Explosion sound classification using machine learning method based on audio features

Nguyen Van Tuan*, Nguyen Dang Tuan

Control, Automation in Production and Improvement of Technology Institute (CAPITI), Academy of Military Science and Technology, 89 Ly Nam De, Hoan Kiem, Hanoi, Vietnam.
*Corresponding author: tnv8084@gmail.com
Received 19 Oct. 2024; Revised 05 Dec. 2024; Accepted 04 Apr. 2025; Published 15 Apr. 2025.
DOI: https://doi.org/10.54939/1859-1043.j.mst.102.2025.133-140

ABSTRACT

This study focuses on the classification of gunshot sounds using multiple audio features and machine learning methods. The gunshot sound samples are converted into spectrograms and processed using Support Vector Machine (SVM) for classification. The model was trained on a dataset of 851 audio files from 8 different gun types. Using a combination of audio features along with data preprocessing techniques, our SVM model achieved 95.32% accuracy in classifying different types of gunshots. The model also demonstrated good performance with real-world data, though with lower confidence levels due to environmental noise. This study provides an effective method for gunshot classification in defense security surveillance systems and sound forensics applications.

Keywords: Audio classification; Spectrogram; Machine learning.

1. INTRODUCTION

In the current context, ensuring public security and safety has become a top priority for functional agencies. Gun-related incidents are increasing, requiring rapid and effective surveillance measures. Accurate identification and classification of gunshots can help authorities respond promptly, reducing human and property losses, while providing important data for crime investigation and military applications.

Domestically, previous research in this field has primarily focused on sound source localization [1-3] or general sound classification. Few studies have specifically addressed gunshot classification, despite its importance for security applications. While traditional methods using time-domain or frequency-domain analysis have been employed, they often struggle with distinguishing different gunshot types in complex environments [4]. Various machine learning approaches have been explored for audio classification. K-Nearest Neighbors (k-NN) achieved 94.48% accuracy in gunshot classification [5] but shows limitations with noisy data. While CNNs have demonstrated success in audio classification [6], they require extensive training data and computational resources [7].

This study proposes a novel approach combining multiple audio features with Support Vector Machine (SVM) classification. SVM was chosen for its effectiveness with both small and high-dimensional data, requiring fewer computational resources compared to deep learning methods [4]. Our objectives include feature extraction from gunshot samples, data preprocessing, and SVM model evaluation.

2. DATA PREPARATION AND AUDIO FEATURE EXTRACTION

2.1. Gunshot dataset

The gunshot dataset used [5] includes 851 audio files of 8 different gun types, with 2 s duration for each gun type, and a sampling frequency of 44100 Hz. The audio files have been checked to ensure they are free from noise contamination and repetition. Table 1 lists the details of the number

of sounds for each gun type in the dataset.

Table 1. *Gunshot dataset.*

No.	Gun type	Number of samples
1	AK-47	72
2	IMI Desert Eagle	100
3	AK-12	98
4	M16	200
5	M249	99
6	MG-42	100
7	MP5	100
8	Zastava M92	82

