PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Instituto de Ciências Exatas e de Informática

RACNN no Domínio Urbano: Uma Análise da Transferência de Aprendizado para Classificação de Eventos Sonoros*

Artur Patitucci Sobroza[1]

Resumo

A classificação automática de eventos sonoros ambientais (*ESC*) é um campo de pesquisa em rápida evolução, impulsionado pela crescente demanda em aplicações como cidades inteligentes e monitoramento ambiental. Modelos de aprendizado profundo, especialmente as Redes Neurais Convolucionais (*CNNs*), têm demonstrado alta performance em *ESC*, mas frequentemente demandam recursos computacionais significativos. A transferência de aprendizado emerge como uma estratégia vital para mitigar a necessidade de vastos *datasets* rotulados e otimizar o uso de recursos. Este estudo investiga a capacidade de generalização de uma Rede Neural Convolucional Adaptativa a Recursos (*RACNN*), previamente pré-treinada no *dataset* ESC-50, para a classificação de eventos sonoros no domínio urbano, utilizando o *dataset* UrbanSound8K. A metodologia emprega *fine-tuning* da camada classificadora do modelo *RACNN*, combinado com uma robusta validação cruzada de 10 *folds* para uma avaliação abrangente. *Mel*-espectrogramas foram utilizados como representação de áudio, e técnicas como *Label Smoothing Cross-Entropy* e *SpecAugment* foram aplicadas. Os resultados demonstram que, embora o modelo *RACNN* consiga aprender a categorizar sons urbanos, alcançando uma acurácia média de 0.3862 ± 0.0406, sua performance varia significativamente entre as 10 classes, com *dog_bark* apresentando o melhor *F1-Score* (0.5985 ± 0.0845) e classes como *engine_idling* (0.2330 ± 0.0984) e *street_music* (0.3012 ± 0.0743) revelando desafios substanciais devido à sua alta heterogeneidade intrínseca e complexidade. A análise detalhada por classe e matriz de confusão revelou padrões de erros específicos, indicando a dificuldade de adaptação para certas categorias no novo domínio. Este trabalho destaca a aplicabilidade do *RACNN* em cenários de transferência de aprendizado, ao mesmo tempo em que aponta para a necessidade de estratégias de *fine-tuning* mais avançadas ou aumentos de dados mais específicos para superar os desafios na

*Artigo apresentado ao Instituto de Ciências Exatas e Informática da Pontifícia Universidade Católica de Minas Gerais como pré-requisito para aprovação na matéria Deep Learning.

[1]Aluno do Programa de Pós-graduação em Informática, Brasil – artur.sobroza@gmail.com.

generalização para *datasets* de áudio complexos como o UrbanSound8K.

Palavras-chave: Classificação de Eventos Sonoros, Redes Neurais Convolucionais, Transferência de Aprendizado, *RACNN*, UrbanSound8K, *Deep Learning*, Processamento de Áudio.

Abstract

Automatic Environmental Sound Classification (ESC) is a rapidly evolving research field, driven by the growing demand in applications such as smart cities and environmental monitoring. Deep learning models—particularly Convolutional Neural Networks (CNNs)—have demonstrated high performance in ESC tasks, though they often require substantial computational resources. Transfer learning emerges as a key strategy to mitigate the need for large labeled datasets and to optimize resource utilization. This study investigates the generalization capacity of a Resource-Adaptive Convolutional Neural Network (RACNN), previously pre-trained on the ESC-50 dataset, for classifying sound events in the urban domain using the UrbanSound8K dataset. The proposed methodology involves fine-tuning the classifier layer of the RACNN model, combined with a robust 10-fold cross-validation scheme for comprehensive evaluation. Mel-spectrograms were employed as the audio representation, while techniques such as Label Smoothing Cross-Entropy and SpecAugment were applied. The results show that although the RACNN is able to learn to categorize urban sounds, achieving a mean accuracy of 0.3862 ± 0.0406, its performance varies significantly across the 10 classes. The class *dog_bark* obtained the best F1-Score (0.5985 ± 0.0845), whereas classes such as *engine_idling* (0.2330 ± 0.0984) and *street_music* (0.3012 ± 0.0743) posed substantial challenges due to their high intrinsic heterogeneity and complexity. Per-class analysis and confusion matrix evaluation revealed specific error patterns, indicating difficulty in adapting to certain categories in the target domain. This work highlights the applicability of RACNN in transfer learning scenarios, while also emphasizing the need for more advanced fine-tuning strategies or domain-specific data augmentation techniques to improve generalization on complex audio datasets such as UrbanSound8K.

Keywords: Sound Event Classification, Convolutional Neural Networks, Transfer Learning, *RACNN*, *UrbanSound8K*, *Deep Learning*, Audio Processing.

1 INTRODUÇÃO

A Classificação de Eventos Sonoros Ambientais (ESC) emergiu como uma área de pesquisa proeminente, impulsionada pelo seu vasto potencial de aplicação em domínios diversos. A capacidade de máquinas identificarem e categorizarem sons em seus ambientes naturais é fundamental para o desenvolvimento de sistemas de monitoramento inteligentes em cidades, onde a detecção de sons como buzinas, sirenes ou latidos de cães pode indicar situações de tráfego, emergência ou segurança (PICZAK, 2015; SALAMON; BELLO, 2017). Além disso, a ESC desempenha um papel crítico no monitoramento ambiental, na vigilância de vida selvagem e na recuperação de informações multimídia, evidenciando a necessidade de soluções eficazes e automatizadas (CRAMER; BENETOS, 2020).

Nos últimos anos, o campo da ESC tem sido revolucionado pelo avanço do Aprendizado Profundo (Deep Learning), em particular pelas Redes Neurais Convolucionais (CNNs). A capacidade intrínseca das CNNs de extrair características hierárquicas e discriminativas diretamente de dados brutos ou suas representações espectrais (como mel-espectrogramas) as tornou ferramentas poderosas para tarefas de classificação de áudio. Modelos baseados em CNNs frequentemente superam abordagens tradicionais, atingindo níveis de acurácia antes inatingíveis em diversos benchmarks (HAN et al., 2017; LU; YANG, 2018).

