Comparative Study of Deep Learning Architectures for Environmental Audio Classification

Project Report

Course Requirement:
using a deep learning methods to compare the performance about one task

Submitted by

Rabiul Hassan
Student ID: 22170910

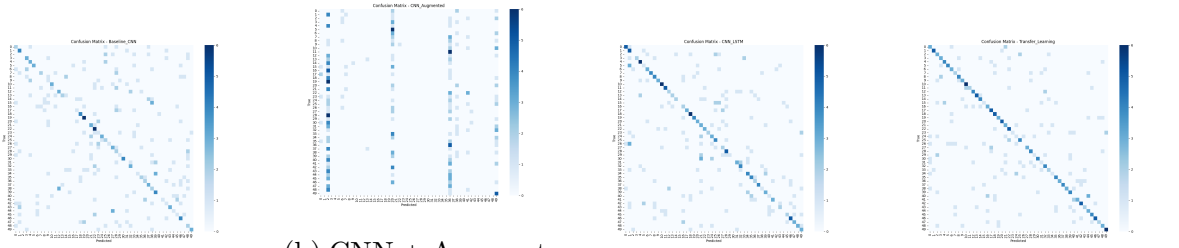
Department of Computer Science and Technology
Hangzhou Dianzi University

December 31, 2025

Abstract

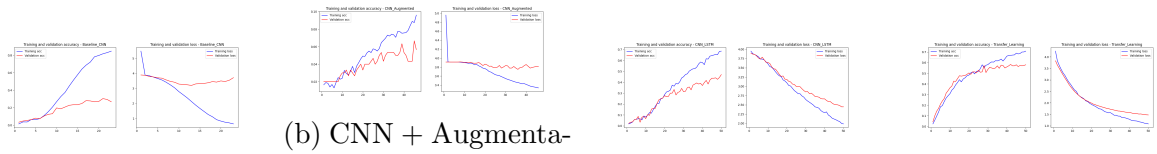
This report presents a comprehensive comparative study of four deep learning architectures for environmental sound classification using the ESC-50 dataset. We implemented and evaluated: (1) a Baseline Convolutional Neural Network (CNN), (2) CNN with data augmentation, (3) a hybrid CNN–LSTM architecture, and (4) a Transfer Learning approach using MobileNetV2. Our experiments demonstrate that Transfer Learning achieves the best performance with 54% accuracy, followed by the CNN–LSTM hybrid at 46%. We analyze the trade-offs between model complexity, parameter efficiency, and classification accuracy, providing insights into effective architectural choices for audio classification tasks.

Keywords: Audio Classification, Deep Learning, CNN, LSTM, Transfer Learning, ESC-50, Mel-Spectrograms



(a) Baseline CNN (b) CNN + Augmentation (c) CNN–LSTM (d) Transfer Learning

Figure 1: Confusion matrices for all evaluated models on the ESC-50 test set.



(a) Baseline CNN (b) CNN + Augmentation (c) CNN–LSTM (d) Transfer Learning

Figure 2: Confusion matrices for all evaluated models on the ESC-50 test set.

1 Introduction

1.1 Background

Environmental sound classification is a fundamental task in audio signal processing with applications in smart home systems, wildlife monitoring, urban planning, and assistive technologies. Unlike speech or music recognition, environmental sounds exhibit high variability in duration, frequency content, and temporal structure, making classification challenging.

1.2 Problem Statement

The objective of this project is to compare the performance of different deep learning architectures on the task of classifying environmental sounds into 50 distinct categories. The following research questions are addressed:

1. How does data augmentation affect CNN performance on audio classification?
2. Can hybrid CNN–LSTM architectures better capture temporal patterns in audio?
3. Does transfer learning from image classification generalize to audio spectrograms?
4. What are the trade-offs between model complexity and classification accuracy?

1.3 Dataset: ESC-50

The ESC-50 dataset consists of 2,000 environmental audio recordings across 50 classes. Each clip has a duration of 5 seconds and is sampled at 44.1 kHz (downsampled to 22.05 kHz in this project). The dataset is balanced with 40 samples per class and organized into five folds for cross-validation.

