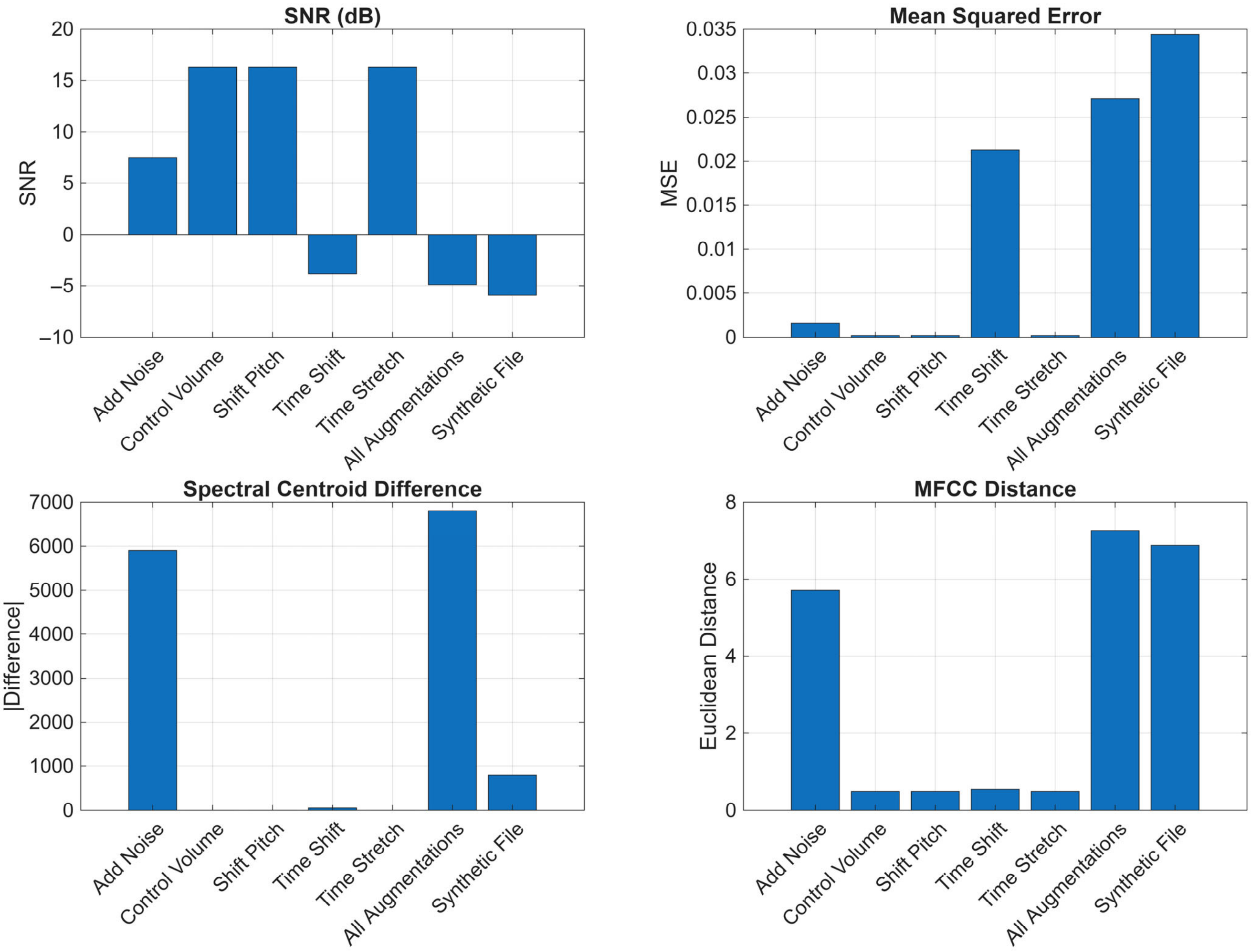

Figure 6. Difference in terms of SNR, MSE, spectral centroid, and MFCC between the original sound file and each of the augmented and two synthetically generated versions.

5.2. Image (Scalogram) Augmentation

The baseline accuracy for ResNet50 and DarkNet53 surpasses that of the sound-based networks (VGGish and YAMNet). As briefly introduced in Section 4, in terms of the effectiveness of the augmentation techniques, shear yields the most significant performance improvement in accuracy, while reflection appears to be the least effective. The simultaneous application of all augmentation techniques results in classification accuracy that reaches the maximum value achieved by the individual methods.