No entanto, apesar dos avanços significativos, a ESC ainda enfrenta desafios consideráveis. A grande variabilidade dos sons em ambientes reais – causada por diferentes fontes sonoras, ruído de fundo, reverberação e múltiplos eventos ocorrendo simultaneamente – exige modelos robustos e adaptáveis (WANG; LIU, 2020). Um dos obstáculos mais notáveis é a necessidade de vastos e meticulosamente anotados datasets de áudio, cuja coleta e rotulagem são processos dispendiosos e demorados (BELLO et al., 2019). Além disso, a capacidade de um modelo treinado em um conjunto de dados específico generalizar seu aprendizado para novos domínios ou ambientes sonoros, onde as características podem diferir, permanece um desafio fundamental, limitando a aplicabilidade prática de muitas soluções (MESAROS et al., 2017; LI; ZHANG, 2021). Este último ponto, a questão da generalização, constitui a principal motivação para o presente estudo.

Para abordar as limitações impostas pela escassez de grandes volumes de dados rotulados e pela demanda computacional do treinamento de modelos profundos do zero, a Transferência de Aprendizado (Transfer Learning) desponta como uma metodologia eficaz e amplamente adotada em diversas áreas do aprendizado de máquina, incluindo a classificação de áudio. O cerne da Transferência de Aprendizado reside na reutilização de um modelo pré-treinado em uma tarefa ou domínio diferente (domínio fonte) para servir como ponto de partida para uma nova tarefa ou domínio (domínio alvo) (PAN; YANG, 2010). Esta abordagem baseia-se na premissa de que os modelos de aprendizado profundo são capazes de aprender características genéricas e de baixo nível nas camadas iniciais da rede, que podem ser úteis para uma variedade de tarefas relacionadas.

No contexto de redes neurais convolucionais, a Transferência de Aprendizado frequen-

temente envolve a técnica de finetuning. Isso consiste em carregar os pesos de um modelo pré-treinado e, então, adaptá-los para a nova tarefa. Dependendo da similaridade entre os domínios fonte e alvo e da quantidade de dados disponíveis no domínio alvo, diferentes estratégias de finetuning podem ser empregadas. As opções variam desde congelar todas as camadas pré-treinadas e treinar apenas as camadas classificadoras finais (uma "cabeça"do modelo), até descongelar algumas ou todas as camadas pré-treinadas para um treinamento adicional com taxas de aprendizado menores (YOSINSKI et al., 2014; ZEILER; FERGUS, 2014).

Os benefícios da aplicação de Transferência de Aprendizado em ESC são multifacetados. Primeiramente, ela permite mitigar o problema da escassez de dados anotados, uma vez que o modelo já adquiriu conhecimento significativo a partir de um grande dataset fonte. Isso é particularmente valioso em domínios como sons urbanos, onde a coleta e anotação manual são trabalhosas. Em segundo lugar, a reutilização de modelos pré-treinados reduz drasticamente o tempo e os recursos computacionais necessários para o treinamento, pois o processo de aprendizado não começa do zero. Por fim, ao aproveitar as representações de alto nível já aprendidas, a Transferência de Aprendizado pode levar a melhores desempenhos no domínio alvo, superando modelos treinados com menos dados ou a partir de inicialização aleatória (WANG et al., 2020; KONG et al., 2020a). Dada a natureza complexa e variada dos sons ambientais, a Transferência de Aprendizado apresenta-se como uma metodologia promissora para o avanço da ESC.

Neste contexto de busca por soluções eficientes e generalizáveis para a Classificação de Eventos Sonoros, o presente estudo foca na avaliação de uma arquitetura específica: a Rede Neural Convolucional Adaptativa a Recursos (RACNN). Proposta por Fang et al. (2022) (FANG et al., 2022), a RACNN foi desenvolvida com o objetivo de conciliar alta acurácia com eficiência computacional. Sua principal inovação reside no módulo convolucional adaptativo a recursos (RAC), que permite a geração de mapas de características com menor custo computacional em comparação com convoluções convencionais, tornando-a particularmente adequada para dispositivos embarcados e cenários com recursos limitados. Originalmente, o modelo RACNN demonstrou desempenho competitivo no dataset ESC-50, o qual serviu como base para seu pré-treinamento.

O ESC-50 (Environmental Sound Classification) é um dataset de referência para a pesquisa em ESC. Criado por Piczak (2015) (PICZAK, 2015), ele consiste em 2000 clipes de áudio de 5 segundos, divididos igualmente em 50 classes ambientalmente relevantes, como sons de animais, ruídos domésticos, e eventos naturais. A uniformidade na duração e o controle na curadoria dos sons tornam o ESC-50 um dataset ideal para o desenvolvimento e pré-treinamento de modelos de ESC.

Para avaliar a capacidade de generalização do modelo RACNN pré-treinado, escolheu-se o UrbanSound8K, um dataset amplamente utilizado para pesquisa em sons urbanos (SALAMON et al., 2014). O UrbanSound8K, introduzido por Salamon et al. (2014) (SALAMON et al., 2014), compreende 8732 clipes de áudio de sons urbanos de 10 classes distintas: air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, gun_shot, jackhammer, siren, e street_music. Uma característica fundamental do UrbanSound8K é a sua

origem em gravações de campo, o que confere aos áudios uma alta variabilidade e a presença de ruídos de fundo inerentes a ambientes urbanos reais. Essa heterogeneidade é particularmente notável em classes como street_music, que pode abranger desde música tocando em alto-falantes até performances de rua com instrumentos diversos. Nossa análise exploratória detalhada do dataset UrbanSound8K confirmou a complexidade intrínseca desta classe e de outras, apresentando um desafio significativo para a classificação precisa. A natureza deste dataset o torna um domínio alvo ideal para investigar a adaptabilidade e os limites de modelos pré-treinados em contextos diferentes do seu treinamento original.

Diante do potencial da Transferência de Aprendizado e dos desafios inerentes à classificação de eventos sonoros em ambientes complexos, o objetivo central deste trabalho é avaliar a capacidade de generalização de uma Rede Neural Convolucional Adaptativa a Recursos (RACNN), previamente pré-treinada no dataset ESC-50, para a tarefa de classificação de eventos sonoros urbanos no dataset UrbanSound8K. Especificamente, busca-se investigar como um modelo otimizado para eficiência e desempenho em um domínio fonte específico (ESC-50) se adapta e performa em um novo e mais heterogêneo domínio alvo (UrbanSound8K). Para tal, será empregada uma estratégia de finetuning da cabeça do modelo, complementada por uma robusta metodologia de validação cruzada K-Fold, a fim de fornecer uma avaliação abrangente e imparcial do desempenho e das limitações do modelo neste novo contexto.

