

Detecção e Classificação de Sons de Disparos: Uma Abordagem com Redes Neurais Convolucionais

Fábio Luiz Júnior, Bruno Eduardo de Oliveira Carvalho, Julio Cesar Santana da Rosa Filho, Paulo Fernando Ferreira Rosa, Julio Cesar Duarte

Instituto Militar de Engenharia (IME)

Praça Gen. Tibúrcio, 80 – 22290-270 – Urca, Rio de Janeiro - RJ – Brazil

{fabioluiz,bruno.carvalho,jcesar,rpaulo,duarte}@ime.eb.br

Abstract. *This work addresses the challenge of detecting and classifying fire-arms, based on the sound emitted during shooting, using a convolutional neural network (CNN). The research was motivated by its applicability in public security, armed forces, and criminal forensics, where the rapid detection and identification of weapons can represent an important differentiator, allowing a quick and effective response from the authorities. To this end, 4 CNN architectures were created to perform the tasks of shot detection, classification by type, caliber, and weapon identification. The results achieved in the work are promising, achieving an accuracy of 100 % in detection, 95.03 % in classification by type, 77.64 % in classification by caliber, and 47.58 % in classification by weapon.*

Resumo. *Este trabalho aborda o desafio da detecção e classificação de armas de fogo, pelo som emitido no disparo, utilizando uma rede neural convolucional (CNN). A pesquisa foi motivada pela sua aplicabilidade em segurança pública, forças armadas e perícia criminal, em que a rápida detecção e identificação do armamento podem representar um diferencial importante, permitindo uma rápida e eficaz resposta das autoridades. Para isso, foram criadas 4 arquiteturas CNN para executar as tarefas de detecção de disparos, classificação por tipo e por calibre e identificação da arma. Os resultados alcançados no trabalho mostram-se promissores, alcançando acurácia de 100 % na detecção, 95,03 % para classificação por tipo, 77,64 % na classificação por calibre e 47,58 % na classificação por armamento.*

1. Introdução

O avanço das tecnologias de análise de áudio, impulsionado por técnicas avançadas de aprendizado de máquina, vem ampliando as possibilidades de interpretação automática de sons em cenários urbanos e operacionais. Em particular, a combinação de redes neurais profundas com dados multimídia tem viabilizado soluções para a compreensão de eventos acústicos em tempo real [1]. Tal tendência se reflete na crescente demanda por sistemas autônomos capazes de integrar, classificar e reagir a informações sensoriais em ambientes dinâmicos, com aplicações que envolvem desde segurança e monitoramento até interação situacional e suporte a decisões em contextos críticos.

Nesse contexto, este trabalho investiga o desafio da detecção e classificação de armas de fogo com base no som do disparo, utilizando arquiteturas de redes neurais convolucionais (CNN). A pesquisa é motivada por sua potencial aplicação em contextos como

segurança pública, forças armadas e perícia criminal. Nas áreas de segurança e defesa, por exemplo, a detecção rápida de disparos e a possível identificação do armamento utilizado podem contribuir para uma resposta mais ágil e eficaz por parte das autoridades [8]. A identificação da arma, por exemplo, pode fornecer subsídios para calibrar o nível de força necessário diante de uma ameaça hostil. No campo da perícia criminal, a detecção e classificação acústica de disparos podem auxiliar na elucidação de crimes em que apenas registros de áudio estão disponíveis, como gravações de câmeras de segurança que não capturaram imagens diretas do ocorrido [2].

Para alcançar esses objetivos, foram desenvolvidas quatro arquiteturas baseadas em CNN, voltadas para: (i) detecção do disparo, (ii) classificação por tipo de arma, (iii) classificação por calibre, e (iv) identificação do modelo de armamento. O treinamento foi realizado a partir da fusão de dois conjuntos de dados públicos, contendo áudios de disparos e sons ambientais.

