Environmental Audio Classification via Residual CNNs and Mathematical Optimization

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, engine noise, footsteps, animal sounds, and mechanical operations. An audio signal is mathematically modeled as a continuous-time pressure function $s(t)$ capturing air pressure fluctuations. For digital processing, the signal is discretized using uniform sampling, guided by the Nyquist–Shannon Sampling Theorem to prevent information loss and aliasing.

However, direct time-domain audio signals lack explicit spatial locality, which is a key requirement for convolutional neural networks. To overcome this limitation, frequency-domain representations are employed. Fourier-based transformations reveal the hidden spectral structure of sounds, allowing temporal and frequency patterns to be represented jointly. Mel spectrograms further incorporate human auditory perception, transforming the signal into a structured two-dimensional matrix suitable for deep convolutional architectures.

2 Problem Statement

The primary objective of this work is to classify environmental audio signals into one of $C = 50$ predefined classes using mathematically grounded deep learning techniques. The challenge lies in the non-stationary nature of real-world acoustic signals, where frequency components evolve dynamically over time.

Let $s[n]$ denote a discrete-time audio signal of length N. The problem reduces to learning a mapping:

$$f : \mathbb{R}^N \to \{1, 2, \dots, C\}$$

which is robust to variations in loudness, temporal shifts, and background noise. This is achieved by embedding the raw signal into a lower-dimensional time–frequency manifold that preserves discriminative information while enabling efficient learning.

3 Methodology

The proposed system follows a structured signal processing and learning pipeline:

- **Data Collection:** Environmental sound recordings are obtained from standardized datasets, with each clip normalized to a fixed duration.
- **Resampling:** All signals are resampled to a uniform frequency of 44.1 kHz to ensure spectral consistency.
- **Framing and Windowing:** The signal is segmented into overlapping frames using a window function to ensure short-term stationarity.
- **STFT Computation:** Short-Time Fourier Transform is applied to capture local frequency content.
- **Magnitude Spectrogram:** Phase information is discarded to focus on spectral energy distribution.
- **Mel Filterbank Processing:** Linear frequency bins are mapped onto the Mel scale using triangular filters.
- **Normalization:** Feature scaling is applied to stabilize training and improve convergence.
- **Deep Feature Extraction:** A Residual CNN extracts hierarchical time–frequency representations.
- **Classification:** High-level features are mapped to class probabilities.

4 Mathematical Foundations of Signal Transformation

Given a discrete signal $s[n]$, the Discrete Fourier Transform (DFT) is defined as:

$$X[k] = \sum_{n=0}^{N-1} s[n] e^{-j2\pi kn/N}. \tag{1}$$

To preserve time localization, the Short-Time Fourier Transform (STFT) is applied:

$$X(m, k) = \sum_{n} s[n] w[n - m] e^{-j2\pi kn/N} \tag{2}$$

where $w[n]$ is a windowing function.

Mel scaling is achieved using a perceptual frequency transformation:

$$f_{\text{mel}} = 2595 \log_{10}(1 + f/700).$$

5 Convolutional Representation and Residual Learning

Convolutional layers perform linear filtering over local regions:

$$Y_{i,j} = \sum_{m,n} X_{i+m,j+n} K_{m,n} + b. \tag{3}$$

Residual networks introduce identity shortcuts:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \mathbf{W}) + \mathbf{x}$$

which ensure stable gradient propagation and enable deeper architectures without degradation.

6 Output Layer Modeling

Let $\mathbf{h}_i \in \mathbb{R}^d$ denote the high-level feature vector extracted by the Residual CNN for the i-th audio sample. Stacking all samples yields:

$$\mathbf{H} \in \mathbb{R}^{N \times d}.$$

Let $\mathbf{T} \in \mathbb{R}^{N \times C}$ represent the target label matrix. The linear output model is:

$$\mathbf{Y} = \mathbf{H}\boldsymbol{\beta}.$$

7 Gradient Descent-Based Optimization

The objective function is the squared error loss:

$$\mathcal{L} = \|\mathbf{H}\boldsymbol{\beta} - \mathbf{T}\|_2^2.$$

The gradient with respect to $\boldsymbol{\beta}$ is:

$$\nabla_{\boldsymbol{\beta}} = 2\mathbf{H}^T(\mathbf{H}\boldsymbol{\beta} - \mathbf{T}).$$

The update rule is:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \eta \nabla_{\boldsymbol{\beta}}.$$