For the ResNet50 network, the highest classification accuracy is observed when the training set is augmented by 50%. Further increase in the number of scalograms leads to reduced performance, potentially even falling below the baseline accuracy. In contrast, the DarkNet53 network has a gradual improvement in classification performance with 50% and 100% augmentation, while excessive augmentation (200%) results in the degradation of accuracy. In both networks, the use of synthetic sounds—converted into scalograms—enhances classification performance across all three sound categories. Specifically, in the sound of things dataset, ResNet50 achieves peak classification accuracy of 90%. For the human sounds and animals categories, the accuracy obtained through synthetic data augmentation is slightly lower than that achieved with the best-performing individual techniques.

Figure 7 presents the maximum classification accuracy achieved with each augmentation technique, applying all techniques in parallel and using synthetic audio files to enhance the training set.

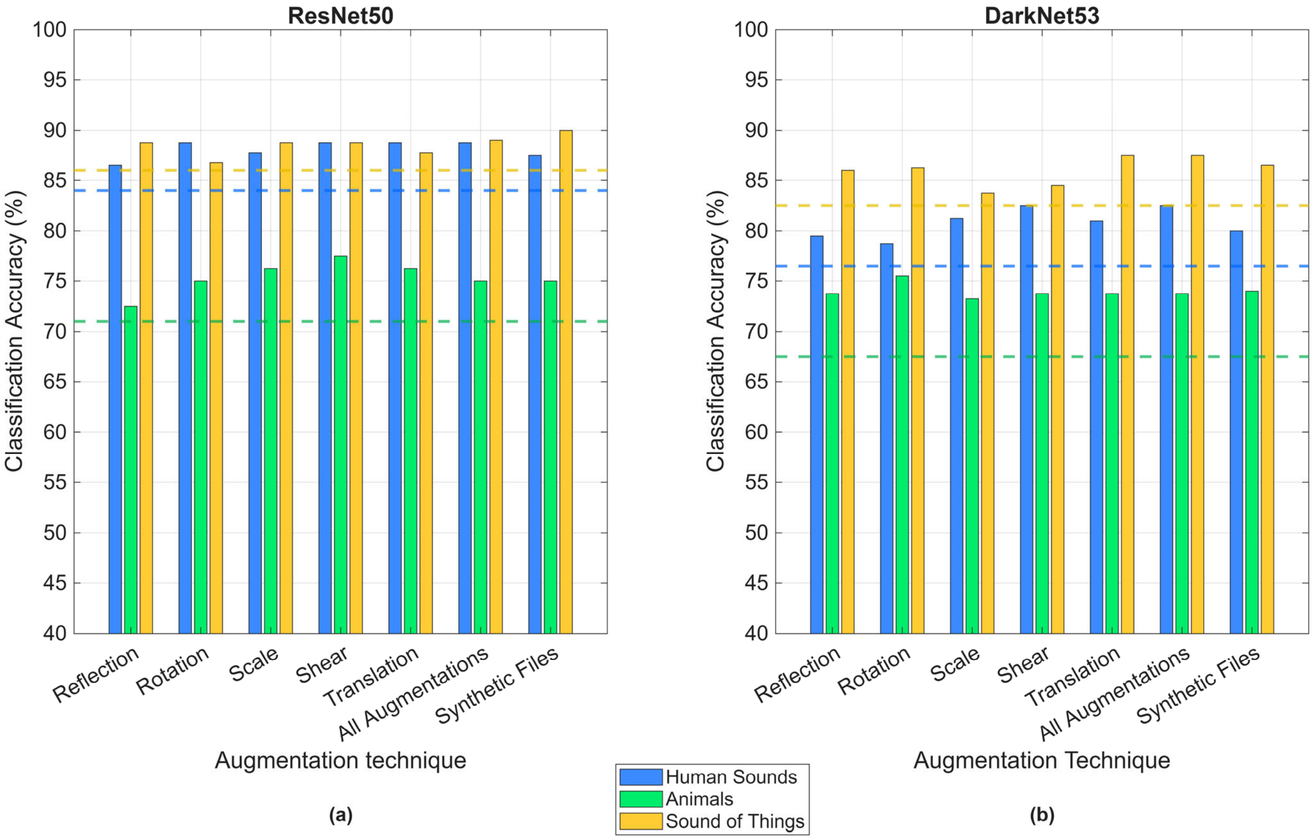

Figure 7. Classification accuracy (%) per augmentation technique and synthetic generation using (**a**) the ResNet50 and (**b**) the DarkNet53 networks. Horizontal dashed lines indicate baseline accuracy without augmentation.

Figure 8 depicts the classification accuracy when all augmentation techniques are applied simultaneously with an augmentation of the datasets by 50, 100, and 200%, for both the ResNet50 and DarkNet53 architectures.