O problema que se pretende solucionar é a detecção e a classificação de armas de fogo por meio do som emitido pelo disparo. Assim, a tarefa de detecção visa indicar se, no áudio analisado, há a presença ou não do som de um disparo de uma arma de fogo. Já a tarefa de classificação visa reconhecer e categorizar, em uma das classes pré-configuradas, abrangendo três níveis de granularidade: (i) o **tipo de arma**, referente à categoria geral do armamento (revólver, pistola ou arma longa); (ii) o **calibre**, referente ao diâmetro nominal da munição utilizada (por exemplo, 9 mm, .45 e 7.62 mm); e (iii) o **modelo de armamento**, referente à identificação específica da arma dentre os 17 modelos do dataset. Os resultados obtidos foram promissores: 100 % de acurácia na detecção de disparos, 95,03 % na classificação por tipo de arma, 77,64 % na classificação por calibre e 47,58 % na identificação do modelo de armamento.

2. Trabalhos Relacionados

A pesquisa bibliográfica realizada sobre o tema abordado incluiu como uma das principais referências o estudo “*Development of Computational Methods for the Audio Analysis of Gunshots*” [5], que fornece o conjunto de dados usado neste trabalho. Este estudo foi financiado pelo Departamento de Justiça dos EUA para o desenvolvimento deste conjunto de dados e para investigar o uso de métodos computacionais na detecção e classificação de armas de fogo por meio do som de disparos. O trabalho [5] explora e compara ainda vários métodos de detecção e classificação, incluindo técnicas convencionais de processamento digital de sinais, como os Coeficientes Espectrais de Mel-Frequência, do inglês *Mel-Frequency Spectral Coefficients* (MFSC), e os Coeficientes Cepstrais de Mel-Frequência, do inglês *Mel-Frequency Cepstral Coefficients* (MFCC), bem como métodos baseados em redes neurais avançadas, tais como CNN e redes de memória de curto e longo prazo, do inglês *Long Short-Term Memory* (LSTM). No entanto, a publicação não oferece muitos detalhes técnicos.

O estudo “*Weapon identification across varying acoustic conditions using an exemplar embedding approach*” [3] emprega técnicas tradicionais de processamento digital de sinais para a detecção e a classificação de armas pelo som de disparos. Neste trabalho, [3] utilizou Modelos de Mistura Gaussiana, do inglês *Gaussian Mixture Models* (GMM), e Modelos de Markov Ocultos, do inglês *Hidden Markov Models* (HMM).

Já o estudo “*Gun type recognition from gunshot audio recordings*” [4] aplicou a

técnica de classificação HMM e decodificação baseada em Viterbi, utilizando um conjunto de dados composto por 372 gravações de disparos de quatro diferentes tipos de armas de fogo. O modelo alcançou uma taxa de acerto de 80 % na classificação das armas.

O trabalho “*Sound of Guns: Digital Forensics of Gun Audio Samples Meets Artificial Intelligence*” [7] utilizou CNN para reconhecer espectrogramas, uma abordagem semelhante à realizada neste trabalho. Segundo os autores, obteve-se acurácia acima de 90 % na classificação por tipo, calibre e modelo do armamento. A rede foi treinada utilizando 3.655 amostras de áudio de mais de 59 armas diferentes. Na pesquisa bibliográfica realizada, foi o trabalho encontrado com as melhores métricas de desempenho.

Em síntese, os trabalhos relacionados confirmam a viabilidade de abordagens baseadas em aprendizado de máquina para classificação de disparos, porém nenhum deles cobre simultaneamente os quatro níveis de granularidade aqui propostos com datasets inteiramente públicos. O trabalho de [7] utiliza maior volume de amostras e armas; o presente trabalho contribui ao avaliar sistematicamente quatro arquiteturas CNN sobre dados reprodutíveis, com protocolo documentado, estabelecendo uma linha de base para pesquisas futuras em Ciência de Dados aplicada à segurança pública.