As principais contribuições deste estudo são múltiplas e buscam preencher lacunas na literatura sobre a generalização de modelos leves de classificação de áudio para domínios urbanos complexos. Primeiramente, apresentamos uma implementação detalhada e validada da arquitetura RACNN, replicando e confirmando o desempenho original do modelo no dataset ESC-50, o que serve como base para a transferência de aprendizado. Em segundo lugar, este trabalho oferece uma avaliação rigorosa da capacidade de generalização do RACNN para o dataset UrbanSound8K, um domínio acústico notoriamente desafiador devido à sua heterogeneidade e variabilidade, utilizando uma metodologia de validação cruzada K-Fold com 10 folds. Esta abordagem garante uma análise estatisticamente robusta do desempenho do modelo em diferentes divisões dos dados. Finalmente, fornecemos uma análise aprofundada do desempenho do modelo por classe, identificando não apenas as classes em que o RACNN obteve sucesso, mas também aquelas que apresentaram maiores desafios, detalhando os padrões de confusão por meio de matrizes de confusão agregadas. Tais insights são cruciais para futuras pesquisas e para o desenvolvimento de modelos mais eficazes em ambientes sonoros reais.

O restante deste artigo está organizado da seguinte forma: A Seção 2 revisa os trabalhos relacionados no campo da classificação de eventos sonoros e transferência de aprendizado. A Seção 3 detalha os Materiais e Métodos empregados, incluindo a descrição dos datasets, o pré-processamento de áudio, a arquitetura do modelo RACNN, e a configuração experimental de finetuning. A Seção 4 apresenta e discute os Resultados obtidos, incluindo o desempenho global do modelo, a análise por fold e por classe, e as matrizes de confusão. Por fim, a Seção 5 conclui o trabalho, resumindo os principais achados, discutindo as limitações e apontando para direções de pesquisa futuras.

2 TRABALHOS RELACIONADOS

A aplicação de Redes Neurais Convolucionais (CNNs) na Classificação de Eventos Sonoros (ESC) tem sido um dos principais motores do avanço da área na última década. Inspiradas no sucesso em tarefas de visão computacional, as CNNs se mostraram particularmente eficazes na extração de características relevantes de representações espectrais de áudio, como os Mel-espectrogramas. Diversas arquiteturas têm sido propostas e refinadas para capturar a complexidade temporal e de frequência dos sons ambientais.

Entre os trabalhos pioneiros e influentes que estabeleceram o uso de CNNs para ESC, destaca-se a abordagem de Piczak (2015) (PICZAK, 2015), que aplicou CNNs a mel-espectrogramas para classificar sons ambientais, incluindo o dataset ESC-50, demonstrando a eficácia dessas redes. Outros estudos iniciais também exploraram o potencial das CNNs para extrair representações discriminativas diretamente dos dados brutos de áudio ou de suas transformações espectrais (LEE et al., 2017; HERSHEY et al., 2017).

Posteriormente, a comunidade de pesquisa explorou arquiteturas mais sofisticadas e profundas, inspiradas no campo da visão computacional. Modelos como variações da VGG-Net (SIMONYAN; ZISSERMAN, 2015), ResNet (HE et al., 2016) e DenseNet (HUANG et al., 2017) foram adaptados com sucesso para tarefas de ESC, demonstrando que o aumento da profundidade da rede, combinado com técnicas para mitigar o problema do gradiente evanescente, pode levar a um desempenho superior na captura de representações de áudio complexas. Além disso, arquiteturas que combinam camadas convolucionais com camadas recorrentes, conhecidas como Redes Recorrentes Convolucionais (CRNNs), também apresentaram resultados promissores em ESC ao modelar dependências temporais de longo alcance em representações espectrais (LIM et al., 2017; ZHANG et al., 2018).

Mais recentemente, a pesquisa tem se voltado para a otimização de modelos em termos de tamanho e eficiência, buscando reduzir o número de parâmetros e a complexidade computacional sem sacrificar significativamente a acurácia. Isso inclui o desenvolvimento de arquiteturas mais leves, muitas vezes inspiradas em modelos como MobileNet (HOWARD et al., 2017) e outras técnicas de convoluções eficientes, visando a implantação em dispositivos com recursos limitados ou cenários de inferência em tempo real (HAN et al., 2016; KIM et al., 2020). Esses avanços sublinham a evolução contínua das CNNs em ESC, pavimentando o caminho para modelos mais práticos e amplamente aplicáveis.

A Transferência de Aprendizado (Transfer Learning) tem se consolidado como uma estratégia poderosa para superar os desafios de escassez de dados rotulados e a alta demanda computacional no treinamento de modelos de aprendizado profundo em diversas aplicações de áudio. A premissa de que características aprendidas em uma tarefa podem ser transferidas para outra tem sido explorada com sucesso tanto em domínios de fala quanto em classificação de sons ambientais.

No campo do processamento de fala, por exemplo, modelos pré-treinados em grandes corpora de fala (como Librispeech) para tarefas de reconhecimento de fala automática (ASR) ou

identificação de locutor demonstraram ser altamente eficazes como extratores de características para tarefas relacionadas, como detecção de emoção ou classificação de sotaque, mesmo com poucos dados rotulados no domínio alvo (LIM et al., 2017; ZHANG et al., 2018). Essa capacidade de transferir "conhecimento"do domínio fonte para o alvo é uma das maiores vantagens do transfer learning.

Especificamente na Classificação de Eventos Sonoros (ESC), a Transferência de Aprendizado tem ganhado destaque. Muitos estudos utilizam modelos pré-treinados em grandes datasets de áudio genéricos ou de música (como AudioSet (GEMMEKE et al., 2017) ou ImageNet para modelos de visão que aceitam espectrogramas como entrada) e os adaptam para datasets menores de sons ambientais. Por exemplo, modelos como VGGish (HERSHEY et al., 2017) (baseado em uma VGG-like treinada no AudioSet) e ResNet-based (KONG et al., 2020b) têm sido amplamente utilizados como extratores de características fixos ou como ponto de partida para finetuning em tarefas de ESC. A eficácia dessas abordagens tem sido comprovada em diversos benchmarks, como o DCASE Challenge (MESAROS et al., 2022), onde a transferência de conhecimento de modelos robustos pré-treinados se tornou uma prática comum para alcançar resultados de ponta.

