Multi-Classificação de Cenários de Violência utilizando Extração de Atributos Acústicos e Redes Neurais Convolucionais

Caléo Meneses Santos
Dep. Engenharia Elétrica e de Computação
Universidade Federal da Bahia
Salvador, Brasil
caleo.santos@ufba.br

Antonio Carlos Lopes Fernandes Junior
Dep. Engenharia Elétrica e de Computação
Universidade Federal da Bahia
Salvador, Brasil
antonio.lopes@ufba.br

***Resumo*—O presente trabalho aborda a necessidade urgente de desenvolver tecnologias que visem reduzir a incidência de eventos violentos em nossa sociedade, particularmente no contexto brasileiro, onde populações marginalizadas como jovens negros enfrentam temores constantes de violência letal. Nesse contexto, a abordagem acústica surge como uma vertente promissora devido às suas vantagens intrínsecas, como a capacidade de detecção automatizada e a não oclusão do ambiente. Este trabalho tem como objetivo principal contribuir para o problema da identificação de cenários de violência treinando um modelo de aprendizado de máquina capaz de identificar e classificar cenários tais como gritos, violência física e disparos de armas de fogo. Foram testadas arquiteturas de redes neurais convolucionais, em particular as redes ResNet152 e MobileNet. Os resultados demonstram que ambos os modelos alcançaram precisões similares na tarefa de multi-classificação, com precisões de 84,9% e 84,1%, respectivamente. Esses achados destacam a viabilidade e eficácia da abordagem proposta, mostrando o potencial das redes neurais convolucionais na classificação de cenários violentos utilizando apenas o áudio.**

***Palavras-chave*—Classificação sonora, Análise de cena acústica, Redes neurais convolucionais, Cenários violentos, Extração de atributos.**

***Abstract*—The present study addresses the pressing need to develop technologies aimed at reducing the incidence of violent events in our society, particularly within the Brazilian context, where marginalized populations such as young Black individuals face constant fears of lethal violence. In this scenario, the acoustic approach emerges as a promising avenue due to its inherent advantages, such as automated detection capabilities and non-obtrusiveness to the environment. The primary objective of this work is to contribute to the problem of identifying violent scenarios by training a machine learning model capable of recognizing and classifying scenarios such as screams, physical violence, and gunshots. Convolutional neural network architectures were tested, specifically the ResNet152 and MobileNet networks. The results show that both models achieved similar accuracies in the multi-classification task, with accuracies of 84.9% and 84.1%, respectively. These findings underscore the feasibility and effectiveness of the proposed approach, highlighting the potential of convolutional neural networks in classifying violent scenarios using only audio.**

***Palavras-chave*—Sound classification, Acoustic scene analysis, Convolutional neural networks, Violent scenarios, Feature extraction.**

I. Introdução

A. Contextualização

A violência é um problema persistente em diversas sociedades, incluindo o Brasil, onde os índices de violência ainda são alarmantes [1].

Além das formas convencionais de violência, como assaltos e homicídios, a violência perpetrada por agentes de segurança pública também é uma questão de extrema importância no contexto brasileiro. A atuação de policiais que resulta em mortes violentas é um fenômeno preocupante, representando 12,9% de todas as MVI no país [21]. O perfil das vítimas de intervenções policiais no país tem se mantido constante ao longo dos anos, com predominância de homens negros, adolescentes e jovens correspondendo a 84,1% das vítimas de mortes decorrentes de intervenções policiais no ano de 2021 [21].

É nesse contexto que a utilização de tecnologias inovadoras, como a vigilância por meio de câmeras corporais, vem ganhando destaque como uma alternativa promissora. Essas câmeras, quando usadas por agentes de segurança pública, registram em tempo real as interações entre os policiais e os cidadãos durante o exercício de suas funções. A implementação dessas câmeras tem se mostrado efetiva na redução da letalidade provocada por policiais, como evidenciado por estudos recentes que apontam uma diminuição significativa na taxa de mortalidade em batalhões que aderiram ao programa de câmeras corporais em comparação com aqueles que ainda não o implementaram [4].

Embora os métodos tradicionais de vigilância sejam importantes e amplamente utilizados, é fundamental explorar e adotar abordagens inovadoras que possam complementar e aprimorar a eficácia desses métodos. Nesse contexto, o uso de técnicas de processamento de áudio e aprendizado de máquina tem ganhado destaque como uma abordagem promissora para identificar e classificar eventos de violência por meio de análise acústica [22], [23] por não sofrerem com problemas como obstrução visual ou baixa iluminação.

B. Definição do Problema

O problema que este trabalho busca explorar é a detecção automática de cenários de violência de maneira automática utilizando apenas o áudio. Apesar de estudos anteriores terem explorado a eficácia do áudio nessa tarefa de detecção, existe uma lacuna significativa na literatura quando se trata de identificar qual o indicativo de violência está acontecendo no evento para além de sua presença de maneira geral.

A resolução deste problema tem o potencial de ser uma ferramenta tecnológica adicional na detecção de eventos violentos, permitindo uma resposta rápida e eficiente para prevenir a violência [2], [5].

C. Organização do Texto

Este trabalho está estruturado em quatro partes principais. Primeiramente, é apresentado uma revisão da literatura enfocando a abordagem acústica na detecção automática de violência, bem como a aplicação do aprendizado profundo na tarefa de classificação de cenas acústicas. Em seguida, é detalhada a metodologia adotada, que engloba as estratégias de seleção e processamento de dados, as fases de treinamento, avaliação, testes e a escolha dos modelos. É descrita também a pipeline de dados desenvolvida. Por fim, apresenta-se os resultados obtidos e as respectivas conclusões.

