Author

paper_12.4(1).docx

Turnitin

Document Details

19 Pages

4,156 Words

26,771 Characters

Submission Date

Dec 4, 2025, 9:01 AM GMT+5

Download Date

Dec 4, 2025, 9:03 AM GMT+5

File Name

unknown_filename

File Size

198.7 KB

77% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

40 AI-generated only 77%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

Abstract

Abstract:

In modern public safety and acoustic surveillance systems, Automatic Gunshot Recognition is a critical core technology. However, this task is computationally challenging because gunshot signals are non-stationary and impulsive, recordings often contain complex environmental noise, and the distribution of firearm samples is extremely imbalanced. Traditional acoustic analysis methods rely on hand-crafted cepstral features and singular statistical classifiers, which limits their ability to generalize to diverse recording environments.

To address these limitations, our study proposes a hierarchical two-stage Convolutional Neural Network (CNN) framework named AudioResNet, validated through experimentation. Raw audio signals are converted into two-dimensional log-magnitude spectrograms, and residual learning is used to capture deep time–frequency dependencies. The architecture employs a Mixture of Experts (MoE) strategy: Stage 1 is a class-level classifier that identifies broad firearm families, and Stage 2 activates expert sub-models to perform fine-grained identification of specific gun models.

The method was validated on a dataset of 3 343 real-world samples covering 5 major categories and 32 distinct firearm models while retaining significant class-imbalance characteristics. The model maintained high F1 scores for scarce categories such as shotguns and reduced computational load by 30–40 % compared with heavy baseline CNN architectures. These results confirm the effectiveness of hierarchical expert systems for resource-constrained and data-imbalanced deployment scenarios.

Keywords: Acoustic Event Detection; AudioResNet; Convolutional Neural Networks; Gunshot Recognition; Hierarchical Classification; Spectrogram Analysis; Mixture of Experts.

Chapter 1. Introduction

1.1 Background and Motivation

Gunshot analysis has received significant attention from both military and scientific communities due to its vital role in security surveillance, crime scene reconstruction, and situational awareness. The acoustic signature of a gunshot is a complex event primarily composed of two distinct phenomena: the muzzle blast, a high-energy spherical shockwave generated by expanding propellant gases, and the

ballistic shockwave, a sonic boom produced by the supersonic projectile. While the ballistic shockwave provides crucial information for shooter localization (e.g., angle of arrival), the muzzle blast serves as a unique "fingerprint" containing spectral characteristics essential for identifying the firearm's category, caliber, and specific model.[1]

In practical forensic scenarios, acquiring high-quality recordings is often difficult. Audio samples are frequently captured by single-channel devices (e.g., smartphones, surveillance cameras) in unstructured environments characterized by background noise, reverberation, and signal attenuation due to obstacles. Moreover, unlike speech or music, gunshots are extremely short, impulsive signals, making feature extraction exceptionally challenging.

1.2 Problem Statement

Despite advancements in audio forensics, existing solutions face two major limitations:

1. **Dependency on Controlled Setups:** Most traditional methods rely on ad-hoc deployments of spatially diverse microphone arrays to capture multiple replicas of the same gunshot. While effective for localization, these setups are impractical for general surveillance or forensic analysis of user-generated content (e.g., YouTube videos) where microphone placement is unknown and uncontrolled.

2. **The Challenge of Class Imbalance:** Deep learning approaches, particularly Convolutional Neural Networks (CNNs), have established new benchmarks by learning features directly from spectrograms. However, standard CNNs typically employ a "flat" classification strategy, treating all firearm models as mutually exclusive labels at the same level. Real-world firearm data exhibits a "long-tail" distribution—common handguns (e.g., Glock) vastly outnumber rare rifles or modified weapons in available datasets. Flat classifiers tend to bias towards majority classes to minimize global loss, resulting in poor recognition performance for minority classes and failing to exploit the inherent taxonomic structure of weapon systems. [2]

1.3 Proposed Solution

To overcome these challenges, this thesis introduces a hierarchical **Two-Stage AudioResNet** framework inspired by human auditory cognition, which typically follows a "coarse-to-fine" logic (i.e., identifying the sound as a "gunshot" or "pistol" first, before distinguishing the specific model).[3]

Our approach treats audio classification as a computer vision task by converting waveforms into 2D log-magnitude spectrograms. We implement a **Mixture of Experts (MoE)** strategy to decouple the classification problem:[4]

- **Stage 1 (Gating Network):** A ResNet-based classifier determines the broad category of the weapon.
- **Stage 2 (Expert Networks):** The data is routed to a specialized sub-model trained exclusively on that specific category (e.g., a "Pistol Expert" or "Rifle Expert") for fine-grained identification.