No entanto, a aplicação da Transferência de Aprendizado em áudio não está isenta de desafios. A diferença na distribuição dos dados entre o domínio fonte e o domínio alvo (conhecido como domain shift) pode limitar a eficácia da transferência (LONG et al., 2016). Por exemplo, um modelo treinado extensivamente em dados de música pode não capturar características ideais para sons de máquinas, e vice-versa. Além disso, a estratégia de finetuning (quantas camadas descongelar, taxa de aprendizado) precisa ser cuidadosamente ajustada para evitar o "catastrophic forgetting"(perda do conhecimento pré-treinado) e garantir que o modelo se adapte eficientemente ao novo domínio sem perder as representações úteis já aprendidas (KIRKPATRICK et al., 2017). Nosso estudo, ao transferir um modelo de ESC pré-treinado no ESC-50 para o UrbanSound8K, explora diretamente esses aspectos de adaptabilidade e desafios da generalização entre datasets de sons ambientais com características distintas.

O dataset UrbanSound8K (SALAMON et al., 2014) tornou-se um dos benchmarks mais importantes para a Classificação de Eventos Sonoros (ESC) em ambientes urbanos, devido à sua composição de gravações de campo que refletem a complexidade e a variabilidade dos sons reais. Consequentemente, uma vasta gama de estudos tem explorado este dataset utilizando diversas metodologias, principalmente baseadas em aprendizado profundo.

Pesquisas iniciais no UrbanSound8K frequentemente empregavam abordagens tradicionais de processamento de sinal combinadas com classificadores de aprendizado de máquina, como Máquinas de Vetores de Suporte (SVMs) ou Florestas Aleatórias, utilizando características extraídas manualmente como MFCCs (Mel-Frequency Cepstral Coefficients) e GMMs (Gaussian Mixture Models) (SPRENGEL et al., 2017; FONVILLE; STURM, 2016). Embora essas abordagens tenham fornecido resultados de linha de base importantes, elas demonstraram as limitações em lidar com a alta dimensionalidade e a complexidade das representações sonoras.

Com o advento do Deep Learning, a performance em UrbanSound8K melhorou significativamente. Diversas arquiteturas de CNN, algumas inspiradas em modelos de visão, foram aplicadas. Por exemplo, trabalhos utilizaram CNNs profundas em mel-espectrogramas ou outras representações tempo-frequência para alcançar acurácias superiores, explorando a capacidade das redes de aprender características diretamente dos dados (LU; YANG, 2017; ZHANG; CAO, 2019). Além disso, abordagens que incorporam atenção ou módulos de rede mais complexos para capturar relações contextuais e temporais têm sido investigadas no UrbanSound8K (KO; LEE, 2019; HAN et al., 2019), visando aprimorar a capacidade discriminativa do modelo em classes intrinsecamente heterogêneas como "street_music"ou "children_playing".

Apesar dos avanços, a performance no UrbanSound8K ainda apresenta desafios. A variabilidade entre as instâncias de uma mesma classe, a presença de múltiplos eventos sonoros e o ruído de fundo são fatores que dificultam a obtenção de acurácias muito elevadas. Estudos recentes continuam a propor modelos mais otimizados, utilizando técnicas como data augmentation avançada, diferentes estratégias de otimização e a exploração de modelos pré-treinados de áudio em larga escala, como VGGish ou PANNs (HERSHEY et al., 2017; KONG et al., 2020a), para melhorar a robustez e a generalização dos classificadores neste dataset (LEE; KIM, 2020; ZHANG; XU, 2021). A avaliação rigorosa com validação cruzada (como a K-Fold proposta por Salamon et al. no artigo original do dataset) é uma prática comum para garantir a generalização dos resultados (SALAMON et al., 2014). Nosso estudo se alinha a essa linha de pesquisa ao reavaliar um modelo otimizado (RACNN) pré-treinado em um dataset complementar (ESC-50) e adaptá-lo ao UrbanSound8K, fornecendo uma análise detalhada de sua generalização.

No contexto da busca por modelos de Classificação de Eventos Sonoros (ESC) que não apenas alcancem alta acurácia, mas também sejam eficientes em termos de recursos computacionais, a Rede Neural Convolucional Adaptativa a Recursos (RACNN) (FANG et al., 2022) emerge como uma arquitetura promissora. Proposta por Fang et al. (2022), a RACNN foi especificamente projetada para o reconhecimento de sons ambientais, com um foco particular em aplicações que exigem leveza e rapidez, como dispositivos embarcados e sistemas em tempo real.

A principal inovação da RACNN reside no seu Módulo Convolucional Adaptativo a Recursos (RAC). Diferente das convoluções tradicionais que geram um grande número de mapas de características com custo computacional fixo, o módulo RAC utiliza uma abordagem mais eficiente. Ele consegue gerar o mesmo número de mapas de características que as operações convolucionais convencionais, mas de maneira mais "barata"computacionalmente, através de uma combinação inteligente de operações que minimizam a redundância e otimizam o fluxo de dados (FANG et al., 2022). Isso permite à RACNN extrair eficientemente características tanto no domínio do tempo quanto da frequência dos áudios de entrada, sem o aumento exponencial de parâmetros e Operações de Ponto Flutuante (FLOPs) tipicamente associado a redes mais profundas ou largas.

A arquitetura RACNN é construída a partir de blocos RAC (RAC blocks), que se em-

pilham para formar a rede completa. Esses blocos são projetados para serem leves e flexíveis, permitindo que a RACNN seja mais enxuta e mais rápida na inferência em comparação com muitas CNNs profundas convencionais (FANG et al., 2022). Essa característica a posiciona favoravelmente em relação a outros modelos leves de áudio, como MobileNets adaptadas para áudio (HOWARD et al., 2017; KIM et al., 2020), ao oferecer uma solução otimizada especificamente para ESC que balança desempenho e eficiência de forma eficaz.