2.2. Audio features

2.2.1. Audio spectrogram

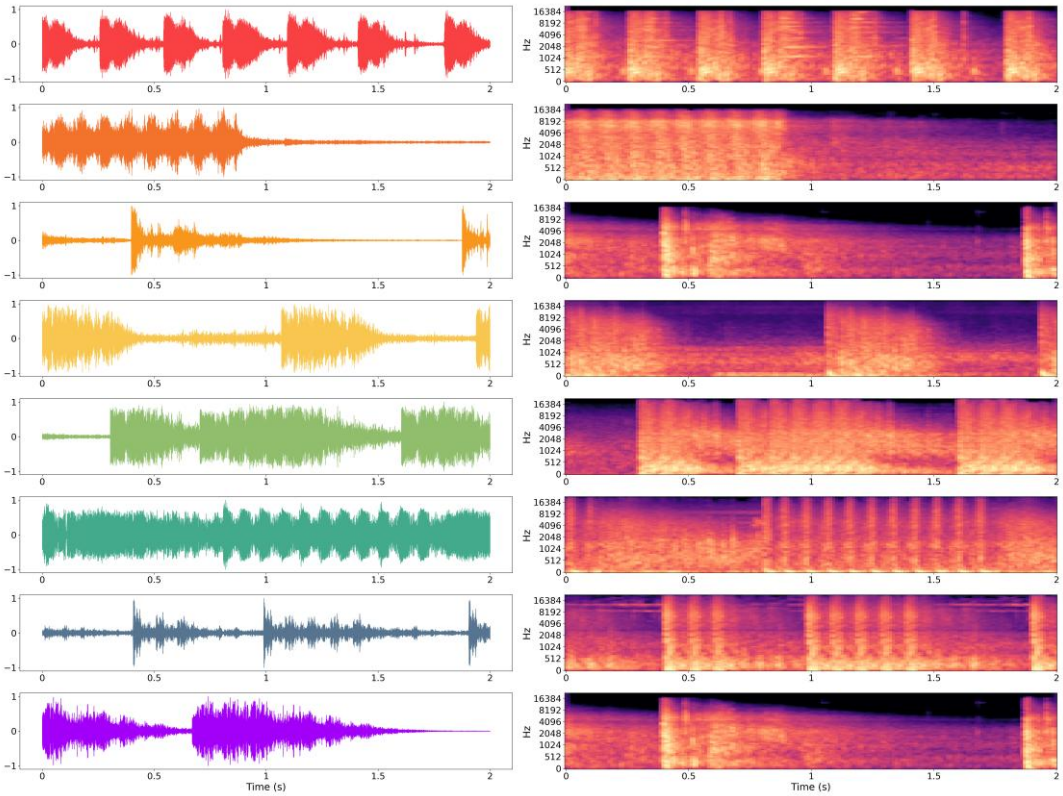

Figure 1. *Time-domain waveform and log-mel spectrogram of some audio samples.*

We apply Short-Time Fourier Transform (STFT) to obtain time-frequency representation of the audio signals. The log-mel spectrogram is generated through the equation (1) [8].

$$S(t,f) = \sum_{n=0}^{N-1} y[n] \cdot w[n-t] \cdot e^{-\frac{j2\pi fn}{N}}, \quad (1)$$

where y is the audio signal, w is the window function, t is the time frame, f is the frequency, and N

is the FFT length. Then, through power spectrum calculation and mel-scale mapping, we obtain the log-mel spectrogram:

$$\log - \mathrm{Mel}(t,m) = \log(M(t,m) + \epsilon), \quad (2)$$

where ε is a small value to avoid taking the logarithm of zero.

Figure 1 shows the time-domain waveform (left, time vs normalized amplitude [-1,1]) and log-mel spectrogram (right, time vs frequency [0-22050 Hz]) of the 8 gun types. The spectrogram's color intensity indicates the energy of frequencies (dB), where brighter bands represent higher energy levels. The AK-47 gunshot exhibits characteristic features of gunshot sounds with distinctive short, high-amplitude pulses and broad frequency spectrum.

2.2.2. Preprocessing

To mitigate overfitting, where the model performs very well on training data but poorly on test data or new data, training audio samples will be augmented with noise using random noise from a Gaussian distribution with mean 0 and variance 0.01. The low variance ensures that the audio information is not lost, and the total number of audio samples is doubled to 1702 samples. Figure 2 compares the log-mel spectrogram of the original AK-47 gunshot with the noise-added gunshot. The log-mel spectrogram of the noise-added gunshot appears less distinct than the original spectrogram, making discrimination more challenging.

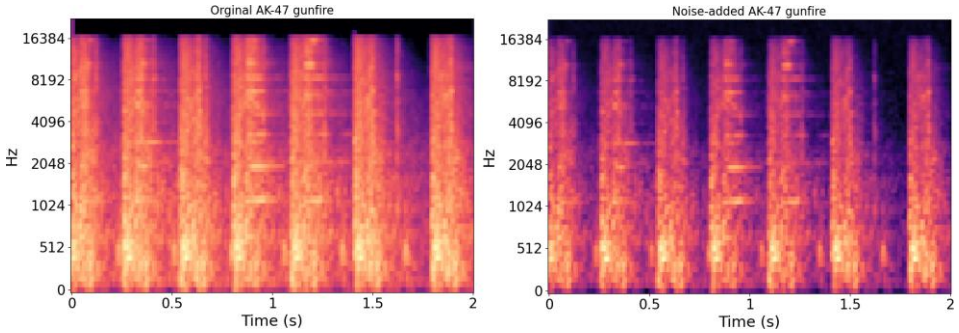

Figure 2. *Spectrogram of noise-added gunshot audio sample.*

2.2.3. Features used for training

Seven key acoustic features were extracted:

