IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 3, June 2026, pp. 2166∼2180
ISSN: 2088-8708, DOI: 10.11591/ijai.v15.i3.pp2166-2180

From audio to image: gunshot classification using Mel spectrogram convolutional neural networks

Peerapol Khunarsa[1], Pafan Doungpaisan[2]
[1]Department of Data Science, Faculty of Science and Technology, Uttaradit Rajabhat University, Uttaradit, Thailand
[2]Department of Information Technology, Faculty of Industrial Technology and Management,
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

Article Info

Article history:

Received Nov 23, 2024
Revised Feb 26, 2026
Accepted May 7, 2026

Keywords:

Audio classification
Convolutional neural networks
Deep learning
Gunshot detection
Machine learning
Mel spectrogram
Spectrogram analysis

ABSTRACT

Accurate identification of firearm types from acoustic signals is essential for modern public safety and forensic applications. Traditional gunshot analysis methods often rely on physical evidence or handcrafted audio features, which can be unreliable under noisy and reverberant conditions. This study presents a systematic investigation of gunshot sound classification using Mel spectrogram representations and convolutional neural networks (CNNs). Raw audio signals are transformed into Mel spectrogram images, enabling firearm classification to be formulated as an image recognition problem. Thirteen CNN architectures, ranging from lightweight to deep models, are evaluated under a unified experimental protocol to analyze both classification performance and computational efficiency. Experiments are conducted on a publicly available multi-firearm dataset recorded in semi-controlled real-world environments. The results demonstrate that Mel spectrogram–based CNN models achieve classification accuracy exceeding 94%, while moderate-complexity architectures provide a favorable balance between accuracy and efficiency. The findings highlight the importance of representation–architecture alignment and offer practical design guidelines for selecting deployable CNN models in real-time gunshot detection systems.

Corresponding Author:

Pafan Doungpaisan
Department of Information Technology, Faculty of Industrial Technology and Management
King Mongkut's University of Technology North Bangkok
1518 Pibulsongkram Road, Bangsue, Bangkok 10800, Thailand
Email: pafan.d@itm.kmutnb.ac.th

1. INTRODUCTION

The global increase in firearm-related violence has intensified the demand for rapid and reliable gunshot identification systems. Artificial intelligence (AI)-based approaches provide a non-invasive and real-time alternative to traditional ballistic analysis by exploiting distinctive acoustic characteristics of gunfire, including frequency content, amplitude, and temporal dynamics. These capabilities are particularly valuable for law enforcement and public safety applications that require immediate situational awareness.

In the United States alone, more than 600 mass shootings were reported in 2022, motivating the widespread deployment of acoustic gunshot detection systems in major cities. Technologies such as ShotSpotter employ distributed acoustic sensors to detect and localize gunfire events, significantly reducing response times in urban environments. Similar challenges are observed globally, particularly in Latin

America, where countries such as Brazil and Mexico account for a substantial proportion of firearm-related homicides [1]–[3]. Unlike conventional firearm identification methods that rely on physical evidence such as bullets or shell casings, AI-based gunshot detection systems operate directly on acoustic signals captured by distributed sensors. Recent advances in deep learning and signal processing have further improved the robustness and scalability of these systems across diverse environments, enabling rapid firearm identification, localization, and timely law enforcement response [4]–[8].

Early AI-based approaches relied on handcrafted audio features, including spectral descriptors and Mel-frequency cepstral coefficients (MFCCs). Although effective to some extent, these features often struggled under noisy and reverberant conditions due to the highly impulsive and broadband nature of gunshot sounds [9]. Consequently, recent research has increasingly adopted time–frequency representations, particularly Mel spectrograms, as primary inputs for deep learning models. Mel spectrograms provide perceptually motivated time–frequency representation that effectively captures rapid onset transients and broadband energy characteristics of gunfire. By transforming raw audio signals into two-dimensional representations, gunshot detection can be formulated as an image classification problem, enabling the effective use of convolutional neural networks (CNNs) [10]. Recent studies have shown that CNNs trained on Mel spectrograms consistently outperform traditional feature-based approaches in terms of accuracy and robustness. For instance, Aggarwal *et al.* [11] combined Mel spectrograms with a convolutional–gated recurrent unit (GRU) architecture to enhance the classification performance in acoustically complex environments, while Goldwater *et al.* [12] demonstrated reliable discrimination between gunshots and other impulsive sounds such as fireworks and thunder.

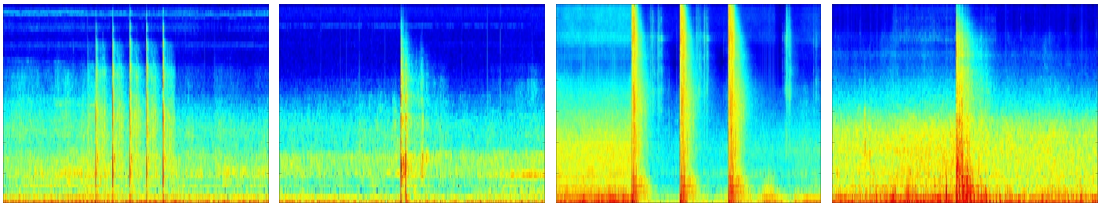
Despite these advances, accurate gunshot detection in real-world environments remains challenging due to environmental noise, reverberation, and variability in firearm acoustic signatures across diverse deployment scenarios, including urban public spaces, schools, and outdoor or conservation environments. A key difficulty lies in extracting discriminative features from highly impulsive and non-stationary gunshot signals, for which traditional time-domain descriptors and handcrafted spectral features often exhibit limited robustness and poor generalization under noisy and reverberant conditions [13]–[15]. In contrast, Mel spectrograms provide a perceptually grounded time–frequency representation that effectively captures rapid onset transients and broadband energy distributions characteristic of firearm discharges, thereby offering improved robustness and suitability for deep learning–based gunshot classification. As illustrated in Figure 1, different firearm types exhibit distinct Mel spectrogram patterns, highlighting their suitability for fine-grained gunshot classification. The horizontal axis denotes time (s), and the vertical axis represents frequency (Hz) on Mel scale. The spectrograms were generated using FFT size of 512, a hop length of 256, and 128 Mel bands.

As illustrated in Figure 1(a), the Smith and Wesson (.38) exhibits concentrated low-frequency energy patterns. Figure 1(b) presents the Glock 17 Gen3 (9mm) Mel spectrogram with sharper transient structures. Figure 1(c) shows the Remington 870 (12-G) with broader spectral distributions. Figure 1(d) illustrates the Ruger AR-556 (.223) with stronger high-frequency components.

By transforming raw audio signals into image-like Mel spectrogram representations, gunshot detection can be reformulated as an image recognition problem. This enables CNNs to exploit spatial and hierarchical feature learning, improving robustness across acoustically complex environments. However, the effectiveness of Mel spectrogram-based classification is strongly influenced by the choice of CNN architecture, as different models vary in their capacity to learn meaningful spectro-temporal patterns. Unlike our previous work, which compared multiple time–frequency representations, this study fixes the Mel spectrogram as the input feature and instead focuses on a systematic evaluation of CNN architectures. By jointly analyzing classification performance and computational scalability through a unified Big-O complexity formulation, the proposed framework provides insights into accuracy–efficiency trade-offs and supports informed architecture selection for deployable gunshot detection systems.