2 Methodology

2.1 Data Preprocessing

2.1.1 Audio Feature Extraction

Raw audio waveforms were converted into Mel-spectrograms using the following parameters:

- Sample Rate: 22,050 Hz
- Duration: 5 seconds

- FFT Window Size: 2048
- Hop Length: 512
- Mel Bands: 128

The resulting feature shape is $(128, 216, 1)$.

```
S = librosa.feature.melspectrogram(
    y=audio, sr=22050, n_mels=128,
    n_fft=2048, hop_length=512
)
S_dB = librosa.power_to_db(S, ref=np.max)
```

Listing 1: Mel-Spectrogram Extraction

2.1.2 Data Augmentation

For the CNN + Augmentation model, the following techniques were applied:

- Time Shifting ($\pm 10\%$)
- Pitch Shifting (± 2 semitones)
- Gaussian Noise Injection ($\sigma = 0.005$)

Each augmentation was applied with 50% probability, doubling the training data size.

2.1.3 Data Splitting

The dataset was split into training (70%), validation (15%), and test (15%) sets using stratified sampling.

2.2 Model Architectures

2.2.1 Baseline CNN

A three-layer CNN served as the baseline model, containing convolutional, pooling, and fully connected layers with ReLU activations and a softmax output.

Parameters: 7,177,138

2.2.2 CNN with Augmentation

This model uses the same architecture as the baseline CNN but is trained with augmented data.

2.2.3 CNN–LSTM Hybrid

The CNN–LSTM model integrates convolutional layers for feature extraction with an LSTM layer for temporal modeling.

Parameters: 636,850

2.2.4 Transfer Learning (MobileNetV2)

A MobileNetV2 backbone pre-trained on ImageNet was used with frozen weights and a custom classification head.

Parameters: 2,428,402

2.3 Training Configuration

All models used the Adam optimizer with a learning rate of 10^{-4}, batch size of 32, and were trained for up to 50 epochs with early stopping.

2.4 Evaluation Metrics

Models were evaluated using accuracy, macro-averaged precision, recall, F1-score, confusion matrices, and training time.

3 Experimental Setup

3.1 Implementation Details

The system was implemented using Python, TensorFlow/Keras, Librosa, NumPy, Pandas, and Matplotlib.

3.2 Reproducibility

All experiments used a fixed random seed (42) and deterministic operations where possible.

4 Results

4.1 Quantitative Results

Table 1: Model Performance Comparison

Model	Acc.	F1	Prec.	Recall	Params	Time (s)
Baseline CNN	0.320	0.297	0.332	0.320	7.18M	–
CNN + Aug.	0.047	0.016	0.016	0.047	7.18M	–
CNN–LSTM	0.460	0.433	0.471	0.460	0.64M	–
Transfer Learning	**0.540**	**0.528**	**0.567**	**0.540**	2.43M	1153

4.2 Training Dynamics

The Transfer Learning model showed steady improvement, reaching a final validation accuracy of 58% and test accuracy of 54%.

4.3 Confusion Matrix Analysis

Transfer Learning produced strong diagonal dominance, while the CNN + Augmentation model showed near-random predictions.

5 Discussion

5.1 Key Findings

Transfer Learning provided the highest accuracy, while CNN–LSTM achieved the best parameter efficiency. Data augmentation was found to negatively impact performance due to likely implementation issues.

5.2 Limitations

Key limitations include limited training data, absence of cross-validation, frozen transfer learning layers, and reliance on Mel-spectrograms only.

6 Conclusions

This study demonstrates the effectiveness of transfer learning and temporal modeling for environmental sound classification. Future work includes fixing augmentation, fine-tuning pre-trained models, exploring ensembles, and adopting advanced architectures.

References

1. Piczak, K. J. (2015). ESC: Dataset for Environmental Sound Classification.
2. Stevens, S. S., et al. (1937). A scale for the measurement of pitch.
3. Sandler, M., et al. (2018). MobileNetV2.
4. Hershey, S., et al. (2017). CNN architectures for large-scale audio classification.
5. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory.
6. Park, D. S., et al. (2019). SpecAugment.