Mel-Frequency Cepstral Coefficients (MFCC), an audio feature originating from human auditory mechanisms, this feature has high recognition rates and good noise resistance. It is defined based on the discrete cosine transform as follows [8]:

$$c_n = \sqrt{\frac{2}{K}} \cdot \sum_{k=1}^{K} (\log S_k) \cdot \cos\left(\frac{1}{K} \cdot (n(k-0.5)\pi)\right), \quad (3)$$

where, K represents the number of bandpass filters, S_k is the spectral power at filter k.

Chroma and Spectral Contrast: Represent pitch distribution and frequency energy variation respectively, crucial for distinguishing structural differences in sounds [9].

Spectral Centroid, indicates frequency distribution balance. The spectral centroid is determined by the formula:

$$C_t = \frac{\sum_{n=1}^{N} M_t[n]^* n}{\sum_{n=1}^{N} M_t[n]}, \quad (4)$$

where, $M_t[n]$ is the amplitude of frequency n in the frequency spectrum corresponding to window t.

Zero Crossing Rate (ZCR), measures the number of times an audio signal crosses the zero axis.

For a given frame, the ZCR of an audio signal x of length *N* is defined as the number of times the audio transitions from positive to negative as follows:

$$\mathrm{ZCR} = \frac{1}{2}\sum_{n=0}^{N-1}|\,\mathrm{sgn}(x[n]) - \mathrm{sgn}(x[n-1])\,| \tag{5}$$

Energy, measures signal strength. The energy of audio signal x is calculated by the formula:

$$E = \sum_{i=1}^{N} x[i]^2, \tag{6}$$

where, *x[i]* is the amplitude value of the ith sample in the audio signal; *N* is the total number of samples in the signal.

Spectral Bandwidth, the difference between upper and lower frequencies in a continuous frequency band. This feature helps distinguish gunshots from other sounds based on spectral width. The spectral bandwidth at frame t can be calculated by the formula [6]:

$$bandwidth[t] = (\sum_k S[k,t] \cdot |\mathrm{freq}[k,t] - \mathrm{centroid}[t]|^p)^{1/p}, \tag{7}$$

where, $S[k,t]$ is the magnitude of the spectrogram at frequency portion *k* and frame *t*; $\mathrm{freq}[k,t]$ is the frequency value at portion *k* and frame *t*; $\mathrm{centroid}[t]$ is the spectral centroid at frame t; p is a parameter, usually set to 2 (Euclidean distance).

These features were chosen because they provide rich and multidimensional information about the audio signal, including intensity, frequency, melodic structure, and spectral characteristics, helping the SVM model classify gunshots more accurately and efficiently.

3. SUPPORT VECTOR MACHINE MODEL TRAINING AND EVALUATION

The model implementation follows two main steps as shown in figure 3: training phase to determine optimal parameters and classification phase to identify new samples.

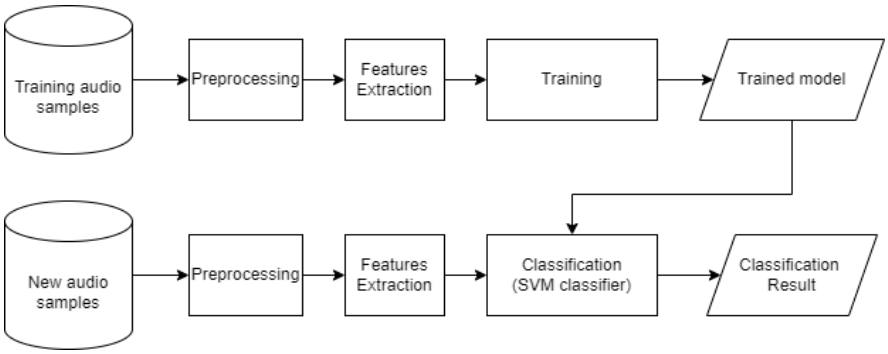

Figure 3. *Workflow of classification model.*

3.1. Feature vector construction and data normalization

A 108-dimensional feature vector x_i is constructed for each audio sample by combining statistical measures (mean, standard deviation, and median) of seven acoustic features:

