Universidade Federal do ABC
Centro de Matemática, Computação e Cognição
Programa de Graduação em Ciência da Computação

Cidades Inteligentes: Aplicação de Inteligência Artificial na Identificação e Classificação de Sons Urbanos

Guilherme Klinkerfuss Guimarães Pereira

Santo André - SP, Abril de 2025

Guilherme Klinkerfuss Guimarães Pereira

Cidades Inteligentes: Aplicação de Inteligência Artificial na Identificação e Classificação de Sons Urbanos

Projeto de Graduação em Computação apresentada ao Programa de Graduação em Ciência da Computação (área de concentração: Redes Neurais), como parte dos requisitos necessários para a obtenção do Título de Bacharel em Ciência da Computação.

Universidade Federal do ABC – UFABC

Centro de Matemática, Computação e Cognição

Programa de Graduação em Ciência da Computação

Orientador: Hugo Puertas de Araújo

Santo André - SP

Abril de 2025

Guilherme Klinkerfuss Guimarães Pereira

Cidades Inteligentes: Aplicação de Inteligência Artificial na Identificação e Classificação de Sons Urbanos/ Guilherme Klinkerfuss Guimarães Pereira. – Santo André - SP, Abril de 2025-

Orientador: Hugo Puertas de Araújo

Projeto de Graduação (Bacharelado) – Universidade Federal do ABC – UFABC
Centro de Matemática, Computação e Cognição
Programa de Graduação em Ciência da Computação, Abril de 2025.

1. Cidade Inteligentes. 2. Inteligência Artificial. 3. Redes Neurais. 4. Sons Urbanos. 5. Classificação de Sons. I. Hugo Puertas de Araújo. II. Universidade Federal do ABC. III. Centro de Matemática, Computação e Cognição. IV. Cidades Inteligentes: Aplicação de Inteligência Artificial na Identificação e Classificação de Sons Urbanos

Resumo

O avanço da urbanização intensificou os desafios enfrentados por grandes centros urbanos, especialmente nas áreas de mobilidade, segurança pública, saúde e eficiência energética. Nesse contexto, o conceito de cidades inteligentes surge como uma alternativa estratégica, integrando tecnologias como a Internet das Coisas (IoT) e a Inteligência Artificial (IA) para promover ambientes urbanos mais eficientes e sustentáveis. Com isso, este trabalho propõe o desenvolvimento de uma solução de baixo custo para a análise e classificação de sons urbanos, visando identificar eventos relevantes como colisões, disparos de arma de fogo, tráfego intenso e sirenes de emergência. Para realizar essa tarefa, é proposta uma abordagem que utiliza sensores acessíveis e bases de dados públicas, aplicando redes neurais convolucionais (CNNs) com fusão de evidências via Teoria de Dempster-Shafer, por meio do modelo TSCNN-DS.

Palavras-chaves: Cidade Inteligentes, Inteligência Artificial, Redes Neurais, Sons Urbanos, Classificação de Sons.

Abstract

The advancement of urbanization has intensified the challenges faced by large urban centers, especially in the areas of mobility, public safety, health, and energy efficiency. In this context, the concept of smart cities emerges as a strategic alternative, integrating technologies such as the Internet of Things (IoT) and Artificial Intelligence (AI) to promote more efficient and sustainable urban environments. Therefore, this work proposes the development of a low-cost solution for the analysis and classification of urban environments, which identifies relevant events such as collisions, gunshots, heavy traffic, and emergency sirens. To accomplish this task, an approach is proposed that uses accessible sensors and public databases, applying convolutional neural networks (CNNs) with evidence fusion via Dempster-Shafer Theory, through the TSCNN-DS model.

Keywords: Smart Cities, Artificial Intelligence, Neural Networks, Urban Sounds, Sound Classification.

Lista de ilustrações

Lista de abreviaturas e siglas

ABNT	Associação Brasileira de Normas Técnicas
WAV	Waveform Audio File Format
CNN	Convolutional Neural Network (Rede Neural Convolucional)
MFCC	Mel-Frequency Cepstral Coefficients
DCT	Discrete Cosine Transform
DS	Dempster-Shafer
STFT	Short-Time Fourier Transform
TSCNN-DS	Two-Stream Convolutional Neural Network with Dempster-Shafer
AI	Artificial Intelligence (Inteligência Artificial)
IoT	Internet of Things (Internet das Coisas)
ESP32	Microcontrolador com conectividade Wi-Fi e Bluetooth

Sumário

Introdução

O conceito de cidades inteligentes vem sendo amplamente estudado e ganhando notoriedade nas últimas duas décadas. Apesar de não haver uma definição clara, cidades inteligentes podem ser interpretadas como um ambiente tecnológico que visa melhorar a vida dos cidadãos. As características principais das cidades inteligentes incluem economia, mobilidade, governança, meio ambiente e qualidade de vida inteligentes. Com isso, diversas áreas vêm ganhando destaque, tais quais a gestão de energia, transporte, saúde, educação e segurança pública (NASTJUK; TRANG; PAPAGEORGIOU, 2022)

No Brasil, os estados pertencentes às regiões Sudeste e Sul possuem o maior número de Smart Cities, em que o destaque vai para São Paulo. Este estado, por sua vez, concentra o maior número de cidades inteligentes do Brasil. No país, as principais áreas de desenvolvimento de soluções estão nas áreas de mobilidade, gestão de resíduos, segurança, infraestrutura urbana, planejamento e gestão, qualidade de vida e soluções ecológicas respectivamente (SANTOS, 2021).

Para uma abordagem mais efetiva, principalmente em países em desenvolvimento, soluções mais baratas podem trazer melhores resultados comparados a outros tipos de soluções que envolvem um poder aquisitivo maior (WEISS; PEREZ, 2024). Desta forma, o intuito deste projeto é a criação de uma solução inspirada pelo trabalho de Su et al. (2019), que utilizou redes neurais convolucionais (CNNs) e fusão de dados para classificação de sons ambientais (como buzinas e tiros no dataset UrbanSound8K), este projeto propõe uma adaptação direcionada ao contexto brasileiro, transformando sons urbanos em indicadores estratégicos para quatro áreas críticas.

Na mobilidade, a classificação de buzinas e ruídos de tráfego permitirá otimizar semáforos e rotas, combatendo os R$ 267 bilhões perdidos anualmente com congestionamentos (RAMALHO, 2018). Para a gestão de energia, a detecção de ruídos de geradores e máquinas ajudará a reduzir desperdícios que, além dos custos financeiros, representam riscos à saúde pública (ANDRADINA; SANTOS, 2024). Na saúde, a identificação de sirenes de ambulâncias possibilitará priorizar corredores de emergência, crucial para um país que ocupa a terceira posição em acidentes de trânsito (CARDOSO, 2024). Já na segurança pública, o monitoramento de tiros e ruídos violentos permitirá acionamento rápido de autoridades em um dos países mais violentos do mundo (PEREIRA, 2025).

A solução proposta baseia-se em três etapas principais: seleção de amostras adequadas do UrbanSound8K e possivelmente captação própria através de equipamento de baixo custo, como ESP32 equipado com microfone e circuitaria de apoio; (2) classificação adaptada usando um modelo TSCNN-DS modificado com fusão de dados via Teoria de

Dempster-Shafer; e (3) sugestão para integração com sistemas urbanos existentes.

Nesse âmbito, a presente pesquisa articula uma abordagem inovadora que integra: custo significativamente menor que soluções tradicionais, utilizando DataSets públicos e possível inclusão de hardware aberto; acompanhado de tempo de resposta em segundos para eventos crítico e perfeito alinhamento com políticas públicas como a Lei Brasileira de Cidades Inteligentes (2023) e os Objetivos de Desenvolvimento Sustentável da ONU, particularmente os ODS 11 (Cidades Sustentáveis) (ONU, 2015a) e ODS 3 (Saúde e Bem-Estar) (ONU, 2015b).

Esta solução não apenas aborda problemas urbanos prementes, mas representa um avanço significativo na construção de cidades mais justas e inclusivas. Ao desenvolver uma tecnologia acessível e adaptável às diferentes realidades brasileiras, o projeto promove a democratização da inovação, permitindo que municípios com menos recursos também possam usufruir dos benefícios das Smart Cities. A abordagem proposta vai além da eficiência técnica - ela reconhece que a verdadeira inteligência urbana deve estar a serviço das pessoas, transformando dados em melhorias concretas para o cotidiano dos cidadãos.

Na prática, isso significa que os sons das cidades, antes considerados meros ruídos, passam a ser valiosas fontes de informação para tomada de decisões que impactam diretamente a qualidade de vida. Desde o morador de periferia que sofre com a violência até o trabalhador que perde horas no trânsito, todos se beneficiam quando a tecnologia é aplicada para resolver problemas reais. O projeto se alinha assim com uma visão de sociedade onde o progresso tecnológico anda de mãos dadas com a redução das desigualdades e o fortalecimento dos serviços públicos essenciais.

Esta perspectiva é particularmente relevante no contexto brasileiro, marcado por profundas disparidades regionais e sociais (BARROS; MENDONÇA, 1996). Quando são priorizadas soluções de baixo custo e alta eficiência, a pesquisa contribui para reduzir o abismo tecnológico entre grandes centros urbanos e cidades do interior, entre bairros ricos e comunidades carentes. Portanto, adiante, veremos que este projeto insere mais do que uma ferramenta de gestão urbana, a solução proposta se configura como um instrumento de transformação social, demonstrando como a inteligência artificial pode ser colocada a serviço do bem comum e do desenvolvimento sustentável.