II. Revisão da Literatura

A. Abordagem Acústica na Detecção de Violência

A abordagem acústica na detecção de violência se baseia na análise de sinais de áudio para identificar características distintivas de eventos violentos. O som é um sinal rico em informações e pode fornecer indícios importantes sobre a ocorrência de ações violentas. A análise acústica permite detectar variações nos padrões de som, como gritos, discussões acaloradas, sons de luta ou outros eventos sonoros associados à violência. Essa abordagem oferece a vantagem de ser menos dependente de condições visuais e pode ser aplicada em ambientes com baixa iluminação ou com obstruções visuais [5].

Tendo em vista as oportunidades oferecidas pela abordagem acústica na detecção de violência, o trabalho de Crocco et al. [5] realizou uma revisão sistemática no campo da vigilância automática utilizando métodos áudio-sensoriais. Similarmente ao campo da vídeo-vigilância, o estudo propôs uma descrição sistemática dos métodos abrangidos, incluindo subtração de fundo, classificação de eventos, rastreamento de objetos e análise de situação.

Nesse segmento, em 2006, Rouas et al. [7] propuseram a análise automática de áudio para uma aplicação de vigilância no transporte público, com o objetivo de detectar gritos em um vagão de metrô em casos de eventos de violência. Devido ao ambiente extremamente ruidoso de um transporte público, foi desenvolvida uma técnica de extração de eventos relevantes utilizando um algoritmo de segmentação automática e detecção de atividade. O estudo utilizou duas técnicas de classificação, SVMs e o Modelo de Mistura Gaussiana - GMM, para analisar cenários de brigas entre duas pessoas ou mais, brigas entre um homem e uma mulher, cenas de roubos e cenários de roubo de bolsa ou celular. Esse trabalho resultou no desenvolvimento de um algoritmo capaz de identificar eventos de grito, porém ainda apresentava desafios relacionados a falsos positivos.

Em 2019, Souto et al. [6] propôs a classificação e detecção de cenas acústicas de violência doméstica utilizando o classificador SVM. Foram selecionados três parâmetros de sinais de áudio nos domínios de frequência e tempo: Coeficientes Mel-Cepstrais - MFCC, energia dos sinais no domínio da frequência e taxa de cruzamento por zero - ZCR no domínio do tempo. Essas características foram extraídas dos áudios por meio de um sequenciamento de janelas de tempo de curta duração, e posteriormente agrupadas por média e desvio padrão em janelas de tempo maiores. Essas características foram então utilizadas nos classificadores, alcançando uma acurácia máxima de 73,14% com o uso do MFCC em um conjunto de dados próprio.

Em 2021, Lacerda et al. [10] desenvolveu e avaliou quatro classificadores para a detecção automática de violência no áudio ambiente. Em vez de processar diretamente os sinais de áudio, Redes Neurais Convolucionais (CNN) pré-treinadas foram utilizadas para classificar as imagens geradas a partir de mel-espectrogramas convertidos em imagens com a utilização do HEAR Dataset, um conjunto de dados criado especificamente para a pesquisa alcançando uma acurácia de 78,9%.

B. Deep Learning na Classificação de Cenas Acústicas

O domínio do aprendizado de máquina, particularmente o *Deep Learning*, tem emergido como um componente indispensável no âmbito do reconhecimento de áudio e detecção de eventos sonoros.

Uma abordagem frequentemente empregada consiste em adaptar redes neurais profundas, previamente consolidadas no campo de processamento de imagem, para o contexto acústico. Isso é realizado através da conversão do sinal unidimensional de áudio em uma representação bi-dimensional de tempo-frequência. Um exemplo notável é a aplicação da Transformada de Fourier no áudio, resultando em um tensor 2D, onde as diversas linhas correspondem a frequências distintas e as colunas representam diferentes instantes temporais [19]. Alternativamente, pode-se também redimensionar e padronizar atributos unidimensionais de áudios, como o centroide espectral e a largura de banda, para que assumam um formato análogo ao de uma imagem [24].

Abeßer [9] proporciona uma revisão extensiva acerca da incorporação do aprendizado profundo neste contexto. Arquiteturas de redes neurais reconhecidas, tais como AlexNet, VGG16, Xception, DenseNet, GoogLeNet e ResNet, têm sido largamente utilizadas na Classificação de Cena Acústica (ASC). Adicionalmente, modelos pré-treinados são frequentemente refinados por meio de transferência de aprendizado para tarefas especificamente relacionadas à classificação de áudio.

Tabela I
DISTRIBUIÇÃO DE EXEMPLOS POR CLASSE NOS CONJUNTOS DE DADOS UTILIZADOS

Dataset	Nada	Grito	Violência Física	Tiro
Audioset	–	883	620	3149
GunshotForensic	–	–	–	1990
HEAR Train Dataset	29974	–	30000	–
MIVIA	–	245	–	790
Sound Events S. App.	275	–	–	101
Violent Scenes Dataset	15523	6158	–	265
Total	45872	7241	30620	5355

III. METODOLOGIA

A. Seleção e Preparação dos Dados

Nesta seção, descreveremos os conjuntos de dados utilizados neste estudo, bem como o processo de preparação dos dados para análise.

Os conjuntos de dados selecionados na Tabela I foram escolhidos devido à sua disponibilidade pública e à presença predominante de eventos sonoros relacionados às classes utilizadas neste estudo: Sem Violência (Nada), Grito, Violência Física e Disparo de Arma de Fogo (Tiro).