3. Metodologia

O reconhecimento de armamento é uma tarefa importante para diversas áreas, como segurança pública, forças armadas e perícia criminal. Por exemplo, descobrir antecipadamente o tipo de armamento, calibre e, possivelmente, a arma utilizada pela força hostil pode fornecer uma vantagem tática e ajudar no planejamento da ação e força necessária a ser empregada.

Nesse contexto, o problema a ser resolvido por este trabalho é a detecção e classificação de armamento por meio do som emitido durante o disparo. Para isso, tem-se como premissa que cada tipo de munição e calibre possui uma energia de disparo diferente. Por exemplo, um calibre .38 possui uma energia aproximada de 300 J, enquanto uma munição de 7,62 mm possui uma energia de 3.000 J. Além disso, cada tipo de armamento possui uma estrutura física diferente, como comprimento do cano e mecanismo de trancamento. Essas diferenças existentes em cada arma resultam em um som de disparo emitido diferente, possuindo uma assinatura que pode ser detectada por algoritmos computacionais. A solução proposta neste trabalho foi a utilização de uma CNN para detectar e classificar armamentos pelo som emitido no disparo. Para isso, os seguintes passos foram utilizados:

- **Passo 1:** O áudio do disparo é capturado por um microfone;
- **Passo 2:** O áudio é pré-processado por meio do cálculo da STFT;
- **Passo 3:** O espectrograma STFT é fornecido à rede neural convolucional para análise;
- **Passo 4:** A rede realiza a detecção e a classificação do armamento.

Para o treinamento da CNN, foram utilizados dois conjuntos de dados distintos, consolidados em um único dataset. O primeiro e principal é o “*Gunshot Audio Forensics Dataset*” [5] que é composto por sons de disparos de diversos armamentos e calibres, capturados em diferentes distâncias e ângulos da fonte do sinal. A Figura 1 ilustra os ângulos e distâncias de captura, nos quais o triângulo vermelho representa o armamento

em análise. À direita tem-se o alvo do disparo e as estrelas amarelas representam os pontos de captura do áudio. Os áudios desse dataset consistem em gravações isoladas de disparos individuais, sem sobreposição de ruídos externos, o que representa uma limitação quanto à generalização para condições acústicas reais.

Figura 1. Posições da captura do áudio emitido pelo disparo do armamento.

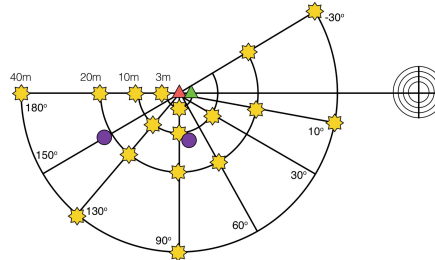


Tabela 1. Tipo, Nome da Arma, Calibre e quantidade de amostras do *dataset* com som de disparos utilizados no treinamento.

Tipo	Arma	Calibre	#
Revólver	S&W 34-1	22LR	109
	S&W 10-8	.38	102
	Ruger Blackhawk	.357	109
Pistola	High Standard Sport King	22LR	115
	Glock 19	9mm	112
	Sig P225	9mm	102
	M&P 40	.40	110
	HK USP Compact	.40	109
	Glock 21	.45	111
	Colt 1911	.45	102
	Kimber Tactical Custom	.45	104
	Ruger 10/22	22LR	102
Arma Longa	Remington 33 Bolt-Action	22LR	108
	M16A1 AR15	5.56	105
	WASR 10/63 AK47	7.62	90
	Winchester M14	7.62	50
	Remington 700	7.62	48
Total de amostras:			1.688

O conjunto completo contém cerca de 10.000 arquivos com aproximadamente 2 segundos de duração, porém, por motivo de complexidade computacional, foi utilizado um subconjunto deste, com 1.688 arquivos de áudio. A Tabela 1 ilustra os tipos, armas, calibres e quantidades de amostras do conjunto utilizado no treinamento.