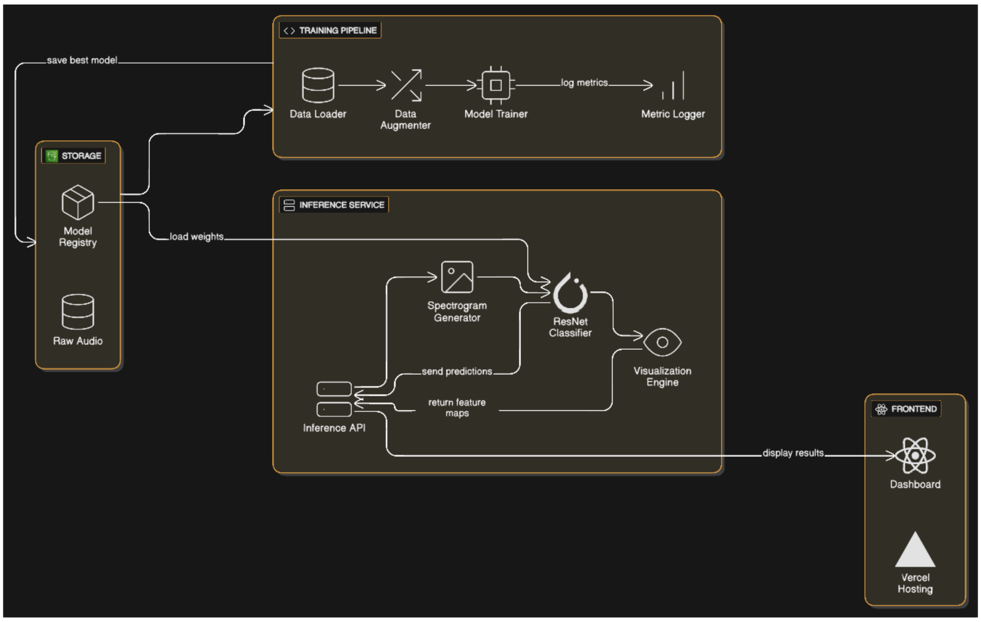

Figure 1: End-to-end system flow for Environmental Audio Classification.

8 Pseudo-Inverse Based Closed-Form Learning

Instead of iterative gradient descent, an analytical solution is obtained via the Moore–Penrose pseudo-inverse:

$$\boldsymbol{\beta} = \mathbf{H}^{\dagger}\mathbf{T}.$$

If $\mathbf{H}^T\mathbf{H}$ is invertible:

$$\mathbf{H}^{\dagger} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T.$$

This approach ensures faster convergence, numerical stability, and removes dependency on hyperparameters.

9 Toy Numerical Example: Pseudo-Inverse Solution

Consider a simple regression problem with $N = 3$ samples and $d = 2$ hidden features.

$$\mathbf{H} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$$

The objective is to compute the output weight vector $\boldsymbol{\beta}$ such that:

$$\mathbf{H}\boldsymbol{\beta} \approx \mathbf{T}.$$

Step 1: Compute $\mathbf{H}^T\mathbf{H}$

$$\mathbf{H}^T\mathbf{H} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}$$

Step 2: Invert $\mathbf{H}^T\mathbf{H}$

The determinant is:

$$\det(\mathbf{H}^T\mathbf{H}) = (3)(14) - (6)(6) = 42 - 36 = 6$$

Thus,

$$(\mathbf{H}^T\mathbf{H})^{-1} = \frac{1}{6} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix}$$

Step 3: Compute $\mathbf{H}^T\mathbf{T}$

$$\mathbf{H}^T\mathbf{T} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 5 \\ 11 \end{bmatrix}$$

Step 4: Compute Output Weights $\boldsymbol{\beta}$

$$\boldsymbol{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{T} = \frac{1}{6} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} \begin{bmatrix} 5 \\ 11 \end{bmatrix} = \begin{bmatrix} \frac{4}{3} \\ \frac{1}{6} \end{bmatrix}$$

Step 5: Verification

Predicted outputs:

$$\hat{\mathbf{T}} = \mathbf{H}\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \frac{4}{3} \\ \frac{1}{6} \end{bmatrix} = \begin{bmatrix} \frac{3}{2} \\ \frac{5}{3} \\ \frac{11}{6} \end{bmatrix}$$

These predictions closely approximate the target values, validating the pseudo-inverse solution.

10 Conclusion

This work demonstrates a mathematically grounded framework for environmental audio classification by combining signal processing theory, Residual CNN architectures, and linear algebraic optimization techniques. While deep residual networks extract robust time–frequency features, pseudo-inverse-based learning provides an efficient and analytically sound alternative to gradient descent for the output layer, strengthening both theoretical clarity and computational efficiency.