1) HEAR Dataset: O HEAR Dataset é um conjunto de dados utilizado para a detecção de violência em áudio, utilizado no estudo de Tiago Lacerda et al. em 2021 [10]. Esse *dataset* contém 30 mil exemplos de áudios, com duração fixa de dez segundos, que incluem eventos de agressão física, como socos e tapas, acompanhados de áudio de fundo. Além disso, o *dataset* também contém a quantidade similar de exemplos de áudios que não contêm qualquer tipo de violência.

No HEAR Dataset, os eventos de áudio relacionados à violência são anotados apenas como presentes ou ausentes, o que impossibilitou a segmentação específica desses eventos. Por essa razão, foi estabelecida a duração fixa de dez segundos para integrar o HEAR Dataset aos demais conjuntos de dados utilizados neste estudo.

2) Violent Scenes Dataset: O Violent Scenes Dataset (VSD) é um conjunto de dados amplamente utilizado para a detecção de cenas violentas em filmes, conforme desenvolvido por Schedi et al. [11]. Esse dataset fornece informações sobre características visuais e sonoras extraídas de vinte e quatro filmes que contêm cenas de violência. No entanto, o *dataset* não disponibiliza o material audiovisual original, a fim de evitar problemas relacionados à distribuição não autorizada de cópias.

Neste estudo, foram selecionados apenas os atributos dos eventos sonoros relevantes das anotações presentes no VSD. Essa seleção abrange tanto os trechos de áudio que contêm eventos sonoros de violência quanto os recortes que não possuem nenhum evento sonoro relevante. Essa abordagem permite uma análise abrangente e comparativa das características acústicas presentes nos filmes violentos, mesmo sem acesso direto ao áudio original dos filmes.

3) Sound Events for Surveillance Applications: O Sound Events for Surveillance Applications (SESA) é um conjunto de dados utilizado em aplicativos de vigilância, criado por Spadini et al. em 2019 [12]. Ele engloba diversos eventos sonoros, incluindo gritos, disparos de arma de fogo, explosões e sirene. O conjunto de dados foi obtido a partir de arquivos do *Freesound* [8] e está dividido em pastas de treinamento e teste. A duração dos arquivos varia, podendo chegar a até 33 segundos.

4) Mivia Audio Events Dataset: O Mivia Audio Events Dataset é um conjunto de dados utilizado para análise de eventos de áudio, criado por Foggia et al. em 2015 [13]. O *dataset* engloba uma variedade de eventos sonoros, incluindo gritos, disparos de arma de fogo e som de vidro quebrando. Os sons eram distribuídos em faixas de áudio com ocorrência e duração dos sons anotadas que foram lidas e utilizadas para segmentação dos áudios de interesse.

5) Gunshot Audio Forensics Dataset: O Gunshot Audio Forensics Dataset consiste em gravações de áudio de tiros de vinte e um tipos de armas, coletadas por Ram Mettu et al. em 2017 [14]. O conjunto de dados abrange uma variedade de armas de fogo, ambientes de gravação e diversos dispositivos de gravação. Para este estudo, foi selecionado o subconjunto de dados gravados com os dispositivos Samsung Edge S7, que fornecem uma amostra representativa dos tiros em diferentes condições de gravação.

Cada gravação de áudio no Gunshot Audio Forensics Dataset corresponde a um único disparo da arma, resultando em durações curtas de áudio. Essa característica permite uma análise precisa e focada nos eventos sonoros de tiros, proporcionando informações valiosas para a detecção e classificação desses eventos.

6) Google Audioset: O Google Audioset é um conjunto de dados desenvolvido por Gemmeke et al. em 2017 [15]. Ele contém uma grande quantidade de áudio coletado de vídeos do YouTube, classificados em diferentes categorias de eventos sonoros. Neste estudo, foram selecionadas as categorias relevantes, como gritos, disparos de arma de fogo e sons de tapas, para a coleta dos trechos de áudio correspondentes a essas categorias.

B. Processamento Digital de Áudio

No Processamento de Áudio Digital foram extraídas oito atributos para a integração de todos os conjuntos de dados de áudio. O objetivo principal era igualar aos atributos fornecidas pelo conjunto de dados Violent Scenes Dataset, que não continha os áudios originais dos filmes devido a questões de direitos autorais, mas apenas os atributos já extraídos em uma base por quadro de vídeo.

No Violent Scenes Dataset, oito características extraídas foram: envoltória de amplitude, valor quadrático médio ou valor eficaz, taxa de cruzamentos por zero, razão de energia de banda, centroide espectral, largura de banda de frequência, fluxo espectral e coeficientes cepstrais de frequência Mel MFCC. Considerando a taxa de amostragem de áudio de 44.100 Hz e a codificação dos vídeos com 25 quadros por

segundo, foram utilizadas janelas de comprimento de 1.764 amostras de áudio. Para cada janela, foram calculados 22 MFCCs, enquanto todas as outros atributos são unidimensionais, ou seja, possuem apenas um valor para cada janela de áudio [11].

Para os outros conjuntos de dados, os áudios foram amostrados em uma taxa de 16kHz e foram utilizadas janelas de comprimento de 640 amostras. Essa configuração foi adotada para garantir a consistência e comparabilidade dos atributos extraídos entre os diferentes conjuntos de dados.

Outra estratégia adotada para manter a consistência dos dados no conjunto de dados foi a padronização da duração das amostras de áudio, por meio da repetição ou corte dos áudios. Para isso, foram implementadas duas abordagens:

- Áudios menores que 10 segundos foram repetidos até atingir a duração de 10 segundos.
- Áudios maiores que 10 segundos tiveram apenas o trecho central de 10 segundos selecionado.

