PHAS0056: Practical Machine Learning for Physicists
Mini-Project Report: Identifying Sounds

Abstract

This project investigates the use of deep learning techniques for environmental sound classification using the ESC-50 dataset. Two types of audio representations—Mel spectrograms and raw waveforms—are evaluated alongside five data augmentation strategies: pitch shifting, time stretching, white noise addition, pink noise addition, and their combination. Convolutional neural network (CNN) architectures are developed and progressively refined to align model complexity with task granularity. In the 50-class task, augmentation is demonstrated to significantly improve model generalisation, with combined methods yielding the highest accuracy of 72% using spectrograms, and 63% using raw audio. The best-performing single augmentation technique is established to be pink noise for waveforms (61%) and pitch shifting for spectrograms (62%). To evaluate the effect of task simplification, the 50 classes are then grouped into five super-categories. A simplified CNN architecture demonstrates smoother convergence and improved generalisation, supporting the notion that reduced model capacity may be better suited to lower task complexity. Confusion matrix analyses revealed consistent misclassifications in acoustically ambiguous categories and directional biases between class pairs. These findings underscore the importance of co-optimising model architecture and augmentation strategies, and suggest directions for future work including class-specific regularisation and spectro-temporal fusion.

Introduction

Sound classification is the process of analysing audio signals and automatically assigning them to predefined categories, such as speech, environmental noises, or machine sounds. Traditionally, this involves data gathering and labelling, preprocessing techniques like denoising or time stretching, feature extraction—commonly using Mel spectrograms, and the application of machine learning or deep learning architectures (e.g., Convolutional Neural Networks). Recent studies emphasize the benefits of these architectures in domains such as healthcare (analysing respiratory or cardiac sounds [1]), environmental conservation (identifying bird calls and other wildlife [2]), and security (surveillance systems capable of detecting anomalies in audio streams [3]).

However, noise in real-world settings, scarcity of labelled data, and highly overlapping acoustic features across categories often hinder accurate sound classification. Modern solutions increasingly rely on data augmentation, such as pitch shifting, time stretching and adding background noise to improve generalisation and mitigate overfitting [4]. Recent advances in network architectures—such as raw-waveform convolutional networks [5] and transformer-based models [6]—have shown strong potential in learning robust audio representations across diverse recording environments. The ESC-50 dataset has become a widely adopted benchmark in this space, supporting ongoing research into both novel model architectures and data augmentation strategies [7]. As these techniques continue to evolve, sound classification systems are moving closer to real-time, scalable deployment, with impactful applications in healthcare, smart cities, and environmental monitoring.

Methodology

1 Data Augmentation

a. Mel Spectrograms

This study utilises the ESC-50 dataset, which provides 2000 short audio segments distributed across 50 classes and 5 super-categories [7]. Each file is loaded in Python (using Librosa), resampled at 44.1 kHz (“CD quality”) to capture a wide frequency range with manageable file size, and subsequently converted into Mel spectrograms. Given the relatively small size of this dataset, data augmentation techniques are commonly employed for enhancing model robustness [4,7,8]. In this study, five augmentation methods are evaluated, using an 80/20 train–test partition (resulting in 1600 and 400 training and testing samples, respectively) at random state = 42 to ensure methodological consistency across different experiments.

Method 1 – Pitch Shifting: Pitch shifting is applied to all training samples, altering their frequency content while preserving temporal structure. This doubles the training set size to 3200 samples, with the testing set held constant at 400. The method aims to improve generalisation to frequency variations commonly encountered across acoustic classes.

Method 2 – Time Stretching: Time stretching modifies the duration of each training signal without affecting its frequency distribution. The training set is similarly expanded to 3200 samples. This approach is intended to enhance robustness to temporal distortions such as speed variation or signal compression.

Method 3 – White Noise Addition: White noise is added to training data at random signal-to-noise ratios between 10 dB and 17 dB, simulating high-frequency environmental interference.

The resulting dataset again comprises 3200 and 400 training and testing samples, respectively. This method targets improved resilience in noisy, unpredictable conditions.

Method 4 – Pink Noise Addition: Pink noise, with stronger energy at lower frequencies, is added using the same SNR range. The resulting augmentation introduces naturalistic low-frequency background conditions to the data. Dataset size matches previous methods.

Method 5 – Combined Augmentation: All four augmentation types are applied to each training sample, generating four augmented variants per original file. This results in a training set of 8000 samples, while the testing set remains at 400. The goal is to maximise acoustic diversity and model robustness by combining spectral, temporal, and noise-based perturbations.

b. Audio Waveforms

Another set of augmentations is applied directly to the raw audio signals, paralleling the five methods used for Mel spectrogram processing. Each waveform is truncated or padded to a standardised duration (e.g., 2.5 seconds) and normalised to maintain its peak amplitude at or below unity. This unified approach streamlines the integration of both original and augmented recordings into the deep learning workflow, aligning with recognised best practices for raw waveform classification [4,5]. Detailed scripts can be retrieved from Project_Sound_Identification.ipynb file.

