Environmental Audio Classification via Residual CNNs

22MAT220 MATHEMATICS FOR COMPUTING 4

Team 10

Aparna Bharani	(CB.SC.U4AIE24304)
Jaswanth Saravanan	(CB.SC.U4AIE24324)
Parkavi R	(CB.SC.U4AIE24338)
Rajashree T	(CB.SC.U4AIE24346)

Amrita School of Artificial Intelligence,
Amrita Vishwa Vidyapeetham, Coimbatore, India

December 21, 2025

1 Introduction

Environmental audio classification focuses on identifying real-world sounds such as sirens, rain, and engine noise. An audio signal is mathematically modeled as a continuous-time function $s(t)$ representing air pressure variations. Digital representation requires sampling according to the Nyquist-Shannon Sampling Theorem to ensure that the original signal can be reconstructed without aliasing. Since time-domain signals lack the spatial structure required for standard convolutional neural networks, frequency-domain analysis is employed to reveal hidden patterns using Fourier-based transformations. Mel spectrograms effectively convert these signals into a structured two-dimensional matrix suitable for visual processing architectures.

2 Problem Statement

The objective is to accurately identify and categorize environmental sounds from raw acoustic waveforms into one of 50 discrete categories. The primary mathematical difficulty lies in the non-stationary nature of environmental audio, where frequency content evolves rapidly over time. Traditional feature extraction methods often fail to capture the local translation invariance—the property where a feature is recognized regardless of its position—necessary for robust classification. We aim to solve this by mapping the highdimensional 1D signal space onto a discrete label space using a low-rank approximation within a 2D time-frequency manifold.

3 Methodology

The methodology involves a structured pipeline designed to transform raw air pressure waves into class probabilities:

- **Data Collection:** Audio samples are collected as fixed-length clips from a standardized environmental sound dataset.
- **Preprocessing:** All signals are resampled to a uniform sampling rate of 44.1 kHz to ensure frequency consistency.
- **STFT Analysis:** Short-Time Fourier Transform (STFT) is applied to capture time-varying frequency content by windowing the signal into stationary segments.
- **Magnitude Extraction:** Phase information is discarded to obtain magnitude spectrograms, focusing on energy distribution.
- **Mel Scaling:** Mel filterbanks map linear frequencies to the perceptual Mel scale, emphasizing frequencies more relevant to human auditory perception.
- **Input Preparation:** The resulting Mel spectrogram is represented as a matrix $X \in \mathbb{R}^{F \times T}$, which is then treated as a multi-channel input for the CNN.
- **Deep Learning Architecture:** A ResNet-based architecture extracts hierarchical time-frequency features. Residual connections are used to preserve identity mappings and prevent the vanishing gradient problem during backpropagation.
- **Classification:** Global Average Pooling reduces feature dimensionality, and a fully connected layer maps these features to class probabilities via a Softmax function.

4 Mathematical Implementation

4.1 Fourier Transformation and Spectrogram Logic

The transformation of the discrete signal $s[n]$ into the frequency domain is governed by the Discrete Fourier Transform (DFT):

$$X[k] = \sum_{n=0}^{N-1} s[n] \cdot e^{-i2\pi kn/N} \tag{1}$$

By applying this transform across sliding time windows, we construct the 2D matrix representing the signal's energy over time and frequency.

4.2 Convolution as a Linear Operator

The feature extraction process is mathematically defined by the convolution of an input matrix X with a trainable kernel K. The resulting feature map y is the Frobenius inner product of the kernel and the local input patch:

$$y = \sum_{i,j} X_{i,j} K_{i,j} + b \tag{2}$$

4.3 Residual Learning and Identity Shortcuts

- **Signal Stability:** As networks get deeper, training signals can vanish. Residual learning ensures information flows freely across layers.
- **Learning Objectives:** Instead of learning an entire transformation, layers only learn the small adjustments (residuals) needed.
- **Skip Connections:** Identity shortcuts allow the original input to bypass specific layers, merging directly with the final output.

5 System Flow Diagram

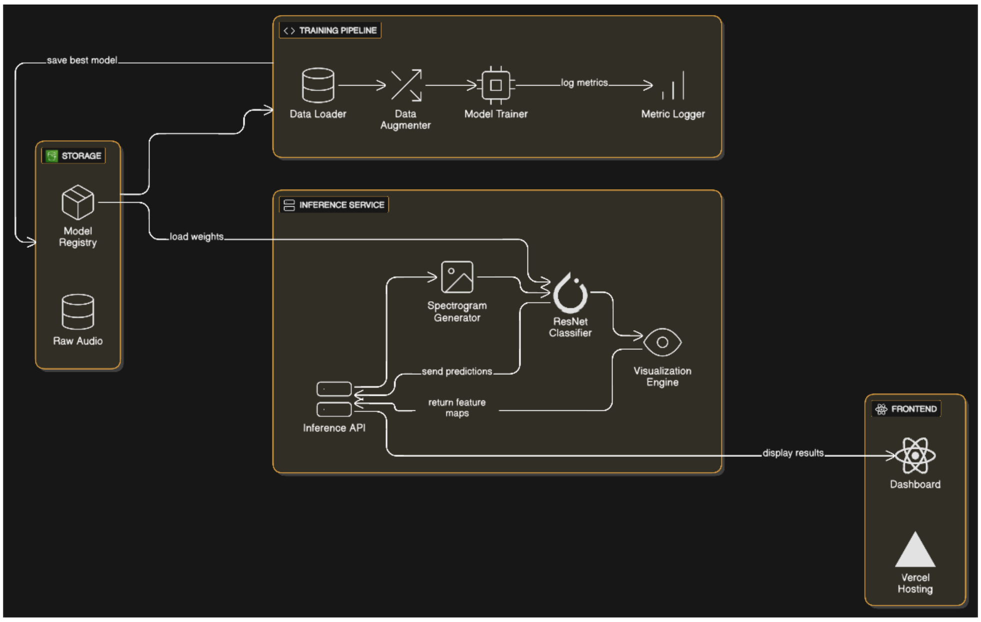

Figure 1: End-to-end system flow for Environmental Audio Classification.

6 Conclusion

By integrating signal processing theory with Residual CNNs, we demonstrate a robust method for classifying complex acoustic environments. The transformation to the Mel-frequency domain provides the necessary spatial structure for convolutional kernels to extract discriminative features, while residual connections ensure the mathematical stability required for deep feature learning.