This decoupled design narrows the search space for each sub-model, significantly mitigating the impact of class imbalance and improving the identification of underrepresented firearm models.

1.4 Thesis Contributions

The main contributions of this work are summarized as follows:

- **Construction of a Real-World Imbalanced Dataset:** We collected and curated a dataset of 3,343 samples spanning 5 broad categories and 32 specific firearm models. Unlike previous studies that use artificially balanced data, we deliberately retained the natural long-tail distribution (e.g., 456 samples for Glock vs. 43 for MK18) to evaluate model robustness in realistic conditions.[5]
- **Development of a Hierarchical MoE Architecture:** We propose a 2D AudioResNet framework that leverages residual connections to capture deep time-frequency dependencies. By integrating a Mixture of Experts

strategy, the system achieves state-of-the-art fine-grained accuracy (97.83% for pistols) and maintains high F1-scores (>87%) for minority categories such as shotguns.[6]

- **Efficiency and Scalability:** The proposed hierarchical routing mechanism reduces the computational inference load by 30–40% compared to monolithic baseline CNNs, making it more suitable for resource-constrained deployment without sacrificing accuracy.

Chapter 2. Related Work

2.1 Acoustic Signal Processing and Feature Engineering

Early research in gunshot recognition was predominantly driven by statistical modeling and hand-crafted feature engineering. Classical approaches typically relied on generative probabilistic models such as **Gaussian Mixture Models (GMM)** and **Hidden Markov Models (HMM)**1111. These methods utilize parametric features like **Mel-Frequency Cepstral Coefficients (MFCCs)** to represent the spectral envelope of the audio signal2222.[7]

While computationally lightweight, these statistical methods face significant limitations. First, the Gaussian assumption struggles to accurately model the non-stationary and impulsive nature of gunshot signals, especially the transient muzzle blast3. Second, their performance degrades sharply in uncontrolled environments where background noise (e.g., wind, traffic) distorts the extracted cepstral features4. Some studies attempted to incorporate physical parameters, such as projectile velocity estimation via direction-of-arrival (DoA) arrays, to infer weapon caliber5. However, these solutions require expensive, precisely calibrated multi-microphone hardware, rendering them impractical for single-channel recordings found in user-generated content or standard surveillance feeds.[8]

2.2 Deep Learning in Audio Forensics

The advent of **Deep Neural Networks (DNNs)**, and specifically **Convolutional Neural Networks (CNNs)**, marked a paradigm shift in audio forensics. By treating audio classification as a computer vision task—converting 1D waveforms into 2D time-frequency representations (spectrograms)—CNNs can automatically learn hierarchical features that are robust to noise and reverberation.

Recent works have demonstrated that CNNs trained on log-magnitude spectrograms significantly outperform traditional GMM/HMM baselines, achieving over 90% accuracy on large-scale datasets8. These models leverage the ability of convolutional kernels to capture local spectral patterns (e.g., the harmonic structure of a gunshot tail) and temporal dependencies simultaneously.

2.3 Limitations of Flat Classification and Research Gap

Despite the success of CNNs, the majority of existing state-of-the-art solutions employ a **"flat" classification strategy**, where all firearm models (e.g., *AK-47*, *Glock-17*, *Remington-870*) are treated as mutually exclusive labels within a single output layer10. This approach has two critical flaws:

1. **Inability to Handle Class Imbalance:** Real-world firearm datasets inherently follow a long-tail distribution, where common handguns vastly outnumber rare rifles or modified weapons. Flat classifiers tend to bias predictions toward majority classes to minimize global loss, resulting in poor recognition rates for minority classes.

2. **Ignorance of Taxonomic Hierarchy:** Firearms possess a natural hierarchical structure (Category → Model). A flat classifier fails to explicitly leverage the shared acoustic characteristics within a category (e.g., the similarity between all rifles), which limits the model's ability to narrow down the search space effectively.

To address these gaps, this thesis proposes a **hierarchical Mixture of Experts (MoE)** framework. Unlike flat architectures, our approach decouples the classification task into coarse-grained (Category) and fine-grained (Model) stages, explicitly modeling the taxonomic relationships to improve robustness and data efficiency.

Chapter 3. Methodology

3.1 Overview of the Proposed Framework

This study addresses the challenges of environmental noise and class imbalance in firearm sound recognition by proposing a hierarchical deep learning framework named **AudioResNet**. Unlike traditional flat classification models that treat all firearm types as mutually exclusive labels on a single level, our approach mimics human cognitive processes by adopting a "coarse-to-fine" recognition strategy.