2 Convolutional Neural Network (CNN) Model Architecture

a. Mel Spectrograms

All models in this study use a softmax output layer for multi-class classification, are compiled with the sparse categorical crossentropy loss (suitable for integer labels), and are evaluated using the accuracy metric. This consistent setup allows for fair comparison across all experiments.

The initial CNN architecture, defined in CNN(), comprised four convolutional blocks with ReLU activations and max-pooling layers, followed by two fully connected layers. While effective for baseline classification, this structure lacked the complexity required to capture the variability introduced by data augmentation. As a result, the model was restructured into OptimisedCNN(), which incorporated batch normalisation after each convolutional layer to stabilise and accelerate training, introduced dropout regularisation to mitigate overfitting, and adjusted filter sizes to better balance model capacity and generalisation.

To further improve generalisation and training stability, the FurtherOptimisedCNN() architecture was introduced. Dropout rates were increased in both convolutional and dense layers to provide stronger regularisation and reduce overfitting. Additionally, a learning rate scheduler was implemented to lower the learning rate after 20 epochs, enabling finer tuning in later epochs.

For the 5-category classification task, the FurtherOptimisedCNN() was simplified to better reflect the reduced output complexity, and also minimise overfitting and computational load. The model's depth and parameter count were reduced by using fewer convolutional layers, smaller kernels, and GlobalAveragePooling2D in place of flattening. Regularisation was applied through L2 penalties and dropout, improving generalisation while reducing overfitting and matching model capacity to task complexity.

b. Audio Waveforms

For raw waveform classification, the architecture RawAudio_FurtherOptimisedCNN() mirrored the design of the spectrogram-based FurtherOptimisedCNN(), incorporating batch normalisation, dropout regularisation, and three Conv1D blocks to capture temporal features. While effective in learning temporal patterns from waveform data, this configuration incurred excessive computational cost, with training times reaching approximately six minutes per epoch. To address this inefficiency, the network was restructured into RawAudio_VeryFurtherOptimisedCNN(), which halved the number of convolutional filters and also reduced the batch size from 64 to 32. Combined with reduced input length and the use of lower-precision (float16) data, these architectural and preprocessing modifications resulted in a substantial decrease in training time—down to approximately 30 seconds per epoch—without a significant compromise in model performance.

Results

1 Mel Spectrograms

a. 50 Sound Categories

To evaluate the effectiveness of data augmentation and architectural modifications, each model was trained under controlled conditions and its training history was recorded. Performance was assessed primarily through test accuracy and test loss. First, the contribution of model complexity to classification performance was isolated by holding the augmentation method constant (pitch shifting, i.e., Method 1) and varying the network architecture. This method allowed for a systematic comparison of generalisation ability across configurations. The results of this experiment are demonstrated in **Fig. 1** below.

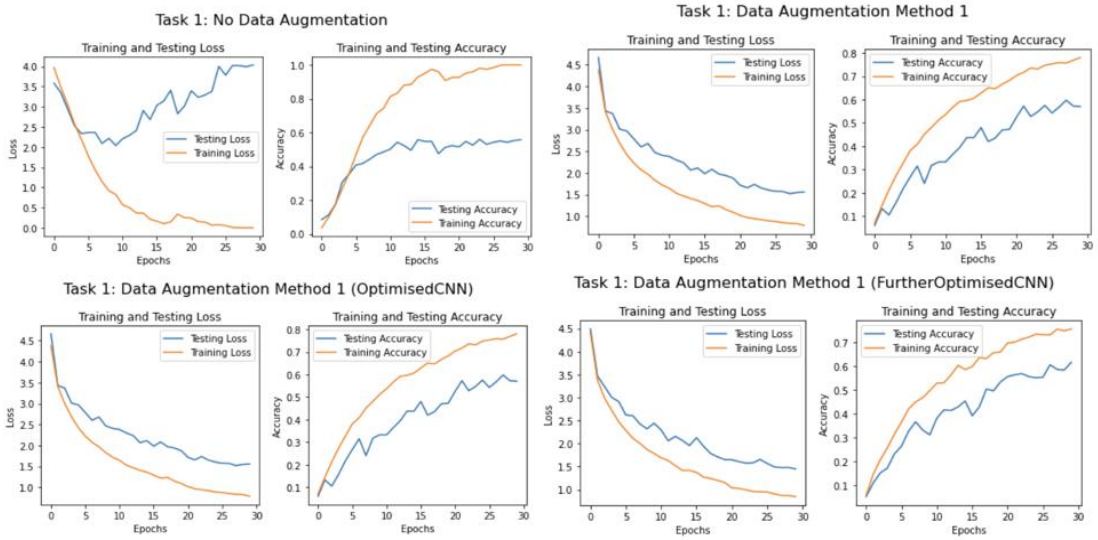


Fig 1. Training and testing loss and accuracy curves for four model configurations: (top left) baseline CNN without augmentation, (top right) baseline CNN with pitch shifting, (bottom left) OptimisedCNN with pitch shifting, and (bottom right) FurtherOptimisedCNN with pitch shifting. Improvements in performance and generalisation are observed with increasing architectural complexity.