O segundo conjunto utilizado é o *ESC-50: Dataset for Environmental Sound Classification* [6]. Este *dataset* possui sons ambientais encontrados na natureza, como, por exemplo, som de chuva, vento, trovão, folhas, animais, etc. Possui também sons produzidos por humanos, como sirenes, motores de carro, aviões, helicópteros, entre outros. O conjunto completo possui cerca de 2.000 arquivos de áudio, porém, por motivos de

complexidade computacional, foi utilizado um subconjunto homogeneamente distribuído com 673 arquivos.

Neste trabalho, adotaram-se as proporções de 70 %, 15 % e 15 % na divisão do conjunto para compor, respectivamente, os conjuntos de treinamento, validação e teste. Destaca-se que os arquivos dos dois datasets unidos são randomizados antes da distribuição nos três conjuntos mencionados, para que não ocorram vieses no treinamento da rede.

Foram utilizadas quatro arquiteturas diferentes de redes convolucionais. As arquiteturas foram nomeadas como CNN 1 a 4, com aumento progressivo de complexidade e número de camadas. A Tabela 2 mostra a quantidade de camadas, o tipo de cada camada e o número de parâmetros treináveis de cada rede.

Tabela 2. Arquiteturas das quatro CNN utilizadas nos experimentos.

Camada	Arq. CNN 1	Arq. CNN 2	Arq. CNN 3	Arq. CNN 4
1	Normalization	Normalization	Normalization	Normalization
2	MaxPooling2D (2 × 2)	MaxPooling2D (2 × 2)	MaxPooling2D (2 × 2)	MaxPooling2D (2 × 2)
3	Conv2D (32 filtros, 3 × 3, ReLU)	Conv2D (32 filtros, 3 × 3, ReLU)	Conv2D (32 filtros, 3 × 3, ReLU)	Conv2D (32 filtros, 3 × 3, ReLU)
4	MaxPooling2D (2 × 2)	MaxPooling2D (2 × 2)	Conv2D (32 filtros, 3 × 3, ReLU)	Conv2D (32 filtros, 3 × 3, ReLU)
5	Conv2D (64 filtros, 3 × 3, ReLU)	Conv2D (64 filtros, 3 × 3, ReLU)	MaxPooling2D (2 × 2)	MaxPooling2D (2 × 2)
6	MaxPooling2D (2 × 2)	MaxPooling2D (2 × 2)	Conv2D (64 filtros, 3 × 3, ReLU)	Conv2D (64 filtros, 3 × 3, ReLU)
7	Dropout (0,25)	Conv2D (128 filtros, 3 × 3, ReLU)	Conv2D (64 filtros, 3 × 3, ReLU)	Conv2D (64 filtros, 3 × 3, ReLU)
8	Flatten	MaxPooling2D (2 × 2)	MaxPooling2D (2 × 2)	MaxPooling2D (2 × 2)
9	Dense (32 neurônios, ReLU)	Flatten	Conv2D (128 filtros, 3 × 3, ReLU)	Conv2D (128 filtros, 3 × 3, ReLU)
10	Dropout (0,50)	Dense (20 neurônios, ReLU)	Conv2D (128 filtros, 3 × 3, ReLU)	Conv2D (128 filtros, 3 × 3, ReLU)
11	Dense (C neurônios, saída linear)	Dropout (0,25)	MaxPooling2D (2 × 2)	MaxPooling2D (2 × 2)
12	–	Dense (C neurônios, saída linear)	Flatten	Conv2D (256 filtros, 3 × 3, ReLU)
13	–	–	Dense (20 neurônios, ReLU)	Conv2D (256 filtros, 3 × 3, ReLU)
14	–	–	Dropout (0,25)	MaxPooling2D (2 × 2)
15	–	–	Dense (C neurônios, saída linear)	Flatten
16	–	–	–	Dense (20 neurônios, ReLU)
17	–	–	–	Dense (C neurônios, saída linear)
Parâmetros	17.308.196	5.171.816	5.048.433	3.035.761

Nota: C representa o número de classes da tarefa de classificação.

