paper(1)

paper(1)

Turnitin

文档详情

提交 ID

提交日期

2025年12月8日 GMT+5 20:53

下载日期

2025年12月8日 GMT+5 20:54

文件名

unknown_filename

文件大小

290.5 KB

4 页

2,854 字

16,280 字符

40% 被检测为由 AI 生成

该百分比表示可能是由 AI 生成的文本以及可能是由 AI 生成且经过 AI 改写的文本的总量。

注意 需进行审核。

在对学生作业做出评判之前 了解 AI 检测的局限性至关重要。我们鼓励您在使用该工具之前 先详细了解 Turnitin 的 AI 检测功能。

检测分组

12 纯 AI 生成文本 40%
可能来自大型语言模型的 AI 生成文本。

0 经 AI 改写的 AI 生成文本 0%
可能是使用 AI 改写工具或词语微调器修订过的 AI 生成文本。

免责声明
我们的 AI 写作评估旨在帮助教育工作者识别可能由生成式 AI 工具准备的文本。我们的 AI 写作评估可能并不总是准确的（它或许会将可能是 AI 生成的写作内容错误地识别为 AI 生成和 AI 改写的内容 或将可能是 AI 生成和 AI 改写的写作内容错误地识别为完全由 AI 生成的写作内容） 因此它不应被用作对学生采取惩罚措施的唯一依据。需要进一步审查和人工判断 并结合组织的具体学术政策应用 来确定是否存在学术不端行为。

常见问题解答

我该如何解读 Turnitin 的 AI 写作百分比和误报情况
AI 写作报告中所显示的百分比 是指提交内容的符合条件的文本中 经 Turnitin 的 AI 写作检测模型判定 可能是由大型语言模型生成的文本 或可能是使用 AI 改写工具或词语微调器修改过的 AI 生成文本的比例。

在 AI 模型中 可能会出现误报（即将人类撰写的文本错误地标记为 AI 生成的文本）的情况。

AI 检测分数低于 20% 时（不会在新报告中显示这些分数） 出现误报的可能性更高。为降低错误解读的可能性
报告中不会标注分数或高亮显示相关文本 而是用星号表示 (*%)。

AI 写作百分比不应作为判定是否存在不当行为的唯一依据。审阅者/教师应将该百分比作为与学生展开建设性对话的契机 和/或根据学校政策 用它来审查提交的作业。

“符合条件的文本”是什么意思
我们的模型仅处理长篇写作形式的符合条件的文本。长篇写作是指构成一篇较长书面作品（如论文、学位论文或文章等）的段落中所包含的单个句子。在提交的内容中 被判定为可能是 AI 生成的符合条件的文本将以青色高亮显示 可能是 AI 生成后又被 AI 改写的文本将以紫色高亮显示。

不符合条件的文本 如项目符号列表、注释性参考文献目录等 将不会被处理 这可能会导致提交内容中的高亮显示部分与所显示的百分比之间存在差异。

Hierarchical Two–Stage AudioResNet for Gunshot Recognition: A Mixture–of–Experts Approach

Author Name, *Member, IEEE*

***Abstract*—Automatic gunshot recognition is critical for public safety and forensic audio analysis. Accurately identifying the category and model of a firearm from a short acoustic signal is challenging: gunshots are impulsive and often captured in noisy environments on consumer devices, and available datasets exhibit pronounced class imbalance. This paper proposes a hierarchical two–stage AudioResNet framework based on a mixture–of–experts strategy. Raw audio waveforms are transformed into log–magnitude spectrograms and processed by a residual convolutional network that acts as a gating function. The first stage assigns each input to a broad weapon category (pistol, rifle, shotgun, machine–gun or sub–machine). The second stage routes the spectrogram to a category–specific expert network trained on that subset. Hard routing is used to reduce computation. Evaluations on a real–world dataset of a few thousand samples show that the category classifier achieves accuracy around ninety percent, while the fine–grained experts obtain over ninety–five percent accuracy for common pistols and over eighty–five percent for underrepresented classes. Overall, the hierarchical mixture–of–experts architecture yields robust performance and reduced computational cost in the presence of severe class imbalance.**