The baseline CNN trained without augmentation achieves a test accuracy of 55.75% and a test loss of 4.04. While the model converges rapidly, the limited diversity in training data results in overfitting and restricted generalisation. Introducing pitch-shifting augmentation without modifying the architecture leads to a substantial drop in performance, with test accuracy falling

to 24.00% and loss increasing to 10.08. This indicates that the original architecture lacks the capacity to handle increased variability introduced through data augmentation.

Replacing the base model with the OptimisedCNN—featuring batch normalisation and dropout—yields significant improvements under the same augmentation regime. Test accuracy increases to 51.75%, and loss decreases to 2.05, suggesting enhanced training stability and improved feature generalisation. The FurtherOptimisedCNN, incorporating higher dropout rates and a learning rate scheduler, achieves the best performance with 61.75% test accuracy and a test loss of 1.45, demonstrating that deeper regularisation and adaptive learning contribute meaningfully to both convergence and generalisation.

In summary, data augmentation alone is insufficient when model complexity is inadequate. However, when combined with appropriate architectural enhancements, it leads to substantial performance gains, confirming the importance of co-optimising data and network design in environmental sound classification tasks.

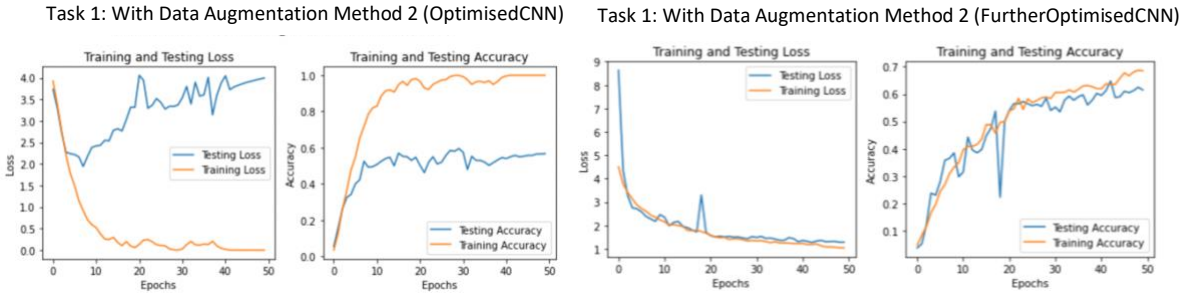


Fig 2. Training and testing performance of OptimisedCNN and FurtherOptimisedCNN on time-stretched data. Improved convergence and generalisation are observed with increased regularisation.

When evaluating the time stretching augmentation, i.e., Method 2 (results shown in **Fig. 2**), the OptimisedCNN achieves a test accuracy of ~58% but shows clear overfitting, with training accuracy near 100% and unstable test loss around 4.0. In contrast, the FurtherOptimisedCNN improves performance substantially, reaching 62% test accuracy and a stable loss of near 1.2 — comparable to the best results from pitch shifting. These results confirm that, like pitch shifting, time stretching is effective only when combined with sufficient regularisation, further underscoring the importance of aligning augmentation strategies with model capacity.

Further evaluations of the FurtherOptimisedCNN architecture using three additional augmentation strategies reveal consistent gains in generalisation. The corresponding results are shown in **Fig. 3** below. With white noise addition, the model achieves a solid test accuracy of 56.00% and a reduced test loss of 1.72, reflecting improved robustness to high-frequency perturbations. Training with pink noise yields slightly lower performance, with 52.50% test accuracy and a test loss of 1.75. While the model retains some resilience to low-frequency noise, the wider gap between training and testing metrics suggests limited generalisation and a tendency to overfit under this augmentation type. In contrast, the highest performance is achieved with combined augmentation, incorporating pitch shifting, time stretching, and both noise types, which results in 72.25% test accuracy and a test loss of 0.93. This result surpasses all single-method augmentations, confirming that multi-strategy augmentation enhances data diversity and model robustness most effectively when paired with strong regularisation.