Motivação

A oportunidade de aplicar métodos de machine learning a problemas urbanos reais motivou fortemente o desenvolvimento deste projeto. Ao adaptar o trabalho de (SU et al., 2019) sobre classificação de sons ambientais, propomos uma solução acessível a diferentes contextos brasileiros, contribuindo para a democratização de tecnologias e para um impacto mais amplo na sociedade e no uso da infraestrutura urbana.

Objetivos

Objetivo Geral

Desenvolver uma solução baseada em inteligência artificial, utilizando redes neurais convolucionais e a teoria da evidência de Dempster-Shafer, para classificar sons urbanos em cidades inteligentes, visando apoiar decisões em áreas críticas como mobilidade, segurança pública, saúde e gestão energética.

Objetivos Específicos

- Investigar o estado da arte sobre análise de sons ambientais aplicados ao contexto urbano e ao conceito de cidades inteligentes;
- Adaptar o modelo TSCNN-DS para a classificação de sons urbanos, utilizando bases de dados públicas como o UrbanSound8K;
- Implementar um pipeline de coleta, pré-processamento e inferência de dados sonoros, com foco em eventos urbanos relevantes (como buzinas, colisões, sirenes e disparos);
- Avaliar o desempenho do modelo proposto quanto à acurácia, tempo de resposta e viabilidade em contextos de baixo custo;
- Propor formas de integração da solução com sistemas urbanos existentes ou dispositivos embarcados acessíveis, como microcontroladores com microfone (e.g., ESP32);
- Discutir os benefícios sociais e as possibilidades de ampliação da proposta para diferentes cenários urbanos no Brasil.

Parte I

Proposta

1 Sistema Proposto

1.1 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNAs) são modelos computacionais inspirados no funcionamento do cérebro humano, projetados para processar informações de maneira paralela e distribuída. Em essência, uma RNA é composta por unidades de processamento chamadas de *neurônios artificiais*, que se organizam em camadas e são interconectadas por *pesos sinápticos*, os quais representam a intensidade das conexões entre as unidades (HAYKIN, 2001; FLECK; RIEDER; LIMA, 2016).

O conceito de neurônio artificial surgiu em 1943, com o modelo de McCulloch e Pitts, que representava matematicamente a atividade de um neurônio biológico. Posteriormente, o modelo *Perceptron*, proposto por Rosenblatt em 1958, introduziu a ideia de aprendizado supervisionado, permitindo que a rede ajustasse seus pesos a partir de exemplos. Com o avanço da capacidade computacional e o desenvolvimento de novos algoritmos de aprendizado, especialmente o de retropropagação do erro (*backpropagation*), as RNAs tornaram-se amplamente aplicáveis a problemas complexos de classificação, regressão e reconhecimento de padrões.

Uma RNA típica é organizada em camadas de entrada, camadas ocultas e camadas de saída. Os sinais de entrada são processados pelas camadas intermediárias, nas quais ocorrem transformações não lineares por meio de *funções de ativação* — como a sigmoide, ReLU ou tangente hiperbólica — que permitem à rede modelar relações complexas entre variáveis. O processo de ajuste dos pesos sinápticos é denominado *treinamento*, e pode ser supervisionado, quando há um rótulo esperado para cada entrada, ou não supervisionado, quando a rede busca autonomamente padrões nos dados (BISHOP, 1995).

Entre as principais vantagens das RNAs estão a capacidade de aprender com exemplos, generalizar para dados não vistos, modelar sistemas não lineares e adaptar-se a novos contextos. No entanto, essas redes também apresentam limitações, como a necessidade de grandes volumes de dados para treinamento, o alto custo computacional e o comportamento de "caixa-preta", que dificulta a interpretação das decisões (FLECK; RIEDER; LIMA, 2016).

No contexto deste trabalho, as redes neurais são aplicadas para a análise e classificação de sons urbanos, aproveitando sua habilidade de reconhecer padrões complexos em espectrogramas de áudio. Essa abordagem é especialmente relevante em ambientes de cidades inteligentes, nos quais a interpretação automática de sons pode fornecer informações valiosas sobre o trânsito, a segurança pública e o ambiente urbano em tempo real.

1.2 Convolução e Redes Neurais Convolucionais (CNNs)

A convolução é uma operação matemática fundamental utilizada em diversos campos da ciência e engenharia, especialmente no processamento de sinais e imagens. Em termos gerais, a convolução permite combinar duas funções — uma representando o sinal de entrada e outra denominada *kernel* ou *filtro* — para produzir uma terceira função que expressa como a forma de uma é modificada pela outra. No contexto de redes neurais, a convolução é aplicada sobre matrizes de dados (como imagens ou espectrogramas), de modo que o filtro percorre a entrada extraindo características locais relevantes, como bordas, texturas e padrões estruturais (GOODFELLOW; BENGIO; COURVILLE, 2016; LECUN et al., 1998).

Matematicamente, para uma entrada bidimensional I e um filtro K de dimensões menores, a convolução discreta é definida por:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot K(m,n)$$

onde (i, j) representam as coordenadas da saída, e o somatório é realizado sobre a vizinhança do ponto de entrada. O resultado S constitui um novo mapa de características (*feature map*) que realça os padrões reconhecidos pelo filtro.

O principal benefício dessa operação é a capacidade de **preservar a relação espacial** entre os dados, reduzindo o número de parâmetros quando comparada às redes totalmente conectadas. Além disso, a convolução explora duas propriedades essenciais: **localidade**, pois cada neurônio é conectado apenas a uma pequena região da entrada, e **compartilhamento de pesos**, que garante que o mesmo filtro seja aplicado a toda a imagem, promovendo invariância a deslocamentos (LECUN; BENGIO; HINTON, 2015).

As Redes Neurais Convolucionais (CNNs, do inglês *Convolutional Neural Networks*) são arquiteturas projetadas especificamente para processar dados que possuem estrutura de grade, como imagens bidimensionais ou sinais de áudio convertidos em espectrogramas. Uma CNN é composta por múltiplas camadas de convolução, seguidas geralmente por camadas de *pooling* e, ao final, camadas totalmente conectadas. As camadas convolucionais atuam como extratoras de características, enquanto as camadas de *pooling* reduzem a dimensionalidade dos mapas de características, concentrando as informações mais relevantes e reduzindo o custo computacional.

Ao longo das últimas décadas, as CNNs tornaram-se o principal paradigma em visão computacional e reconhecimento de padrões, graças à sua capacidade de aprender automaticamente representações hierárquicas. As primeiras camadas aprendem padrões simples (como bordas e contornos), enquanto as camadas mais profundas identificam combinações complexas desses padrões, permitindo o reconhecimento de objetos, faces e sons (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

No contexto do processamento de áudio, as CNNs são particularmente úteis quando aplicadas a **espectrogramas** — representações visuais do sinal sonoro ao longo do tempo e das frequências. Nesses casos, as operações de convolução permitem capturar tanto a **estrutura temporal** quanto as **características espectrais** dos sons, possibilitando a distinção entre diferentes fontes sonoras. Essa abordagem tem sido amplamente utilizada em tarefas como reconhecimento de fala, detecção de eventos acústicos e classificação de sons ambientais, mostrando-se altamente eficaz para aplicações em cidades inteligentes, onde a detecção automática de padrões sonoros pode auxiliar na segurança e mobilidade urbana.

Assim, as redes neurais convolucionais se configuram como uma das ferramentas mais poderosas do aprendizado profundo, oferecendo uma combinação de desempenho, eficiência e generalização que as torna ideais para o desenvolvimento de sistemas de análise e interpretação de sons urbanos.

1.3 Sistema Proposto

O sistema proposto tem como objetivo o desenvolvimento de um modelo de aprendizado profundo capaz de classificar sons urbanos relevantes para o contexto de cidades inteligentes. A abordagem adotada baseia-se na utilização de Redes Neurais Convolucionais (CNNs), amplamente empregadas em tarefas de reconhecimento de padrões e classificação de sinais estruturados, como imagens e espectrogramas (GOODFELLOW; BENGIO; COURVILLE, 2016).

No processamento de áudio, representações espectrais como Mel-Spectrograms podem ser tratadas como imagens bidimensionais, permitindo que redes convolucionais identifiquem padrões característicos associados a diferentes eventos sonoros (SU et al., 2019). Dessa forma, a utilização de CNNs torna-se adequada para a tarefa de classificação de sons urbanos.

O pipeline de processamento adotado neste trabalho foi inspirado em implementações existentes na plataforma Kaggle, adaptadas para atender aos objetivos específicos desta pesquisa (BANSAL, 2020). O modelo é treinado diretamente a partir das representações acústicas extraídas dos sinais de áudio, sem a utilização de técnicas de *transfer learning*.

Como referência metodológica para o desenvolvimento do sistema, foi analisado o trabalho de Su et al. (SU et al., 2019), que propõe o modelo TSCNN-DS (*Two-Stream Convolutional Neural Network with Dempster–Shafer Evidence Theory*). Nesse modelo, duas redes neurais convolucionais são utilizadas para extrair diferentes representações do sinal acústico, sendo suas saídas posteriormente combinadas por meio da Teoria da Evidência de Dempster–Shafer. Essa abordagem permite integrar evidências provenientes

de diferentes representações do áudio, aumentando a robustez do processo de decisão em tarefas de classificação de sons ambientais.

Inspirado nessa abordagem, o sistema desenvolvido neste trabalho utiliza uma arquitetura baseada em duas redes neurais convolucionais independentes, cada uma responsável por processar uma representação distinta do sinal acústico. Especificamente, são utilizadas as representações *Mel-Spectrogram* e *Mel-Frequency Cepstral Coefficients* (MFCC), amplamente empregadas em tarefas de análise e classificação de áudio. Após o treinamento independente das duas redes, as previsões geradas por cada modelo são combinadas utilizando a regra de combinação da Teoria da Evidência de Dempster–Shafer, caracterizando um processo de fusão no nível de decisão (*decision-level fusion*).