This study addresses the aforementioned challenges by proposing an enhanced gunshot classification framework that integrates Mel spectrogram–based feature extraction with a systematic evaluation of multiple state-of-the-art CNN architectures. In contrast to prior studies that assess only one or a limited number of models, this work conducts a comprehensive comparison of thirteen CNN architectures, ranging from lightweight networks such as SqueezeNet and EfficientNet-B0 to deeper models including ResNet50, DenseNet-201, and InceptionResNetV2. A publicly available dataset comprising gunshot recordings from multiple firearm types captured in a semi-controlled real-world environment is employed, thereby enhancing the ecological validity and reproducibility of the experimental results [16]. Related studies on gunshot sound

classification, including CNN-based and hybrid deep learning approaches, are summarized in Table 1.



(a) (b) (c) (d)

Figure 1. Mel spectrogram images of different firearm types: (a) Smith and Wesson (.38), (b) Glock 17 Gen3 (9mm), (c) Remington 870 (12-G), and (d) Ruger AR-556 (.223)

Table 1. Comparison of representative works on gunshot sound classification

Reference	Dataset	Audio representation	Model / method	Key outcome
Aggarwal *et al*. [11], 2023	Gunshot	Mel spectrogram	CNN–GRU	Improved classification in acoustically complex environments
Bajzik *et al*. [10], 2020	Gunshot	Spectrogram	CNN	High accuracy for impulsive sound detection
Singh *et al*. [9], 2021	Gunshot-like sounds	Handcrafted features	Machine learning	Robust performance under noisy recordings
Raponi *et al*. [4], 2022	Gunshots	Raw audio samples	Deep learning	Application to digital forensics of firearm audio
This work	Multi-firearm dataset	Mel spectrogram	CNN (13 architectures)	Accuracy $>$ 94% with systematic architectural evaluation

The main contributions of this study are summarized as follows:

- This paper provides a systematic analysis of Mel spectrogram-based gunshot sound classification, focusing on the interaction between time–frequency representation design and CNN architectural capacity rather than proposing a new task-specific model.
- It demonstrates that classification performance is primarily driven by representation–architecture alignment, and that increasing network depth or complexity alone leads to performance saturation under fixed Mel spectrogram representations.
- A unified evaluation framework is established that jointly considers classification performance and computational complexity, enabling explicit analysis of accuracy–efficiency trade-offs across a wide range of CNN architectures.
- Through extensive experiments on a multi-firearm dataset recorded under realistic acoustic conditions, the study provides empirical evidence that moderate-complexity CNNs can achieve competitive performance when paired with perceptually grounded representations.
- The findings yield practical and reproducible design guidelines for selecting CNN architectures in deployable gunshot detection systems under real-world computational constraints.

The remainder of this paper is organized as follows. Section 2 presents the literatur review of this study. Section 3 presents the problem formulation, research objectives, dataset characteristics, and associated constraints. Section 4 reports the experimental results and provides a detailed discussion and comparative analysis of proposed method. Finally, section 5 concludes paper and outlines directions for future research.

2. LITERATURE REVIEW

Recent research on gunshot detection and firearm-type classification has increasingly leveraged deep learning to improve robustness under diverse acoustic conditions. Prior works can be broadly grouped into: i) audio representation design for discriminative feature extraction and ii) modeling choices that balance accuracy and computational efficiency for deployment. Audio representations for gunshot analysis: a

common strategy is to transform raw waveforms into time–frequency representations, such as short-time fourier transform (STFT)-based spectrograms and Mel spectrograms, which expose transient and broadband gunfire patterns as image-like inputs suitable for convolutional learning. Studies using spectrogram-based representations report improved discrimination between gunshots and background noise sources in real-world settings [6], [10], [12]. In parallel, MFCCs and handcrafted descriptors remain competitive for certain datasets, but their performance can degrade under strong noise and reverberation due to the impulsive and non-stationary nature of gunshot signals [9]. More recent approaches also consider efficient training and inference for rare-event detection and practical deployments [17], [18].

Deep learning models for gunshot detection and classification: CNNs are widely adopted for gunshot detection and firearm-type classification because they effectively learn spatial and hierarchical patterns from time–frequency representations, such as spectrogram images, while offering favorable accuracy–latency trade-offs for real-time and near–real-time applications. By exploiting local receptive fields and weight sharing, CNNs are particularly well suited for capturing the transient and broadband spectral characteristics of impulsive gunshot sounds. Beyond conventional CNN-based models, several studies have explored hybrid architectures that combine convolutional feature extractors with temporal modeling components, such as recurrent neural networks (RNNs) or gated recurrent units (GRUs). These CNN–RNN and CNN–GRU frameworks aim to better capture the temporal evolution of gunshot impulses across consecutive time frames, often improving robustness in acoustically complex environments at the cost of increased computational complexity and latency [11].

More recently, transformer-based models have been investigated for gunshot-related audio tasks due to their ability to model long-range dependencies and global contextual relationships in sequential data. While these models can achieve competitive performance, they typically require substantially greater computational resources and larger training datasets, which may limit their applicability in low-power or resource-constrained deployment scenarios [6]. Beyond gunshot-specific applications, broader studies on deep learning–based audio classification have analyzed common architectural patterns, representation choices, and learning paradigms across diverse sound domains [19]. These studies provide general insights into model design trends for audio data but are not tailored to impulsive firearm sound classification. In addition to task-specific models, general methodological studies provide broader perspectives on deep learning model taxonomies, architectural design principles, and transfer learning strategies for classification problems [20], [21]. Such works are referenced here solely to provide background context on deep learning architectures, rather than as direct solutions for gunshot detection or firearm classification tasks.

Summary and motivation of this study: evidence from music information retrieval (MIR) demonstrates that Mel spectrograms combined with CNNs enable hierarchical feature learning and strong performance across audio classification tasks, including chord recognition and pitch-related estimation [22], [23]. Additional MIR studies show that deeper CNNs can improve discriminative performance for isolated or monophonic sounds, although architecture choice and evaluation protocols strongly influence the observed gains [24]–[26]. However, systematic evaluations of modern CNN architectures with explicit accuracy–efficiency considerations remain limited in firearm audio domains. Motivated by these gaps, this study fixes Mel spectrograms as the input representation and systematically evaluates thirteen CNN architectures under a unified and reproducible protocol, with an emphasis on accuracy–efficiency trade-offs for deployable firearm-type classification.

Optimization trends and open challenges: beyond model selection, recent work highlights optimization strategies for practical deployment, including data augmentation, robustness improvements, and efficiency-oriented training to reduce false positives and latency in noisy environments. Remaining challenges include improving robustness under overlapping impulsive sounds, reducing false alarms, and achieving reliable performance on resource-constrained edge devices [27]. Table 1 summarizes representative gunshot sound classification studies in terms of datasets, audio representations, modeling approaches, and reported outcomes [4], [9]–[11]. Compared with prior works that often evaluate a small number of models, this study provides a broader cross-architecture comparison using a unified Mel spectrogram representation and reports both classification performance and computational considerations.