MFCC: 39 dimensions (13 coefficients × 3 measures); Chroma: 36 dimensions (12 semitones × 3); Spectral Contrast: 21 dimensions (7 bands × 3); Spectral Centroid, ZCR, Energy, and Spectral Bandwidth: 3 dimensions each.

While MFCC can generate more coefficients, only the first 13 are used as they capture the most significant information about the spectral envelope; higher-order coefficients typically represent fine spectral details that are less crucial for classification. The librosa library [10], an open-source audio library, is used to extract features. The dataset was split into training and validation sets (8:2 ratio) with 1360 and 342 samples respectively. Features were normalized using standard scaling:

$$X_{scaled} = \frac{X - mean}{std}, \tag{8}$$

where X_{scaled} is the normalized feature vector, *mean* is the average value, and *std* is the standard deviation. The result is a normalized dataset with mean 0 and standard deviation 1.

3.2. SVM model training

Assuming the training samples are defined as (x_i, y_i), and $i \in \{1,2,\cdots,l\}$, $x_i \in R^n$, $y_i \in \{1,-1\}$ are satisfied. The SVM optimizes the following problem [8]:

$$\begin{gathered} \min_{w,b,\xi} \left(\frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i\right) \\ y_i(w^T z_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i \in \{1,2,\cdots,l\} \end{gathered}, \tag{9}$$

where, w is the weight vector that defines the decision hyperplane, b is the bias term, ξi are slack variables allowing misclassification, $z_i = \varphi(x_i)$ represents the transformation of input vectors x_i into a higher dimensional feature space via the mapping function φ and C is the penalty coefficient. Then, the audio data classification problem can be solved as follows [8], where α_i are the Lagrange multipliers obtained from the dual optimization problem:

$$\text{sgn}(w^T \phi(x) + b) = \text{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b\right). \tag{10}$$

Grid search with stratified 5-fold cross-validation was used to find optimal hyperparameters, with results shown in table 2. The data is divided into 5 subsets and shuffled before division.

Table 2. *Results of best hyperparameter search.*

Parameter / Values	C	γ	Multi-class strategy	Kernel
Test	1, 10, 100	0.1, 0.01, 0.001	ovo, ovr	RBF, Linear, Poly
Best	10	0.01	ovo	RBF

A total of 54 different parameter sets were tested, with the SVM model being trained 270 times (54 parameter sets * 5 folds). The Radial Basis Function (RBF) kernel defined by:

$$k(\mathbf{x}, \mathbf{z}) = \exp\left(-\gamma \parallel \mathbf{x} - \mathbf{z} \parallel_2^2\right), \gamma > 0 \tag{11}$$

The model completed after 1098 iterations with 189 support vectors.

3.3. Model evaluation

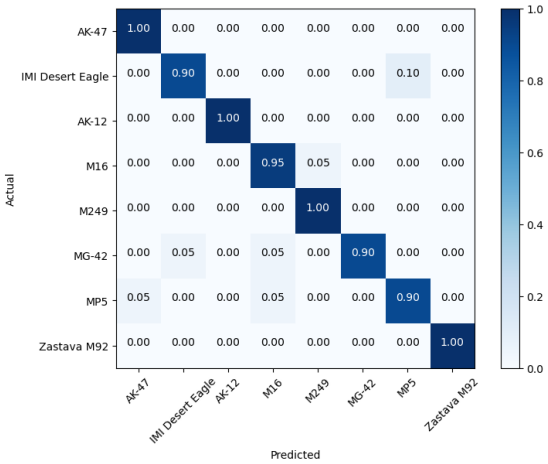

Figure 4. *Confusion matrix.*

The model achieved 95.32% accuracy on the validation set. Figure 4 shows the confusion matrix demonstrating strong classification performance across all gun types. Observing the matrix shows that the model performs well and accurately, with only a few minor confusions between classes.