To investigate the differences between the scalograms of the augmented and synthetically generated sounds and that of the original one (for the ‘baby crying’ class), three image similarity metrics have been quantified, i.e., the Structural Similarity Index Measure (SSIM), MSE, and Peak Signal-to-Noise Ratio (PSNR). SSIM is the more informative metric as it captures the structure and texture, related to formant structure, temporal energy, and harmonics and it is associated with perceived similarity of sounds. MSE and PSNR, designed for signal fidelity but not perception, identify tiny shifts. These do not affect audio meaning (e.g., time or phase shifts) nor do they reflect frequency emphasis or distribution. In this view, as depicted in Figure 9, the difference among scalograms is maximized when rotation is applied and all augmentation types are applied, while the difference between the original and the synthetically generated is limited.

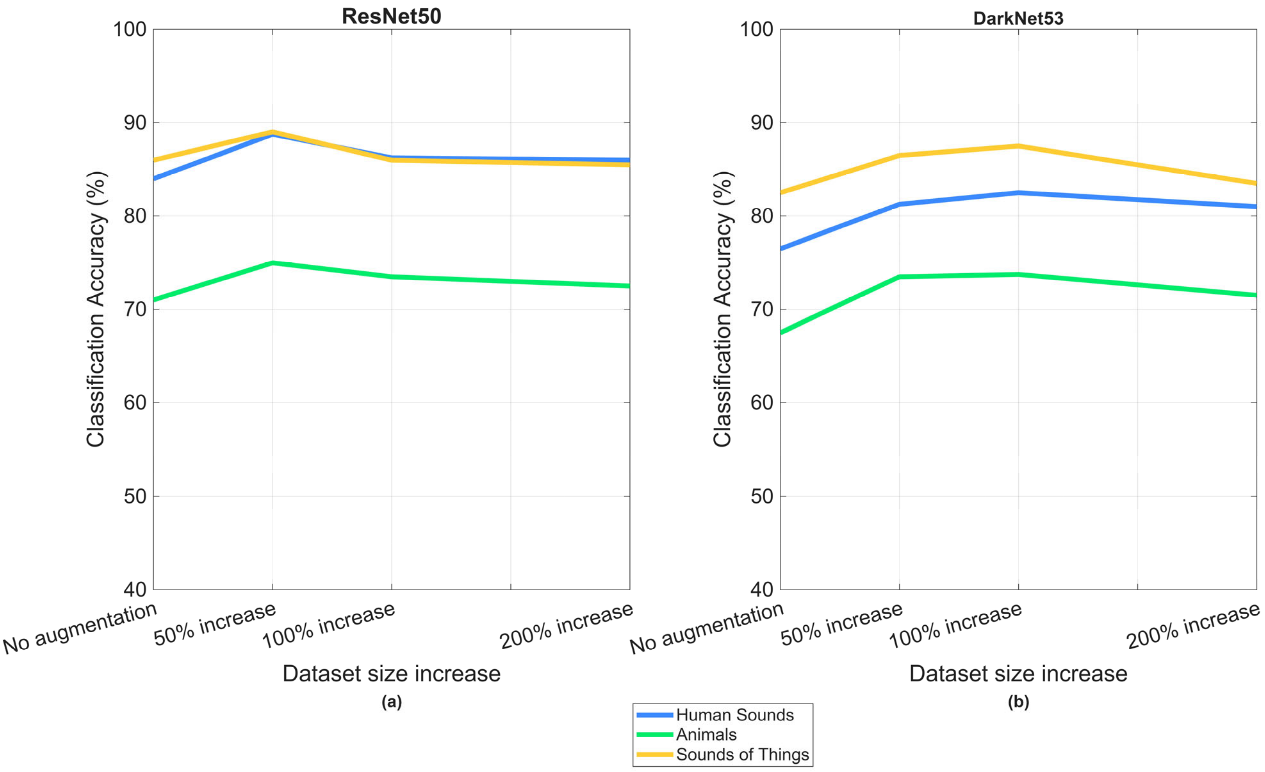

Figure 8. The classification accuracy (%) with a training set increase of 50%, 100%, and 200% applying all augmentation techniques simultaneously for (**a**) the ResNet50 and (**b**) for the DarkNet53 architectures.