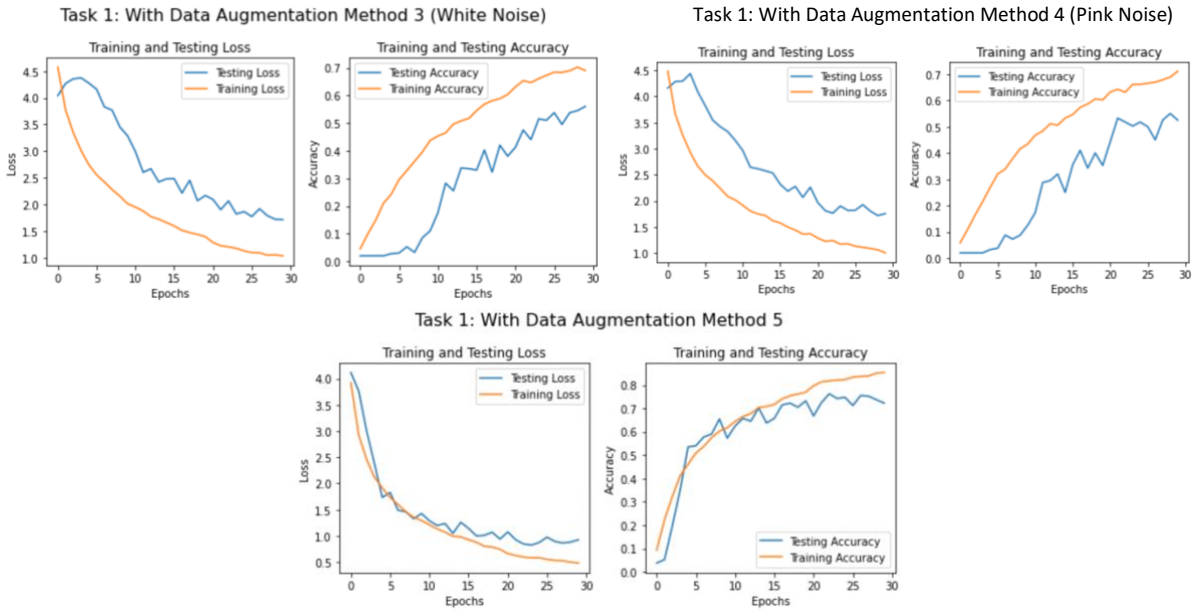


Fig. 3. Training and testing performance of the FurtherOptimisedCNN using white noise (Method 3), pink noise (Method 4), and combined augmentations (Method 5: both noise types, pitch shifting, and time stretching). Method 5 yields the highest accuracy and best generalisation, while pink noise shows moderate overfitting.

b. 5 Sound Categories

To simplify the classification task, the 50 original ESC-50 classes were consolidated into 5 broader super-categories – *animals*, *natural soundscapes*, *human non-speech sounds*, *interior/domestic sounds*, and *exterior/urban noises* – based on the grouping proposed by Karol J. Piczak, the creator of the dataset [7]. The metadata, data processing pipeline, and CNN architecture were adjusted accordingly, including reducing the final output layer to five nodes. Classification was then performed using spectrogram representations with pitch-shift augmentation, chosen due to its previously demonstrated effectiveness in spectrogram-based models. To evaluate the suitability of model complexity for broader class distinctions, both the original (FurtherSimplied_CNN) and simplified (Simple_CNN_5Categories) CNN architectures were applied to the 5-category classification task using spectrogram input with pitch shift augmentation, and their respective performances are presented in **Fig 4**. below.

Fig. 4. Comparison of training and testing performance for the original (left) and simplified (right) CNNs on the 5-category classification task using spectrograms and pitch shift augmentation. The simplified model trades a small drop in accuracy for improved training stability and reduced overfitting.

The original CNN exhibited a higher final test accuracy of approximately 62%, but showed considerable fluctuation during training, reflecting instability and signs of overfitting, as expected. In contrast, the simplified CNN reached a slightly lower accuracy of around 57%,

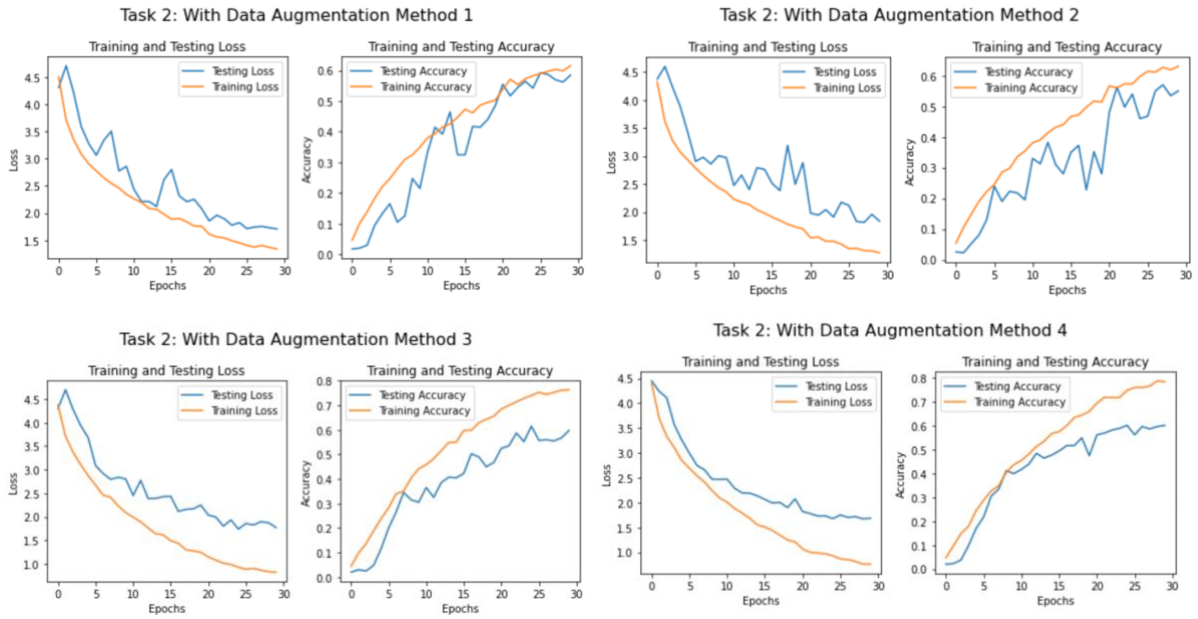
but demonstrated smoother convergence and improved alignment between training and testing curves. This suggests that while the original architecture may capture more complex patterns, its increased capacity might be unnecessary for coarse-grained classification and may lead to overfitting. The simplified model offers a more stable and generalisable solution under constrained label granularity.