3. PROBLEM FORMULATION AND CONSTRAINTS

This study aims to classify firearm types based on gunshot audio recordings by leveraging Mel spectrogram representations and modern CNN architectures. The problem is formulated as a multi-class audio classification task, where each input audio signal is transformed into a time–frequency representation

and mapped to a corresponding firearm category. The primary challenge arises from the subtle acoustic differences between firearm types, which are influenced by factors such as weapon design, ammunition, and recording environment. These variations often overlap in the time–frequency domain, making discrimination difficult for traditional feature-based or shallow learning methods. To address these challenges, this work employs Mel spectrogram–based feature extraction in combination with multiple CNN architectures capable of learning hierarchical spectral patterns. The main constraints of the problem include acoustic similarity between certain firearm classes, environmental variability in real-world recordings, and the need to balance classification accuracy with computational efficiency for practical deployment.

3.1. Research objectives

This study aims to investigate the effectiveness of Mel spectrogram representations and CNN architectures for gunshot sound classification. The primary research objectives are summarized as follows:

- Improve classification accuracy: develop a system to accurately identify firearms from gunshot sounds by leveraging advanced audio features and CNN architectures, with a focus on precise analysis of distinct acoustic characteristics.
- Analyze audio features: utilize Mel spectrograms techniques to extract and analyze key audio features from gunshot recordings, providing valuable insights into the unique properties of each sound.
- Evaluate CNN architectures: train and assess a variety of CNN models, including ResNet-18, ResNet-50, DenseNet-201, EfficientNet-b0, Inception-ResNet-v2, and others, to identify the most effective architecture for gunshot sound classification.
- Ensure model robustness: implement 5-fold cross-validation to validate model performance, reduce overfitting, and ensure that the models can generalize well across different subsets of data.

3.2. Scope of research

This subsection defines the scope and limitations of the proposed study. It includes the dataset characteristics and firearm categories. It also covers the selected feature extraction and deep learning models considered in the experimental evaluation.

- Dataset: the dataset used in this study was originally provided by Kabealo *et al*. [16] and consists of a comprehensive collection of 2,443 gunshot audio recordings captured from multiple firearm types and orientations. The data were collected in a semi-controlled outdoor firing range environment using edge devices strategically deployed at different locations to record gunshots from various angles, thereby ensuring diverse and realistic acoustic signatures. All recordings were captured in stereo format at a sampling rate of 44.1 kHz with a bit rate of 128 kbps, providing high-fidelity audio suitable for detailed acoustic analysis. The distribution of audio files across the four firearm categories considered in this study is summarized in Table 2.
- Firearm types: this research focuses on four specific firearm types, namely Smith and Wesson .38 caliber, Glock 17 Gen3 9 mm, Remington Model 870 12-gauge, and Ruger AR-556 .223 caliber. These firearms represent distinct weapon categories, including revolvers, semi-automatic pistols, shotguns, and rifles, which exhibit different acoustic characteristics.
- Feature and model selection: the study concentrates on Mel spectrogram–based audio representations in combination with selected CNN architectures. While other audio features and deep learning models may also be applicable to gunshot classification, they are beyond the scope of this work and are not considered in the present analysis.

Table 2. Types of guns and number of audio files in dataset 1

Gun type	Number of files
Smith and Wesson .38 caliber	519
Glock 17 Gen3 9 mm	730
Remington model 870 12-gauge	466
Ruger AR-556 .223 caliber	728

3.3. Proposed method overview

The overall workflow of the proposed gunshot classification system is illustrated in Figure 2. The workflow illustrates the main stages from raw audio input, preprocessing, Mel spectrogram generation,

CNN-based training and inference, to performance evaluation. The pipeline begins with raw gunshot audio signals recorded in waveform format, which are first subjected to preprocessing steps including resampling, normalization, and framing with windowing to ensure consistency across all audio samples. Subsequently, Mel spectrograms are generated to transform the one-dimensional audio signals into two-dimensional time–frequency representations that effectively capture perceptually meaningful acoustic patterns. These Mel spectrogram images serve as input features for training and evaluating multiple CNN architectures, including both lightweight and deep models. Finally, the trained CNN models are evaluated using standard performance metrics derived from the confusion matrix, including accuracy, precision, recall, F1-score, specificity, false positive rate (FPR), and area under the curve (AUC). These metrics provide a comprehensive assessment of the model's classification performance by jointly capturing correctness, sensitivity, and robustness to false alarms.

Importantly, the use of Mel spectrograms as input representations plays a critical role in achieving reliable evaluation outcomes. By mapping the frequency components of gunshot signals onto a perceptually motivated Mel scale, the spectrograms emphasize acoustically meaningful patterns while reducing sensitivity to minor spectral variations and background noise. This representation enhances the discriminative capability of CNN models, enabling clearer separation between firearm classes and more stable decision boundaries. As a result, the evaluation metrics reflect not only high overall accuracy but also balanced precision and recall, indicating effective mitigation of misclassification and false positive errors in acoustically complex scenarios. The implementation of the proposed pipeline was carried out using MATLAB R2024a, leveraging the deep learning toolbox and audio toolbox. These tools provide efficient support for audio preprocessing, Mel spectrogram generation, CNN training, and GPU-accelerated computation, ensuring reliable and reproducible experimental results under consistent technical constraints.

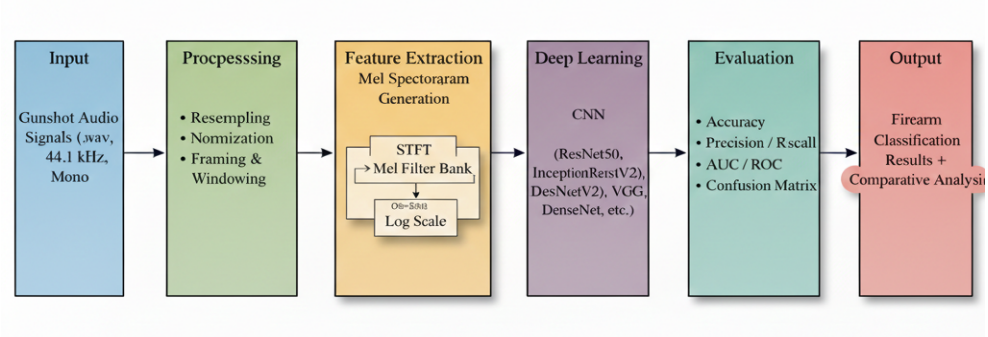


Figure 2. Overview of proposed end-to-end gunshot classification pipeline using Mel spectrograms and CNNs

3.4. Analyze audio features: Mel spectrograms

Mel spectrograms are widely adopted for analyzing non-stationary acoustic signals, such as gunshots, due to their ability to represent localized time–frequency characteristics in a perceptually meaningful manner. Gunshot sounds are impulsive, broadband, and highly transient, exhibiting rapid energy variations across a wide frequency range. Conventional time-domain features or linear frequency representations often fail to capture these properties effectively, particularly under noisy or reverberant conditions. In contrast, Mel spectrograms provide a compact and robust representation by aligning spectral resolution with human auditory perception, making them especially suitable for firearm sound analysis.

All Mel spectrograms were generated in MATLAB using the melSpectrogram function with fixed parameters to isolate the effect of CNN architectural design. Each input waveform was loaded at its native sampling rate f_s and converted to mono by channel averaging when stereo recordings were provided. A Hamming analysis window of length $N = 512$ samples was applied with an overlap of 256 samples (hop size $H = 256$), and the STFT was computed with an FFT length of $N_{\mathrm{FFT}} = 512$. The magnitude spectrum was then projected onto a Mel filter bank with $M = 128$ Mel bands spanning the full frequency range $[0, f_s/2]$ (Nyquist limit). Finally, log-compression was applied to obtain image-like inputs using $\log(S)$, where S denotes the Mel spectrogram matrix. Under these settings, the time resolution is determined by the hop size (H/f_s seconds per frame), while the frequency coverage is fixed to the Nyquist range and discretized into 128 Mel bins.