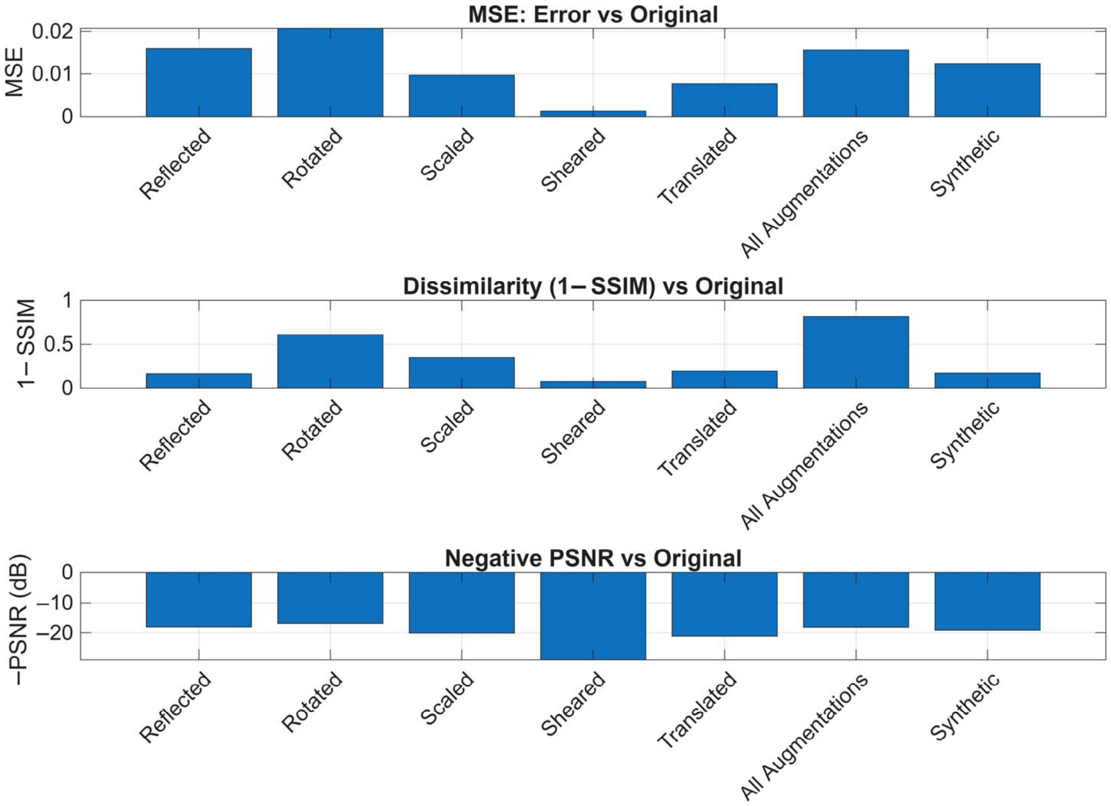

Figure 9. Comparison of scalograms of the original, augmented, and synthetic sounds using the MSE, the SSIM, and PSNR metrics.

Figure 10 investigates the average contribution of the augmentation techniques to the classification accuracy achieved per each CNN. The more significant increase is observed with VGGish, showing an improvement of approximately 9% across all three sound categories. YAMNet benefits the most from augmentation techniques particularly in the case of human sounds (6.6%), and the least for animal sounds (2.9%). While ResNet50 exhibits the highest accuracy among all models, its average accuracy improvement remains around 2%. Similarly, DarkNet53 demonstrates an average increase of approximately 3.4%.