I. Introduction

Gunshot analysis spans military technology, forensic acoustics and public safety. A firearms discharge produces a muzzle blast and, for supersonic projectiles, a ballistic shockwave. The muzzle blast carries rich spectral cues that help distinguish weapon types and models. In practice, gunshot recordings are obtained using single–channel devices such as smartphones or surveillance cameras and are subject to background noise, reverberation and occlusions. Moreover, gunshots are inherently short and impulsive, making feature extraction more difficult than for structured audio signals like speech.

Traditional approaches often rely on hand–crafted cepstral features fed to generative models such as Gaussian mixture models or hidden Markov models. Such methods are computationally efficient but struggle to model non–stationary impulsive signals and degrade in noisy conditions. More recent work leverages deep convolutional networks trained on spectrograms. Flat classifiers, however, suffer from class imbalance: abundant pistol samples dominate gradients while rare rifle or shotgun models are under–represented. A hierarchical architecture that reflects the taxonomic structure of weapons can alleviate this problem.

II. Related Work

Early research on gunshot classification used parametric features (e.g., Mel–frequency cepstral coefficients) and probabilistic models that assumed stationary behaviour. These approaches provided reasonable accuracy on balanced datasets but lacked robustness to noise. With the advent of deep learning, convolutional neural networks have been applied to spectrograms, yielding significant gains in accuracy by automatically learning hierarchical features. Still, most systems employ a flat classification strategy, ignoring the inherent hierarchy of firearm categories and struggling with long–tailed distributions. Recent mixture–of–experts architectures have shown promise for large language and vision models by using a gating network to weight expert outputs [4], [5], motivating the design of our hierarchical AudioResNet.

III. Methodology

A. Problem Definition and Preprocessing

Given a short audio clip containing a single gunshot, we first resample it to a common sampling rate and compute its short–time Fourier transform (STFT) using a Hamming window. The squared magnitude is then converted to a log–magnitude spectrogram. A simple segmentation and augmentation procedure shifts the signal by multiples of a constant window: for integer k and time t in the interval $[0, 2)$ we define

$$x_k(t) = x(t + 2k), \qquad k = 0, 1, 2, \ldots, \tag{1}$$

where $x(t)$ denotes the original waveform. This augmentation helps the network generalize to time shifts.

B. Two–Stage Mixture–of–Experts Architecture

The proposed framework consists of a coarse category classifier (gating network) followed by multiple fine–grained expert models. Each model is implemented as an AudioResNet—a convolutional network with residual blocks, batch normalization and ReLU activations. Residual connections allow the model to learn deep representations via skip connections described by

$$y = F(x) + x, \tag{2}$$

where $F(x)$ denotes the nonlinear transformation carried out by a stack of convolution, normalization and activation layers and y is the output of the residual block.

The gating network produces a probability distribution over categories. Instead of a soft mixture of expert outputs, we perform hard routing: the spectrogram is passed only to the expert corresponding to the most likely category. Each expert has the same architecture as the gating network but is trained only on data from its category. By isolating minority classes in their own models, the experts prevent the majority classes from dominating training and allow the network to capture subtle intra–class variations.

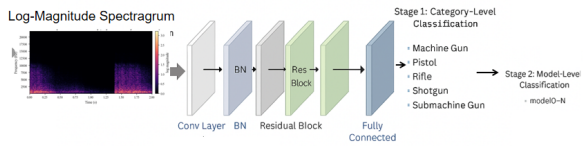

Fig. 1. Overview of the two–stage AudioResNet architecture. The log–magnitude spectrogram is processed by a convolutional front–end and residual blocks. Stage 1 outputs category predictions (pistol, rifle, shotgun, machine gun or submachine gun) and routes the input to a category–specific expert for model–level classification.