A escolha de utilizar 10 segundos como amostra de áudio teve como base o maior conjunto de dados disponível, o HEAR Dataset, que continha o maior número de exemplos de áudio, incluindo sons de socos e tapas. Essas amostras poderiam estar presentes em qualquer parte dos 10 segundos de duração, justificando a escolha desse tamanho como padrão para as demais amostras.

Essa abordagem assegurou a consistência e comparabilidade dos dados no conjunto de dados, possibilitando uma análise precisa e coerente das características e eventos de violência presentes nos áudios.

C. Atributos

1. Envoltória de Amplitude (AE): Essa característica é obtida calculando a envoltória de amplitude do sinal de áudio usando a transformada de Hilbert. a envoltória de amplitude representa a variação da amplitude do sinal ao longo do tempo. É calculada a média da envoltória em janelas para capturar características temporais do sinal [16].

$$AE = \frac{1}{N}\sum_{i=1}^{N}|H(x_i)| \tag{1}$$

Onde:

- $\frac{1}{N}$: é um fator de normalização, onde N representa o número total de amostras no sinal de áudio.
- $|H(x_i)|$: representa o módulo da transformada de Hilbert aplicada à amostra x_i do sinal de áudio. A transformada de Hilbert é usada para calcular o envoltória de amplitude, que é a magnitude da componente analítica do sinal [16]. O módulo é tomado para eliminar a informação de fase e obter apenas a amplitude.

2. Root Mean Square (RMS): O RMS é uma medida do valor eficaz do sinal de áudio. Ele representa a média das amplitudes ao quadrado do sinal em janelas. Essa característica fornece informações sobre a energia do sinal sonoro. O RMS é um dos recursos de intensidade mais comuns e às vezes é referido diretamente como a intensidade do som [17].

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2} \tag{2}$$

3. Zero Crossing Rate (ZCR): O ZCR é uma medida da taxa de cruzamentos por zero do sinal de áudio. Ele representa a quantidade de vezes que o sinal muda de polaridade em um determinado período de tempo. Essa característica é útil para distinguir entre sons com diferentes características de frequência. Quanto mais frequentemente o sinal muda seu sinal, mais conteúdo de alta frequência pode ser considerado no sinal [17].

$$ZCR = \frac{1}{2N}\sum_{i=1}^{N-1}|sgn(x_i) - sgn(x_{i+1})| \tag{3}$$

Onde:

- $\frac{1}{2N}$: representa o fator de normalização para obter a média correta. Nesse caso, N representa o número total de amostras no sinal de áudio.
- $sgn(\cdot)$: é a função sinal que retorna o sinal de um número. Se o número for positivo, $sgn(\cdot)$ retorna 1; se for zero, retorna 0; e se for negativo, retorna -1.

4. Band Energy Ratio (BER): O BER fornece a relação entre as bandas de frequência mais baixas e mais altas. Pode ser entendido como a medida de quão dominantes são as frequências baixas. Essa característica tem sido amplamente utilizada na discriminação entre música e fala, classificação de música, entre outros [18].

A equação para calcular o BER é dada por:

$$BER_t = \frac{\sum_{n=1}^{F} m_t(n)^2}{\sum_{n=F+1}^{N} m_t(n)^2} \tag{4}$$

Onde:

- BER_t é a razão de energia da banda (BER) no frame t.
- $\sum_{n=1}^{F} m_t(n)^2$ representa a soma dos quadrados das magnitudes do sinal (potência) nos bins de frequência mais baixos até a frequência de corte F.
- $\sum_{n=F+1}^{N} m_t(n)^2$ representa a soma dos quadrados das magnitudes do sinal (potência) nos bins de frequência a partir da frequência de corte $F+1$ até o número total de bins N.

5. Spectral Centroid (SC): O centroide espectral é uma medida da localização média do espectro de frequência do sinal. Ele representa o ponto médio das frequências presentes no sinal. Essa característica fornece informações sobre a "cor" ou "timbre" do som. É definido como a soma ponderada em frequência do espectro de potência normalizado por sua soma não ponderada [17].

$$SC = \frac{\sum_{i=1}^{N} f_i X(f_i)}{\sum_{i=1}^{N} X(f_i)} \tag{5}$$

onde:

- $\sum_{i=1}^{N} f_i X(f_i)$: essa parte da equação representa a soma ponderada das frequências f_i multiplicadas pelas amplitudes espectrais $X(f_i)$ correspondentes.
- $\sum_{i=1}^{N} X(f_i)$: denota a soma das amplitudes espectrais não ponderadas.
- $X(f_i)$: é a amplitude espectral correspondente à frequência f_i.

6. Bandwidth (BW): A largura de banda instantânea, também conhecida como dispersão espectral, descreve a concentração do espectro de potência em torno do centróide espectral e é uma descrição técnica da forma espectral. Pode ser interpretada como o desvio padrão do espectro de potência em relação ao centro espectral [17].

$$BW = \sqrt{\frac{\sum_{i=1}^{N}(f_i - \bar{f})^2 X(f_i)}{\sum_{i=1}^{N} X(f_i)}} \tag{6}$$

- $\sum_{i=1}^{N}(f_i - \bar{f})^2 X(f_i)$: representa a soma dos desvios quadráticos entre a frequência f_i e a média das frequências $\bar{f}$, ponderados pelas amplitudes espectrais $X(f_i)$ correspondentes.
- $\sum_{i=1}^{N} X(f_i)$: denota a soma das amplitudes espectrais não ponderadas.