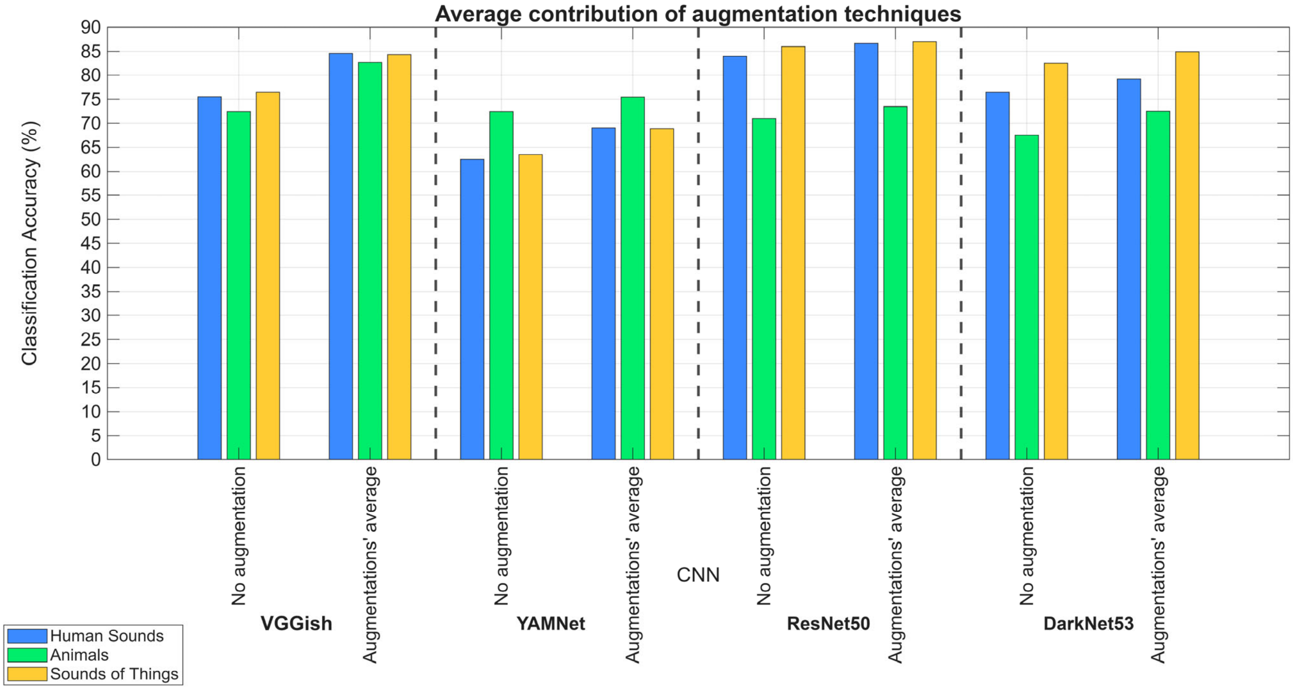

Figure 10. The classification accuracy (%) of each dataset by each CNN without any augmentation and the average classification accuracy after the application of augmentation techniques.

6. Conclusions

Sound datasets may include clear, discrete, laboratory, sounds which lack variety in acoustics, context and speaker differences, or include overlapping, real-life sounds which may be less appropriate for training. We have investigated the impact of sound augmentation and synthetic generation techniques upon classification accuracy using three sound categories, i.e., sounds of humans, animals, and things, created as subsets of the ESC-50 dataset. Augmentation has been performed upon raw sound (wav files) and scalograms (images). Sound augmentation techniques include time stretch, shift pitch, add noise, control volume, and time shift, while image augmentation involves reflection, rotation, scale, shear, and translation. Synthetic generation has been based on the AudioGen neural network, with custom prompts. The classification has been performed with pretrained sound deep learning networks VGGish and YAMNet, using audio spectrograms, and the image classifiers ResNet50 and DarkNet53 using the scalograms.

Sound augmentation and synthetic generation lead to classification accuracy improvement, particularly for underrepresented sound classes and classes with high intra-class variability supporting generalization to real-world environments. Specifically, dataset expansion up to a threshold of doubling the initial dataset has positively contributed to classification accuracy, with an average of approximately 7% for VGGish and YAMNet, and of 2.7% for ResNet50 and DarkNet53. For VGGish the maximum accuracy boost has been 14.5% in the human sounds dataset through synthetic data augmentation and YAMNet yielded 10% improvement under a 50% increase in the training dataset. The maximum increase with ResNet50 was 4.75% on the human sound dataset, while for DarkNet53 it was 6.25% on the animal sounds dataset.

A key observation has been that accuracy boost is possible with the training set expansion up to 100%, whereas further expansion (i.e., 200%) causes accuracy to decrease. Overly expanding the dataset leads to artifact overfitting, as the classifier focuses on synthetic samples, overwhelming the real signal. In addition, as data augmentation acts as a normalized mechanism protecting from real-world varieties, excessive normalization may focus on coarse-grained patterns and prevent the model from spotting subtle but substantial differences.

Classification accuracy exceeds 80%, reaching in case 90% depending on the sound categories and classifiers. The performance is associated with the complexity of the model, as VGGish achieves the highest results. The neural networks classifiers involve computational cost to be considered for deployments at edge devices, in comparison with more lightweight approaches of traditional machine learning techniques.

Future work involves considering larger datasets more diverse than ESC-50, as well as focusing on realistic conditions with noisy and overlapping sounds. Further investigation of overfitting can lead to the fine-grained quantification of reliance on synthetic patterns and model generalization thresholds.