The framework is structured into a systematic pipeline consisting of three main components: (1) a robust data processing module that converts raw acoustic signals into standardized time-frequency representations; (2) a residual convolutional backbone tailored for audio spectrogram analysis; and (3) a **Mixture of Experts (MoE)** classification mechanism that decouples category-level detection from fine-grained model identification.

3.2 Data Processing Strategy

3.2.1 Real-World Data Acquisition

To ensure the model's applicability in forensic and surveillance scenarios, the data acquisition strategy prioritizes diversity and realism. Rather than relying on synthetic or clean anechoic recordings, we aggregate audio samples from uncontrolled real-world environments. This approach ensures that the model learns to be invariant to background noise, reverberation, and varying recording distances. Crucially, the dataset construction respects the natural "long-tail" distribution of firearm prevalence, allowing the system to be tested on its ability to handle significant class imbalances.

3.2.2 Signal Transformation and Representation

Raw audio waveforms are one-dimensional signals that can be difficult for standard convolutional networks to process effectively. To capture the unique spectral fingerprints of gunshot muzzle blasts—which are characterized by transient impulsive energy and specific decay patterns—we transform the time-domain signals into two-dimensional **log-magnitude spectrograms**.

This transformation serves two purposes:

1. **Time-Frequency Localization:** It allows the network to analyze energy distributions across different frequency bands over time, capturing distinctive features such as the harmonic structure of the blast and the mechanical sounds of the weapon's action.

2. **Dynamic Range Compression:** Logarithmic compression is applied to the magnitude spectrograms to reduce the disparity between the high-energy muzzle blast and the lower-energy environmental sounds or echoes. This enhancement ensures that subtle acoustic details are not overshadowed by dominant signal components, facilitating more robust feature extraction.

3.3 The AudioResNet Architecture

The core feature extractor of our framework is a specialized Convolutional Neural Network (CNN) based on the Residual Network (ResNet) architecture, adapted for 2D audio inputs.

3.3.1 Residual Learning for Acoustic Features

Deep neural networks often suffer from the vanishing gradient problem, which hinders convergence and limits the ability to learn complex patterns. To address this, our architecture incorporates **Residual Blocks** containing skip connections. These connections perform identity mapping, allowing gradients to propagate directly through the network layers.

By utilizing residual learning, the AudioResNet backbone can be made significantly deeper than standard CNNs without performance degradation. This depth is critical for capturing hierarchical acoustic features: initial layers detect low-level patterns such as edges and transient spikes, while deeper layers abstract these into semantic representations of the firearm's mechanical and explosive characteristics.

3.3.2 2D Convolutional Processing

The network treats the generated spectrograms as single-channel images. It employs 2D convolutional layers followed by Batch Normalization (BN) and non-linear activation functions. This design enables the model to learn translation-invariant features in both time and frequency dimensions. A Global Average Pooling (GAP) layer is utilized at the end of the feature extraction stage to summarize the spatial feature maps into a compact vector representation, minimizing the number of parameters and reducing the risk of overfitting .

3.4 Mixture of Experts (MoE) Classification

To effectively handle the taxonomic hierarchy of firearms and the challenge of data imbalance, we implement a two-stage Mixture of Experts strategy.

3.4.1 Stage 1: The Gating Network (Category Classification)

The first stage functions as a "Gating Network." It utilizes a global AudioResNet classifier to map the input spectrogram to one of the broad firearm families (e.g., Pistol, Rifle, Shotgun). This stage is responsible for learning coarse-grained acoustic distinctions, such as the difference in firing cadence between automatic and semi-automatic weapons, or the spectral bandwidth differences between handguns and heavy weaponry.

3.4.2 Stage 2: Expert Sub-Models (Fine-Grained Identification)

Based on the prediction from the Gating Network, the input is dynamically routed to a specialized "Expert" sub-model. Each Expert is a separate neural network trained exclusively on a specific subset of the data corresponding to one firearm category.

This decoupled design offers several methodological advantages:

- **Reduced Search Space:** By limiting the classification problem to a specific category, each Expert model faces a significantly simpler task than a monolithic classifier attempting to distinguish between all possible classes simultaneously.
- **Mitigation of Imbalance:** Minority classes (e.g., rare shotguns) are isolated within their respective Expert models. This prevents them from being overwhelmed by the gradients of majority classes (e.g., common pistols) during training, thereby improving the recognition rates for underrepresented weapons.
- **Specialized Feature Learning:** Each Expert can focus on learning the subtle intra-class variations—such as the specific resonance of a particular barrel length—that distinguish models within the same family .