The generation of a Mel spectrogram follows a structured signal-processing pipeline, as illustrated in

Figure 3. First, the raw audio signal is loaded and preprocessed through normalization and, when necessary, resampling to ensure consistency across recordings. The signal is then segmented into short, overlapping frames and windowed to capture local temporal dynamics. Subsequently, the STFT is applied to obtain a time-dependent frequency representation. The resulting power spectrum is projected onto the Mel scale using a bank of overlapping triangular filters, followed by logarithmic amplitude compression. This sequence preserves discriminative spectral cues while suppressing insignificant variations caused by background noise or recording conditions, thereby enhancing feature stability for subsequent learning stages.

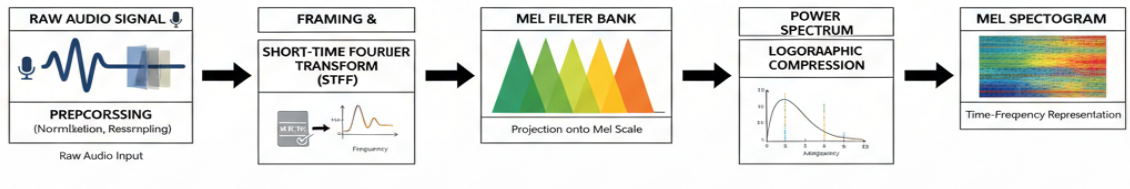


Figure 3. Processing pipeline for generating a Mel spectrogram from an audio signal

The Mel scale is a nonlinear frequency transformation designed to approximate the human auditory system's perception of pitch. It provides higher resolution at lower frequencies, where perceptual sensitivity is greater, and progressively coarser resolution at higher frequencies. The relationship between a linear frequency f (in Hertz) and its Mel-scale representation m is defined in (1).

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

Formally, the computation of a Mel spectrogram consists of the following steps:

– STFT: the audio signal is divided into short, overlapping frames to capture local temporal variations. A Fourier transform is then applied to each frame, producing a time-frequency representation. The STFT of a signal $x(t)$ is defined in (2).

$$X(t,f) = \int_{-\infty}^{\infty} x(\tau)\, w(t-\tau)\, e^{-j2\pi f\tau}\, d\tau \tag{2}$$

Where $w(t)$ denotes the window function that controls spectral leakage and the time–frequency trade-off.

– Projection onto the Mel scale: the linear-frequency spectrum obtained from the STFT is mapped onto the Mel scale using a bank of overlapping triangular filters. This step emphasizes perceptually relevant frequency bands while reducing redundancy in higher-frequency regions that are less informative for discriminating firearm sounds.

– Logarithmic amplitude compression: the Mel-filtered power spectrum is converted to a logarithmic scale to compress dynamic range variations and approximate the nonlinear loudness perception of the human auditory system. This operation improves numerical stability and enhances low-energy spectral components that may carry discriminative information.

– Resulting representation: the final Mel spectrogram is a two-dimensional time-frequency representation, where the horizontal axis corresponds to time, the vertical axis represents Mel-scaled frequency bins, and the intensity encodes signal energy.

From a machine learning perspective, Mel spectrograms enable the transformation of one-dimensional audio signals into image-like representations that are well suited for CNNs. The localized spectro-temporal patterns of gunshot sounds, such as rapid onset transients and broadband energy distributions, are captured as spatial features within the spectrogram. CNNs can exploit these patterns through hierarchical convolutional filters, leading to improved discrimination between firearms with similar acoustic signatures.

Computational complexity analysis: In addition to their representational advantages, Mel spectrograms offer a tractable computational profile. For an audio signal of length L, using an STFT window size of N with overlap O, the computational complexity of the STFT stage is given in (3).

$$T_{\text{STFT}} = O\left(\frac{L}{O} \cdot N \log N\right) \tag{3}$$

Where $\frac{L}{O}$ denotes the number of analysis frames and $N \log N$ corresponds to the computational cost of the Fast Fourier Transform (FFT) per frame. As shown in (3), the overall computational cost of the STFT grows linearly with the signal length and logarithmically with the window size.

The projection onto the Mel scale using M Mel filters has a computational complexity of (4).

$$T_{\text{Mel-Filter}} = O(M \cdot N) \tag{4}$$

Where M denotes the number of Mel filters and N represents the number of frequency bins. As shown in (4), the computational cost of the Mel filterbank grows linearly with both M and N.

In addition, the logarithmic amplitude compression introduces an additional computational cost given by (5).

$$T_{\text{Log}} = O(M \cdot N) \tag{5}$$

As indicated in (5), the logarithmic scaling step also incurs linear computational overhead with respect to the Mel spectrogram dimensions. Consequently, the overall computational complexity for generating a Mel spectrogram can be expressed as (6).

$$T_{\text{Mel spectrogram}} = O\left(\frac{L}{O} \cdot N \log N + M \cdot N\right) \tag{6}$$

Where L is signal length, N is FFT window size, O is overlap, and M denotes the number of Mel frequency bands. As summarized in (6), the overall computational cost is dominated by the STFT operation, while the Mel filterbank and logarithmic scaling introduce an additional linear overhead. This complexity demonstrates that Mel spectrogram generation is computationally efficient and scalable, making it suitable for both offline analysis and real-time or near-real-time firearm sound classification systems. As a result, Mel spectrograms form a critical bridge between raw acoustic signals and deep learning models in this study, supporting accurate firearm classification and reliable performance evaluation in realistic acoustic environments.

4. RESULTS AND DISCUSSION

This study presents an effective approach for gunshot sound classification by combining Mel spectrogram-based feature extraction with multiple CNN architectures. The experimental results demonstrate that the proposed framework achieves high classification performance across various evaluation metrics, including precision, recall, F1-score, specificity, FPR, and AUC. The experiments were conducted using a comprehensive firearm dataset reported in [9], [16], which includes gunshot recordings from multiple firearm types. The results indicate that classification performance is strongly influenced by the selected CNN architecture, with deeper models generally achieving superior accuracy and robustness. These findings confirm the effectiveness of Mel spectrograms and CNN-based models for reliable gunshot detection in realistic acoustic environments.

4.1. Performance comparison of convolutional neural network architectures using Mel spectrograms

This subsection presents a comparative evaluation of multiple CNN architectures using Mel spectrograms for gunshot sound classification. Performance was assessed using standard metrics, including accuracy, precision, recall, F1-score, specificity, FPR, and AUC, as summarized in Table 3. The results indicate that deeper architectures consistently outperform lightweight models. In particular, ResNet50 and InceptionResNetV2 achieved the highest classification accuracies of 94.83% and 94.51%, respectively, along with high precision, recall, and AUC values. These architectures benefit from residual connections and hybrid inception modules, which enhance feature reuse and enable effective learning of fine-grained spectro-temporal patterns present in Mel spectrograms.