Author Contributions: Conceptualization E.T. and A.P.; methodology, E.T., A.P. and G.P.; software, E.T.; validation, G.P. and M.S.; formal analysis, A.P.; data curation, E.T.; writing—original draft preparation, A.P. and E.T.; writing—review and editing, M.S.; supervision, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Supporting data (e.g. the full set of prompts) are available upon request. The sound datasets including the original, the augmented, and synthetically generated files are available upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AST	Audio Spectrogram Transformer
CNN(s)	Convolutional Neural Network(s)
CWT	Continuous Wavelet Transformation
DCASE	Detection and Classification of Acoustic Scenes and Events
GAN(s)	Generative Adversarial Network(s)
HITL	Human-In-The-Loop
LLM(s)	Large Language Models
LPCNet	Linear Prediction Coding Network
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
MIR	Music Information Retrieval
ML	Machine Learning
MSE	Mean Square Error
(P)SNR	(Peak) Signal-to-Noise Ratio
RNN	Recurrent Neural Network
RVQ	Residual Vector Quantization

SER	Sound Event Recognition
SSIM	Structural Similarity Index Measure
STFT	Short-Time Fourier Transform
TSM	Time-Scale Modification
(VQ-)VAE	(Vector Quantized-) Variational AutoEncoder

Appendix A

Table A1. The classification accuracy (%) for each dataset, achieved by VGGish and YAMNet, enhancing the training set by 50%, 100%, and 200%, with augmentation techniques and generative synthesis. For each dataset the metrics in left sub-column refer to VGGish and those of the right sub-column to YAMNet.

Sound Augmentations	**Human Sounds**		**Animals**		**Sound of Things**	
Baseline (with No Augmented Files)	**VGGish** **75.50**	**YAMNet** **62.50**	**VGGish** **72.50**	**YAMNet** **72.50**	**VGGish** **76.50**	**YAMNet** **63.50**
Add Noise						
50% increase	81.50	71.25	82.73	78.75	85.00	71.00
100% increase	86.50	67.50	83.33	75.00	89.50	68.75
200% increase	83.50	67.50	80.70	75.00	81.60	68.55
Control Volume						
50% increase	84.50	72.50	83.75	75.00	82.50	73.75
100% increase	85.00	71.25	83.75	73.75	86.50	67.50
200% increase	84.00	65.00	78.75	73.00	83.50	66.75
Shift Pitch						
50% increase	82.50	70.00	80.00	73.75	82.00	70.00
100% increase	83.00	68.75	84.50	73.75	84.50	70.00
200% increase	82.00	67.50	81.70	73.50	80.50	66.25
Time Shift						
50% increase	84.00	68.75	81.70	75.00	82.00	70.00
100% increase	85.50	67.25	81.70	73.75	88.50	63.75
200% increase	84.00	65.00	80.80	73.75	84.50	65.75
Time Stretch						
50% increase	83.50	69.50	81.70	73.75	81.50	68.75
100% increase	85.00	67.75	83.50	73.50	84.00	63.75
200% increase	81.50	67.50	82.00	73.00	81.50	62.50
All augmentations						
50% increase	84.50	72.50	81.00	78.75	82.50	73.75
100% increase	86.50	70.00	85.50	77.50	89.00	72.50
200% increase	83.00	67.50	82.00	73.75	80.50	70.00
Synthetic Sounds						
50% increase	88.33	72.50	85.50	82.50	85.50	72.50
100% increase	90.00	70.00	86.50	80.00	90.00	71.00
200% increase	88.33	67.50	85.00	77.50	86.50	70.00

Table A2. The classification accuracy (%) by ResNet50 and DarkNet53, enhancing the training set by 50%, 100%, and 200%, with image-based augmentation techniques and generative synthesis.

Image Augmentations	**Human Sounds**		**Animals**		**Sound of Things**	
Baseline (with No	**ResNet50**	**DarkNet53**	**ResNet50**	**DarkNet53**	**ResNet50**	**DarkNet53**
Augmented Files)	**84.00**	**76.50**	**71.00**	**67.50**	**86.00**	**82.50**
Reflection [X, Y]						
50% increase	86.50	77.00	72.50	72.50	88.75	85.00
100% increase	85.75	79.50	71.25	73.75	86.50	86.00
200% increase	83.50	75.25	68.50	71.25	85.50	85.50
Rotation [−45 45]						
50% increase	88.75	77.50	75.00	73.50	86.75	84.00
100% increase	87.50	78.75	73.75	75.50	85.50	86.25
200% increase	86.25	77.00	71.50	71.00	84.00	85.50
Scale [0.8 1.2]						
50% increase	87.75	77.50	76.25	75.00	88.75	83.50
100% increase	86.50	81.25	73.75	73.25	86.25	83.75
200% increase	86.25	77.50	72.50	72.50	85.75	84.00
Shear [−30 30]						
50% increase	88.75	77.75	77.50	70.00	88.75	83.75
100% increase	87.50	82.50	76.25	73.75	87.50	84.50
200% increase	86.25	81.00	73.75	72.00	86.00	83.75
Translation [−50 50]						
50% increase	88.75	78.75	76.25	72.50	87.75	83.50
100% increase	86.25	81.00	75.00	73.75	87.00	87.50
200% increase	86.00	78.75	72.50	71.25	86.25	86.25
All augmentations						
50% increase	88.75	81.25	75.00	73.50	89.00	86.50
100% increase	86.25	82.50	73.50	73.75	86.00	87.50
200% increase	86.00	81.00	72.50	71.50	85.50	83.50
Synthetic Sounds						
50% increase	87.50	78.50	75.00	72.50	90.00	85.50
100% increase	86.25	80.00	73.50	74.00	89.50	86.50
200% increase	81.00	78.50	68.00	67.50	85.75	81.50