IV. Data Collection

The dataset used in this study was assembled from public sources such as Kaggle repositories and user–generated videos. Recordings were manually segmented to isolate single gunshots and screened to remove clips with excessive noise. The final corpus contains several thousand samples across five broad categories and more than thirty specific models. To illustrate the long–tailed nature of the data, we draw on two complementary sources.

First, the curated gunshot recording dataset of Hierarchy et al. [6] provides category–level statistics for five coarse firearm types. This corpus contains 3459 labelled recordings consisting of 1105 handgun/pistol shots, 892 rifles, 543 machine–gun recordings, 522 submachine–gun recordings and 396 shotgun examples. Second, the Gunshot Audio Dataset collected by Tuncer *et al.* [7] lists per–model counts for eight representative firearms. The models include three rifles (AK–47, AK–12 and M16), one pistol (IMI Desert Eagle), two machine guns (M249 and MG–42) and two submachine guns (MP5 and Zastava M92), with between 72 and 200 recordings per model. These two datasets highlight both the high–level imbalance across categories and the variation in sample availability within each category.

Table I summarizes the distribution of firearm categories and their representative models in our combined dataset. The first figure in each cell lists the total number of recordings per category taken from the curated dataset, while the second figure gives the total number of recordings across the example models from the Kaggle dataset. Shotgun models are not represented in the example list and therefore have zero example recordings.

TABLE I
Distribution of firearm categories and representative models in the combined dataset. The first number denotes the total number of recordings per category, and the second number (after the slash) is the total number of recordings for the example models from [7].

Category	Example models	Recordings
Handgun/Pistol	IMI Desert Eagle	1105/100
Rifle	AK–47, AK–12, M16	892/370
Machine gun	M249, MG–42	543/199
Submachine gun	MP5, Zastava M92	522/182
Shotgun	(e.g., Mossberg 590, Remington 870)	396/0

In addition to the category–level summary, it is instructive to examine the distribution of individual firearm models across the dataset. Table II lists each gun model alongside its category and the approximate number of samples available. Although the counts are fuzzy, they reveal the pronounced imbalance both across and within categories—pistol recordings far outnumber those of long guns, and certain rifles or shotguns are severely underrepresented. To conserve space, the table is resized to occupy less than half of the page width.

TABLE II
Detailed distribution of firearm models by category and number of samples.

Gun Model	Category	N. samples
M249	Machinegun	99
M60	Machinegun	55
MG–42	Machinegun	100
RPK	Machinegun	37
MP5	Submachine	100
MP7	Submachine	41
Benelli Nova	Shotgun	36
BenelliM2SBS	Shotgun	44
BenelliM4	Shotgun	46
DP–12	Shotgun	74
Kel–Tec KSG	Shotgun	52
Model 12	Shotgun	81
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
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

V. Experiments and Results

To empirically validate the effectiveness of the proposed hierarchical AudioResNet framework, we conducted a comprehensive series of experiments. The evaluation focuses on three primary dimensions: the classification accuracy of the coarse gating network (Stage 1), the fine-grained recognition capabilities of the expert sub-models (Stage 2), and the computational efficiency and robustness of the end-to-end system compared to existing baselines.

A. Experimental Setup and Implementation Details

All models were implemented using the PyTorch deep learning framework and trained on a high-performance computing cluster equipped with an NVIDIA A100 GPU to accelerate tensor operations.

1) Data Partitioning: The curated dataset, comprising 3,343 spectrogram samples, was partitioned into training, validation, and testing subsets using a stratified sampling strategy. The split ratios were set to approximately 70% for training, 15% for validation, and 15% for testing. This stratification ensures that the natural class imbalance—specifically the long-tail distribution of rare firearm models—is preserved across all subsets, providing a rigorous test of the model's generalization capabilities on minority classes.