Originalmente, a RACNN demonstrou um desempenho competitivo no dataset ESC-50, validando sua capacidade de classificar sons ambientais com alta acurácia, mesmo com sua arquitetura otimizada para recursos. A escolha da RACNN para este estudo reflete o interesse em investigar a transferibilidade dessas vantagens de eficiência e precisão para um novo domínio, o UrbanSound8K, que apresenta características de dados distintas e mais heterogêneas. Ao focar na RACNN, este trabalho visa contribuir para a compreensão de como modelos de "áudio consciente de recursos"podem ser efetivamente generalizados para cenários do mundo real.

3 MATERIAIS E MÉTODOS

3.1 Dataset ESC-50 (Fonte e Pré-treinamento)

O dataset ESC-50 (Environmental Sound Classification) (PICZAK, 2015) serve como o ponto de partida fundamental para o presente estudo, funcionando como o domínio fonte para o pré-treinamento do modelo RACNN. Criado por Piczak em 2015, o ESC-50 é um conjunto de dados amplamente reconhecido e utilizado na pesquisa de classificação de eventos sonoros. Ele é composto por 2000 clipes de áudio, cada um com uma duração fixa de 5 segundos, totalizando aproximadamente 2.78 horas de áudio. Os arquivos são gravados a uma taxa de amostragem de 44.1 kHz e estão organizados em 50 classes distintas, abrangendo uma vasta gama de sons ambientais. Estas classes são agrupadas em cinco categorias principais: Animais, Sons Naturais, Sons de Veículos, Sons Domésticos e Sons Humanos não-vocais, oferecendo uma diversidade representativa de cenários acústicos. A curadoria e a estrutura balanceada do ESC-50 o tornam um dataset ideal para o desenvolvimento e avaliação de modelos de ESC.

Para validar a reprodutibilidade da arquitetura RACNN e estabelecer uma base de comparação sólida para o processo de transferência de aprendizado, nossa implementação do modelo RACNN foi avaliada no próprio dataset ESC-50, utilizando a metodologia de validação cruzada K-Fold com 5 folds. A implementação da arquitetura RACNN seguiu o descrito no artigo (FANG et al., 2022) da forma mais fiel possível, replicando a estrutura dos blocos RAC, as camadas convolucionais, de pooling e as camadas densas finais, buscando manter a integridade do design original. Os resultados detalhados da nossa implementação no ESC-50, conforme registrado no log de treinamento (ESC-50_R11.ipynb), indicaram uma acurácia média de validação de 91,25% para o melhor fold. Em contraste, o artigo original da RACNN por Fang et al. (2022) (FANG et al., 2022) reportou uma acurácia média de 85,65% (e a acurácia mais alta de

86%) para o modelo RACNN no ESC-50, conforme apresentado na Tabela 6 do referido artigo.

A diferença notável de aproximadamente 5,6 pontos percentuais a favor da nossa implementação requer uma análise detalhada e transparente. Embora a arquitetura base da RACNN tenha sido replicada fielmente, a divergência de desempenho pode ser atribuída a múltiplos fatores e à introdução de técnicas estado da arte no processo de treinamento, que não foram explicitamente detalhadas ou não se sabe se foram empregadas com a mesma configuração pelos autores originais (FANG et al., 2022). Os principais elementos que explicam essa diferença são:

1. **Parâmetros de Pré-processamento e Normalização:** Nossa implementação utilizou parâmetros específicos para a geração dos Mel-espectrogramas (n_fft = 2048, hop_length = 512, n_mels = 128, f_min = 20, f_max = 22050), seguidos por conversão para decibéis (AmplitudeToDB) e, crucialmente, normalização global dos dados (média e desvio padrão calculados no conjunto de treino). O artigo original menciona o uso de "Log-Mel spectrogram", mas omite os detalhes exatos desses parâmetros e da estratégia de normalização. A normalização robusta é uma prática padrão que pode estabilizar e otimizar o treinamento de redes neurais.

2. **Agendador de Taxa de Aprendizado (Learning Rate Scheduler):** A nossa implementação empregou o OneCycleLR (SMITH, 2018) como agendador de taxa de aprendizado. Este é um método consagrado que varia a taxa de aprendizado e o momentum de forma cíclica ao longo das épocas de treinamento, permitindo que o modelo explore o espaço de pesos de forma mais eficaz e se assente em mínimos mais promissores. O artigo original apenas menciona o uso do otimizador Adam, sem detalhar se algum scheduler foi utilizado, e o OneCycleLR é conhecido por frequentemente superar estratégias mais simples de ajuste de taxa de aprendizado.

3. **Função de Perda:** Adicionalmente, nossa implementação utilizou a função de perda LabelSmoothingCrossEntropy. Esta técnica de regularização suaviza os rótulos alvo, desencorajando o modelo de se tornar excessivamente confiante em suas previsões e, por consequência, melhorando sua capacidade de generalização para dados não vistos. Esta é uma técnica que pode levar a ganhos de acurácia em validação, e não foi explicitamente mencionada no artigo original como parte de sua configuração de treinamento.

4. **Uso de Data Augmentation (SpecAugment):** O artigo original (SALAMON; BELLO, 2017) sugere que "métodos de data augmentation, como adição de ruído, pitch shifting e time stretching, podem ser usados para melhorar ainda mais a acurácia de classificação". No entanto, não é explicitado se essas técnicas foram aplicadas para obter os resultados de 85.65% reportados na Tabela 6 para o modelo RACNN. Nossa implementação, para o pré-treinamento no ESC-50, incluiu a aplicação de SpecAugment diretamente nos Mel-espectrogramas. Esta técnica, que mascara regiões de tempo e frequência do espectrograma, atua como uma poderosa forma de regularização, forçando o modelo a aprender

características mais robustas e generalizáveis, contribuindo significativamente para o desempenho superior observado.

Em suma, a superioridade da acurácia alcançada pela nossa implementação do RACNN no ESC-50 não sugere um erro, mas sim a aplicação de um conjunto de práticas e tecnologias de treinamento avançadas, e que não foram explicitamente detalhadas como parte da metodologia original dos autores para seus resultados baseline. Essa performance aprimorada no dataset fonte estabelece uma base robusta e otimizada para o estudo de Transfer Learning subsequente para o dataset UrbanSound8K.