O treinamento e a avaliação do modelo são realizados utilizando o conjunto de dados **UrbanSound8K**, amplamente utilizado em pesquisas na área de classificação de sons ambientais. Esse conjunto contém milhares de amostras de áudio rotuladas pertencentes a diferentes categorias de eventos sonoros urbanos.

Para alinhar o estudo com o contexto de cidades inteligentes, as classes originais do conjunto de dados foram agrupadas em duas macro-categorias temáticas de interesse:

- **Mobilidade urbana**: inclui sons relacionados ao tráfego e ao transporte urbano, como buzinas, motores em marcha lenta e ruídos de ferramentas utilizadas em infraestrutura urbana. Esses eventos sonoros podem fornecer informações relevantes sobre condições de tráfego e atividades urbanas.
- **Segurança pública**: inclui eventos sonoros potencialmente associados a situações de risco ou alerta, como sirenes, disparos de arma de fogo e latidos de cães, que podem ser utilizados em sistemas de monitoramento e resposta a incidentes.

O sistema proposto segue três etapas principais:

1. Pré-processamento dos sinais de áudio e extração de características acústicas, incluindo MFCC e Mel-Spectrogram;
2. Treinamento de duas Redes Neurais Convolucionais independentes, cada uma especializada em uma das representações acústicas;
3. Combinação das previsões geradas pelos modelos utilizando a regra de Dempster–Shafer e avaliação do desempenho por meio de métricas de classificação, como acurácia, matriz de confusão e *classification report*.

Com isso, o sistema desenvolvido foi capaz de identificar padrões acústicos relevantes presentes no ambiente urbano. A identificação automática desses eventos sonoros contribui

para aplicações de monitoramento em cidades inteligentes, auxiliando na análise de mobilidade urbana e no suporte a sistemas de segurança pública, fornecendo informações úteis para a gestão e o planejamento urbano.

1.4 Teoria da Evidência de Dempster–Shafer

A Teoria da Evidência de Dempster–Shafer (DS), também conhecida como Teoria das Funções de Crença, foi introduzida inicialmente por Dempster e posteriormente formalizada por Shafer. Essa teoria propõe um arcabouço matemático para representar e combinar incertezas, sendo amplamente utilizada como uma alternativa e extensão aos modelos probabilísticos tradicionais, como o Bayesiano.

Diferentemente da probabilidade clássica, que exige que toda a incerteza seja distribuída exclusivamente entre classes mutuamente exclusivas, a teoria DS permite a representação explícita da ignorância. Isso significa que, além de indicar o grau de crença em uma determinada hipótese, o modelo pode expressar a ausência de evidência suficiente para tomar uma decisão clara. Essa característica torna a teoria DS particularmente adequada para problemas reais de classificação, nos quais há ambiguidade ou sobreposição entre classes.

Essa abordagem tem sido explorada com sucesso em tarefas de classificação de sons ambientais. Em especial, Su et al. (SU et al., 2019) propõem o uso da teoria de Dempster–Shafer para realizar fusão em nível de decisão entre diferentes redes neurais convolucionais, demonstrando ganhos de robustez e confiabilidade no reconhecimento de sons ambientais.

1.4.1 Quadro de discernimento

A base da teoria DS é o chamado *quadro de discernimento*, denotado por Θ, que representa o conjunto de todas as hipóteses possíveis do problema. Em tarefas de classificação, Θ corresponde ao conjunto de todas as classes consideradas.

No contexto deste trabalho, o conjunto de dados UrbanSound8K possui 10 classes de sons ambientais. Assim, o quadro de discernimento original pode ser definido como:

$$\Theta = \{A_1, A_2, \ldots, A_{10}\}$$

onde cada hipótese A_i representa uma classe distinta de som, como *car horn*, *dog bark*, *siren*, entre outras.

Entretanto, neste trabalho, essas classes foram agrupadas em macro-categorias com base em sua relevância para aplicações em cidades inteligentes. Dessa forma, o problema de classificação é reformulado em um espaço reduzido de hipóteses.

As macro-classes consideradas são:

- **Mobilidade Urbana**: *car_horn*, *engine_idling*, *jackhammer*
- **Segurança Pública**: *gun_shot*, *siren*, *dog_bark*

Assim, o quadro de discernimento adotado na aplicação da teoria de Dempster-Shafer é definido como:

$$\Theta = \{\text{mobilidade\_urbana}, \text{seguranca\_publica}\}$$

1.4.2 Função de massa

A principal ferramenta da teoria DS é a *função de massa*, também chamada de atribuição básica de probabilidade. Essa função é definida como:

$$M : \mathcal{P}(\Theta) \rightarrow [0, 1]$$

onde $\mathcal{P}(\Theta)$ representa o conjunto potência de Θ. A função de massa atribui um valor de crença a cada subconjunto de Θ, indicando o quanto a evidência disponível sustenta aquela hipótese.

A função de massa deve obedecer a duas restrições fundamentais:

- A soma das massas atribuídas a todos os subconjuntos de Θ deve ser igual a 1:

$$\sum_{A \subseteq \Theta} M(A) = 1$$

- Nenhuma massa pode ser atribuída ao conjunto vazio:

$$M(\varnothing) = 0$$

Quando essas condições são satisfeitas, a função de massa é considerada normalizada. Na prática, isso permite que o modelo atribua crença tanto a classes específicas quanto a conjuntos de classes, representando explicitamente situações de incerteza ou ambiguidade quando a evidência não é suficiente para uma decisão precisa.

1.4.3 Combinação de evidências

Um dos principais diferenciais da teoria DS é a possibilidade de combinar evidências provenientes de diferentes fontes. Em sistemas de aprendizado de máquina, essas fontes podem ser modelos distintos, diferentes representações dos dados ou arquiteturas independentes.

No trabalho de Su et al. (SU et al., 2019), diferentes redes neurais convolucionais são treinadas a partir de representações complementares do áudio, e suas saídas são interpretadas como funções de massa. Essas evidências são então combinadas utilizando a *regra de combinação de Dempster*, que permite integrar informações consistentes e atenuar conflitos entre as fontes.

Formalmente, dadas duas funções de massa M_1 e M_2, a combinação resulta em uma nova função de massa $M_{1\oplus2}$, definida como:

$$M_{1\oplus2}(A) = \alpha \sum_{B\cap C=A} M_1(B)M_2(C), \quad \forall A \subseteq \Theta, A \neq \varnothing$$

$$M_{1\oplus2}(\varnothing) = 0$$

onde A, B e C são subconjuntos do quadro de discernimento Θ, representando hipóteses ou conjuntos de hipóteses do problema.

Intuitivamente, B e C correspondem às hipóteses suportadas pelas fontes de evidência M_1 e M_2, respectivamente, enquanto A representa a hipótese resultante da interseção entre essas evidências.

O fator de normalização α é dado por:

$$\alpha = \frac{1}{\sum_{B\cap C\neq\varnothing} M_1(B)M_2(C)}$$

Esse fator garante que a função de massa resultante permaneça normalizada.

1.4.4 Aplicação no contexto deste trabalho

Seguindo a abordagem apresentada por Su et al. (SU et al., 2019), neste trabalho as saídas das camadas *softmax* de diferentes redes neurais convolucionais são interpretadas como funções de massa. Cada CNN é treinada utilizando uma representação específica do sinal de áudio, como MFCCs ou Mel Spectrograms, de forma a capturar características complementares do som.

A fusão dessas evidências por meio da regra de Dempster permite obter uma decisão final mais robusta, integrando múltiplas fontes de informação e fornecendo, além da classe prevista, uma medida explícita de incerteza. Essa estratégia contribui para aumentar a confiabilidade do sistema de classificação de sons ambientais, especialmente em cenários com sobreposição acústica ou ambiguidade entre classes.

A fusão dessas evidências por meio da regra de Dempster permite obter uma decisão final mais robusta, integrando múltiplas fontes de informação. Além da classe prevista, a

teoria de Dempster-Shafer possibilita a representação explícita da incerteza, por meio da atribuição de massa de crença a subconjuntos de hipóteses.

Por exemplo, considere duas classes A e B. Um modelo pode atribuir:

$$m(\{A\}) = 0.6, \quad m(\{B\}) = 0.3, \quad m(\{A, B\}) = 0.1$$

Nesse caso, a massa atribuída ao conjunto $\{A, B\}$ representa incerteza do modelo, indicando que não há evidência suficiente para distinguir entre as duas classes. Diferentemente da probabilidade tradicional, essa incerteza não é forçada a ser distribuída entre classes individuais, sendo representada explicitamente.

Essa estratégia contribui para aumentar a confiabilidade do sistema de classificação de sons ambientais, especialmente em cenários com sobreposição acústica ou ambiguidade entre classes.

No presente trabalho, as saídas das camadas softmax são interpretadas como funções de massa associadas apenas a hipóteses individuais (classes singleton). Nesse caso, a combinação das evidências é realizada por meio do produto elemento a elemento das distribuições, seguido de normalização.

Essa abordagem corresponde a uma forma simplificada da regra de Dempster, na qual não são consideradas massas atribuídas a subconjuntos compostos de hipóteses. Consequentemente, a incerteza explícita da teoria de Dempster-Shafer não é modelada diretamente, sendo refletida apenas na distribuição das probabilidades.

1.5 Metodologia

Esta seção descreve a metodologia adotada para o desenvolvimento do sistema de identificação e classificação de sons urbanos. O processo envolve a preparação do conjunto de dados, a extração de características acústicas, o treinamento de modelos baseados em redes neurais convolucionais e a combinação das previsões utilizando teoria de evidência.