2 Audio Waveforms

The VeryFurtherOptimisedCNN, when applied to raw waveform data, demonstrates strong performance across all four augmentation strategies, as can be seen in **Fig. 5**. With pitch shifting (Method 1) and time stretching (Method 2), the model achieves stable convergence, with test accuracies just above 60%. Despite slight fluctuations in test accuracy—particularly with time stretching—both methods exhibit reliable learning and generalisation, suggesting the architecture is well-suited to capturing temporal features directly from raw input.

The model also responds well to noise-based augmentations. White noise addition (Method 3) results in a respectable test accuracy of 59.75%, indicating resilience to high-frequency perturbations. Pink noise (Method 4) yields slightly better results, with a test accuracy of 61.00%, outperforming all other single augmentation techniques. This suggests that pink noise, which more closely resembles natural acoustic environments, may provide a more effective form of regularisation in the waveform domain. Nonetheless, both methods exhibit moderate overfitting, as evidenced by a consistent gap between training and testing accuracies, highlighting the challenge of maintaining generalisation with single noise-based inputs.

As with the spectrogram images before, the best results are obtained with combined augmentation (Method 5), integrating pitch shifting, time stretching, and both noise types. This configuration achieves a test accuracy of 63.00% and a notably reduced test loss of 1.28, with tightly aligned accuracy curves and smooth convergence. These findings, again, reinforce the value of augmentation diversity in raw waveform modelling.



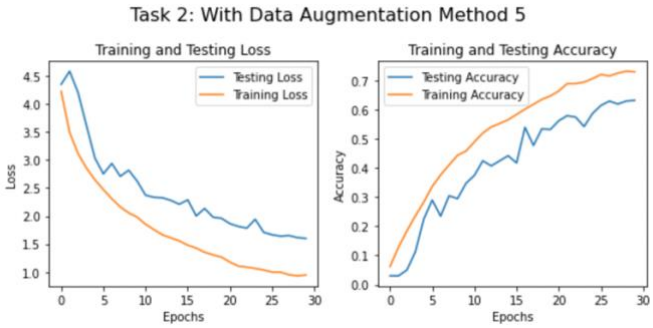


Fig. 5. Performance of the VeryFurtherOptimisedCNN on raw audio using five augmentation methods. Pink noise (Method 4) achieves the best individual accuracy (61%) but shows signs of overfitting. Combined augmentation (Method 5) yields the highest overall accuracy (63%) with improved generalisation.

3 Category-Based Evaluation

a. 50 Sound Categories

To assess the classification performance of the model and its ability to distinguish between the 50 categories and 5 super-categories in the ESC-50 dataset, three standard evaluation metrics were used: accuracy, precision, and recall. Accuracy reflects the proportion of correctly predicted samples, precision measures the proportion of correct predictions among all predicted instances of a class, and recall quantifies the proportion of actual class instances that were correctly identified. Together, these metrics provide a more nuanced view of model performance than overall accuracy alone.