3.2 Dataset UrbanSound8K (Domínio Alvo)

O dataset UrbanSound8K (SALAMON et al., 2014) constitui o domínio alvo para o experimento de Transferência de Aprendizado e, consequentemente, o principal foco de avaliação da generalização do modelo RACNN neste estudo. Lançado por Salamon et al. em 2014, o UrbanSound8K é um dos maiores e mais desafiadores conjuntos de dados para classificação de eventos sonoros urbanos. Ele é composto por 8732 clipes de áudio, com durações variando de menos de 1 segundo a 4 segundos, e uma taxa de amostragem de 44.1 kHz. Os clipes são extraídos de gravações de campo feitas na cidade de Nova Iorque, refletindo a complexidade, a heterogeneidade e a presença de ruídos de fundo característicos de ambientes urbanos reais.

Os dados do UrbanSound8K são categorizados em 10 classes principais de sons urbanos: air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, gun_shot, jackhammer, siren e street_music. A tabela abaixo detalha a contagem de amostras por classe, evidenciando o desbalanceamento existente no dataset:

Tabela 1 – Dataset UrbanSound8K: contagem de amostras por classe

Classe	Contagem de Amostras
air_conditioner	1000
car_horn	429
children_playing	1000
dog_bark	1000
drilling	1000
engine_idling	1000
gun_shot	374
jackhammer	1000
siren	929
street_music	1000
Total	8732

Como demonstrado, as classes car_horn e gun_shot possuem significativamente menos exemplos do que as demais, o que pode influenciar o desempenho do modelo e a necessidade de técnicas de balanceamento ou data augmentation específicas. O dataset é pré-dividido em 10 folds (fold1 a fold10), permitindo uma validação cruzada padronizada para comparações entre pesquisas. Esta estrutura de folds foi seguida rigorosamente em nossa metodologia para garantir a comparabilidade dos resultados.

A transição do modelo RACNN pré-treinado no ESC-50 para o UrbanSound8K envolveu uma estratégia de Transferência de Aprendizado via *finetuning*. O modelo RACNN treinado no ESC-50 (conforme detalhado na Seção 3.1) foi carregado e sua camada classificadora final (projetada para 50 classes) foi substituída por uma nova camada densa configurada para as 10 classes do UrbanSound8K. Durante todo o processo de *finetuning* no UrbanSound8K, todas as camadas convolucionais e os blocos RAC pré-treinados foram mantidos congelados. Apenas a recém-adicionada camada classificadora final foi treinada para adaptar as características extraídas do domínio fonte (ESC-50) às classes específicas do UrbanSound8K.

O pré-processamento dos clipes de áudio do UrbanSound8K seguiu as mesmas configurações de Mel-espectrogramas utilizadas para o ESC-50 (n_fft = 2048, hop_length = 512, n_mels = 128, etc.) para garantir a consistência das características de entrada esperadas pelo modelo. Para a normalização, no entanto, as estatísticas (média e desvio padrão) foram recalculadas especificamente a partir do conjunto de treino do UrbanSound8K e aplicadas aos dados. Essa abordagem otimiza a representação para a distribuição intrínseca do domínio alvo, que pode diferir significativamente do ESC-50, potencializando a adaptação do modelo.

Assim como no pré-treinamento, técnicas de treinamento estado da arte foram emprega-

das durante o *finetuning* no UrbanSound8K. A função de perda LabelSmoothingCrossEntropy e o otimizador AdamW com o agendador de taxa de aprendizado OneCycleLR foram mantidos. Adicionalmente, a técnica de SpecAugment foi aplicada aos Mel-espectrogramas dos dados de treino do UrbanSound8K. A utilização consistente dessas técnicas visa maximizar o desempenho e a robustez do modelo na tarefa de classificação em um dataset mais complexo.

A avaliação do modelo no UrbanSound8K foi conduzida utilizando a validação cruzada K-Fold com 10 folds, seguindo as divisões pré-definidas no próprio dataset (SALAMON et al., 2014). Os resultados do *finetuning* do RACNN no UrbanSound8K, conforme os logs do experimento (UrbanSound8K_R6.ipynb), indicaram uma acurácia média de validação de 38.62% (± 4.06%) ao longo dos 10 folds. Esta acurácia, embora aparentemente modesta em comparação com o ESC-50, reflete a maior complexidade e o desafio inerente ao dataset UrbanSound8K, que possui variabilidade sonora significativa dentro das classes e sobreposição de eventos.

3.3 Pré-processamento de Áudio

O pré-processamento de áudio é uma etapa crítica para a eficácia de modelos de Deep Learning em tarefas de classificação de eventos sonoros, transformando os sinais brutos de áudio em representações que as redes neurais podem efetivamente aprender. Para este estudo, um pipeline consistente de pré-processamento foi aplicado tanto ao dataset de pré-treinamento (ESC-50) quanto ao dataset alvo (*finetuning*, UrbanSound8K), garantindo que o modelo RACNN recebesse *inputs* com características espectrais padronizadas.

3.3.1 Transformações para Mel-Espectrogramas

Todos os clipes de áudio foram convertidos em Mel-espectrogramas Logarítmicos, uma representação tempo-frequência que se alinha mais com a percepção auditiva humana e tem se mostrado altamente eficaz para tarefas de classificação sonora. As seguintes configurações de parâmetros foram utilizadas para a geração dos Mel-espectrogramas:

- **Frequência de Amostragem (*sample_rate*)**: 44.1 kHz (44100 Hz). Este valor corresponde à frequência de amostragem original dos arquivos de áudio em ambos os datasets (ESC-50 e UrbanSound8K), garantindo que não haja perda de informação ou artefatos introduzidos por *resampling*.
- **Tamanho da Janela FFT (*n_fft*)**: 2048 amostras. Define o tamanho da janela utilizada na Transformada Rápida de Fourier (FFT), influenciando a resolução de frequência do espectrograma.
- **Comprimento do Salto (*hop_length*)**: 512 amostras. Determina o número de amostras entre quadros sucessivos da FFT, impactando a resolução temporal e a sobreposição das

janelas.

- **Número de Bandas de Mel (*n_mels*)**: 128. Especifica o número de filtros de Mel utilizados para mapear as frequências lineares para a escala Mel, que é mais próxima da percepção humana. Um número maior de bandas geralmente captura mais detalhes espectrais.
- **Frequência Mínima (*f_min*)**: 20 Hz. Define o limite inferior de frequência para os filtros Mel, excluindo ruídos de baixa frequência irrelevantes.
- **Frequência Máxima (*f_max*)**: 22050 Hz (ou *sample_rate* / 2). Corresponde ao limite superior de frequência (*Nyquist frequency*), abrangendo todo o espectro audível relevante.
- **Intervalo Dinâmico (*top_db*)**: 80 dB. Após a geração do Mel-espectrograma, a intensidade é convertida para decibéis (Log-Mel) e um limiar é aplicado para cortar o espectro em um intervalo dinâmico de 80 dB, removendo ruídos de fundo de baixa intensidade e focando nos componentes mais significativos.