7. Spectral Flux (SF): O fluxo espectral é uma medida da mudança do espectro de frequência ao longo do tempo. É uma medida que quantifica a quantidade de mudança na forma espectral ao longo do tempo. É calculado como a diferença média entre quadros consecutivos na Transformada de Fourier de Tempo Curto - STFT. Essa medida pode ser interpretada como uma aproximação rudimentar da sensação de "rugosidade" percebida no som, que pode ser modelada como uma variação quase periódica ou uma modulação nos níveis do padrão de excitação [17].

$$SF(t) = \sum_{k=1}^{N} \left(X(k,t) - X(k,t-1)\right)^2 \tag{7}$$

Onde:

- $SF(t)$ é o spectral flux no instante t.
- $X(k,t)$ é o valor da magnitude do espectro no bin k no instante t.

8. Mel-Frequency Cepstral Coefficients (MFCC): Os coeficientes cepstrais de frequência mel são uma representação compacta do espectro de frequência do sinal de áudio. Eles capturam informações sobre as características espectrais do som, levando em consideração a percepção auditiva humana. Essa característica é amplamente utilizada em tarefas de classificação e reconhecimento de eventos sonoros [17]. Os MFCCs têm sido amplamente utilizados no campo do processamento de sinal de fala desde sua introdução em 1980 e também são úteis em aplicações de processamento de sinal de música. No contexto da classificação de sinais de áudio, foi demonstrado que um pequeno subconjunto dos MFCCs resultantes já contém a informação principal - na maioria dos casos, o número de MFCCs usados varia na faixa de 4 a 20 [17].

$$MFCC = \mathbf{DCT}\left(\log\left(\sum_{k=1}^{K} H_k(f) \cdot X(f)\right)\right) \tag{8}$$

Onde:

- $\mathbf{DCT}$ representa a Transformada Discreta de Cosseno, que é aplicada à sequência resultante do logaritmo da soma ponderada dos valores do espectro.
- $H_k(f)$ representa a resposta em frequência do filtro Mel k na frequência f. Essa resposta em frequência é tipicamente triangular e é usada para ponderar as componentes espectrais.
- $X(f)$ é o espectro de magnitude do sinal de áudio na frequência f, obtido por meio da STFT. O espectro é uma representação das componentes de frequência presentes no sinal de áudio.

Esses oito atributos fornecem informações complementares sobre o conteúdo e as propriedades do sinal de áudio. Ao analisar e comparar esses atributos em diferentes eventos sonoros, é possível identificar padrões distintos e realizar classificações precisas dos eventos com base em seus atributos acústicas.

Na Figura 1, são apresentados exemplos das características extraídas de áudios de cada dataset e suas respectivas classes.

D. Escolha dos Modelos

A escolha dos modelos ResNet-152 e MobileNetV2 para este trabalho foi baseada em estudos anteriores como o de [9] que demonstraram sua eficácia em tarefas de classificação de áudio e em particular, o estudo realizado por Lacerda [10] destacou a aplicação dessas arquiteturas em um contexto semelhante ao deste trabalho.

E. Treinamento

No caso das redes ResNet152V2 e MobileNetV2, com o objetivo de usar o aprendizado por transferência foi carregado os modelos pré-treinados no dataset ImageNetV2 um conjunto de dados com milhões de imagens em milhares de categorias [25].

Para adaptar as redes ao tamanho dos atributos de entrada foi então modificado a camada de entrada para aceitar imagens de 29x251 *pixels* e a camada de saída para categorizar as imagens em uma de 4 classes. Estas redes foram então treinadas por 15 épocas com um tamanho de lote de 32, utilizando a biblioteca *PyTorch*. Durante o treinamento, o critério de *Early Stopping* foi aplicado com uma paciência de 2 épocas. Então foi feito o ajuste fino na última camada no novo conjunto de dados para melhor ajustar o modelo às especificidades do problema em questão.

F. Avaliação das métricas

Foram utilizadas métricas comumente empregadas na classificação multi-classe, isso é,acurácia (*accuracy*), precisão (*precision*), revocação (*recall*), *F1-score* e suporte (*support*).