A implementação inicial do pipeline de classificação foi baseada em um notebook público disponibilizado na plataforma Kaggle (BANSAL, 2020). A partir dessa base foram realizadas diversas modificações e extensões, incluindo a redefinição das classes para o contexto de cidades inteligentes, a utilização de duas representações espectrais distintas (MFCC e Mel-Spectrogram) e a aplicação de fusão de evidências utilizando a teoria de Dempster-Shafer.

De forma geral, o sistema desenvolvido segue quatro etapas principais:

1. preparação do conjunto de dados;
2. extração de características acústicas;

3. treinamento de modelos de classificação baseados em redes neurais convolucionais;
4. fusão das previsões utilizando teoria de evidência.

Antes da descrição detalhada das etapas, a Figura 1 apresenta uma visão geral do pipeline proposto neste trabalho.

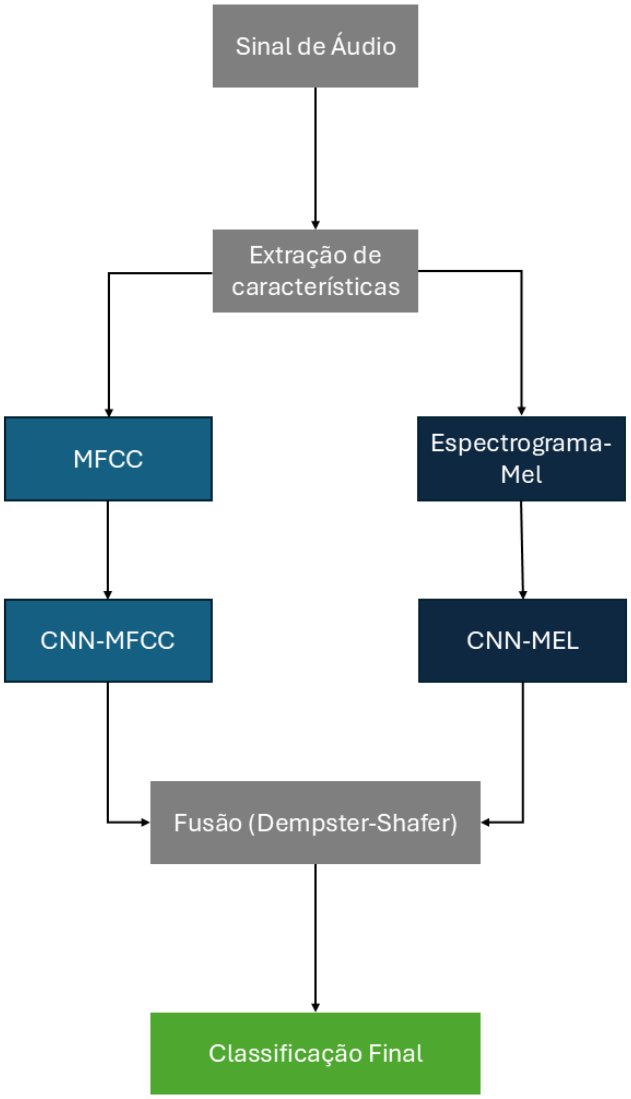

Figura 1 – Pipeline do sistema proposto para classificação de sons urbanos utilizando duas representações espectrais (MFCC e Mel-Spectrogram) e fusão baseada na teoria de Dempster-Shafer.

1.5.1 Base de Dados e Pré-processamento

Para o treinamento e avaliação do modelo foi utilizado o conjunto de dados **UrbanSound8K**, amplamente empregado em pesquisas de classificação de sons ambientais. Esse dataset contém aproximadamente 8732 amostras de áudio anotadas manualmente, distribuídas em 10 classes de eventos sonoros urbanos: latido de cachorro, crianças brincando, buzina de carro, ar-condicionado, música de rua, tiro, sirene, motor em marcha lenta, britadeira, perfuração.

Os arquivos de áudio são disponibilizados no formato WAV e possuem duração máxima aproximada de quatro segundos. O conjunto de dados está organizado em dez subconjuntos (*folds*), que permitem a realização de experimentos de treinamento e validação de forma estruturada.

Cada amostra possui metadados associados contendo informações como o nome do arquivo, a classe do evento sonoro e o subconjunto ao qual pertence.

Com o objetivo de alinhar o estudo com aplicações em cidades inteligentes, as classes originais foram agrupadas em duas macro-categorias funcionais relacionadas a domínios de interesse para monitoramento urbano:

- **Mobilidade urbana**, que inclui eventos sonoros associados ao tráfego e atividades urbanas, como buzinas de veículos, motores em marcha lenta e ruídos de ferramentas utilizadas em obras urbanas;
- **Segurança pública**, que inclui eventos potencialmente associados a situações de risco ou alerta, como sirenes, disparos de arma de fogo e latidos de cães.

Após o mapeamento das classes, as amostras que não pertenciam a essas categorias foram removidas do conjunto de dados utilizado nos experimentos.

1.5.2 Extração de Características Acústicas

Para permitir que os modelos de aprendizado de máquina identifiquem padrões relevantes nos sinais de áudio, foram extraídas duas representações espectrais distintas: *Mel-Frequency Cepstral Coefficients* (MFCC) e *Mel-Spectrogram*. Essas representações são amplamente utilizadas em tarefas de reconhecimento de áudio por capturarem diferentes aspectos do espectro sonoro.

Os MFCC foram introduzidos por Davis e Mermelstein (DAVIS; MERMELSTEIN, 1980) e consistem em uma representação cepstral[1] baseada na escala Mel, que busca

[1] O termo "cepstral"refere-se a uma representação obtida a partir da Transformada do logaritmo do espectro de um sinal. Essa abordagem permite separar componentes relacionados à envoltória espectral (timbre) das características de excitação do sinal, sendo amplamente utilizada em processamento de fala e áudio.

aproximar a percepção humana de frequências sonoras. O processo de extração envolve a transformação do sinal para o domínio da frequência, a aplicação de um banco de filtros na escala Mel, o cálculo do logaritmo da energia em cada banda e a aplicação da Transformada Discreta do Cosseno para obter os coeficientes cepstrais.

Já o Mel-Spectrogram representa a distribuição de energia do sinal ao longo do tempo e das frequências utilizando a escala Mel. Essa representação é obtida a partir da aplicação da Transformada de Fourier de Curto Prazo (STFT), seguida da projeção das frequências em um banco de filtros Mel e da conversão da energia para escala logarítmica em decibéis. O formato bidimensional dessa representação permite que redes neurais convolucionais explorem padrões locais de tempo e frequência de forma semelhante ao processamento de imagens (GOODFELLOW; BENGIO; COURVILLE, 2016).

Neste trabalho foram extraídos 40 coeficientes MFCC e 40 bandas Mel para cada amostra de áudio.

A extração das características acústicas foi implementada utilizando a biblioteca Librosa, conforme ilustrado no Código 1.1.

Listing 1.1 – Extração de MFCC e Mel-Spectrogram

```
1 mfcc = librosa.feature.mfcc(y=raw, sr=sr, n_mfcc=40)
2
3 mel = librosa.feature.melspectrogram(y=raw, sr=sr, n_mels=40)
4 mel = librosa.power_to_db(mel, ref=np.max)
```

1.5.3 Padronização das Entradas

Os espectrogramas gerados a partir dos sinais de áudio podem apresentar dimensões variáveis, especialmente no eixo temporal, devido às diferenças de duração entre as amostras. No entanto, modelos baseados em redes neurais convolucionais requerem entradas com dimensões fixas para o correto funcionamento do processo de treinamento (GOODFELLOW; BENGIO; COURVILLE, 2016).

Dessa forma, todas as representações espectrais foram redimensionadas para um formato padrão de 40×173, onde 40 corresponde ao número de bandas de frequência e 173 à dimensão temporal. Esse processo de padronização permite tratar os espectrogramas como imagens bidimensionais, possibilitando a aplicação de operações convolucionais para a extração de padrões no domínio tempo-frequência (SU et al., 2019).

A escolha dessa dimensão foi baseada na adaptação de um pipeline previamente proposto em um notebook público da plataforma Kaggle (BANSAL, 2020), no qual essa configuração apresentou bom desempenho na tarefa de classificação de sons urbanos. Ressalta-se que essa escolha não é única, podendo variar conforme o conjunto de dados e a

aplicação, sendo definida neste trabalho de forma a equilibrar a preservação de informações relevantes e a eficiência computacional.

Listing 1.2 – Redimensionamento dos espectrogramas utilizando interpolação linear

```
mfcc = cv2.resize(mfcc, (up_width, up_height), interpolation=cv2.
    INTER_LINEAR)
mel  = cv2.resize(mel,  (up_width, up_height), interpolation=cv2.
    INTER_LINEAR)
```

Como mostrado no Código 1.2, o redimensionamento foi realizado utilizando interpolação linear.

1.5.4 Divisão dos Dados em Treinamento e Teste

Após a extração e padronização das características acústicas, os dados foram divididos em conjuntos de treinamento e teste, utilizando uma proporção de 80% para treinamento e 20% para teste. Essa divisão é amplamente adotada em tarefas de aprendizado de máquina por permitir que o modelo seja treinado com uma quantidade significativa de dados, ao mesmo tempo em que reserva um subconjunto independente para avaliação de desempenho e generalização (GOODFELLOW; BENGIO; COURVILLE, 2016).

A divisão foi realizada de forma aleatória, utilizando uma semente fixa (*random state*) para garantir a reprodutibilidade dos experimentos.

Um aspecto importante dessa etapa está relacionado ao uso de duas representações distintas do mesmo sinal de áudio (MFCC e Mel-Spectrogram). Para que a fusão de decisões baseada na teoria de Dempster-Shafer seja válida, é fundamental que as amostras correspondentes em ambas as representações permaneçam alinhadas.