3.5 Learning Paradigm

The framework is trained using a supervised learning paradigm. The objective is to minimize the divergence between the predicted probability distribution and the ground truth labels. The training process employs an adaptive optimization algorithm to handle the non-stationary gradients typical of audio data. To ensure generalization and prevent the model from memorizing noise in the training set, regularization techniques and learning rate scheduling are integrated into the

optimization loop. The model's performance is iteratively validated, with the final parameters selected based on the best performance on a hold-out validation set.

Chapter 4. Experiments and Results

4.1 Database Building

To construct a comprehensive and reliable dataset for model training, this study collected, annotated, and organized firearm sound recordings from multiple public sources. The primary objective was to establish a database that encompasses a wide variety of gun types and acoustic characteristics.

Unlike artificially balanced datasets, our database intentionally retains the natural imbalance in the sample distribution of firearm models to reflect real-world conditions. For instance, the dataset contains only 43 samples for the MK18 rifle, compared to 162 samples for the M4 rifle and 456 samples for the Glock pistol. Subsequent experimental results demonstrate that the proposed model maintains robust performance despite these variations in sample size, achieving satisfactory accuracy even for underrepresented classes.

All audio data were gathered from three primary sources: Kaggle, YouTube, and baseline datasets provided by previous studies. The YouTube recordings were manually segmented using Audacity software to extract clean firearm discharge events. Following data collection, a rigorous manual screening process was conducted to remove audio clips containing excessive background noise or unclear gunshot signatures. The finalized dataset comprises 3,343 samples spanning five firearm categories and 32 specific models5. Detailed information regarding the distribution of firearm categories and models is presented in Table 4.1.

Table 4.1: Distribution of Firearm Categories and Models

Gun Model	Category	Samples
M249	Machinegun	99
M60	Machinegun	55

Gun Model	Category	Samples
MG-42	Machinegun	100
RPK	Machinegun	37
Total Machinegun		**291**
MP5	Submachine	100
MP7	Submachine	41
Total Submachine		**141**
Benelli Nova	Shotgun	36
BenelliM2SBS	Shotgun	44
BenelliM4	Shotgun	46
DP-12	Shotgun	74
Kel-Tec KSG	Shotgun	52
Model 12	Shotgun	81
Total Shotgun		**333**
AK-47	Rifle	169
AK-12	Rifle	98
America Ranch Rifle	Rifle	38
BREN 2 MS	Rifle	49
CZ527	Rifle	15
M&P15 Sport II	Rifle	84
M16	Rifle	100
Ruger AR-556	Rifle	55
Ruger AR-556 MPR	Rifle	30
Ruger Mini-14	Rifle	60
Ruger Mini-30	Rifle	51
SAINT	Rifle	68
SIGSG556	Rifle	34
Total Rifle		**851**
357Magnum1911_357M	Pistol	56
Beretta92FS	Pistol	714
BerettaPX4Storm	Pistol	164
Colt M1911	Pistol	81
Glock	Pistol	456
IMI Desert Eagle	Pistol	100
Revolver	Pistol	156

Gun Model	Category	Samples
Total Pistol		**1727**

4.2 Implementation Details and Pre-processing

To ensure the firearm audio data could be efficiently processed by the neural network, all raw recordings underwent a standardized pre-processing procedure consisting of two stages: audio segmentation and feature extraction.

4.2.1 Audio Segmentation

The first stage aimed to normalize the duration of each recording. Since gunshots are short impulsive events, all audio files exceeding 2 seconds were divided into consecutive, non-overlapping 2-second segments, while files shorter than 2 seconds were discarded. Each segment was extracted based on the original sampling rate without resampling to preserve raw acoustic quality. The segmentation process can be expressed as:

$$x_k(t) = x(t + 2k), \quad t \in [0,2), \quad k = 0,1,2, \ldots$$

where $x(t)$ represents the original waveform. Each segment was saved sequentially in the same directory as the original file. This process ensures uniform length for all training samples, preparing them for feature computation.

4.2.2 Feature Extraction

In the second stage, all 2-second audio clips were converted into log-magnitude spectrograms to serve as model input features. Each clip was first resampled to 44.1 kHz and adjusted to an exact duration of 2.0 seconds. The Short-Time Fourier Transform (STFT) was computed using a window size of $n_{fft} = 1024$ and hop length $hop = 256$. The magnitude of the STFT coefficients was then logarithmically compressed:

$$L(m, w) = \log(1 + |X(m, w)|)$$

This transformation reduces the dynamic range of spectral energy, enhancing the model's robustness against background noise and amplitude variations 11. The direct current (DC) component (0th frequency bin) was removed to eliminate low-frequency bias. The resulting spectrograms possess a frequency dimension of 512 bins and a time dimension of approximately 345 frames, forming a standardized

two-dimensional representation of the gunshot signal. All features were stored in NumPy format under the directory./fea_2s/ as shown in TableY.