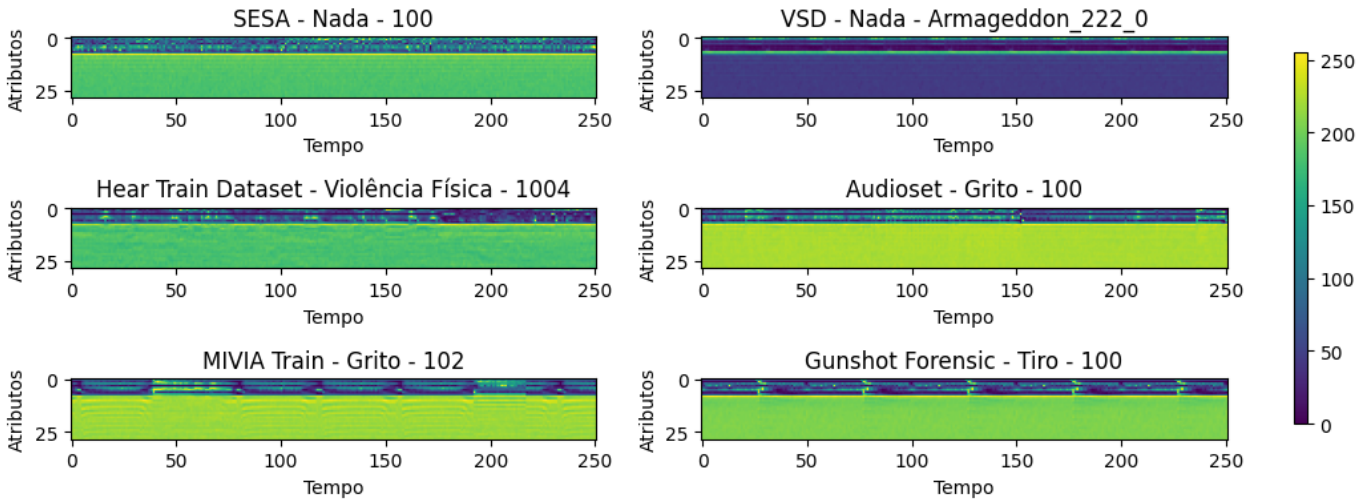

Figura 1. Imagens geradas pela concatenação dos oito atributos extraídos de áudios de cada dataset e suas classes. Cada imagem apresenta as sete primeiras linhas representando os atributos AE, RMS, ZCR, BER, SC, BW e SF e a 22 últimas linhas com os 22 MFCCs. Os valores dos atributos foram normalizados para intensidades de 0 a 255, formando imagens de um canal com dimensões de 29x251 pixels.

Foi utilizado o teste T pareado para comparação dos modelos ResNet-152 e MobileNetV2. O teste T pareado é uma técnica estatística amplamente utilizada para comparar duas médias e determinar se há diferença estatisticamente significativa entre elas. Para conduzir o teste T pareado, é necessário calcular a diferença entre as métricas dos dois modelos em avaliada em cada *fold*. Em seguida, aplica-se o teste T pareado às diferenças calculadas para verificar se a média das diferenças é significativamente diferente de zero [20], como mostrado na eq. 9

$$t = \frac{\sqrt{N} * \overline{m}}{s_d} \tag{9}$$

onde $\overline{m}$ é a média das diferenças entre cada fold, s_d é o desvio padrão das diferenças, e N é o número de observações.

A hipótese nula do teste T pareado afirma que não há diferença significativa no desempenho dos modelos na métrica avaliada. Um valor-p menor que 0.05 foi escolhido como critério para rejeitar a hipótese nula, indicando uma diferença estatisticamente significativa entre os modelos para a métrica específica em consideração.

G. Validação

Durante a etapa de validação, adotou-se a técnica de validação cruzada *k-fold* com o objetivo de avaliar a capacidade de generalização dos modelos. Foram utilizados 5 *folds* para dividir o conjunto de dados em conjuntos de treinamento e validação.

IV. Desenvolvimento do Programa

O programa destinado à detecção de cenas acústicas violentas foi desenvolvido empregando a linguagem de programação *Python*. Como ilustrado na Figura 2, o programa opera seguindo uma sequência estruturada de etapas que envolve: coleta de dados, ingestão, pré-processamento, carregamento, estratificação e, finalmente, o treinamento do modelo.

O código-fonte do programa desenvolvido está disponibilizado no GitHub violence-detection-acoustic-scenes[1].

V. Resultados

As métricas de precisão, revocação e F1-Score estão apresentadas na Tabela II e mostra os valores da métrica para cada classe e modelo.

Tabela II
Resultados das métricas de precisão, revocação e F1-Score para os modelos

Classes	Modelo	Precisão	Revocação	F1-Score
Nada	ResNET152	0.870	0.859	0.864
	MobileNet	0.866	0.842	0.854
Grito	ResNET152	0.666	0.774	0.716
	MobileNet	0.665	0.756	0.707
V. Física	ResNET152	0.866	0.861	0.863
	MobileNet	0.845	0.865	0.855
Tiro	ResNET152	0.870	0.804	0.835
	MobileNet	0.881	0.815	0.846

Analisando as métricas de precisão, revocação e F1-Score para cada classe, pode-se observar que ambos os modelos apresentam resultados variados, dependendo da classe em questão.

A classe denominada "grito" apresentou resultados notavelmente discrepantes em relação às outras classes durante a avaliação. Essa disparidade pode ser atribuída, em parte, à própria variabilidade inerente ao processo humano de emissão de gritos. Os gritos, por natureza, possuem características acústicas distintas, tais como diferentes durações e timbres, e também podem expressar uma ampla gama de emoções humanas, como medo, surpresa ou esforço físico. Essa diversidade