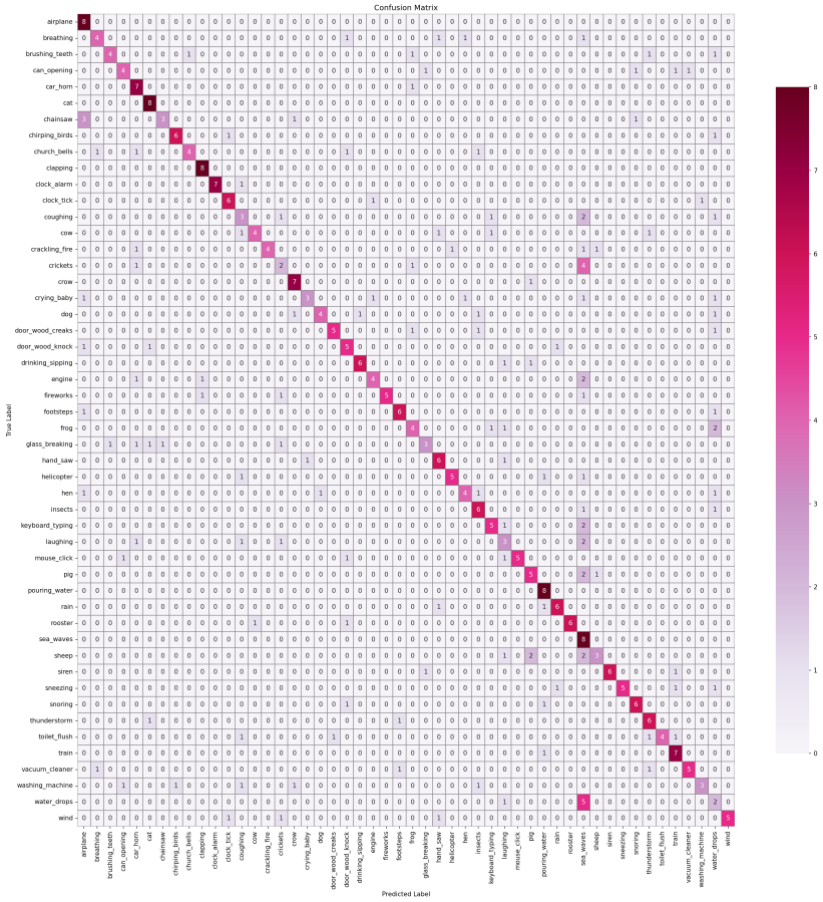


Fig. 6. Confusion Matrix showing performance of VeryFurtherOptimisedCNN trained on raw audio waveforms using Method 5.

The class-wise performance of the model trained on raw audio using Method 5 (combined augmentations), represented by its confusion matrix, is shown in **Fig. 6.** Classes such as *airplane*, *car horn*, *coughing*, and *fireworks* exhibit perfect or near-perfect recall, suggesting the model effectively captures distinctive temporal characteristics in the raw audio signals. Other strongly performing classes include *helicopter*, *rain*, and *speech*, which also maintain consistent accuracy and precision across validation runs.

However, some categories, such as *glass breaking*, *mouse click*, and *pouring water*, exhibit lower precision or recall—often below 0.40—highlighting continued difficulty in distinguishing between similar or acoustically ambiguous events. The category *pouring water*, for instance, suffers from both low recall (0.25) and low precision (0.14), indicating high confusion with other classes, likely due to its broadband and less distinctive acoustic profile. Targeted strategies, such as class-specific augmentation or feature enhancement, may be required to improve its separability.

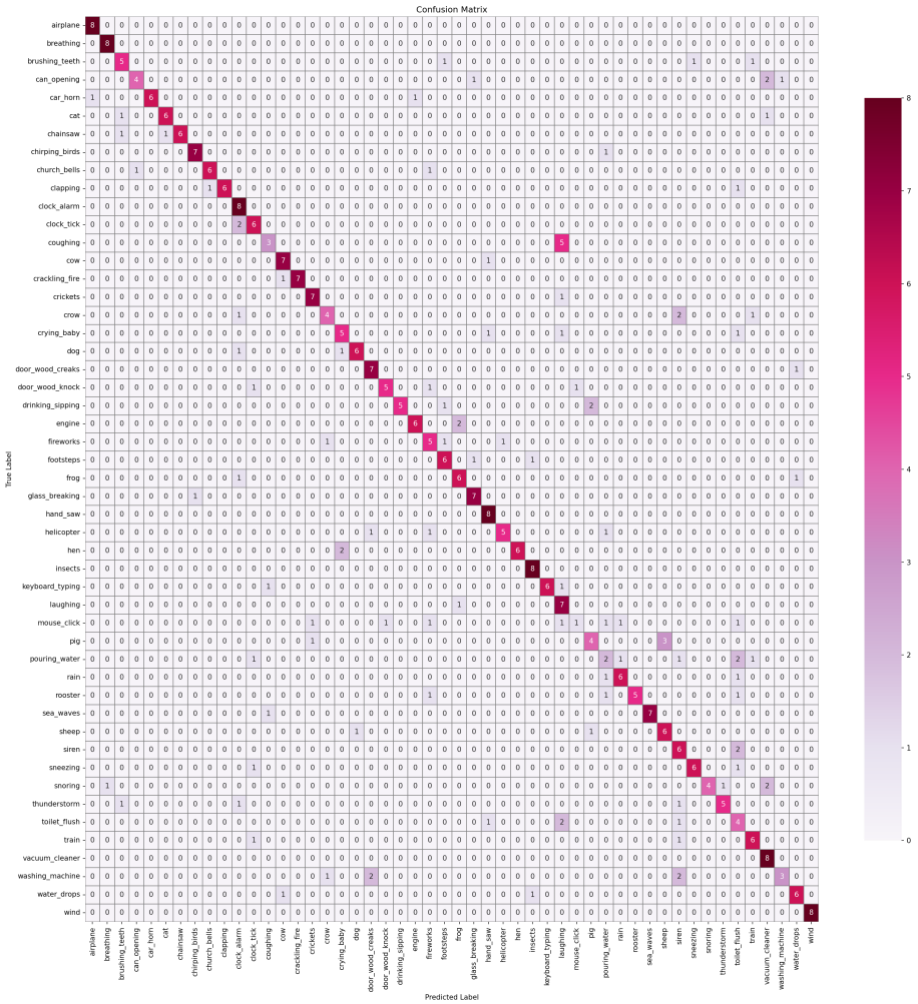


Fig. 7. Confusion Matrix showing performance of FurtherOptimisedCNN trained on Mel spectrograms using Method 5.