Os experimentos relacionados ao treinamento das redes neurais e ao processamento dos dados de áudio foram conduzidos utilizando a linguagem Python 3.9.1, com o suporte de bibliotecas como TensorFlow 2.12.0, NumPy 1.23.4 e Matplotlib 3.3.4. No que se refere ao ambiente de hardware, os experimentos foram executados em uma máquina equipada com um processador Intel(R) Core(TM) i7-7500U CPU @ 2.70 GHz e 16 GB de memória RAM.

4. Análise dos Resultados Obtidos

Os resultados obtidos nos experimentos realizados foram divididos em: detecção do disparo, classificação por tipo de arma, classificação por calibre e classificação por modelo de armamento, e estão dispostos a seguir.

A tarefa de detecção envolve verificar a presença de sons de disparos de armas de fogo. O resultado é binário, indicando se um tiro foi detectado ou não. Nesta tarefa, todas as redes apresentaram desempenho máximo de 100 %. A Figura 2 ilustra a matriz de confusão da arquitetura CNN 4, que convergiu mais rapidamente, em poucas épocas.

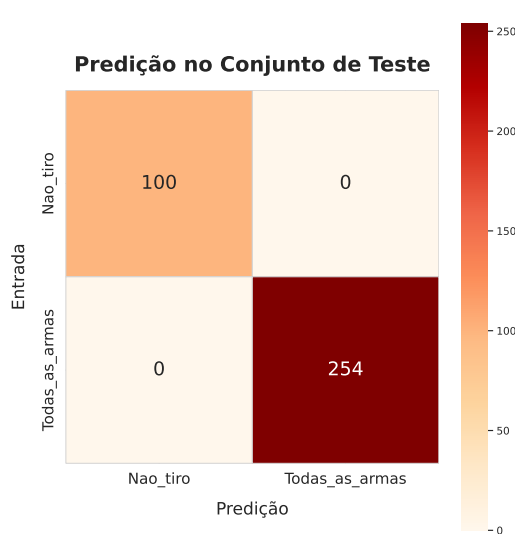


Figura 2. Matriz de confusão na tarefa de detecção do disparo.

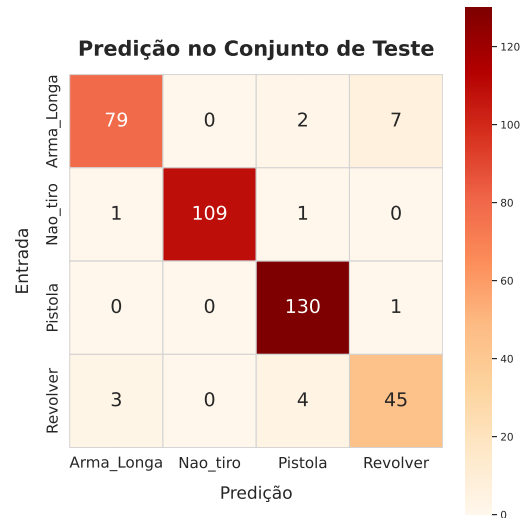


Figura 3. Matriz de confusão na tarefa de classificação por tipo de arma (4 classes).

A tarefa de classificação por tipo de armamento é realizada em quatro categorias distintas: revólver, pistola, arma de cano longo e tiro não detectado. Nesta tarefa, a arquitetura CNN 2 apresentou os melhores resultados, com acurácia de 95,03 %, precisão de 93,74 %, revocação de 93,44 % e *F1-score* de 93,55 %. A Figura 3 ilustra a matriz de confusão da arquitetura CNN 2.

A tarefa de classificação por calibre do armamento é realizada em 9 classes distintas: 22LR; .38; .357; 9mm; .40; .45; 5.56mm e 7.62mm e não_tiro. Nesta tarefa, a arquitetura CNN 4 apresentou os melhores resultados, com acurácia de 77,64 %, precisão de 79,42 %, revocação de 69,56 % e *F1-score* de 71,69 %. A Figura 4 ilustra a matriz de confusão da arquitetura CNN 4.