Dessa forma, foi utilizado o mesmo procedimento de divisão e a mesma semente aleatória para ambas as representações, garantindo que cada amostra no conjunto de teste possua correspondência direta entre suas versões em MFCC e Mel-Spectrogram.

Esse alinhamento é essencial para permitir a combinação coerente das previsões geradas pelos dois modelos, assegurando que ambas as evidências estejam associadas ao mesmo evento sonoro.

A implementação dessa divisão é apresentada no Código 1.3.

Listing 1.3 – Divisão dos dados em treino e teste

```
X_train_mfcc, X_test_mfcc, y_train, y_test = train_test_split(
    X_mfcc, y, test_size=0.2, random_state=42
)

X_train_mel, X_test_mel, _, _ = train_test_split(
```

```
    X_mel, y, test_size=0.2, random_state=42
)
```

1.5.5 Construção do Modelo

Para a arquitetura da rede neural convolucional utilizada neste trabalho foi inicialmente baseada em um pipeline disponível no notebook público da plataforma Kaggle (BANSAL, 2020). A partir dessa implementação, foram realizadas adaptações com o objetivo de tornar o modelo mais modular, reutilizável e adequado ao cenário de múltiplas representações de entrada. A escolha desse pipeline foi motivada, entre outros fatores, pela utilização da biblioteca Keras, integrada ao framework TensorFlow. O Keras fornece uma API de alto nível para o desenvolvimento de modelos de aprendizado profundo, permitindo a construção, treinamento e avaliação de redes neurais de forma modular e eficiente (TEAM, 2024).

Uma das principais modificações realizadas foi a generalização do processo de construção do modelo por meio da definição de uma função parametrizável, denominada `build_model`. Essa função permite instanciar diferentes modelos com a mesma arquitetura, variando parâmetros como o formato da entrada e o número de classes.

Essa abordagem foi essencial para a implementação da arquitetura de duas correntes (*two-stream*), permitindo a criação de dois modelos independentes: um treinado com MFCC e outro com Mel-Spectrogram, mantendo consistência estrutural entre ambos.

Listing 1.4 – Função para construção do modelo CNN

```
def build_model(input_shape=(40, 173, 1), n_classes=2,
   learning_rate=0.001):

    model = models.Sequential([
        layers.Conv2D(32, (3,3), activation='relu', input_shape=
           input_shape),
        layers.MaxPooling2D(2),

        layers.Conv2D(128, (3,3), activation='relu'),
        layers.MaxPooling2D(2),
        layers.Dropout(0.3),

        layers.Conv2D(128, (3,3), activation='relu'),
        layers.MaxPooling2D(2),

        layers.GlobalAveragePooling2D(),
        layers.Dense(256, activation='relu'),
```

```
16          layers.Dense(n_classes, activation='softmax')
17      ])
18
19      optimizer = keras.optimizers.Adam(learning_rate=learning_rate
            )
20
21      model.compile(
22          loss='categorical_crossentropy',
23          optimizer=optimizer,
24          metrics=['accuracy']
25      )
26
27      return model
```

Como apresentado no Código 1.4, a função encapsula tanto a definição da arquitetura quanto a etapa de compilação do modelo. Essa estratégia favorece a reutilização do código e reduz a redundância na implementação, além de garantir que ambos os modelos (MFCC e Mel-Spectrogram) compartilhem exatamente a mesma configuração estrutural.

A utilização de uma arquitetura idêntica para ambas as representações permite que as diferenças de desempenho observadas sejam atribuídas às características das representações de entrada, e não a variações no modelo, contribuindo para maior rigor experimental.

Além disso, a função foi parametrizada com `n_classes = 2`, refletindo a proposta deste trabalho de agrupar as classes originais do dataset UrbanSound8K em duas macro-categorias no contexto de cidades inteligentes: mobilidade urbana e segurança pública. Essa redefinição do espaço de classes, conforme apresentado anteriormente, permite direcionar o modelo para tarefas de maior relevância prática no monitoramento urbano, ao invés de uma classificação puramente taxonômica dos sons.

1.5.6 Treinamento do Modelo

Após a definição da arquitetura das redes neurais convolucionais, foi realizada a etapa de treinamento dos modelos utilizando o framework TensorFlow com a interface Keras (TEAM, 2024).

O treinamento foi conduzido de forma independente para cada uma das representações espectrais utilizadas: MFCC e Mel-Spectrogram. Para ambos os modelos, foram utilizados os mesmos hiperparâmetros, permitindo uma comparação justa entre as abordagens. De forma que os principais parâmetros de treinamento adotados foram:

- Tamanho do lote (*batch size*): 8

- Número máximo de épocas: 15
- Função de perda: *categorical cross-entropy*
- Otimizador: Adam, com taxa de aprendizado de 0.001

Com o objetivo de evitar o problema de *overfitting*, foi utilizada a técnica de *Early Stopping*, que interrompe o treinamento caso não haja melhora no erro de validação após um número definido de épocas consecutivas. Neste trabalho, foi adotado um valor de paciência igual a 8 épocas, monitorando a métrica de perda no conjunto de validação (*val_loss*).

Além disso, foi configurado o parâmetro *restore_best_weights=True*, garantindo que o modelo final utilize os pesos correspondentes ao melhor desempenho observado durante o treinamento.

O processo de treinamento para os dois modelos é apresentado no Código 1.5.

Listing 1.5 – Treinamento das redes neurais convolucionais

```
batch_size = 8

callback_mfcc = tf.keras.callbacks.EarlyStopping(
    monitor='val_loss',
    patience=8,
    restore_best_weights=True
)

callback_mel = tf.keras.callbacks.EarlyStopping(
    monitor='val_loss',
    patience=8,
    restore_best_weights=True
)

# Treinamento CNN - MFCC
history_mfcc = model_mfcc.fit(
    X_train_mfcc,
    y_train,
    validation_data=(X_test_mfcc, y_test),
    epochs=15,
    batch_size=batch_size,
    callbacks=[callback_mfcc],
    verbose=1
)

```

```
# Treinamento CNN - Mel-Spectrogram
history_mel = model_mel.fit(
    X_train_mel,
    y_train,
    validation_data=(X_test_mel, y_test),
    epochs=15,
    batch_size=batch_size,
    callbacks=[callback_mel],
    verbose=1
)
```

Durante o treinamento, foram monitoradas as métricas de acurácia e perda tanto no conjunto de treinamento quanto no conjunto de validação. Essa análise permite avaliar a capacidade de generalização dos modelos e identificar possíveis sinais de sobreajuste.

Com as adaptações realizadas na construção do modelo, obtêm-se duas redes treinadas de forma independente, o que constitui um passo fundamental para a etapa posterior de fusão de evidências. Nessa etapa, as previsões dos modelos são combinadas utilizando a teoria de Dempster-Shafer, visando aumentar a robustez da classificação final.

1.5.7 Fusão de Decisão com Teoria de Dempster-Shafer

Após o treinamento dos modelos baseados em MFCC e Mel-Spectrogram, foi realizada a fusão das previsões utilizando a Teoria da Evidência de Dempster-Shafer. Essa teoria permite combinar evidências provenientes de diferentes fontes (SU et al., 2019).

No contexto deste trabalho, cada rede neural produz um vetor de probabilidades para as classes, representando a confiança do modelo em cada possível classificação. Esses vetores foram interpretados como funções de massa associadas a hipóteses individuais (classes singleton), sendo posteriormente combinados por meio de uma versão simplificada da regra de Dempster.

A fusão foi implementada considerando o produto elemento a elemento entre as probabilidades das duas redes, seguido de uma etapa de normalização. Esse processo pode ser descrito da seguinte forma:

1. Multiplicação elemento a elemento entre os vetores de probabilidade;
2. Cálculo de um fator de normalização;
3. Normalização do vetor resultante para garantir que a soma das probabilidades seja igual a 1.

Essa abordagem corresponde a uma forma simplificada da regra de Dempster, na qual apenas hipóteses individuais são consideradas. Dessa forma, não há atribuição explícita de massa a subconjuntos de hipóteses, e a incerteza não é modelada diretamente, sendo refletida apenas na distribuição das probabilidades.

A implementação da função de fusão é apresentada no Código 1.6.

Listing 1.6 – Função de fusão baseada na regra de Dempster-Shafer

```
def dsempster_shafer(pred1, pred2, eps=1e-12):
    combined = []

    for p1, p2 in zip(pred1, pred2):
        # Produto elemento a elemento
        prod = p1 * p2

        # Normalizacao (fator alpha)
        alpha = np.sum(prod) + eps

        # Distribuicao combinada
        m_comb = prod / alpha
        combined.append(m_comb)

    return np.array(combined)
```

Após a aplicação da função de fusão, a classe final foi obtida por meio da seleção do índice de maior probabilidade:

Listing 1.7 – Predição final após fusão das evidências

```
pred_mfcc = model_mfcc.predict(X_test_mfcc)
pred_mel  = model_mel.predict(X_test_mel)

final_pred = dsempster_shafer(pred_mfcc, pred_mel)
y_final = np.argmax(final_pred, axis=1)
```

Essa abordagem reduz a influência de erros individuais de cada modelo, reforçando evidências consistentes entre as redes e penalizando previsões divergentes. Como resultado, observa-se um aumento na robustez e na confiabilidade das classificações finais, especialmente em cenários com ambiguidade nos dados.

Parte II

Parte Final

2 Resultados e Discussão

2.1 Análise do Processo de Treinamento

Para avaliar o comportamento dos modelos durante o treinamento, foram analisadas as métricas de acurácia e função de perda ao longo das épocas, tanto para o modelo baseado em MFCC quanto para o modelo baseado em Mel-Spectrogram.