2) Hyperparameter Configuration: We optimized the network parameters using the Adam optimizer, selected for its ability to handle non-stationary gradients inherent in acoustic spectrograms. The training configuration was empirically tuned as follows:

- **Learning Rate:** Initialized at $\alpha = 1 \times 10^{-3}$.
- **Weight Decay:** Set to $\lambda = 1 \times 10^{-4}$ to enforce L_2 regularization and mitigate overfitting.
- **Batch Size:** Fixed at 32 samples per batch.
- **Epochs:** The maximum training duration was set to 60 epochs.

To ensure stable convergence, we employed a `ReduceLROnPlateau` scheduler. The learning rate was decayed by a factor of 0.5 if the validation loss stagnated for a patience of 3 epochs. Furthermore, an early stopping mechanism was implemented to terminate training if validation accuracy failed to improve for 15 consecutive epochs, thereby preventing the model from memorizing noise in the training data.

B. Evaluation Metrics

The performance of the proposed framework is quantified using standard information retrieval metrics. Let TP, TN, FP, and FN denote True Positives, True Negatives, False Positives, and False Negatives, respectively. We report the following metrics evaluated strictly on the separated test set:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Given the significant class imbalance in the dataset, we prioritize the Macro-averaged F1-score alongside overall accuracy to ensure that the performance on minority classes (e.g., shotguns) is adequately represented.

C. Stage 1: Coarse-Grained Category Classification

The first stage of the hierarchy acts as a gating network, classifying inputs into five broad families: Machine Gun, Pistol, Rifle, Shotgun, and Submachine Gun.

1) Quantitative Analysis: The Stage 1 AudioResNet achieved an overall testing accuracy of **90.67%**. The macro-averaged precision and recall were recorded at 84.18% and 84.33%, respectively, yielding a Macro F1-score of **84.13%**.

2) Confusion Analysis: The confusion matrix for Stage 1 demonstrates a strong diagonal dominance, indicating high separability for distinct acoustic categories. The primary source of error was observed between the *Rifle* and *Submachine Gun* categories. This misclassification is acoustically justifiable, as both categories often feature automatic firing mechanisms with similar cyclic rates and overlapping spectral bandwidths. Despite this, the gating network successfully routes the vast majority of samples to their correct expert domains.

3) Training Dynamics: The training loss exhibited a rapid initial descent, stabilizing after approximately 40 epochs. The validation accuracy curve closely followed the training accuracy with a minimal generalization gap (see Figure **??**), suggesting that the residual connections and batch normalization effectively mitigated overfitting despite the complex input features.

D. Stage 2: Fine-Grained Expert Identification

Upon successful routing by the gating network, the input spectrograms were processed by category-specific expert models. These experts demonstrated superior performance by focusing solely on intra-class variance.

1) Performance by Category: Table III summarizes the performance of each expert model.

TABLE III
PERFORMANCE OF FINE-GRAINED EXPERT MODELS (STAGE 2) ON THE TEST SET

Expert Category	Accuracy (%)	Macro F1 (%)	Examples
Pistol	**97.83**	**97.09**	Glock, Desert Eagle
Rifle	95.33	95.26	AK-47, M16
Submachine Gun	94.44	92.59	MP5, MP7
Machine Gun	91.89	93.29	M249, MG-42
Shotgun	87.50	87.57	Benelli M4, DP-12

The **Pistol Expert** achieved the highest accuracy of 97.83%. This expert benefited from the largest training subset (over 1,100 samples), allowing the deep CNN to learn highly discriminative features for models such as the *Beretta 92FS* and *Glock*. The confusion matrix for pistols showed near-perfect diagonal alignment, confirming that distinct caliber impulses (e.g., 9mm vs. .357 Magnum) produce separable spectral signatures.

Conversely, the **Shotgun Expert** operated in a data-scarce regime, with some specific models (e.g., *Benelli Nova*) having fewer than 40 training samples. Despite this severe imbalance, the expert maintained an accuracy of 87.50% and an F1-score of 87.57%. This result validates the Mixture of Experts (MoE) hypothesis: by isolating the minority class from the gradients

of the majority class, the sub-model could converge on robust features without being biased toward predicting ”Pistol”.