[1]GitHub: https://github.com/caleo-hub/violence-detection-acoustic-scenes

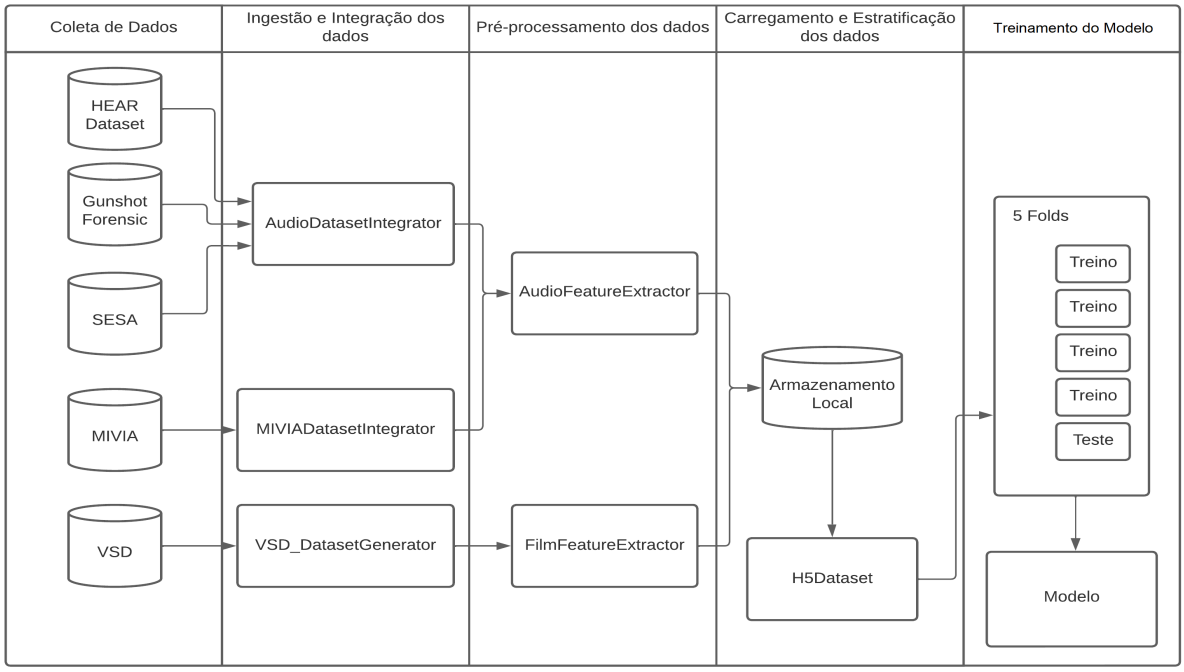

Figura 2. *Pipeline* do fluxo de dados do programa

intrínseca aos gritos pode dificultar a generalização adequada dos modelos em relação às métricas de avaliação utilizadas.

Ao analisar a acurácia média dos modelos nos 5 *folds* na Tabela III, observa-se que ambos apresentam um desempenho semelhante, com as médias totais de cada modelo próximas de 85%. É possível fazer uma comparação livre, mas relevante, com estudos recentes que empregaram técnicas de *Deep Learning* em combinação com mel-espectrogramas visto na Tabela IV.

Tabela III
RESULTADOS DE ACURÁCIAS DOS MODELOS RESNET152 E MOBILENET NOS 5 FOLDS

Modelo	Fold1	Fold2	Fold3	Fold4	Fold5
ResNet152	0.848	0.830	0.842	0.863	0.863
MobileNet	0.850	0.831	0.845	0.847	0.831

Tabela IV
COMPARAÇÃO COM OS ESTUDOS DE DETECÇÃO DE VIOLÊNCIA

Estudo	Arquitetura	Acurácia
Souto [6]	SVM	73,14%
Lacerda [10]	MobileNetV2	78,9%
Santos, C.	MobileNetV2	84,1%
Santos, C.	ResNet152	84,9%

A. Seleção do Modelo

Inicialmente, realizou-se o teste T pareado para a métrica de acurácia. O resultado do teste T para a acurácia indica que não há diferença significativa entre os modelos (valor-p $= 0.197$), considerando um valor-p escolhido de 0.05, o que corresponde a um nível de confiança de 95%.

Em seguida, aplicamos o teste T pareado para as métricas de classificação de cada classe. Os resultados dos testes T pareados para as métricas de precisão, revocação e F1-Score de cada classe estão apresentados na Tabela V.

Tabela V
RESULTADOS DO TESTE T PAREADO PARA AS MÉTRICAS DE CLASSIFICAÇÃO

Métrica	Classe	Valor-p
Precisão	Nada	0.402
	Grito	0.855
	V. Física	0.300
	Tiro	0.440
Revocação	Nada	0.218
	Grito	0.153
	V. Física	0.675
	Tiro	0.582
F1-Score	Nada	0.214
	Grito	**0.011**
	V. Física	0.419
	Tiro	0.0743

Ao analisar os resultados da Tabela V, observamos que a métrica 'F1-Score do Grito' apresenta uma diferença estatisticamente significativa no desempenho de classificação para a classe 'Grito'. Isso sugere que o modelo ResNet152 possui uma melhor aptidão na tarefa de detectar gritos em comparação ao modelo MobileNetV2.

Considerando que a acurácia entre os modelos não apresentou diferença significativa, a escolha entre os modelos dependerá das necessidades e prioridades do problema em

questão. Se a detecção de gritos for uma prioridade importante, o modelo ResNet152, apesar de ser maior e mais complexo, pode ser preferível devido ao seu desempenho superior na classificação dessa classe específica. Por outro lado, se a eficiência computacional ou a limitação de recursos forem fatores relevantes, o modelo MobileNetV2, sendo mais leve e com menos parâmetros, pode ser uma opção mais adequada.

VI. Conclusão

Em suma, este trabalho apresenta uma contribuição para a pesquisa de multi-classificação de eventos de violência utilizando o áudio. Essa investigação contribuí para o avanço da detecção de violência em áudio e sua aplicação em diversos contextos, como segurança pública e monitoramento de ambiente e mostra que é possível treinar um modelo de aprendizado de máquina para que seja capaz de identificar e classificar quatro cenários: Sem violência (Nada), Grito, Violência Física e Tiros, utilizando atributos de frequência, de energia e do espectro do áudio observado.

Este trabalho apresenta diversas possibilidades de estudos futuros que podem aprimorar e expandir a pesquisa sobre a detecção de eventos de violência em áudio. Algumas das direções promissoras são:

- **Desenvolvimento de um *dataset* maior e mais abrangente:** A criação de um conjunto de dados expandido, especialmente para as classes de "grito" e "tiro", permitiria um treinamento mais robusto e uma avaliação mais precisa do desempenho do sistema em diferentes situações. A inclusão de amostras sintéticas geradas por técnicas de *data augmentation* também pode contribuir para aumentar a diversidade do conjunto de dados e melhorar a capacidade de generalização do sistema.
- **Testes em diferentes ambientes e condições:** Realizar testes e avaliações em diversos ambientes e condições é fundamental para verificar a robustez e a eficácia do sistema em cenários reais. Coletar dados em ambientes diversos, considerando a presença de ruído, variações na duração dos áudios e outros fatores que possam impactar a detecção de eventos de violência em áudio, pode fornecer insights importantes sobre o desempenho do sistema em situações mais desafiadoras.