References

1. Abayomi-Alli, O.O.; Damaševičius, R.; Qazi, A.; Adedoyin-Olowe, M.; Misra, S. Data augmentation and deep learning methods in sound classification: A systematic review. *Electronics* **2022**, *11*, 3795. [CrossRef]
2. Donahue, C.; McAuley, J.; Puckette, M. Adversarial Audio Synthesis. *arXiv* **2018**, arXiv:1802.04208.
3. Mehri, S.; Kumar, K.; Gulrajani, I.; Kumar, R.; Jain, S.; Sotelo, J.; Courville, A.; Bengio, Y. SampleRNN: An Unconditional End-to-End Neural Audio Generation Model. *arXiv* **2016**, arXiv:1612.07837.
4. Wang, H.; Zou, Y.; Wang, W. SpecAugment++: A Hidden Space Data Augmentation Method for Acoustic Scene Classification. *arXiv* **2021**, arXiv:2103.16858.
5. Sarris, A.L.; Vryzas, N.; Vrysis, L.; Dimoulas, C. Investigation of Data Augmentation Techniques in Environmental Sound Recognition. *Electronics* **2024**, *13*, 4719. [CrossRef]
6. Nanni, L.; Maguolo, G.; Paci, M. Data Augmentation Approaches for Improving Animal Audio Classification. *Ecol. Inform.* **2020**, *57*, 101084. [CrossRef]
7. Ramires, A.; Serra, X. Data Augmentation for Instrument Classification Robust to Audio Effects. *arXiv* **2019**, arXiv:1907.08520. [CrossRef]
8. Bian, W.; Wang, J.; Zhuang, B.; Yang, J.; Wang, S.; Xiao, J. Audio-Based Music Classification with DenseNet and Data Augmentation. In Proceedings of the Pacific Rim International Conference on Artificial Intelligence (PRICAI 2019), Cham, Switzerland, 26–30 August 2019; Springer: Cham, Switzerland, 2019; pp. 56–65. [CrossRef]
9. Salamon, J.; Bello, J.P. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Process. Lett.* **2017**, *24*, 279–283. [CrossRef]