A tarefa de classificação por armamento visa encontrar o modelo do armamento utilizado no disparo. São 18 classes distintas: BoltAction22; Colt1911; Glock45; Glock9; HKUSP; Kimber45; M16; MP40; Rem700; Ruger22; Ruger357; SIG9; SpKing22; SW22; SW38SP; WASR, WinM14 e Não disparo. Nesta tarefa, a arquitetura CNN 4 apresentou os melhores resultados, com acurácia de 47,58 %, precisão de 52,56 %, revocação de 50,24 % e *F1 score* de 47,21 %. A Figura 5 apresenta a matriz de confusão obtida pela arquitetura CNN 4 na tarefa de classificação por armamento.

É importante destacar que, apesar dos resultados serem promissores, a tarefa permanece desafiadora devido ao elevado número de classes envolvidas e à variabilidade intrínseca dos sons de disparos, que podem ser influenciados por fatores como o ambiente, a qualidade da gravação e a similaridade acústica entre armas do mesmo tipo ou de características próximas, como, por exemplo, a Ruger22 e a Ruger357.

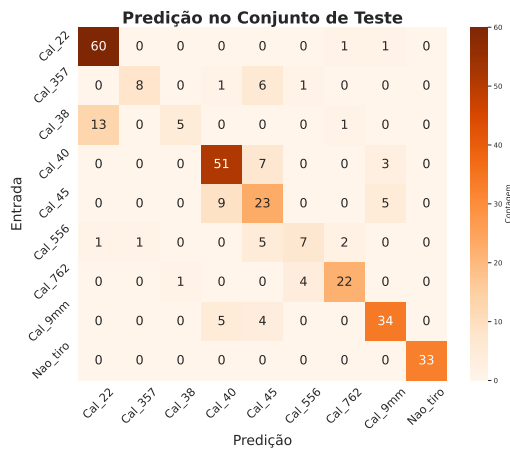


Figura 4. Matriz de confusão na tarefa de classificação por calibre (9 classes).

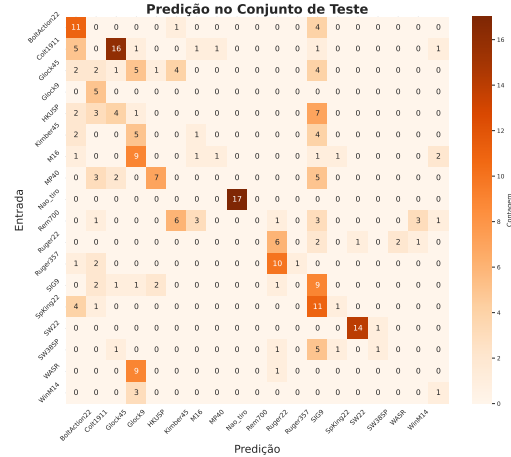


Figura 5. Matriz de confusão na tarefa de classificação por arma (18 classes).

Por fim, a Tabela 3 compila todos os resultados encontrados nos experimentos com as 4 configurações de CNN utilizadas. Os valores em negrito representam as estatísticas da arquitetura escolhida com os melhores resultados.

Cabe registrar limitações metodológicas relevantes: a divisão treino/validação/teste foi feita por randomização no nível do arquivo, sem agrupamento por sessão de gravação (*group split*), o que pode inflar as métricas — sobretudo a detecção perfeita (100 %); o desbalanceamento entre a classe “não disparo” e os modelos individuais de arma não foi tratado com técnicas de ponderação ou reamostragem; e a arquitetura de melhor desempenho foi selecionada com base no conjunto de teste.