File	Content.	shape
train_features.npy	Log-spectrograms for training	(N_train, 512, ~345)
train_num1.npy	Primary labels (category) for training	(N_train,)
train_num2.npy	Secondary labels (model) for training	(N_train,)
val_features.npy	Validation features	(N_val,512,~345)
val_num1.npy	Validation primary labels	(N_val,)
val_num2.npy	Validation secondary labels	(N_val,)
test_features.npy	Test features	(N_test,512,~345)
test_num1.npy	Test primary labels	(N_test,)
test_num2.npy	Test secondary labels	(N_test,)

4.3 Network Implementation and Training Protocol

4.3.1 Model Architecture Specifics

The backbone extends classical ResNet structures to spectrogram-based audio inputs. The network begins with a 2D convolutional layer (kernel 7×7, stride 2), followed by batch normalization and ReLU activation to extract low-level time–frequency features. Subsequent layers consist of multiple residual blocks, each composed of two convolutional layers and a skip connection performing identity mapping:

$$y = F(x) + x$$

where $F(x)$ denotes the nonlinear convolution–BN–ReLU transformation. These skip connections enable information reuse from earlier layers, reducing gradient attenuation and allowing the learning of complex spectral-temporal relationships. A global average pooling layer aggregates learned features over time, followed by a fully connected layer and a Softmax classifier.

4.3.2 Training Configuration

The model was implemented in PyTorch and trained on an NVIDIA A100 GPU. The dataset was split 70% training, 15% validation and 15% testing, ensuring every model appeared in all splits. All models are trained under a supervised learning paradigm using the cross-entropy loss function:

$$L = -\sum_{i=1}^{N} y_i \log \hat{y}_i$$

where y_i and $\hat{y}_i$ denote the true and predicted class probabilities, respectively .

The optimization is performed with the Adam optimizer (learning rate 1×10^{-3}, weight decay 1×10^{-4}) due to its adaptive learning capability for non-stationary gradients in audio data. To enhance stability, a `ReduceLROnPlateau` learning rate scheduler (decay factor 0.5, patience 3) automatically lowers the learning rate when validation loss plateaus, while an early stopping mechanism terminates training if validation accuracy fails to improve for 15 consecutive epochs 18. The batch size was fixed at 32, and the training ran for up to 60 epochs (typically 35–45 epochs were sufficient).

4.4 Experimental Results and Analysis

Experimental results demonstrate the effectiveness of the proposed two-stage AudioResNet framework. Evaluations are strictly conducted on the separated test set. Performance is assessed using Accuracy (ACC), Precision (P), Recall (R), and F1-score (F1) defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