In comparison to the spectrogram-based model, trained using the same augmentation strategy (corresponding confusion matrix shown in **Fig. 7**) the raw waveform model yields more balanced performance across classes. The spectrogram model, while generally effective, displays more sharply defined performance extremes, with certain categories overfitting and others underperforming. Visually, the raw-audio confusion matrix appears less fragmented,

with fewer strong off-diagonal entries and more consistent diagonal dominance. Classes such as *door knock, engine,* and *hand saw* are better distinguished in the waveform-based model, likely because some temporal patterns may be blurred during spectrogram conversion.

Nonetheless, for classes with subtle spectral cues—such as *keyboard typing* or *clock tick*—the spectrogram representation may provide an edge, as these distinctions are more apparent in frequency than in time alone. This suggests that while raw waveform models can outperform spectrogram-based models in overall generalisation and robustness, certain fine-grained acoustic categories may still benefit from frequency-domain features.

Interestingly, asymmetric misclassifications observed in both tasks indicate bias in the learned feature space. In Task 2, *water drops* are often misclassified as *sea waves*, but not vice versa – likely because the broader acoustic profile of sea waves encompasses features present in water drops. Similarly, in Task 1, *coughing* is frequently confused with *laughing*, yet *laughing* is never predicted as coughing. These directional errors suggest certain classes act as default attractors, possibly due to their acoustic generality or dominance in the training data.

In summary, applying combined augmentation techniques to raw audio, alongside a strongly regularised architecture, yields more stable and generalisable performance across diverse sound categories. Nonetheless, persistent confusion among spectrally similar or acoustically ambiguous classes suggests that further gains may be achieved through class-specific augmentation strategies or tailored architectural refinements to enhance class-level discrimination.

b. 5 Sound Categories

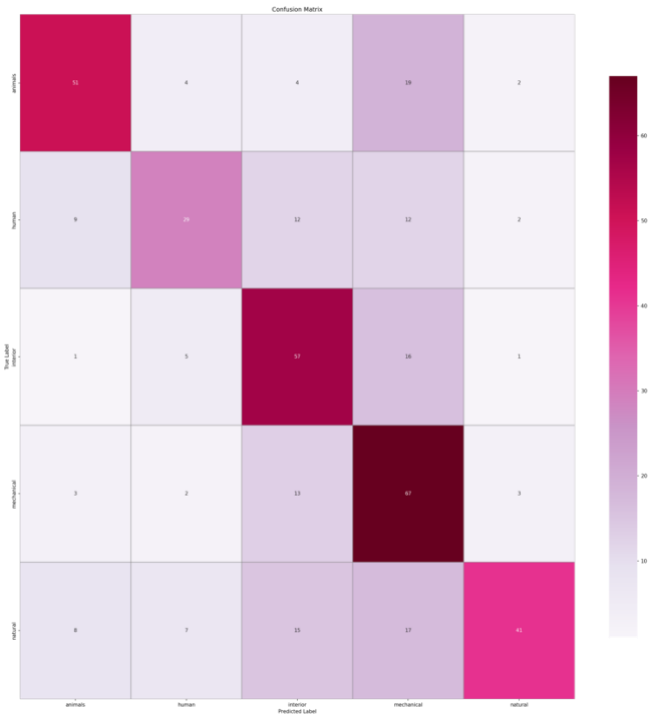


Fig. 8. Confusion matrix for the original, FurtherOptimised_CNN. Best separation is seen in the "natural" and "animals" categories.

To assess the effect of architectural complexity on performance, the original CNN (with four convolutional blocks and dense layers) was compared to a simplified version designed specifically for the five super-category classification task. The confusion matrices reveal notable differences in how the original and simplified CNN architectures handle the five super-categories. The original CNN (**Fig. 8**) achieves stronger

diagonal dominance, indicating higher per-class accuracy, particularly for the “animal” and “natural” classes. However, it also exhibits more pronounced confusion between “interior” and “human” sounds, and between “exterior” and “natural,” suggesting less generalisation across ambiguous boundary cases.

In contrast, the simplified CNN (**Fig. 9**) produces a more balanced distribution of errors. While the diagonal elements are slightly weaker overall—especially for the “interior” and “human” classes—the model avoids the sharp bias seen in the original network. Misclassifications are more symmetrically spread, and inter-class confusion is less concentrated, suggesting that the simplified model learns broader, more transferable features. This shift implies that reducing model complexity helped prevent overfitting to specific patterns while slightly compromising on peak class-specific accuracy.