Other deep models, including VGG-16, VGG-19, DenseNet-201, and ResNet-101, also demonstrated strong performance, achieving accuracies above 93%. DenseNet-201 benefits from dense feature propagation, while VGG-based architectures effectively capture spatial structures through deep stacks of convolutional layers. These models further exhibited high specificity and low FPR, indicating reliable discrimination between firearm classes and reduced false alarms. Moderate-complexity architectures such as AlexNet and EfficientNet-B0 achieved competitive performance, with accuracies exceeding 91%. Notably, AlexNet reached an accuracy of 93.62%, highlighting that well-designed Mel spectrogram representations can compensate for

reduced architectural depth. Such models present a favorable balance between computational efficiency and classification accuracy, making them suitable for near–real-time applications.

In contrast, lightweight architectures primarily optimized for computational efficiency, including SqueezeNet and the proposed custom CNN, exhibited comparatively lower classification performance. SqueezeNet achieved an accuracy of 88.55%, while the custom CNN attained an accuracy of 81.73%. This performance gap is mainly attributed to the limited representational capacity of lightweight and shallow architectures in capturing subtle spectral variations between acoustically similar firearm classes. A detailed analysis and discussion of the custom CNN performance, including a direct comparison with existing gunshot detection models, is provided in subsection 4.3.

Overall, the results confirm that classification performance improves with increasing architectural depth and connectivity. While high-capacity models such as ResNet50 and InceptionResNetV2 are preferable for applications requiring maximum accuracy and robustness, lighter models remain viable alternatives in resource-constrained or low-latency deployment scenarios. This trade-off highlights the importance of selecting CNN architectures based on both performance requirements and computational constraints.

Table 3. Performance metrics for Mel spectrogram with CNN architectures

CNN architecture	Accuracy (%)	Precision	Recall	F1-score	Specificity	FPR	AUC
AlexNet	93.62	0.89908	0.98	0.93780	0.96657	0.03344	0.99507
DenseNet-201	93.62	0.95960	0.94059	0.95000	0.98784	0.01216	0.99553
EfficientNet-b0	91.20	0.89320	0.92	0.90640	0.96657	0.03344	0.98629
GoogLeNet	89.67	0.92632	0.88	0.90256	0.97879	0.02121	0.98531
InceptionResNetV2	94.51	0.95000	0.95	0.95000	0.98485	0.01515	0.99402
Inception-v3	92.32	0.90385	0.93069	0.91707	0.96951	0.03049	0.98929
ResNet18	93.06	0.92079	0.92079	0.92079	0.97576	0.02424	0.98993
ResNet50	94.83	0.92381	0.97	0.94634	0.97561	0.02439	0.99647
ResNet101	93.95	0.96809	0.90099	0.93333	0.99088	0.00912	0.99451
SqueezeNet	88.55	0.94318	0.82178	0.87831	0.98485	0.01515	0.97755
VGG-16	93.06	0.97917	0.94	0.95918	0.99390	0.00610	0.99565
VGG-19	93.67	0.96040	0.9604	0.96040	0.98780	0.01220	0.99730
Custom architecture	81.73	0.80582	0.80208	0.80192	0.93772	0.06227	0.94940

4.2. Confusion matrix analysis

This subsection focuses on the four most representative and high-performing CNN architectures, namely ResNet50, InceptionResNetV2, VGG19, and EfficientNet-B0. Their class-wise confusion matrices are illustrated in Figure 4, while the overall classification accuracy comparison across CNN architectures is shown in Figure 5. Overall, all four models exhibit strong diagonal dominance, indicating reliable class-wise discrimination when trained on Mel spectrogram representations. The results demonstrate that deeper CNN architectures can effectively capture discriminative spectro-temporal patterns associated with different firearm categories. Figure 4(a) illustrates the confusion matrix of ResNet50, which achieves the most balanced classification performance across firearm categories with minimal off-diagonal errors. Figure 4(b) presents the results of InceptionResNetV2, demonstrating strong robustness against inter-class confusion through multi-scale feature extraction and residual learning. Figure 4(c) shows the VGG19 confusion matrix with stable classification behavior across most firearm types due to its deep convolutional feature hierarchies. Figure 4(d) illustrates the EfficientNet-B0 results, highlighting a favorable trade-off between classification accuracy and computational efficiency while maintaining a compact model size.

The remaining CNN architectures, including AlexNet, GoogLeNet, Inception-v3, DenseNet-201, ResNet18, ResNet101, and SqueezeNet, exhibit varying degrees of classification performance. While these models generally maintain dominant diagonal elements, they show higher levels of inter-class confusion, particularly between acoustically similar handgun categories. Lightweight and shallower architectures tend to struggle in fully exploiting fine-grained Mel spectrogram details, whereas deeper models without optimal architectural design do not consistently outperform the selected top-performing networks. Consequently, these architectures are considered supplementary for comparative analysis rather than primary candidates for high-reliability firearm sound classification.

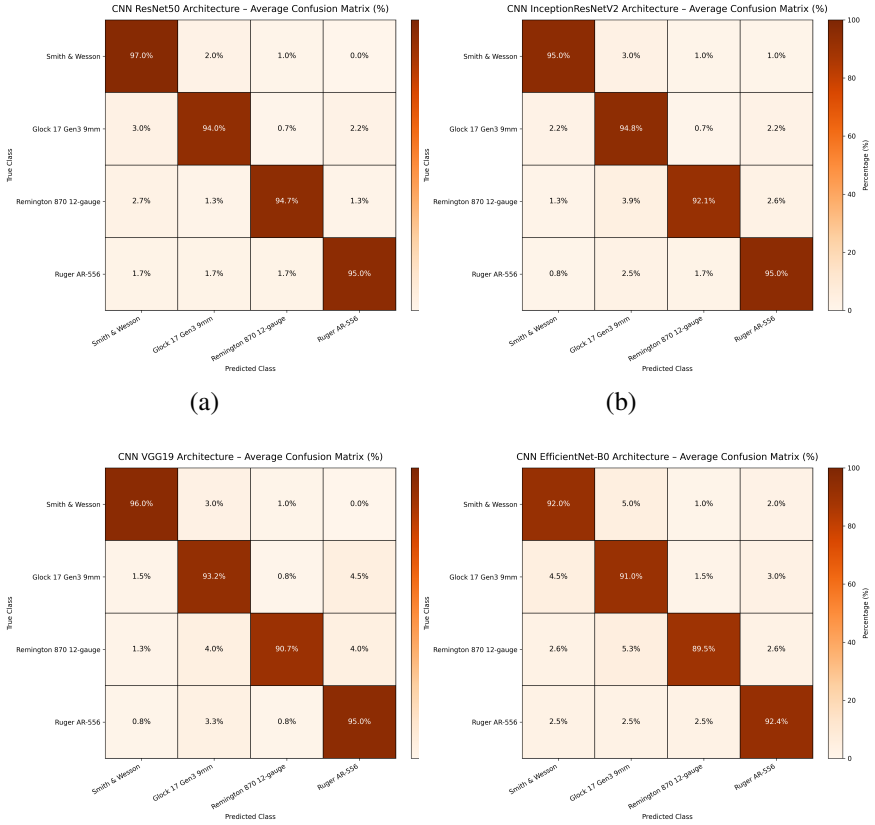


Figure 4. Average confusion matrices of the four best-performing CNN architectures trained on Mel spectrogram representations: (a) ResNet50, (b) InceptionResNetV2, (c) VGG19, and (d) EfficientNet-B0