2.1.1 Modelo baseado em MFCC

A Figura 2 apresenta a evolução da acurácia para o modelo baseado em MFCC. Observa-se um crescimento rápido da acurácia nas primeiras épocas, indicando que a rede é capaz de aprender padrões relevantes a partir dos dados de entrada. A acurácia de validação acompanha a tendência da acurácia de treinamento, o que sugere boa capacidade de generalização.

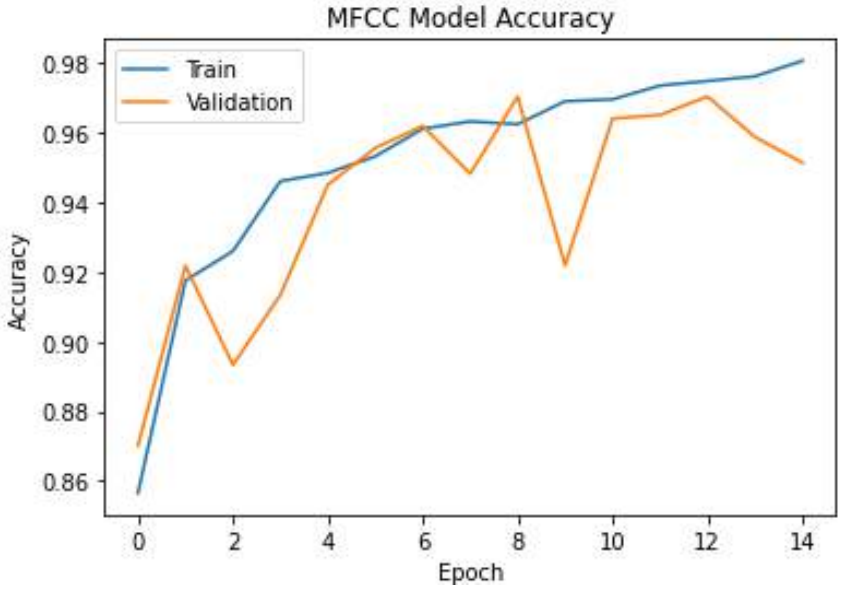

Figura 2 – Evolução da acurácia do modelo baseado em MFCC

A Figura 3 apresenta a evolução da função de perda. Observa-se uma redução consistente ao longo das épocas, com pequenas oscilações na validação, o que é esperado devido à variabilidade dos dados.

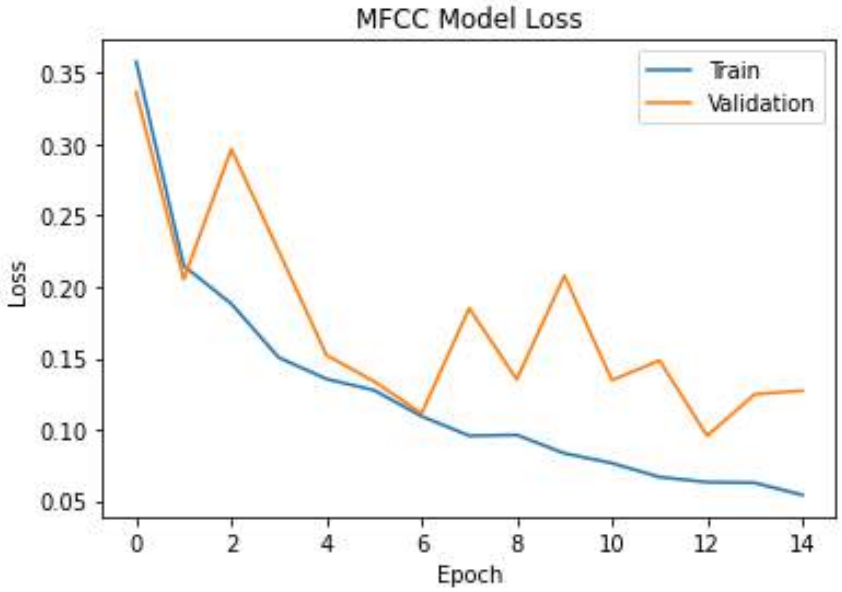

Figura 3 – Evolução da função de perda do modelo baseado em MFCC

2.1.2 Modelo baseado em Mel-Spectrogram

A Figura 4 apresenta a evolução da acurácia do modelo baseado em Mel-Spectrogram. Assim como observado no modelo MFCC, a acurácia de treinamento cresce rapidamente nas primeiras épocas. A acurácia de validação acompanha esse comportamento, atingindo valores próximos a 97%, indicando boa capacidade de generalização.

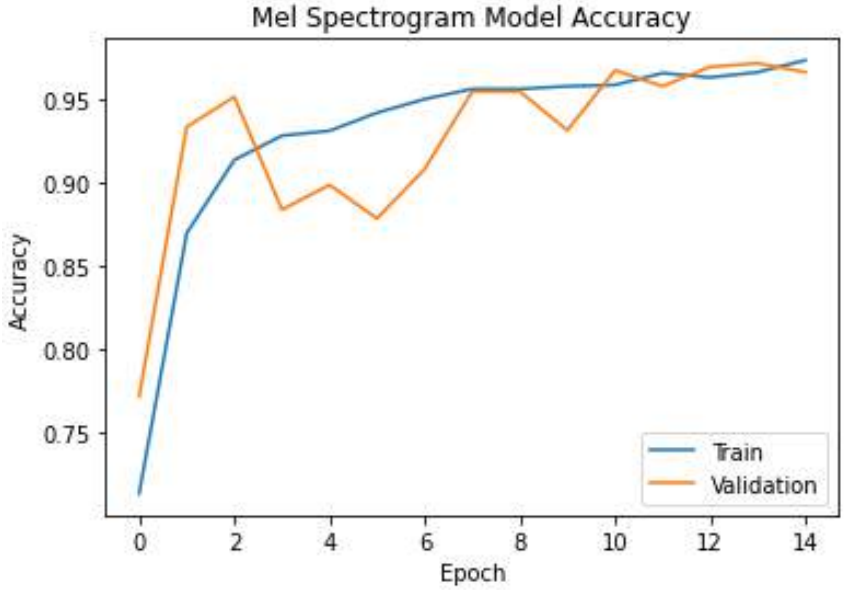

Figura 4 – Evolução da acurácia do modelo baseado em Mel-Spectrogram

A Figura 5 apresenta a evolução da função de perda. Nota-se uma redução consistente da perda ao longo do treinamento, tanto para os dados de treino quanto de validação, sem indícios significativos de sobreajuste (*overfitting*).

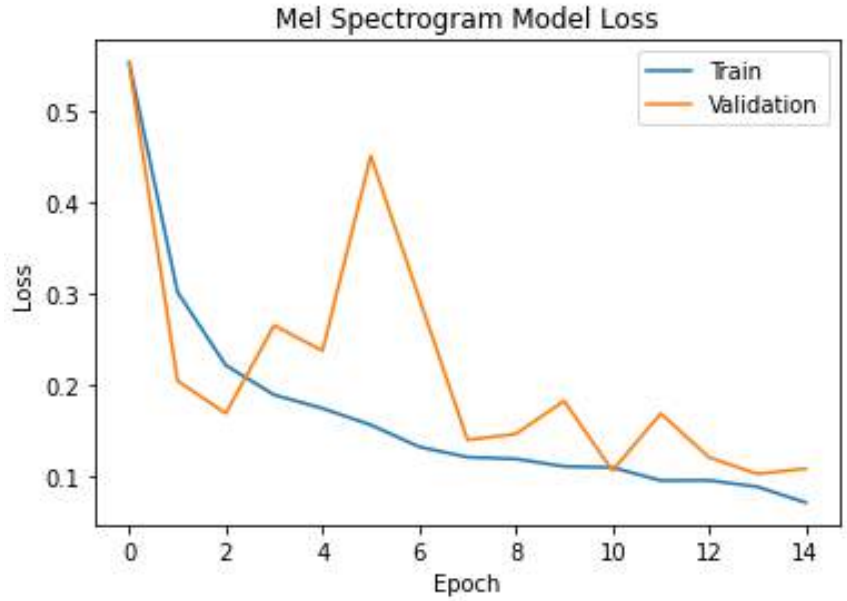

Figura 5 – Evolução da função de perda do modelo baseado em Mel-Spectrogram

2.1.3 Discussão

De forma geral, ambos os modelos apresentaram boa convergência durante o treinamento, com redução consistente da função de perda e aumento da acurácia ao longo das épocas. No entanto, observa-se que o modelo baseado em Mel-Spectrogram apresenta desempenho ligeiramente superior, especialmente em termos de estabilidade da validação, indicando maior capacidade de generalização.

Esse comportamento pode ser explicado pelo fato de que o Mel-Spectrogram preserva de forma mais completa as informações do sinal no domínio tempo-frequência, favorecendo a extração de padrões pelas redes neurais convolucionais.

2.2 Avaliação dos Modelos Individuais

Além da análise das curvas de treinamento, o desempenho dos modelos foi avaliado por meio de métricas clássicas de classificação, incluindo *precision*, *recall*, *f1-score* e acurácia.

2.2.1 Modelo baseado em MFCC

O desempenho do modelo baseado em MFCC é apresentado na Tabela 1.

Tabela 1 – Classification Report do modelo baseado em MFCC

Classe	Precision	Recall	F1-score	Support
0	0.93	0.99	0.96	506
1	0.98	0.91	0.95	441
Acurácia		0.95		

Observa-se que o modelo apresenta alta acurácia geral (95%), com desempenho equilibrado entre as classes. A classe 0 apresenta maior *recall*, indicando maior capacidade de identificação correta dessa categoria, enquanto a classe 1 apresenta maior *precision*, sugerindo menor taxa de falsos positivos.