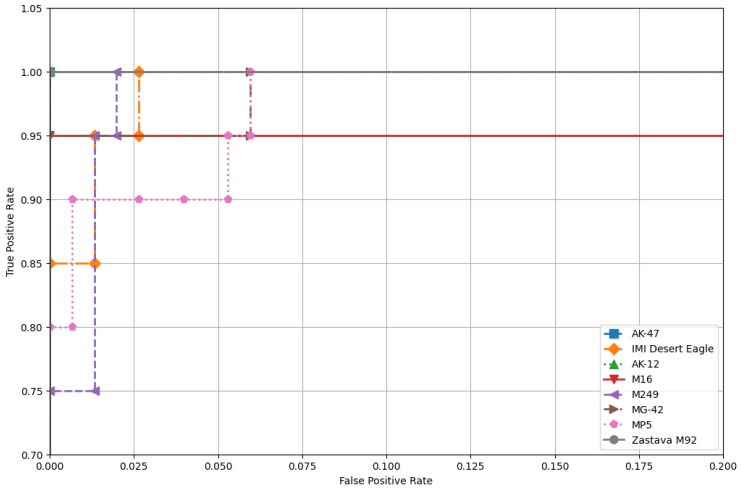

Figure 5. *ROC curve for each class.*

Figure 5 shows the ROC curve for each class with AUC values close to 1, demonstrating strong classification performance. Detailed performance metrics are shown in table 3.

Table 3. *Evaluation metrics for each class.*

Classes	Precision	Recall	F1-score	Samples
AK-47	0.93	1.00	0.97	28
IMI Desert Eagle	0.95	0.90	0.92	40
AK-12	1.00	1.00	1.00	40
M16	0.95	0.95	0.95	80
M249	0.91	1.00	0.95	40
MG-42	1.00	0.90	0.95	40
MP5	0.90	0.90	0.90	40
Zastava M92	1.00	1.00	1.00	34

The model proposed by the authors has higher accuracy than the original paper's model [5] which achieved 94.48% using kNN classifier, showing the effectiveness of combining preprocessing methods and data augmentation with the SVM classifier. Detailed comparisons are shown in table 4.

Table 4. *Comparison of proposed model results.*

Metrics	Proposed SVM model	kNN model from [5]
Accuracy	95.32%	94.48%
Unweighted average recall	95.62%	93.92%
Unweighted average precision	95.50%	94.91%
F1-score	95.48%	94.41%

The model was tested on a computer with AMD Ryzen 5 5600H CPU, 16 GB RAM, Windows 10 operating system. Execution time was evaluated and shown in table 5, total time to produce results on new audio are measured using average of 10 samples of 2 s length.

Table 5. *Model execution time.*

Process	Time
Feature extraction and training	1 minute 39 s
Total time to produce results on new audio	70 ms

The authors also tested the model with real-world data, audio recorded directly at the shooting range, cutting a 2 s audio segment containing AK-47 gunfire, with the waveform shown in figure 6. The model correctly classified it as AK-47 with 44.12% confidence. Due to the limited dataset and enviroment's noises, although the result was correct, the confidence level remains low.

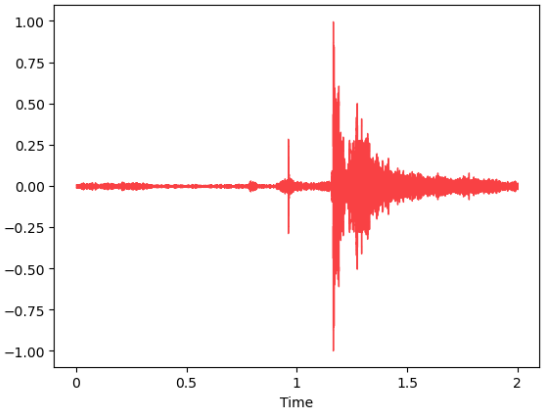

Figure 6. *Waveform of real recorded AK-47 gunfire.*

4. CONCLUSIONS

This study demonstrated the effectiveness of SVM-based gunshot classification using multiple audio features, achieving 95.32% accuracy. While the model performs well on standardized data, real-world testing revealed the need for more diverse training samples and robust noise handling. Future work could explore deep learning approaches for improved performance in noisy environments.