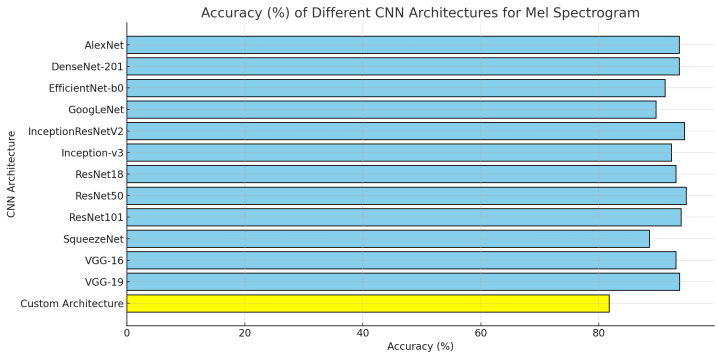


Figure 5. Performance comparison of CNN architectures using Mel spectrogram features for gunshot classification

4.3. Performance comparison between proposed method and existing gunshot detection models

This subsection compares the performance of the proposed custom architecture with an existing gunshot detection model reported by Aggarwal *et al.* [11]. To ensure a fair and reproducible comparison, the same Mel spectrogram generation parameters and CNN configuration described in the reference study were adopted. Mel spectrograms were generated using an FFT size of 512, a hop length of 256 (50% overlap), and 128 Mel frequency bands spanning from 0 Hz to the Nyquist frequency. A Hann window function was applied, and logarithmic amplitude scaling was performed using $\log(1+1000\times S)$, consistent with the baseline

methodology. The custom architecture consisted of four convolutional layers with batch normalization, rectified linear unit (ReLU) activation, and max-pooling, followed by fully connected layers with 128 neurons and dropout regularization. The network was trained using the Adam optimizer with a learning rate of 1×10^{-4}, a batch size of 32, and 20 epochs. Under these conditions, the custom architecture achieved an accuracy of 81.73%, with corresponding precision, recall, and F1-score values of 0.80582, 0.80208, and 0.80192, respectively. While these results demonstrate that the baseline model can effectively learn discriminative patterns from Mel spectrograms, its performance remains noticeably lower than that of more advanced CNN architectures evaluated in this study.

A broader comparison with twelve established CNN models—including AlexNet, DenseNet-201, EfficientNet-B0, GoogLeNet, Inception-v3, InceptionResNetV2, ResNet variants, SqueezeNet, and VGG architectures—reveals that state-of-the-art networks consistently outperform the custom architecture. In particular, ResNet50 and InceptionResNetV2 achieved accuracies exceeding 94%, while VGG-19 and DenseNet-201 also surpassed 93%. The performance gap, ranging from approximately 10% to 13% in accuracy, highlights the limitations of shallow or manually designed architectures when applied to complex gunshot classification tasks.

These findings indicate that while the custom architecture serves as a useful baseline for benchmarking and methodological comparison, deeper CNN architectures with residual, dense, or inception-based connectivity are more effective in exploiting the fine-grained spectro-temporal information encoded in Mel spectrograms. The results further emphasize the trade-off between computational efficiency and classification accuracy. Lightweight models offer reduced computational cost but lower performance, whereas deeper architectures achieve superior accuracy at the expense of increased computational and memory requirements. Overall, this comparison underscores the importance of selecting CNN architectures that balance performance and efficiency according to deployment constraints. For high-accuracy gunshot detection systems, deep residual or hybrid architectures are preferable, while simpler models may be suitable for real-time or resource-limited applications.

4.4. Analysis of time complexity using Big-O notation

This section analyzes the computational complexity of the proposed gunshot classification pipeline using Big-O notation. The analysis focuses on two primary components: i) Mel spectrogram computation as the feature extraction stage, and ii) the CNN architectures employed for classification. By jointly considering these components, the overall computational cost of the system can be systematically characterized and compared across different model configurations.

4.4.1. Computational complexity of convolutional neural network architectures

In this study, multiple CNN architectures with varying depths and structural designs were employed to classify Mel spectrogram representations of gunshot audio signals. These architectures differ significantly in their computational demands. Due to variations in convolutional depth, kernel sizes, feature map dimensions, and connection strategies (e.g., residual or dense connections). To facilitate a clear comparison, the CNN architectures used in this research are summarized and ranked in ascending order of computational complexity, as shown in Table 4. The ranking is based on the dominant convolutional operations, which typically constitute the primary computational cost in CNN-based models. Here, H_i and W_i denote the spatial dimensions of the feature maps at layer i, K_i represents the kernel size, and F_i is the number of output feature maps. For architectures with fully connected layers, N_{FC} and F_L correspond to the number of neurons and input features of the final classification layer, respectively. The term C_{dense} accounts for the additional cost introduced by dense connectivity patterns in DenseNet architectures.

This complexity ranking highlights the trade-off between computational efficiency and representational capacity. Lightweight models such as the custom CNN and SqueezeNet require minimal computational resources, making them suitable for real-time or embedded gunshot detection systems. EfficientNet-B0 offers a favorable balance between accuracy and efficiency by leveraging depth-wise separable convolutions and compound scaling. Moderately complex architectures, including AlexNet and GoogLeNet, provide reliable performance with manageable computational costs. In contrast, deeper models such as ResNet-50, ResNet-101, InceptionResNetV2, and DenseNet-201 exhibit substantially higher computational demands but are capable of extracting more discriminative spectro-temporal features from Mel spectrograms. These architectures are particularly effective in scenarios where high classification accuracy is prioritized, and sufficient computational resources are available.

Table 4. Order of computational complexity for CNN architectures

CNN architecture	Computational complexity (Big-O)	Remarks
Custom CNN	Minimal	Lightweight architecture for low-resource environments
SqueezeNet	$O(\sum H_i W_i K_i^2 F_i)$	Parameter-efficient fire modules
EfficientNet-B0	$O\left(\frac{\sum H_i W_i K_i^2 F_i}{C}\right)$	Depth-wise separable convolutions and compound scaling
AlexNet	$O(\sum H_i W_i K_i^2 F_i + N_{\text{FC}} F_L)$	Shallow model with moderate fully connected cost
GoogLeNet	$O(\sum H_i W_i K_i^2 F_i)$	Inception modules reduce redundant computation
ResNet-18	$O(\sum H_i W_i K_i^2 F_i)$	Shallow residual architecture
VGG-16	$O(\sum H_i W_i \cdot 3^2 \cdot F_i)$	Deep architecture with fixed small kernels
VGG-19	$O(\sum H_i W_i \cdot 3^2 \cdot F_i)$	Increased depth compared to VGG-16
Inception-v3	$O(\sum H_i W_i K_i^2 F_i)$	Deeper multi-branch convolutional structure
InceptionResNetV2	$O(\sum H_i W_i K_i^2 F_i)$	Combined inception and residual connections
ResNet-50	$O(\sum H_i W_i K_i^2 F_i)$	Deeper residual network
ResNet-101	$O(\sum H_i W_i K_i^2 F_i)$	Very deep residual architecture
DenseNet-201	$O(\sum H_i W_i K_i^2 F_i + C_{\text{dense}})$	High complexity due to dense feature reuse

4.4.2. Implications for Mel spectrogram-based gunshot detection

When combined with the Mel spectrogram feature extraction stage, the overall computational complexity of the proposed system is dominated by the convolutional layers of the CNN models rather than the spectrogram generation process itself. As discussed in the previous subsection, Mel spectrogram computation scales as $O\left(\frac{L}{O} N \log N + MN\right)$, which remains tractable for typical audio lengths and parameter settings. Consequently, the selection of CNN architecture plays a decisive role in determining whether the system can operate in real-time or is more suitable for offline analysis. In real-world gunshot detection applications, such as urban surveillance or security monitoring, environmental noise, reverberation, and acoustic variability pose significant challenges. Mel spectrograms provide a perceptually grounded and noise-robust representation, while appropriately selected CNN architectures enable a balance between computational efficiency and classification performance. Therefore, understanding the computational complexity of both feature extraction and classification stages is essential for designing deployable and scalable AI-based gunshot detection systems.