2.2.2 Modelo baseado em Mel-Spectrogram

O desempenho do modelo baseado em Mel-Spectrogram é apresentado na Tabela 2.

Tabela 2 – Classification Report do modelo baseado em Mel-Spectrogram

Classe	Precision	Recall	F1-score	Support
0	0.97	0.96	0.97	506
1	0.96	0.97	0.96	441
Acurácia		0.97		

Os resultados indicam que o modelo baseado em Mel-Spectrogram apresenta desempenho superior ao modelo MFCC, atingindo uma acurácia de 97%. Além disso, observa-se maior equilíbrio entre *precision* e *recall* em ambas as classes, indicando maior consistência nas previsões.

2.2.3 Análise Comparativa

Comparando os dois modelos, observa-se que:

- O modelo baseado em Mel-Spectrogram apresentou melhor desempenho geral;
- O modelo MFCC apresentou maior variação entre *precision* e *recall*;
- O Mel-Spectrogram mostrou maior estabilidade na classificação entre as classes.

Esses resultados estão alinhados com a literatura, que indica que representações baseadas em espectrogramas preservam mais informações do sinal no domínio tempo-frequência, favorecendo o aprendizado por redes neurais convolucionais (GOODFELLOW; BENGIO; COURVILLE, 2016).

2.3 Fusão de Evidências baseada em Dempster-Shafer

2.3.1 Matriz de Confusão

Para uma análise mais detalhada do desempenho dos modelos, são apresentadas as matrizes de confusão dos modelos individuais baseados em MFCC e Mel-Spectrogram, bem como do modelo após a fusão das evidências.

2.3.1.1 Modelo baseado em MFCC

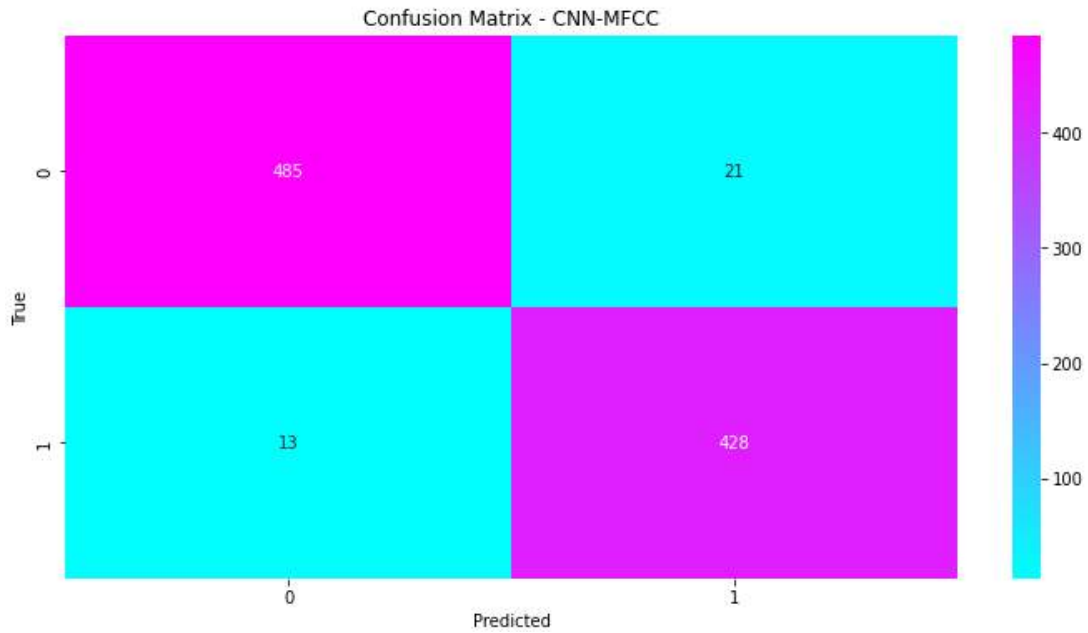

Figura 6 – Matriz de confusão do modelo baseado em MFCC

Observa-se que o modelo MFCC apresenta bom desempenho geral, com 485 classificações corretas para a classe 0 e 428 para a classe 1. No entanto, ainda há uma quantidade significativa de erros, especialmente 21 amostras da classe 0 classificadas incorretamente como classe 1.

2.3.1.2 Modelo baseado em Mel-Spectrogram

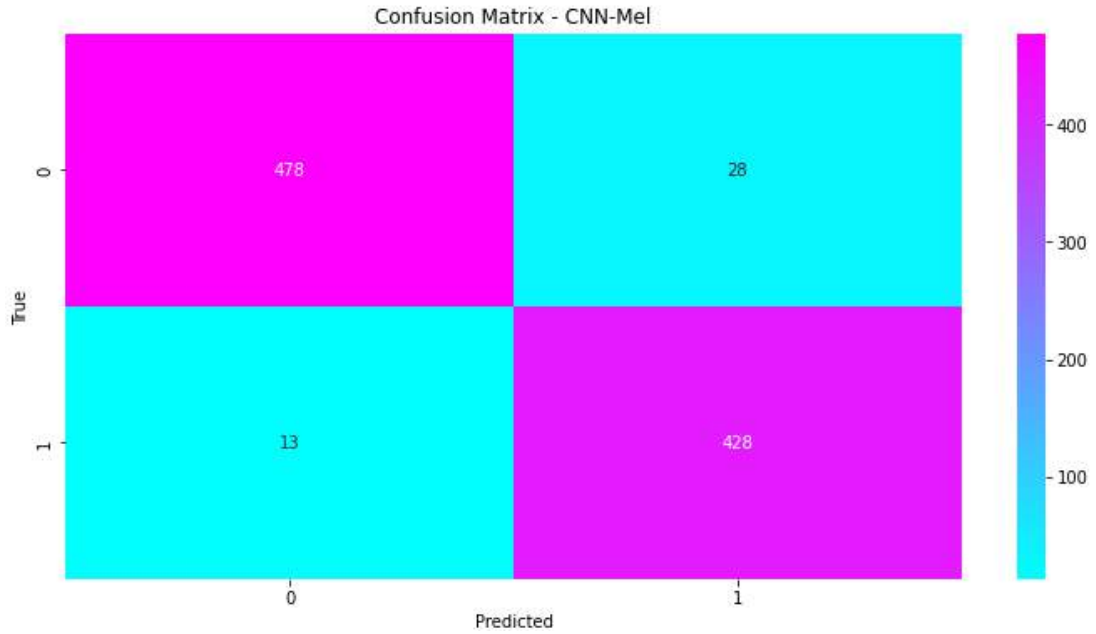

Figura 7 – Matriz de confusão do modelo baseado em Mel-Spectrogram

O modelo baseado em Mel-Spectrogram apresenta desempenho superior ao modelo MFCC, com 478 classificações corretas para a classe 0 e 428 para a classe 1. Entretanto, observa-se um aumento nos erros da classe 0, com 28 amostras classificadas incorretamente, indicando maior confusão nessa categoria.

2.3.1.3 Fusão simplificada

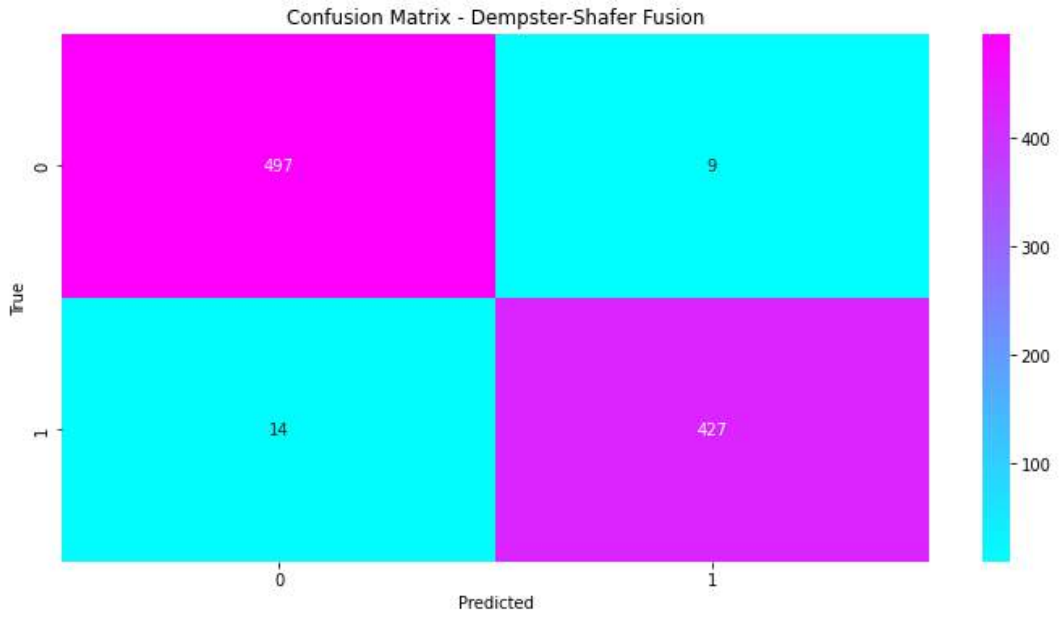

Figura 8 – Matriz de confusão após fusão utilizando Dempster-Shafer

Após a aplicação da fusão de evidências, observa-se uma redução significativa nos erros de classificação. Para a classe 0, foram corretamente classificadas 497 amostras, com apenas 9 erros, representando uma melhoria expressiva em relação aos modelos individuais. Para a classe 1, 427 amostras foram classificadas corretamente, com 14 erros, mantendo desempenho consistente.