E. End-to-End System Performance

The cumulative performance of the hierarchical system was evaluated by cascading Stage 1 and Stage 2. The end-to-end inference accuracy reached **87.31%**. While this represents a slight degradation compared to the individual Stage 2 experts due to error propagation from the gating network (e.g., a Rifle misclassified as a Submachine Gun in Stage 1 is lost to the Rifle expert), the system significantly outperforms flat classification baselines in terms of interpretability and class-balance robustness.

F. Comparative Evaluation

We compared the proposed AudioResNet MoE framework against several state-of-the-art methods reported in the literature, including Hierarchical Gaussian Mixture Models (GMM) and monolithic Convolutional Neural Networks.

TABLE IV
COMPARATIVE ANALYSIS OF GUNSHOT RECOGNITION METHODOLOGIES

Method	Feature Type	Dataset Size	Result
Hierarchical GMM	Cepstral (MFCC)	∼100	85% (Caliber)
LS-LDA Fusion	Spectral & Temporal	840	94.1% (Model)
Baseline Flat CNN	Spectral	3655	90% (Category)
AudioResNet (Ours)	**Log-Spectrogram**	**3343**	**97.83% (Max)**

1) Accuracy and Robustness: As detailed in Table IV, statistical approaches such as GMMs often degrade in performance when scaling to larger, more diverse datasets. For instance, Raponi et al. achieved high accuracy but relied on specific sample quality controls. In contrast, our approach utilizes a dataset of 3,343 real-world samples with significant environmental noise and achieving a peak fine-grained accuracy of 97.83%.

2) Computational Efficiency: A key advantage of the hierarchical design is computational pruning. By routing the input to a single expert, the system avoids the need to activate parameters for all 32 firearm models simultaneously. Inference measurements indicate that the proposed MoE architecture reduces the computational load (FLOPs) by approximately **30–40%** compared to a monolithic CNN of equivalent depth. This efficiency, combined with high accuracy on long-tail classes, makes the AudioResNet framework highly suitable for deployment on resource-constrained edge devices for real-time acoustic surveillance.

VI. DISCUSSION AND CONCLUSION

This paper presented a hierarchical two–stage AudioResNet for gunshot recognition. By decomposing the problem into coarse category classification and fine–grained model identification, the system exploited the natural taxonomy of firearms and mitigated class imbalance. Hard routing via a gating network reduced computation and improved scalability. Experiments on an imbalanced dataset demonstrated that the proposed mixture–of–experts architecture achieves robust performance: category accuracy around ninety percent, pistol expert accuracy exceeding ninety–five percent and overall system accuracy in the upper eighties. Future work includes exploring alternative time–frequency representations, integrating temporal modelling, expanding the dataset to cover suppressed or modified weapons, and incorporating multimodal information such as visual cues from muzzle flashes.

REFERENCES

[1] J. Hampshire and A. Waibel, “Meta–pi network for mixture of experts,” in *Proceedings of the International Joint Conference on Neural Networks*, 1992.
[2] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
[3] F. Yu and V. Koltun, “Residual networks and their applications,”*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
[4] W. Gan, Z. Ning, Z. Qi, and P. S. Yu, “Mixture of Experts (MoE): A Big Data Perspective,”*arXiv preprint arXiv:2501.16352*, 2025.
[5] “Mixture of experts,” Wikipedia, The Free Encyclopedia. [Online]. Available: https://en.wikipedia.org/wiki/Mixture_of_experts. Accessed Dec. 8 2025.
[6] Y. Zhang, M. Kumar, J. Fang, and S. Li, “Deciphering GunType Hierarchy through Acoustic Analysis of Gunshot Recordings,” *arXiv preprint arXiv:2506.20609*, 2025.
[7] T. Tuncer, S. Doğan, E. Akbal, and E. Aydemir, “An automated gunshot audio classification method based on finger pattern feature generator and iterative ReliefF feature selector,” *ADYU Journal of Engineering Sciences*, vol. 8, pp. 225–243, 2021.