Referências

[1] H. Ferreira and M. K. Soares, "Violência e Segurança Pública: uma síntese da produção da Diest nos últimos dez anos" ,*Boletim de Análise Político-Institucional*, pp. 129-144, Dec. 2021.

[2] Dalila Duraes, Francisco S. Marcondes, Filipe Gonçalves, Joaquim Fonseca, José Machado, and Paulo Novais, *Detection Violent Behaviors: A Survey*, pp. 106-116, 2021. 10.1007/978-3-030-58356-9$_1$1

[3] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, article no. 16, 2011.

[4] R. S. de Lima, S. Bueno, I. Sobral, and D. Pacheco, "Câmeras na farda reduzem a letalidade policial?," *GV EXECUTIVO*, vol. 21, no. 2, pp. 13-21, 2022.

[5] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio Surveillance," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, article no. 61, 2016.

[6] H. Souto, R. Mello, and A. Furtado, "An acoustic scene classification approach involving domestic violence using machine learning," in *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2019)*, pp. 705-716, 2019.

[7] J. L. Rouas, J. Louradour, and S. Ambellouis, "Audio events detection in public transport vehicle," *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, pp. 733-738, 2006.

[8] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Proceedings of the 18th ISMIR Conference*, H. X. Cunningham, S. J. Turnbull, Z. Duan, Eds. Suzhou, China: International Society for Music Information Retrieval, 2017, pp. 486-493.

[9] J. Abeßer, "A Review of Deep Learning Based Methods for Acoustic Scene Classification," *Applied Sciences*, vol. 10, no. 6, article no. 2020, 2020. Available: https://www.mdpi.com/2076-3417/10/6/2020. ISSN: 2076-3417.

[10] T. B. Lacerda, P. Miranda, A. Câmara, A. Paula, C. Furtado, "Deep Learning and Mel-spectrograms for Physical Violence Detection in Audio," *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pp. 268-279, 2021.

[11] M. Schedi, M. Sjoberg, I. Mironica, B. Ionescu, V. L. Quang, Y.-G. Jiang, and C.-H. Demarty, "VSD2014: A dataset for violent scenes detection in Hollywood movies and web videos," *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 1-6, Jun. 2015.

[12] T. Spadini, "Sound Events for Surveillance Applications," *Zenodo*, Oct. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3519845

[13] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Reliable detection of audio events in highly noisy environments," *Pattern Recognition Letters*, vol. 65, pp. 22-28, Aug. 2015. doi: 10.1016/j.patrec.2015.06.026

[14] R. Mettu, T. Weller, L. Haag, M. Haag, R. Lilien, and J. Housman, "Gunshot Audio Forensics Dataset," [Online]. Available: http://cadreforensics.com/audio/ doi: 10.1371/journal.pone.0183754

[15] J. Gemmeke, G. Heigold, A. Ewert, O. Puhr, and B. Schuller, "Audioset: An ontology and dataset for audio events," *Proceedings of the 2017 ACM on Multimedia Conference*, pp. 845-854, 2017.

[16] Mathuranathan, "Extract envelope, phase using Hilbert transform: Demo," 2017. [Online]. Available: https://www.gaussianwaves.com/2017/04/extract-envelope-instantaneous-phase-frequency-hilbert-transform/

[17] A. Lerch, *An introduction to audio content analysis: applications in signal processing and music informatics*, John Wiley & Sons, Ltd, 2012.

[18] V. Velardo, "Frequency-Domain Audio Features," *The Sound of AI*, 2020. [Online]. Available: https://www.youtube.com/watch?v=3-bjAoAxQ9o

[19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016. http://www.deeplearningbook.org

[20] M. Politi, *Paired t-test to evaluate Machine Learning classifiers using Python*, Towards Data Science, Jul. 6, 2022. Available at: https://towardsdatascience.com/paired-t-test-to-evaluate-machine-learning-classifiers-1f395a6c93fa.

[21] Brasil, *Anuário Brasileiro de Segurança Pública*, Fórum Brasileiro de Segurança Pública, 2022. https://forumseguranca.org.br/wp-content/uploads/2022/06/anuario-2022.pdf?v=4

[22] C. Javdani, *Detecting Workplace Violence With Sound Detection And Audio Analytics*, Julho, 2017. https://www.securityinformed.com/insights/audio-analytics-reduce-workplace-violence-co-2931-ga.23323.html

[23] K. Doshi, *Audio Deep Learning Made Simple (Part 1): State-of-the-Art Techniques*, Towards Data Science, Fevereiro, 2021.

[24] Papia Nandi, *CNNs for Audio Classification: A primer in deep learning for audio classification using tensorflow*, Towards Data Science, Mar 24, 2021. [Online]. Available: https://towardsdatascience.com/cnns-for-audio-classification-6244954665ab. Accessado em: 21/07/2023.

[25] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar, *Do ImageNet Classifiers Generalize to ImageNet?*, CoRR, vol. abs/1902.10811, 2019. [Online]. Available: http://arxiv.org/abs/1902.10811. eprint: arXiv:1902.10811.