4.4.3. Relationship between computational complexity and classification performance

The empirical results presented in Table 3 provide important insights that complement the theoretical time-complexity analysis discussed in the previous subsection. Although several CNN architectures share similar asymptotic Big-O orders due to convolution-dominated operations, their actual classification performance when applied to Mel spectrogram representations differs noticeably. This observation highlights that Big-O notation captures asymptotic growth behavior but does not fully reflect the constant factors, architectural efficiencies, or feature reuse mechanisms that influence real-world model performance. Lightweight architectures such as the custom CNN and SqueezeNet exhibit the lowest computational complexity, which aligns with their suitability for resource-constrained or real-time applications. However, their classification accuracy remains comparatively limited, achieving only 81.73% and 88.55%, respectively. This performance gap indicates that while reduced computational cost is beneficial, insufficient network depth and representational capacity may limit the ability to capture discriminative spectro-temporal patterns present in Mel spectrograms of gunshot signals.

Mid-range architectures, including AlexNet, EfficientNet-B0, GoogLeNet, and ResNet-18, demonstrate a more balanced trade-off between computational efficiency and classification performance. Despite having similar Big-O complexity orders, EfficientNet-B0 leverages depth-wise separable convolutions to reduce constant factors, while AlexNet benefits from a relatively shallow structure. These models achieve accuracies above 90%, confirming that Mel spectrograms provide sufficiently rich representations even for moderately complex CNNs.

High-capacity architectures such as ResNet-50, ResNet-101, VGG-16, VGG-19, Inception-v3, InceptionResNetV2, and DenseNet-201 consistently yield superior performance, with accuracies exceeding 93% and AUC values approaching unity. Notably, ResNet-50 achieves the highest accuracy of 94.83%, while VGG-19 and DenseNet-201 exhibit strong precision–recall balance and low FPRs. These results confirm that deeper architectures are more effective at exploiting the fine-grained spectro-temporal structures encoded in Mel spectrograms. Nevertheless, their increased depth and connectivity introduce higher computational overhead, which is consistent with the elevated complexity terms identified in the Big-O analysis.

Overall, the combined theoretical and empirical analysis demonstrates that the dominant computational cost in Mel spectrogram-based gunshot classification systems arises from the CNN inference stage rather than the feature extraction process. While Big-O notation provides a principled framework for comparing scalability, the experimental results emphasize the importance of considering performance metrics such as accuracy, AUC, and FPR when selecting an architecture. Consequently, the optimal model choice depends on the intended deployment scenario, balancing computational constraints against the required level of detection accuracy and reliability.

4.4.4. End-to-end computational complexity of the proposed framework

The overall computational complexity of the proposed gunshot classification framework consists of two main stages: i) Mel spectrogram generation for feature extraction, and ii) CNN-based inference for classification. Accordingly, the end-to-end time complexity can be expressed as the sum of the complexities of these two components. As discussed previously, the computational complexity of Mel spectrogram generation for an audio signal of length L, using an STFT window size N with overlap O and M Mel frequency bands, is given by (7).

$$T_{\text{Mel}} = O\left(\frac{L}{O} \cdot N \log N + M \cdot N\right) \tag{7}$$

Which combines the complexity of the STFT stage in (3) and the Mel filterbank projection in (4). As shown in (7), the overall computational complexity of Mel spectrogram generation is dominated by the STFT operation, with an additional linear cost introduced by the Mel filterbank.

The computational complexity of CNN inference is dominated by convolutional operations across all layers and can be generally expressed as (8).

$$T_{\text{CNN}} = O\left(\sum_{i=1}^{D} H_i \cdot W_i \cdot K_i^2 \cdot F_i\right) \tag{8}$$

Where D denotes the total number of convolutional layers, H_i and W_i represent the spatial dimensions of the feature map at layer i, K_i is the convolutional kernel size, and F_i denotes the number of output feature maps. As shown in (8), the inference complexity grows linearly with the number of layers and feature map dimensions, and quadratically with the kernel size.

By combining both stages, the total end-to-end computational complexity of the proposed Mel spectrogram-based CNN framework can be formulated as (9).

$$T_{\text{total}} = O\left(\frac{L}{O} \cdot N \log N + M \cdot N + \sum_{i=1}^{D} H_i \cdot W_i \cdot K_i^2 \cdot F_i\right) \tag{9}$$

Which integrates the Mel spectrogram extraction complexity in (7) and the CNN inference complexity in (8). As shown in (9), the overall computational cost is dominated by convolutional operations in deep CNN layers, while the feature extraction stage introduces a comparatively lower overhead.

In practical deployment scenarios, the Mel spectrogram computation term grows linearly with the signal length and remains relatively small compared to the convolutional inference cost, especially for deep architectures. Consequently, the CNN component dominates the overall runtime, particularly for high-capacity models such as ResNet-50, ResNet-101, DenseNet-201, and InceptionResNetV2. This end-to-end formulation provides a unified theoretical framework for analyzing scalability and efficiency, enabling systematic comparison between different CNN architectures under identical feature extraction settings. Moreover, it clarifies the trade-off between computational cost and classification performance, which is essential for selecting suitable models for real-time or offline gunshot detection applications.

5. CONCLUSION

This study presented a systematic investigation of gunshot sound classification using Mel spectrogram representations in combination with modern CNN architectures. By transforming firearm audio signals into time-frequency images and evaluating thirteen CNN models under a unified experimental protocol,