Este conjunto de parâmetros foi escolhido para equilibrar a riqueza da representação espectral com a complexidade computacional e tem sido amplamente adotado em pesquisas na área.

3.3.2 Normalização

A normalização dos Mel-espectrogramas é crucial para padronizar as entradas da rede neural, acelerar a convergência do treinamento e melhorar a estabilidade do modelo. Duas estratégias de normalização foram empregadas:

- **Pré-treinamento (Dataset ESC-50)**: Para o pré-treinamento no ESC-50, os Mel-espectrogramas foram normalizados usando a média e o desvio padrão calculados globalmente sobre todo o conjunto de treino do próprio ESC-50. Esta abordagem assegura que as características de entrada para o modelo RACNN durante a fase inicial de aprendizado estejam dentro de uma escala consistente, otimizando o treinamento das camadas pré-treinadas.
- **Finetuning (Dataset UrbanSound8K)**: Durante a fase de *finetuning* com o dataset UrbanSound8K, os Mel-espectrogramas foram normalizados utilizando a média e o desvio padrão calculados especificamente a partir do conjunto de treino do UrbanSound8K. Embora o modelo já tenha sido pré-treinado com estatísticas do ESC-50, recalcular e aplicar a normalização baseada no domínio alvo permite uma adaptação mais precisa do modelo à distribuição de dados do UrbanSound8K, que possui características ambientais e espectrais distintas. Esta estratégia visa otimizar a adaptação fina do modelo ao novo domínio.

3.3.3 Aumento de Dados (Data Augmentation) - SpecAugment

Para aumentar a robustez do modelo e sua capacidade de generalização, a técnica de SpecAugment foi aplicada dinamicamente aos Mel-espectrogramas durante a fase de treinamento em ambos os datasets (ESC-50 e UrbanSound8K). O SpecAugment opera diretamente na representação tempo-frequência, introduzindo variabilidade através de duas operações principais:

- **Mascaramento de Frequência (FrequencyMasking)**: Bloqueia uma faixa contínua de canais de frequência, forçando o modelo a aprender a reconhecer sons a partir de informações espectrais incompletas e distribuídas.
- **Mascaramento de Tempo (TimeMasking)**: Bloqueia uma faixa contínua de quadros de tempo, simulando interrupções no sinal ou variações na duração dos eventos, incentivando o modelo a focar em características temporais mais robustas.

A aplicação do SpecAugment é uma prática estado da arte que atua como uma poderosa técnica de regularização, reduzindo o overfitting e melhorando significativamente o desempenho do modelo em dados não vistos.

3.4 Arquitetura do Modelo RACNN

A *Resource Adaptive Convolutional Neural Network* (RACNN) (FANG et al., 2022) é a arquitetura de rede neural convolucional central empregada neste estudo. Proposta por Fang et al. (2022), a RACNN foi desenvolvida com o objetivo de otimizar a classificação de eventos sonoros (ESC) em ambientes com restrição de recursos, como dispositivos embarcados, ao mesmo tempo em que mantém alta performance. Sua inovação reside na introdução de um módulo e um bloco convolucional adaptativos que permitem uma extração eficiente de características tempo-frequência. A arquitetura geral da RACNN é composta por uma sequência de blocos de processamento que atuam sobre os Mel-espectrogramas Logarítmicos de entrada. Os elementos-chave que definem a RACNN são:

3.4.1 Módulo Convolucional Adaptativo a Recursos (RAC Module)

O coração da RACNN é o RAC Module, uma unidade computacional projetada para gerar o mesmo número de mapas de características que as operações convolucionais convencionais, mas com um custo computacional significativamente reduzido. O RAC Module consegue isso através de uma estratégia de paralelização e fusão de características, combinando:

- **Convoluções Ponto a Ponto (*Pointwise Convolutions*)**: Operações 1x1 que reduzem a dimensionalidade do canal e misturam informações entre canais de forma eficiente.

- **Convoluções Agrupadas (*Grouped Convolutions*)**: Dividem os canais de entrada em grupos e aplicam convoluções separadamente a cada grupo, reduzindo o número de operações, mas mantendo a capacidade de extração de características espaciais.
- **Mecanismo de Atenção (*Attention Mechanism*)**: Embora não detalhado como um mecanismo de atenção explícito nos diagramas, o conceito de "adaptação de recursos"implica que o módulo pode focar ou dar pesos diferenciados a certas características, otimizando a representação. No contexto do artigo, isso é alcançado pela forma como as convoluções agrupadas e pontuais são combinadas.

A combinação dessas técnicas permite que o RAC Module extraia características ricas com um número reduzido de parâmetros e *Floating Point Operations* (FLOPs), tornando-o "adaptativo a recursos".

3.4.2 Bloco RAC (RAC Block)

Os RAC Blocks são os componentes fundamentais que constroem a RACNN. Cada RAC Block consiste em uma sequência de operações que incluem:

- **RAC Modules**: Aplicam a lógica de convolução eficiente.
- **Normalização em Lotes (*Batch Normalization*)**: Estabiliza as ativações da rede, acelerando o treinamento e melhorando a generalização.
- **Funções de Ativação (*Activation Functions*)**: ReLU, introduzindo não-linearidade e permitindo que a rede aprenda padrões complexos.
- **Operações de Pooling (*Média ou Máxima*)**: Reduzem a dimensionalidade espacial dos mapas de características, consolidando informações e tornando o modelo mais robusto a pequenas variações na entrada.

A organização sequencial desses blocos permite que a rede aprenda hierarquicamente características cada vez mais abstratas dos dados de áudio.