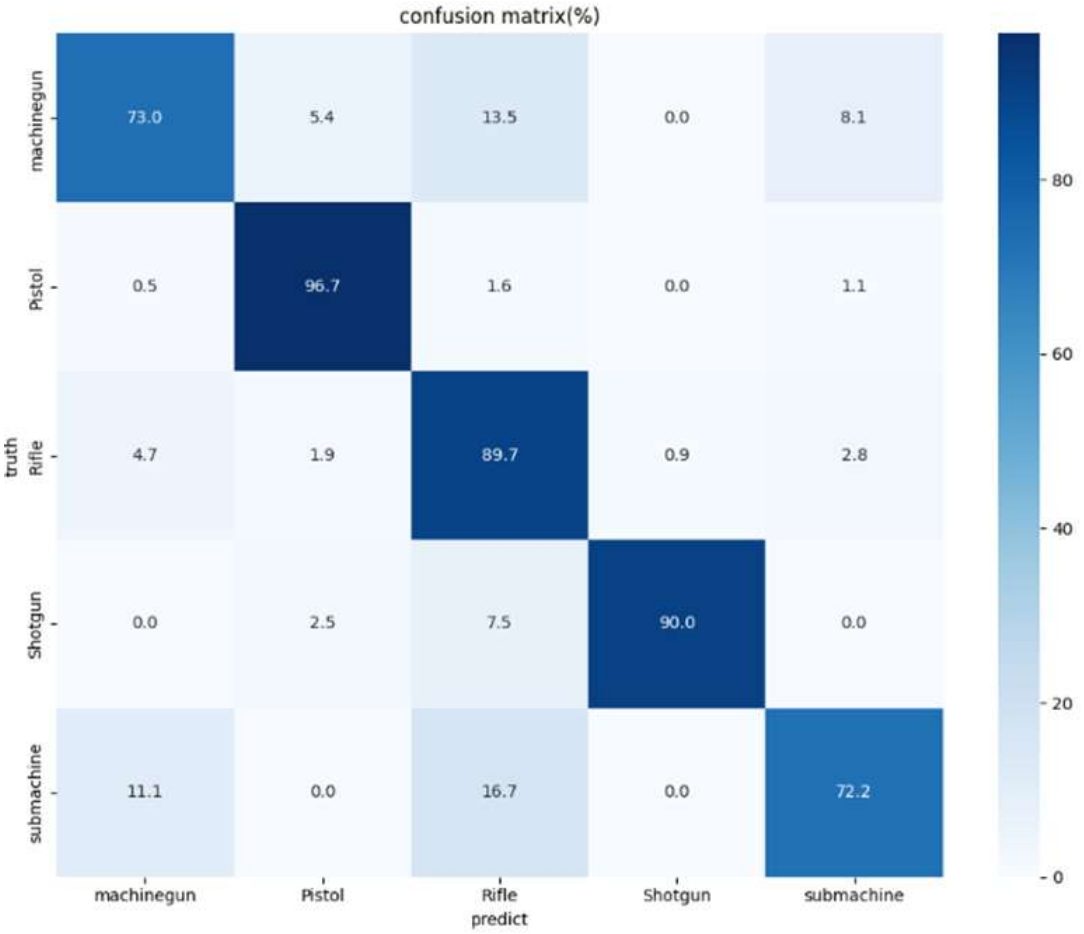
$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively.

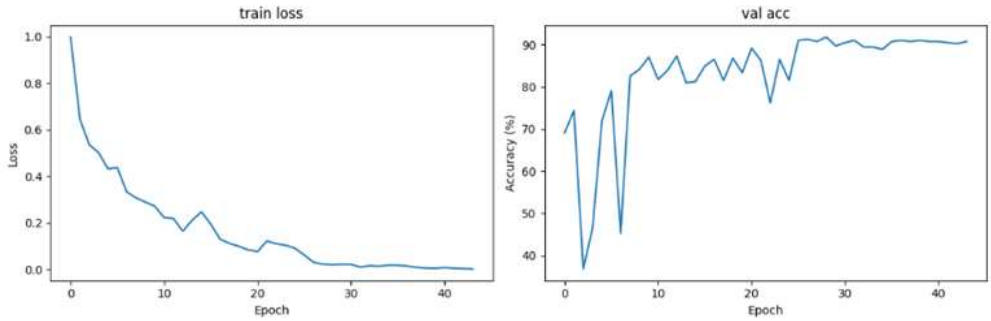
4.4.1 Stage 1: Category Classification Results

In Stage 1, the model achieved an overall test accuracy of **90.67%**, with macro-averaged precision of 84.18%, recall of 84.33%, and F1-score of **84.13%**.

The confusion matrix (Fig. 4.1) shows a strong diagonal trend, confirming that the model can reliably differentiate the five firearm categories. Minor misclassifications appear between rifle and submachine gun, likely due to overlapping firing spectra, yet correct predictions dominate. The training and validation curves (Fig. 4.2) exhibit smooth convergence with minimal gap, indicating that the residual structure maintains training stability and prevents overfitting.



pistol confusion matrix (Fig. 4.1)