the proposed framework demonstrated strong classification performance, with the best-performing models achieving accuracy exceeding 94%. The results confirm that Mel spectrograms effectively capture the impulsive and broadband characteristics of gunshot sounds, while deeper and hybrid CNN architectures provide enhanced discriminative capability. In addition to classification accuracy, this work also analyzed computational aspects, offering insights into accuracy-efficiency trade-offs relevant to real-world deployment. Overall, the findings provide a reproducible and extensible foundation for deep learning–based firearm sound classification and support future research toward scalable and practical acoustic surveillance systems.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Peerapol Khunarsa	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	
Pafan Doungpaisan	✓	✓		✓	✓	✓	✓		✓	✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The dataset used in this study is publicly available from the work of Kabealo *et al*. [16]. Additional derived data supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] M. Naghavi *et al.*, "Global mortality from firearms, 1990–2016," *JAMA Network*, vol. 320, no. 8, pp. 792–814, Aug. 2018, doi: 10.1001/jama.2018.10060.
[2] M. Werbick, I. Bari, N. Paichadze, and A. A. Hyder, "Firearm violence: a neglected 'Global Health' issue," *Global Health*, vol. 17, no. 1, Dec. 2021, doi: 10.1186/s12992-021-00771-8.
[3] S. Zadey, C. C. Branas, and C. N. Morrison, "Gun violence: a global problem in need of local solutions," *The Lancet*, vol. 403, pp. 2783–2784, Jun. 2024, doi: 10.1016/S0140-6736(24)01123-1.
[4] S. Raponi, G. Oligeri, and I. M. Ali, "Sound of guns: digital forensics of gun audio samples meets artificial intelligence," *Multimedia Tools and Applications*, vol. 81, no. 21, pp. 30387–30412, Sep. 2022, doi: 10.1007/s11042-022-12612-w.
[5] J. Park, Y. Cho, G. Sim, H. Lee, and J. Choo, "Enemy spotted: in-game gun sound dataset for gunshot classification and localization," in *2022 IEEE Conference on Games (CoG)*, Aug. 2022, pp. 56–63, doi: 10.1109/CoG51982.2022.9893670.
[6] R. Nijhawan, S. A. Ansari, S. Kumar, F. Alassery, and S. M. El-kenawy, "Gun identification from gunshot audios for secure public places using transformer learning," *Scientific Reports*, vol. 12, no. 1, Aug. 2022, doi: 10.1038/s41598-022-17497-1.
[7] M. L. Doucette, C. Green, J. N. Dineen, D. Shapiro, and K. M. Raissian, "Impact of ShotSpotter technology on firearm homicides and arrests among large metropolitan counties: a longitudinal analysis, 1999–2016," *Journal of Urban Health*, vol. 98, no. 5, pp. 609–621, Oct. 2021, doi: 10.1007/s11524-021-00515-4.
[8] A. Goldenberg *et al*., "Use of ShotSpotter detection technology decreases prehospital time for patients sustaining gunshot wounds," *Journal of Trauma and Acute Care Surgery*, vol. 87, no. 6, pp. 1253–1259, Dec. 2019, doi: 10.1097/TA.0000000000002483.
[9] R. B. Singh, H. Zhuang, and J. K. Pawani, "Data collection, modeling, and classification for gunshot and gunshot-like audio events: a case study," *Sensors*, vol. 21, no. 21, Nov. 2021, doi: 10.3390/s21217320.
[10] J. Bajzik, J. Prinosil, and D. Koniar, "Gunshot detection using convolutional neural networks," in *2020 24th International Conference Electronics*, Jun. 2020, pp. 1–5, doi: 10.1109/IEEECONF49502.2020.9141621.

[11] T. Aggarwal, N. Sharma, and N. Aggarwal, "Gunshot detection and classification using a convolution-GRU based approach," in *Proceedings of Emerging Trends and Technologies on Intelligent Systems*, vol. 1414, Singapore: Springer Nature Singapore, 2023, pp. 95–107, doi: 10.1007/978-981-19-4182-5_8.
[12] M. Goldwater, J. Bonnel, A. Cammareri, D. Wright, and D. P. Zitterbart, "Classification of dispersive gunshot calls using a convolutional neural network," *JASA Express Letters*, vol. 1, no. 10, Oct. 2021, doi: 10.1121/10.0006718.
[13] M. Y. Priya, S. P. Shendre, and P. B. Pati, "Enhanced gunshot sound detection using AlexNet and XGBoost from Fourier spectrograms," in *2023 4th International Conference on Intelligent Technologies (CONIT)*, Jun. 2024, pp. 1–7, doi: 10.1109/CONIT61985.2024.10626951.
[14] N. Morsa, "EDGAR: embedded detection of gunshots by AI in real-time," in *Advanced Analytics and Learning on Temporal Data: 7th ECML PKDD Workshop, AALTD 2022, Grenoble, France*, 2022, pp. 148–166, doi: 10.1007/978-3-031-24378-3_10.
[15] S. Xie, J. Xie, J. Zhang, Y. Zhang, L. Wang, and H. Hu, "MDF-Net: a multi-view dual-attention fusion network for efficient bird sound classification," *Applied Acoustics*, vol. 225, 2024, doi: 10.1016/j.apacoust.2024.110138
[16] R. Kabealo *et al.*, "A multi-firearm, multi-orientation audio dataset of gunshots," *Data in Brief*, vol. 48, Jun. 2023, doi: 10.1016/j.dib.2023.109091.
[17] J. Li, J. Guo, X. Sun, C. Li, and L. Meng, "A fast identification method of gunshot types based on knowledge distillation," *Applied Sciences*, vol. 12, no. 11, May 2022, doi: 10.3390/app12115526.
[18] J. Yin *et al.*, "A joint framework with audio generation for rare gunshot event detection," in *PRICAI 2023: Trends in Artificial Intelligence (PRICAI 2023)*, Singapore: Springer, 2023, pp. 133–144, doi: 10.1007/978-981-99-7022-3_13.
[19] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A survey of audio classification using deep learning," *IEEE Access*, vol. 11, pp. 106620–106649, 2023, doi: 10.1109/ACCESS.2023.3318015.
[20] Z. Chen, H. Zheng, L. Wu, J. Huang, and Y. Yang, "Deep-transfer-learning-based intelligent gunshot detection and firearm recognition using tri-axial acceleration," *IEEE Internet of Things Journal*, vol. 12, no. 5, pp. 5891–5900, Mar. 2025, doi: 10.1109/JIOT.2024.3489963.
[21] T. Perumal, N. Mustapha, R. Mohamed, and F. M. Shiri, "A comprehensive overview and comparative analysis on deep learning models," *Journal on Artificial Intelligence*, vol. 6, no. 1, pp. 301–360, 2024, doi: 10.32604/jai.2024.054314.
[22] E. J. Humphrey and J. P. Bello, "Rethinking automatic chord recognition with convolutional neural networks," in *2012 11th International Conference on Machine Learning and Applications*, Dec. 2012, pp. 357–362, doi: 10.1109/ICMLA.2012.220.
[23] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Deep salience representations for F0 estimation in polyphonic music," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017*, 2017, pp. 63–70.
[24] J. Engel *et al.*, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 1068–1077.
[25] J. J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals," in *13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, 2012, pp. 559–564.
[26] C. Emanuele, D. Ghisi, V. Lostanlen, F. Lévy, J. Fineberg, and Y. Maresz, "TinySOL: an audio dataset of isolated musical notes," *Zenodo*, 2020, doi: 10.5281/zenodo.3685367
[27] A. Morehead, L. Ogden, G. Magee, R. Hosler, B. White, and G. Mohler, "Low cost gunshot detection using deep learning on the Raspberry Pi," in *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, USA, Dec. 2019, pp. 3038–3044, doi: 10.1109/BigData47090.2019.9006456.

BIOGRAPHIES OF AUTHORS

Peerapol Khunarsa received the B.Sc. degree in Computer Science (1999) from the Uttaradit Rajabhat University, Thailand, the M.Sc. degree in Information Technology (2002) from the King Mongkut's University of Technology North Bangkok, Thailand and the Ph.D. degree in Soft Computing from Chulalongkorn University, Bangkok, Thailand. He research interests include pattern recognition, machine learning, and intelligent systems. He can be contacted at email: peerapol@uru.ac.th.

Pafan Doungpaisan received the B.B.A. degree in Business Computer from Dhurakij Pundit University, Thailand, in 1998, the M.Sc. degree in Information Technology from King Mongkut's Institute of Technology North Bangkok, Thailand, in 2002 and the Ph.D. degree in Information Technology from King Mongkut's University of Technology North Bangkok, Thailand, in 2017. Her research interests include artificial intelligence, pattern recognition, and deep learning. She can be contacted at email: pafan.d@itm.kmutnb.ac.th.