10. Davis, N.; Suresh, K. Environmental sound classification using deep convolutional neural networks and data augmentation. In Proceedings of the 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS), Thiruvananthapuram, India, 6–8 December 2018; IEEE: Piscataway, NJ, USA; pp. 41–45. [CrossRef]
11. Mushtaq, Z.; Su, S.F. Environmental sound classification using a regularized deep convolutional neural network with data augmentation. *Appl. Acoust.* **2020**, *167*, 107389. [CrossRef]
12. Lu, Y.; Shen, M.; Wang, H.; Wang, X.; van Rechem, C.; Fu, T.; Wei, W. Machine Learning for Synthetic Data Generation: A Review. *arXiv* **2023**, arXiv:2302.04062. [CrossRef]
13. Goyal, M.; Mahmoud, Q.H. A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics* **2024**, *13*, 3509. [CrossRef]
14. Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; van den Oord, A.; Dieleman, S.; Kavukcuoglu, K. Efficient Neural Audio Synthesis. *arXiv* **2019**. Available online: https://arxiv.org/abs/1802.08435 (accessed on 21 May 2025).
15. Valin, J.M.; Skoglund, J. LPCNet: Improving Neural Speech Synthesis through Linear Prediction. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 5891–5895. [CrossRef]
16. Liu, X.; Singh, S.; Cornelius, C.; Busho, C.; Tan, M.; Paul, A.; Martin, J. Synthetic Dataset Generation for Adversarial Machine Learning Research. *arXiv* **2022**, arXiv:2207.10719. [CrossRef]
17. Van Den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. *arXiv* **2016**, arXiv:1609.03499. [CrossRef]
18. Dhariwal, P.; Jun, H.; Payne, C.; Kim, J.W.; Radford, A.; Sutskever, I. Jukebox: A Generative Model for Music. *arXiv* **2020**, arXiv:2005.00341. [CrossRef]
19. Kreuk, F.; Synnaeve, G.; Polyak, A.; Singer, U.; Défossez, A.; Copet, J.; Parikh, D.; Taigman, Y.; Adi, Y. Textually Guided Audio Generation. *arXiv* **2022**, arXiv:2209.15352.
20. Lhoest, L.; Lamrini, M.; Vandendriessche, J.; Wouters, N.; da Silva, B.; Chkouri, M.Y.; Touhafi, A. Mosaic: A Classical Machine Learning Multi-Classifier Based Approach Against Deep Learning Classifiers for Embedded Sound Classification. *Appl. Sci.* **2021**, *11*, 8394. [CrossRef]
21. Tsalera, E.; Papadakis, A.; Samarakou, M.; Voyiatzis, I. CNN-Based Segmentation and Classification of Sound Streams under Realistic Conditions. In Proceedings of the 26th Pan-Hellenic Conference on Informatics (PCI 2022), Athens, Greece, 25–27 November 2022; pp. 373–378. [CrossRef]
22. Abdul, Z.K.; Al-Talabani, A.K. Mel Frequency Cepstral Coefficient and Its Applications: A Review. *IEEE Access* **2022**, *10*, 122136–122158. [CrossRef]
23. Ren, Z.; Qian, K.; Zhang, Z.; Pandit, V.; Baird, A.; Schuller, B. Deep Scalogram Representations for Acoustic Scene Classification. *IEEE/CAA J. Autom. Sin.* **2018**, *5*, 662–669. [CrossRef]
24. Phan, D.T.; Jakob, A.; Purat, M. Comparison Performance of Spectrogram and Scalogram as Input of Acoustic Recognition Task. *arXiv* **2024**, arXiv:2403.03611. [CrossRef]
25. Chi, P.H.; Chung, P.H.; Wu, T.H.; Hsieh, C.C.; Chen, Y.H.; Li, S.W.; Lee, H.Y. Audio ALBERT: A Lite BERT for Self-Supervised Learning of Audio Representation. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT 2021), Shenzhen, China, 19–22 January 2021; pp. 344–350. [CrossRef]
26. Tsalera, E.; Papadakis, A.; Samarakou, M. Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning. *J. Sens. Actuator Netw.* **2021**, *10*, 72. [CrossRef]
27. Gong, Y.; Chung, Y.A.; Glass, J. AST: Audio Spectrogram Transformer. *arXiv* **2021**, arXiv:2104.01778. [CrossRef]
28. Palanisamy, K.; Singhania, D.; Yao, A. Rethinking CNN Models for Audio Classification. *arXiv* **2020**, arXiv:2007.11154. [CrossRef]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778. [CrossRef]
30. Redmon, J. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
31. Tsalera, E.; Papadakis, A.; Samarakou, M.; Voyiatzis, I. Feature Extraction with Handcrafted Methods and Convolutional Neural Networks for Facial Emotion Recognition. *Appl. Sci.* **2022**, *12*, 8455. [CrossRef]
32. Piczak, K.J. ESC: Dataset for Environmental Sound Classification. In Proceedings of the 23rd ACM International Conference on Multimedia (MM 2015), Brisbane, Australia, 26–30 October 2015; pp. 1015–1018. [CrossRef]
33. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An ontology and human-labeled dataset for audio events. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 776–780. [CrossRef]
34. Driedger, J.; Müller, M. A Review of Time-Scale Modification of Music Signals. *Appl. Sci.* **2016**, *6*, 57. [CrossRef]
35. Morrison, M.; Jin, Z.; Bryan, N.J.; Caceres, J.P.; Pardo, B. Neural Pitch-Shifting and Time-Stretching with Controllable LPCNet. *arXiv* **2021**, arXiv:2110.02360.

36. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
37. Wei, S.; Zou, S.; Liao, F. A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification. *J. Phys. Conf. Ser.* **2020**, *453*, 012085. [CrossRef]
38. Lin, G.; Jiang, J.; Bai, J.; Su, Y.; Su, Z.; Liu, H. Frontiers and developments of data augmentation for image: From unlearnable to learnable. *Inf. Fusion* **2025**, *114*, 102660. [CrossRef]
39. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation: Red Hook, NY, USA, 2012; Volume 25. [CrossRef]
40. Jensen, J.H.; Christensen, M.G.; Ellis, D.P.W.; Jensen, S.H. Quantitative Analysis of a Common Audio Similarity Measure. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 693–703. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.