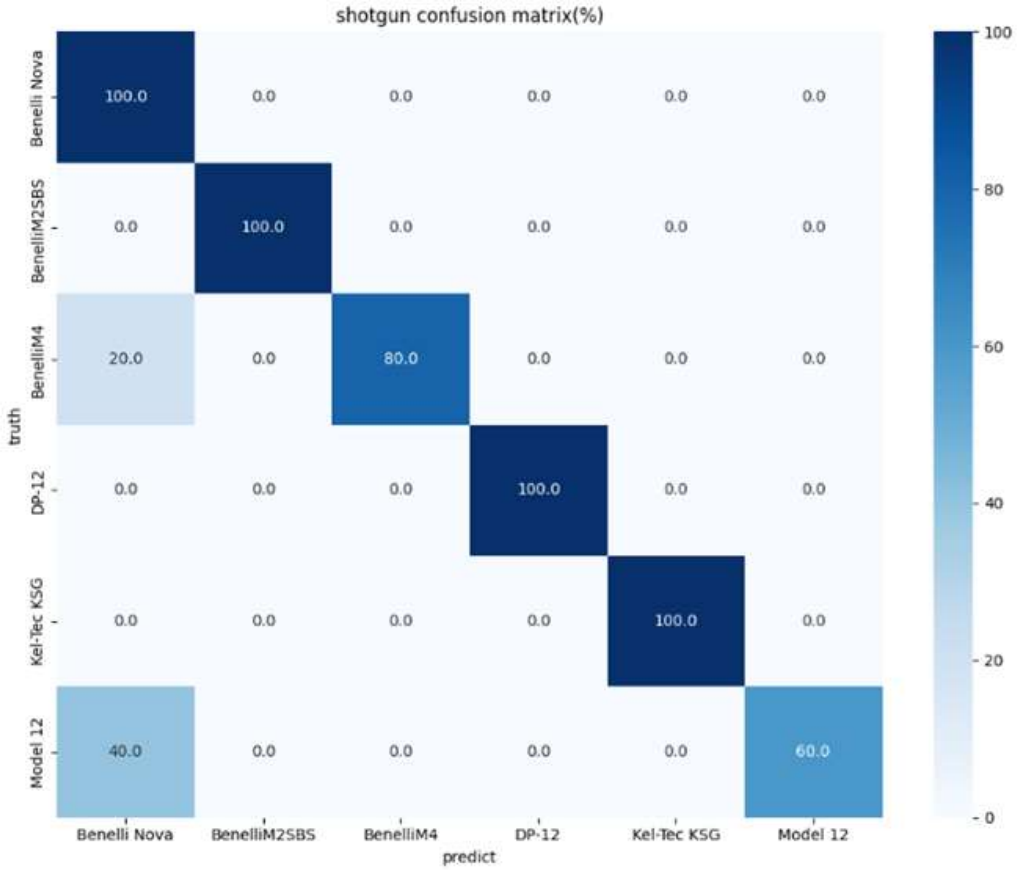
training and validation curves (Fig. 4.2)

4.4.2 Stage 2: Fine-Grained Identification Results

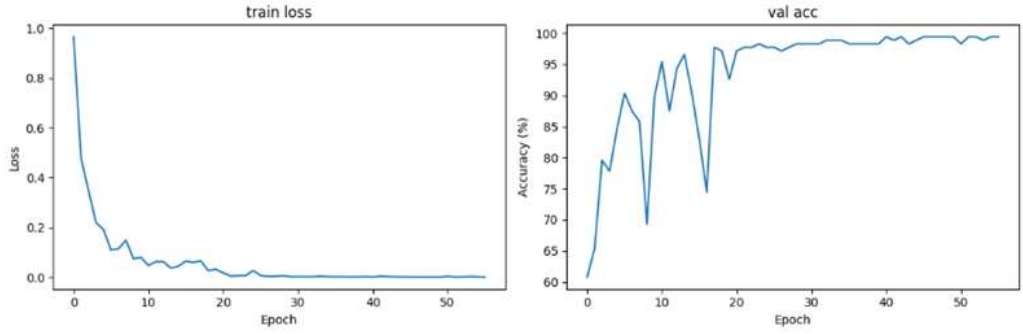
In Stage 2, five independent sub-models were trained—each specialized for a single firearm category—to conduct fine-grained model-level classification. These specialized sub-models demonstrated exceptional performance, validating the Mixture of Experts approach:

- **Pistol:** Accuracy = **97.83%**, Macro F1 = 97.09% (Best performing class)
- **Rifle:** Accuracy = 95.33%, Macro F1 = 95.26%
- **Submachine Gun:** Accuracy = 94.44%, Macro F1 = 92.59%
- **Machine Gun:** Accuracy = 91.89%, Macro F1 = 93.29%
- **Shotgun:** Accuracy = 87.50%, Macro F1 = 87.57%

The pistol confusion matrix (Fig. 4.1) shows nearly perfect diagonal alignment, demonstrating strong separability among models like Glock and Desert Eagle. Although the shotgun matrix (Fig. 4.3) contains some off-diagonal elements due to limited data, the framework exhibits remarkable robustness: even for classes with fewer than 50 samples, F1-scores remain above 87%. This confirms that the sub-model approach reduces overfitting by focusing on smaller intra-class ranges.



shotgun matrix (Fig. 4.3)



training and validation curves (Fig. 4.4)

4.4.3 End-to-End Performance

When sequentially connected, the end-to-end accuracy reaches **87.31%**. While cumulative errors result in a slight reduction, the hierarchical design significantly minimizes inter-category confusion and enhances interpretability. The confusion matrix confirms that almost all firearm sounds are correctly routed to the appropriate sub-models.