REFERENCES

[1]. H. M. Sang and B. T. Duyen, *"Research sound monitoring system using multi sensor for military purposes,"* Journal of Science and Technology, Hanoi University of Industry, vol. 59, no. 3, 2023.

[2]. L. C. Duan, T. V. Kien and N. N. Minh, *"Sound source localization for intrusion warning systems,"* Journal of Military Science and Technology, vol. 59, pp. 90-96, (2019).

[3]. T. C. Thin, N. T. Kien, B. N. My and N. H. Hoang, *"Improving sound event detecting in sound source localization using TDOA method,"* Journal of Military Science and Technology, vol. 80, pp. 60-70, (2022).

[4]. A. Bansal and N. K. Garg, *"Environmental Sound Classification: A descriptive review of the literature,"* Intelligent Systems with Applications, vol. 16, p. 200115, (2022).

[5]. T. Tuncer, Ş. Doğan, E. Akbal and E. Aydemir, *"An automated gunshot audio classification method based on finger pattern feature generator and iterative relieff feature selector,"* ADYU Mühendislik Bilimleri Dergisi, vol. 8, pp. 225-243, (2021).

[6]. S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen and R. C. Moore, *"CNN architectures for large-scale audio classification,"* in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, (2017).

[7]. S. B. Nesar, B. M. Whitaker and R. C. Maher, *"Machine Learning Analysis on Gunshot Recognition,"* in Intermountain Engineering, Technology and Computing (IETC), Logan, UT, USA, (2024).

[8]. U. Zölzer, *"Digital Audio Signal Processing"*, pp. 21-115: Wiley, (2008).

[9]. J. Urbano, D. Bogdanov, P. Herrera, E. Gomez and X. Serra, *"What is the effect of audio quality on the robustness of MFCCS and chroma features?,"* in 15th International Society for Music Information Retrieval Conference, (2014).

[10]. McFee, Brian, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg and O. Nieto, *"Librosa: Audio and music signal analysis in python.,"* Proceedings of the 14th python in science conference, pp. 18-25, (2015).

[11]. A. Chacón-Rodríguez, P. Julián, P. Alvarado and N. Hernández, *"Evaluation of Gunshot Detection Algorithms,"* IEEE Transactions On Circuits And Systems, vol. 58, (2011).

[12]. M. Grandini, E. Bagli and G. Visani, *"Metrics for Multi-Class Classification: an Overview"*, arXiv, (2020).

TÓM TẮT

Phân loại tiếng nổ bằng phương pháp học máy dựa trên các đặc trưng của âm thanh

Nghiên cứu này tập trung vào việc phân loại tiếng súng bằng cách sử dụng nhiều đặc trưng âm thanh và phương pháp học máy. Các mẫu âm thanh tiếng súng được chuyển đổi thành quang phổ và xử lý bằng phương pháp Máy vector hỗ trợ (SVM) để phân loại. Mô hình được huấn luyện trên tập dữ liệu gồm 851 tệp âm thanh từ 8 loại súng khác nhau. Sử dụng kết hợp các đặc trưng âm thanh cùng với kỹ thuật tiền xử lý dữ liệu, mô hình SVM của nhóm tác giả đạt độ chính xác 95,32% trong việc phân loại các loại súng khác nhau. Mô hình cũng thể hiện hiệu suất tốt với dữ liệu thực tế, mặc dù có độ tin cậy thấp hơn do nhiễu môi trường. Nghiên cứu này cung cấp một phương pháp hiệu quả cho việc phân loại tiếng súng trong các hệ thống giám sát an ninh quốc phòng và ứng dụng giám định âm thanh.

Từ khoá: Phân loại âm thanh; Quang phổ; Học máy.