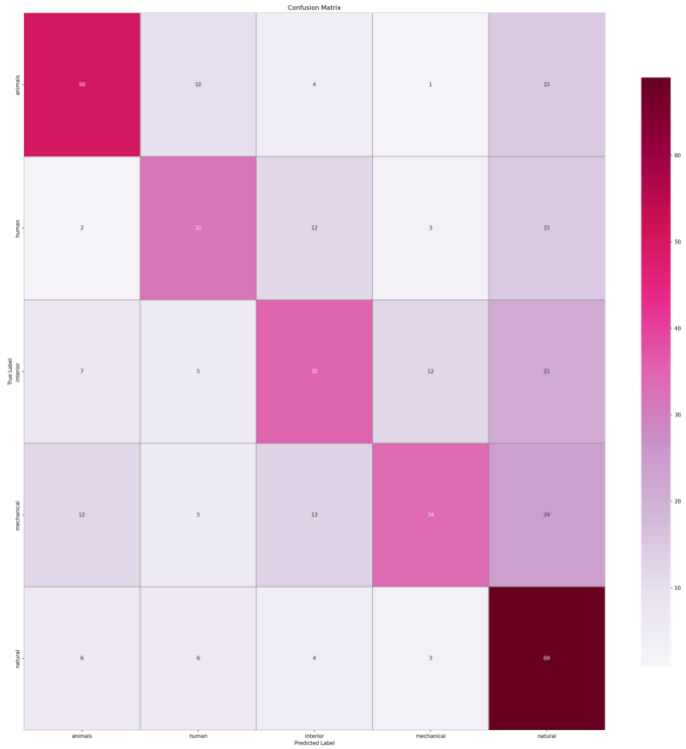


Fig. 9. Confusion matrix for the simplified Simple_CNN_5Categories. Misclassifications are more balanced, with improved stability across classes.

Discussion & Conclusion

This project investigated the use of convolutional neural networks for environmental sound classification, using the ESC-50 dataset and a range of augmentation techniques applied to both Mel spectrograms and raw waveforms. By systematically comparing architectures and preprocessing methods, the study highlighted the importance of aligning model complexity with task granularity and data diversity.

Key findings emerged from evaluating five data augmentation strategies across multiple CNN configurations. In both time–frequency and raw waveform domains, combined augmentation (pitch shift, time stretch, and white and pink noise) consistently delivered the best performance, improving generalisation and reducing overfitting. Particularly, pink noise alone achieved the strongest performance among single-method augmentations in waveform-based models (61%), but still showed signs of overfitting. In contrast, the combined augmentation method achieved the highest test accuracy (63%) with smoother convergence. For spectrogram-based models, pitch shifting emerged as the most effective individual augmentation (near 62%), demonstrating strong generalisation. However, as with waveforms, combining all four

augmentation techniques yielded the best performance (72%), highlighting the advantage of diverse perturbations in enhancing model robustness.

The experiments also explored the impact of task simplification by grouping the 50 ESC-50 categories into 5 broader super-categories. Here, reducing the classification granularity exposed the limitations of the original CNN, which, though capable of high accuracy, demonstrated instability and overfitting. A simplified CNN, designed with fewer parameters, smaller kernel sizes, and stronger regularisation (L2 penalties and dropout), achieved more consistent performance with slightly lower but more stable accuracy of 62%. The associated confusion matrices showed that the simplified model produced more balanced class predictions and avoided over-reliance on dominant features.

Confusion matrix analyses revealed that some classes—particularly those with ambiguous or broadband acoustic characteristics—remained difficult to distinguish. Pouring water, for example, exhibited low precision and recall across experiments, while bias was evident in asymmetric confusions (e.g., coughing misclassified as laughing, but not vice versa). These findings suggest that future models could benefit from class-specific augmentation or targeted loss functions to address imbalanced separability.

Overall, the study demonstrates the value of co-designing augmentation pipelines and model architectures. While deeper, more regularised networks improve generalisation under complex augmentation regimes, simpler models may suffice for coarser tasks, offering efficiency and improved stability. Future work could explore domain-adaptive augmentations, integrate spectro-temporal fusion, and extend evaluation to real-time applications.

References

[1] Raza, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, and B.-W. On, “Heartbeat sound signal classification using deep learning,” *Sensors*, vol. 19, no. 21, 2019. DOI: 10.3390/s19214819

[2] Mohaimenuzzaman, M., Bergmeir, C., West, I., & Meyer, B. “Environmental Sound Classification on the Edge: A Pipeline for Deep Acoustic Networks on Extremely Resource-Constrained Devices.” *Pattern Recognition*, 133, 2023. DOI: 10.1016/j.patcog.2022.109025.

[3] Marchi, E., Vesperini, F., Weninger, F., Eyben, F., Squartini, S., & Schuller, B. “A Novel Approach for Automatic Acoustic Novelty Detection Using a Denoising Autoencoder with Bidirectional LSTM Neural Networks.” *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 1996-2000. DOI: 10.1109/ICASSP.2015.7178320.

[4] Y. Tokozume, Y. Ushiku, and T. Harada, “Learning from Between-Class Examples for Deep Sound Recognition,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018. DOI: 10.48550/arXiv.1711.10282.

[5] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, “Very Deep Convolutional Neural Networks for Raw Waveforms,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 421–425. DOI: 10.1109/ICASSP.2017.7952190.

[6] Y. Gong, Y.-A. Chung, and J. Glass, “AST: Audio Spectrogram Transformer,” in *Proceedings of Interspeech 2021*, Brno, Czech Republic, 2021, pp. 571–575. DOI: 10.21437/Interspeech.2021-698.

[7] Piczak, K. J. “ESC: Dataset for Environmental Sound Classification.” In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, ACM Press, 2015, pp. 1015–1018. DOI: 10.1145/2733373.2806390.

[8] Y. Tokozume and T. Harada, “Learning environmental sounds with end-to-end convolutional neural network,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 2721–2725. DOI: 10.1109/ICASSP.2017.7952651.