Comparando os três modelos, verifica-se que a fusão reduz substancialmente os erros da classe 0, corrigindo casos em que os modelos individuais apresentavam divergência. Esse comportamento evidencia que a combinação das evidências permite mitigar erros específicos de cada modelo, resultando em uma classificação mais robusta.

2.3.2 Discussão

A melhoria observada após a fusão pode ser explicada pela complementaridade entre as representações espectrais utilizadas. Enquanto os MFCC capturam características mais compactas relacionadas à percepção auditiva, o Mel-Spectrogram preserva informações mais detalhadas no domínio tempo-frequência.

Dessa forma, a fusão das previsões permite explorar essas diferentes perspectivas do sinal, reduzindo erros individuais e aumentando a consistência das classificações. Observa-se, em particular, que erros recorrentes nos modelos individuais são corrigidos após a combinação, evidenciando o efeito de consenso entre os modelos.

Os resultados obtidos estão alinhados com a abordagem proposta por Su et al. (SU et al., 2019), que demonstra que a fusão de evidências em arquiteturas multi-representação pode melhorar o desempenho em tarefas de classificação de sons ambientais.

Conclusões e Trabalhos Futuros

Conclusões

Este trabalho teve como objetivo desenvolver um sistema de identificação e classificação de sons urbanos aplicado ao contexto de cidades inteligentes, com foco nas áreas de **mobilidade urbana** e **segurança pública**. Para isso, foi proposta uma abordagem baseada em aprendizado profundo, utilizando redes neurais convolucionais aplicadas a duas representações espectrais distintas: MFCC e Mel-Spectrogram.

Inicialmente, foi realizada a adaptação do conjunto de dados UrbanSound8K, com a redefinição das classes originais em duas macro-categorias alinhadas ao contexto de cidades inteligentes. Essa etapa permitiu direcionar o problema para aplicações práticas de monitoramento urbano.

Os resultados experimentais demonstraram que ambos os modelos foram capazes de aprender padrões relevantes dos dados, apresentando bom desempenho individual. O modelo baseado em MFCC alcançou uma acurácia de 95%, enquanto o modelo baseado em Mel-Spectrogram obteve desempenho superior, com acurácia de 97%. Esses resultados evidenciam a eficácia das representações espectrais na tarefa de classificação de sons urbanos.

Como principal contribuição deste trabalho, destaca-se a aplicação da fusão de evidências utilizando a teoria de Dempster-Shafer. A combinação das previsões dos dois modelos permitiu explorar a complementaridade entre as representações, resultando em uma melhora no desempenho global do sistema, que atingiu uma acurácia de 98%. Além disso, observou-se maior equilíbrio entre as métricas de avaliação, indicando maior robustez na classificação.

Os resultados obtidos reforçam a relevância de abordagens baseadas em múltiplas representações e fusão de decisões para problemas de classificação de áudio. No contexto de cidades inteligentes, o sistema proposto demonstra potencial para aplicação em soluções de monitoramento automático, contribuindo para a análise de mobilidade urbana e apoio a sistemas de segurança pública.

Dessa forma, conclui-se que a integração de técnicas de aprendizado profundo com métodos de fusão de evidências constitui uma abordagem eficaz e promissora para a análise de sons urbanos, podendo ser expandida para diferentes cenários e aplicações no contexto de cidades inteligentes.

Trabalhos Futuros

Como continuidade deste trabalho, diversas melhorias e extensões podem ser exploradas com o objetivo de aproximar o sistema de aplicações reais em cenários de cidades inteligentes.

Uma das principais possibilidades consiste na substituição do conjunto de dados UrbanSound8K por dados coletados em ambiente real, utilizando dispositivos embarcados, como a ESP32. Trabalhos recentes demonstram a viabilidade do uso da ESP32 para aquisição e transmissão de áudio em sistemas de classificação de cenas acústicas, utilizando microfones digitais MEMS e comunicação via WiFi para envio dos dados a um servidor remoto (VIEIRA, 2024).

Nesse contexto, sistemas embarcados podem realizar a captura contínua de áudio e transmitir os dados para processamento externo, permitindo a utilização de modelos de aprendizado profundo mais complexos sem sobrecarregar o hardware local. Conforme apresentado por (VIEIRA, 2024), essa arquitetura combina aquisição local com processamento em nuvem, sendo uma solução eficiente para contornar limitações de memória e processamento em microcontroladores.

A utilização de dados reais capturados em ambiente urbano permite que o sistema seja exposto a condições mais próximas da aplicação final, incluindo ruídos de fundo, interferências e variabilidade temporal, fatores que não estão completamente representados em bases de dados controladas. Dessa forma, espera-se que o modelo desenvolvido possa ser avaliado de maneira mais robusta e adaptado para cenários reais de operação.

Outra direção relevante para trabalhos futuros está relacionada à ampliação do número de macro-classes consideradas. Neste trabalho, a definição das macro-categorias de *mobilidade urbana* e *segurança pública* foi estabelecida como proposta central do projeto. No entanto, devido às limitações do conjunto de dados UrbanSound8K, apenas duas categorias puderam ser exploradas de forma consistente, considerando as classes disponíveis e sua adequação ao contexto proposto.

Nesse sentido, a utilização de bases de dados mais abrangentes ou a coleta de dados próprios pode viabilizar a inclusão de novas macro-classes, permitindo uma representação mais completa dos diferentes domínios de interesse em cidades inteligentes. A expansão do número de categorias tende a aumentar a complexidade do problema, exigindo modelos mais robustos e estratégias de fusão de evidências ainda mais eficazes.

Por fim, a abordagem baseada em múltiplas representações espectrais e fusão por teoria de evidência pode ser explorada em cenários com maior número de classes, onde a presença de incerteza e ambiguidade tende a ser mais significativa. Nesses casos, a teoria de Dempster-Shafer pode contribuir de forma ainda mais relevante para a tomada de decisão.

Referências

ANDRADINA, C. D. N.; SANTOS, B. L. D. A importância do programa nacional de conservação de energia elétrica (procel) para o brasil. 2024. Citado na página 1.

BANSAL, B. *UrbanSound8K CNN Classifier*. 2020. <https://www.kaggle.com/code/bhavybansal/urbansound8k-cnn-classifier>. Kaggle Notebook. Citado 4 vezes nas páginas 7, 12, 15 e 17.

BARROS, R. P. d.; MENDONÇA, R. S. P. d. Os determinantes da desigualdade no brasil. Instituto de Pesquisa Econômica Aplicada (Ipea), 1996. Citado na página 2.

BISHOP, C. M. *Neural Networks for Pattern Recognition*. [S.l.]: Oxford University Press, 1995. Citado na página 5.

CARDOSO, A. O. Mortalidade por acidentes de trânsito no brasil: uma análise de série temporal. Universidade Federal de Uberlândia, 2024. Citado na página 1.

DAVIS, S.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 28, n. 4, p. 357–366, 1980. Citado na página 14.

FLECK, F.; RIEDER, C.; LIMA, A. Redes neurais artificiais: Princípios básicos. *Revista de Informática Teórica e Aplicada*, 2016. Citado na página 5.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. Citado 5 vezes nas páginas 6, 7, 15, 16 e 26.

HAYKIN, S. *Redes Neurais: Princípios e Prática*. [S.l.]: Bookman, 2001. Citado na página 5.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems (NIPS)*. [S.l.: s.n.], 2012. p. 1097–1105. Citado na página 6.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, p. 436–444, 2015. Citado na página 6.

LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, 1998. Citado na página 6.

NASTJUK, I.; TRANG, S.; PAPAGEORGIOU, E. I. Smart cities and smart governance models for future cities: Current research and future directions. *Electronic Markets*, Springer, v. 32, n. 4, p. 1917–1924, 2022. Citado na página 1.

ONU. *Sustainable Development Goal 11: Cidades e comunidades sustentáveis | As Nações Unidas no Brasil*. 2015. Disponível em: <https://brasil.un.org/pt-br/sdgs/11>. Citado na página 2.

ONU. *Sustainable Development Goal 3: Saúde e Bem-Estar | As Nações Unidas no Brasil*. 2015. Disponível em: <https://brasil.un.org/pt-br/sdgs/3>. Citado na página 2.

PEREIRA, F. T. Soluções tecnológicas em segurança pública: contribuições para a administração pública. 2025. Citado na página 1.

RAMALHO, G. *Brasil perde R$ 267 bilhões por ano com congestionamentos*. 2018. Disponível em: <https://g1.globo.com/globonews/noticia/2018/08/07/brasil-perde-r-267-bi-por-ano-com-congestionamentos.ghtml>. Citado na página 1.

SANTOS, L. F. V. dos. A expansão das smart cities e as novas formas de difusão do capital no território brasileiro. *Boletim Campineiro de Geografia*, v. 11, n. 1, p. 59–73, 2021. Citado na página 1.

SU, Y. et al. Environment sound classification using a two-stream cnn based on decision-level fusion. *Sensors*, MDPI, v. 19, n. 7, p. 1733, 2019. Citado 8 vezes nas páginas 1, 2, 7, 9, 11, 15, 20 e 29.

TEAM, K. *Keras Developer Guides*. 2024. Acesso em: 12/2025. Disponível em: <https://keras.io/guides/>. Citado 2 vezes nas páginas 17 e 18.

VIEIRA, J. P. Trabalho de Conclusão de Curso, *Desenvolvimento de um sistema de aquisição de áudio via ESP32 para classificação de cenas acústicas*. 2024. Citado na página 31.

WEISS, M. C.; PEREZ, G. Smart cities: An analysis of information and communication technology capabilities for digital transformation in brazilian cities. *Revista Brasileira de Gestão e Desenvolvimento Regional*, v. 20, n. 1, 2024. Citado na página 1.