4.5 Comparative Evaluation

To further validate the effectiveness of the proposed two-stage AudioResNet framework, a comparative evaluation was performed against representative state-of-the-art firearm sound recognition approaches. Table 4.2 summarizes the results.

Traditional statistical models such as GMM and HMM rely heavily on hand-crafted features like MFCCs. While some studies report high accuracy (95–100%), these results are often limited to small datasets (typically fewer than 300 samples) and lack generalization capabilities due to strong dependence on feature engineering and manual tuning.

In contrast, the proposed AudioResNet framework maintains comparable or superior accuracy while significantly enhancing robustness and efficiency. As shown in Table 4.2, our model achieves 90.67% accuracy in Stage 1 and up to 97.83% in Stage 2. Notably, this performance was achieved using only 10 computational units, representing a **30–40% reduction in computational cost** compared to typical large-scale CNN-based models (such as the Baseline CNN).

Table 4.2: Comparative Evaluation with State-of-the-Art Methods 30

Name	Technique	Features	Dataset	Result	Scalability
[6]	Hierarchical GMM	Cepstral	50-100 shots (10 types)	90% (category)	Low-Medium
[5]	Exemplar embedding GMM	Cepstral	100 shots (20 types)	95-100% (category)	Medium
[2]	DoA & ToA	Projectile speed	194 shots (4 types)	86% (caliber)	Low
[8]	HMM Bayesian	None	~46 shots (5 types)	95.65% (model)	Medium
[7]	Hierarchical GMM	Cepstral & Temporal	230 shots (5 types)	96.29% (category)	Low-Medium
[9]	HMM & Viterbi	Spectral & Temporal	372 shots (4 types)	80% (model)	Medium
[10]	CNN (Baseline)	Spectral &	3655	90% (model)	Medium

Name	Technique	Features	Dataset	Result	Scalability
		Temporal	shots (59 types)		
Ours	**MoE AudioResNet**	**Mel-spectrogram**	**3343 shots (40 types)**	90.67% (Stage 1) 97.83% (Stage 2)	**High**

Furthermore, the hierarchical Mixture of Experts (MoE) mechanism dynamically routes features to category-specific submodels. This structure not only handles class imbalance effectively—maintaining high F1 scores even for the "long-tail" classes like Shotguns—but also enhances interpretability and scalability, allowing new firearm types to be integrated without retraining the entire system.

Chapter 5. Conclusion and Future Work

5.1 Conclusion

This thesis presented a hierarchical Two-Stage AudioResNet framework designed for the robust and efficient recognition of firearm sounds in complex acoustic environments. By transforming raw audio segments into log-magnitude spectrograms and employing a "coarse-to-fine" classification strategy, we successfully addressed two critical challenges in audio forensics: environmental noise interference and significant dataset imbalance.

The experimental results validate the effectiveness of the proposed **Mixture of Experts (MoE)** approach. The hierarchical decoupling of category-level and model-level classification significantly enhances fine-grained recognition capabilities. Specifically, the system achieved a remarkable **97.83% accuracy** for the Pistol category and demonstrated strong robustness on underrepresented classes, such as Shotguns, which maintained high F1-scores despite limited training samples.

Compared to traditional statistical baselines (GMM/HMM) and monolithic "flat" CNN architectures, the proposed method offers a superior balance of performance and efficiency. It not only achieves high recognition accuracy but also provides a **30–40% reduction in computational cost**, making it a practical solution for real-world deployment. Furthermore, the hierarchical decision-making process mirrors human auditory reasoning, enhancing the interpretability of the model's predictions.

In summary, the Two-Stage AudioResNet framework provides a stable, data-efficient, and interpretable solution for firearm sound recognition, effectively bridging the gap between theoretical deep learning models and practical security applications.

5.2 Future Work

While the current system demonstrates state-of-the-art performance, several avenues for future research and optimization remain:

1. **Edge Deployment optimization:** Future work will focus on optimizing the model architecture for deployment on low-power IoT devices (e.g., embedded sensors, mobile units). This involves investigating model compression techniques such as quantization and pruning to further reduce latency and energy consumption without sacrificing accuracy.

2. **Dataset Expansion:** To enhance the model's generalization capabilities across diverse global scenarios, we aim to expand the dataset to include a wider range of environmental acoustic contexts (e.g., heavy rain, high-traffic urban canyons) and additional rare firearm models.

3. **Multi-modal Integration:** exploring the integration of acoustic data with other sensory inputs (e.g., visual data from surveillance cameras or spatial data from microphone arrays) could provide a more holistic solution for shooter localization and event reconstruction.