5. Conclusões

Este trabalho investigou a detecção e a classificação de armas de fogo pelo som do disparo utilizando CNN, desenvolvendo 4 arquiteturas treinadas sobre dois datasets públicos. Os resultados foram promissores: 100 % de acurácia na detecção, 95,03 % na classificação por tipo de arma (CNN 2), 77,64 % por calibre e 47,58 % por modelo de armamento (CNN 4), indicando potencial de aplicação em segurança pública e perícia criminal.

Embora os resultados sejam expressivos, há espaço para aprimoramentos. Trabalhos futuros podem utilizar o dataset completo (10.000 arquivos), explorar estratégias evolucionárias para seleção de arquiteturas e aplicar *data augmentation* com ruídos ambientais. Do ponto de vista metodológico, recomenda-se adotar divisão agrupada por sessão de gravação (*group split*), aplicar técnicas de balanceamento de classes e avaliar a robustez dos modelos em ambientes acústicos ruidosos e em cenários de implantação em tempo real.

Tabela 3. Métricas de desempenho das diferentes arquiteturas CNN nas quatro tarefas propostas. Em negrito, os melhores desempenhos em cada tarefa.

Tarefa	Métrica	Arq. 1	Arq. 2	Arq. 3	Arq. 4
Detecção	Acurácia	100.00%	100.00%	100.00%	100.00%
	Precisão	100.00%	100.00%	100.00%	100.00%
	Revocação	100.00%	100.00%	100.00%	100.00%
	F1 Score	100.00%	100.00%	100.00%	100.00%
Classificação por tipo de arma [4 classes]	Acurácia	91.10%	95.03%	92.67%	94.50%
	Precisão	89.64%	93.74%	90.37%	93.16%
	Revocação	87.19%	93.44%	89.33%	92.28%
	F1 Score	88.17%	93.55%	89.44%	92.62%
Classificação por calibre [9 classes]	Acurácia	61.34%	66.45%	75.08%	77.64%
	Precisão	54.64%	65.08%	78.04%	79.42%
	Revocação	51.51%	60.23%	72.44%	69.56%
	F1 Score	50.94%	61.56%	74.00%	71.69%
Classificação por armamento [18 classes]	Acurácia	32.34%	35.69%	44.24%	47.58%
	Precisão	32.60%	29.61%	40.51%	52.56%
	Revocação	30.89%	32.37%	42.67%	50.24%
	F1 Score	30.49%	28.68%	38.44%	47.21%

Referências

- Gencoglu, O., Virtanen, T., and Huttunen, H. (2014). Recognition of acoustic events using deep neural networks. In *2014 22nd European Signal Processing Conference (EUSIPCO)*, pages 506–510, Lisbon, Portugal. IEEE.
- Gray, R. (2017). Gunfire audio opens new front in crime-fighting. BBC News.
- Khan, S., Divakaran, A., and Sawhney, H. S. (2010). Weapon identification across varying acoustic conditions using an exemplar embedding approach. In *Sensors, and command, control, communications, and intelligence (C3I) technologies for homeland security and homeland defense IX*, volume 7666, pages 494–501, Bellingham, WA USA. SPIE, SPIE.
- Kiktova, E., Lojka, M., Pleva, M., Juhar, J., and Cizmar, A. (2015). Gun type recognition from gunshot audio recordings. In *3rd international workshop on biometrics and forensics (IWBIF 2015)*, pages 1–6, Gjøvik, Norway. IEEE.
- Lilien, R. (2018). *Development of computational methods for the audio analysis of gunshots*. Cadre Research Lab, Chicago, IL.
- Piczak, K. J. (2015). Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, page 1015–1018, New York, NY, USA. Association for Computing Machinery.
- Raponi, S., Oligeri, G., and Ali, I. M. (2022). Sound of guns: digital forensics of gun audio samples meets artificial intelligence. *Multimedia tools and applications*, 81(21):30387–30412.
- Stephens, S. (2023). Knightscope announces automated gunshot detection. <https://www.businesswire.com/news/home/20230714391246/en/Knightscope-Announces-Automated-Gunshot-Detection>.