3.4.3 Estrutura da Rede e Eficiência

A RACNN é construída pelo empilhamento de múltiplos RAC Blocks, seguidos por camadas densas (*Fully Connected*) que realizam a classificação final. A arquitetura é projetada para ser leve, utilizando os módulos RAC para reduzir a contagem de parâmetros e FLOPs em comparação com CNNs convencionais de desempenho similar. Essa característica é particularmente vantajosa para aplicações em tempo real e em dispositivos com capacidade computacional limitada. No contexto deste trabalho, a arquitetura RACNN foi inicialmente pré-treinada

no dataset ESC-50 (50 classes) e, subsequentemente, adaptada via *finetuning* para a tarefa de classificação no UrbanSound8K (10 classes). A camada classificadora final do modelo foi substituída para se adequar ao número de classes do domínio alvo, enquanto as camadas convolucionais pré-treinadas e os blocos RAC, que aprenderam características genéricas de som, foram mantidos congelados para preservar o conhecimento adquirido, e a adaptação às especificidades dos sons urbanos ocorreu através do treinamento da nova camada classificadora, conforme detalhado na Seção 3.2. A fidelidade à arquitetura descrita por Fang et al. (FANG et al., 2022) foi mantida em nossa implementação, garantindo a validade do estudo de *Transfer Learning*.

3.5 Transferência de Aprendizado e Finetuning

A fase de *finetuning* é crucial no processo de Transferência de Aprendizado, permitindo que o modelo pré-treinado no domínio fonte (ESC-50) adapte seu conhecimento ao domínio alvo (UrbanSound8K). A configuração experimental para esta etapa foi cuidadosamente definida para maximizar a adaptação do modelo e garantir uma avaliação robusta de sua generalização.

3.5.1 Validação Cruzada K-Fold

Para avaliar de forma robusta a capacidade de generalização do modelo RACNN no dataset UrbanSound8K, foi empregada a metodologia de Validação Cruzada K-Fold com K=10. Esta escolha é alinhada com a divisão pré-definida de folds no próprio dataset UrbanSound8K (SALAMON et al., 2014), garantindo a comparabilidade dos resultados com outros estudos na literatura. Em cada um dos 10 folds:

- O modelo RACNN, pré-treinado no ESC-50 e com a camada classificadora final substituída (conforme Seção 3.2), foi carregado.
- Nove dos dez *folds* foram utilizados para o treinamento do modelo, enquanto o *fold* restante foi reservado exclusivamente para validação.
- Este processo foi repetido 10 vezes, garantindo que cada *fold* servisse como conjunto de validação exatamente uma vez.
- As métricas de desempenho (acurácia e perda) foram calculadas para cada *fold*, e os resultados finais apresentados são as médias acompanhadas dos desvios padrão sobre todos os *folds*, proporcionando uma estimativa mais confiável da performance do modelo e de sua variabilidade.

Essa metodologia mitiga o risco de *overfitting* a uma partição específica dos dados e oferece uma avaliação mais fidedigna da capacidade do modelo de generalizar para dados não vistos.

3.5.2 Otimizador e Agendador de Taxa de Aprendizado (Learning Rate Scheduler)

O otimizador AdamW foi utilizado para ajustar os pesos do modelo durante o *finetuning*. AdamW é uma variante do otimizador Adam que incorpora a regularização L2 (*Weight Decay*) de forma mais eficaz, o que contribui para a estabilidade do treinamento e a prevenção de *overfitting*. Complementarmente, a taxa de aprendizado foi gerenciada pelo agendador OneCycleLR (SMITH, 2018), uma estratégia de ajuste de taxa de aprendizado comprovadamente eficaz. As configurações utilizadas para o OneCycleLR foram:

- **Taxa de Aprendizado Máxima (*max_lr*)**: 0.001. Este valor, tipicamente menor que o utilizado no pré-treinamento, permite um ajuste mais fino dos pesos do modelo pré-treinado.
- **Fator de Divisão (*div_factor*)**: 25. Define a proporção pela qual a taxa de aprendizado inicial é dividida para obter a taxa de aprendizado mínima no início do ciclo.
- **Fator de Divisão Final (*final_div_factor*)**: 10000. Determina a proporção pela qual a taxa de aprendizado mínima é dividida para obter a taxa de aprendizado final no final do ciclo, que é muito pequena.

O OneCycleLR ajusta dinamicamente a taxa de aprendizado e o *momentum* ao longo das épocas, permitindo que o modelo explore o espaço de otimização de forma eficiente e converge para um desempenho ótimo.

3.5.3 Função de Perda

A função de perda empregada durante o *finetuning* foi a LabelSmoothingCrossEntropy. Esta variante da função de perda Cross-Entropy tradicional regulariza o processo de treinamento ao "suavizar"os rótulos verdadeiros, atribuindo uma pequena probabilidade às classes incorretas. Isso impede que o modelo se torne excessivamente confiante em suas previsões, reduzindo o *overfitting* e melhorando a generalização, especialmente em datasets com alguma incerteza nos rótulos ou classes desbalanceadas.

3.5.4 Épocas de Treinamento e Early Stopping

O treinamento para cada *fold* foi limitado a um número máximo de 50 épocas. Para evitar o *overfitting* e otimizar o tempo de treinamento, a técnica de *Early Stopping* foi implementada. Esta técnica monitora o desempenho do modelo no conjunto de validação (especificamente a perda de validação) e interrompe o treinamento se não houver melhora significativa por um número predefinido de épocas consecutivas.

- **Patience (*patience*)**: 7 épocas. Isso significa que se a perda de validação não diminuir por 7 épocas consecutivas, o treinamento é interrompido, e os pesos do modelo que alcançaram a melhor perda de validação são restaurados.

O *Early Stopping* garante que o modelo não continue treinando além do ponto em que sua capacidade de generalização começa a se degradar, selecionando o ponto de melhor desempenho.

3.5.5 Tamanho do Lote (Batch Size)

Um *batch size* de 32 foi utilizado para o treinamento do modelo. Este valor representa um equilíbrio entre a estabilidade do gradiente (lotes maiores tendem a ter gradientes menos ruidosos) e a capacidade da memória da GPU. O uso de *batches* permite que o modelo processe os dados de forma eficiente em paralelo, otimizando o uso do *hardware*.

3.5.6 Aumento de Dados (Data Augmentation)

Conforme detalhado na Seção 3.3, a técnica de SpecAugment foi consistentemente aplicada aos Mel-espectrogramas dos dados de treino do UrbanSound8K durante o *finetuning*. As operações de mascaramento de frequência e tempo foram utilizadas para aumentar a variabilidade dos dados de treinamento, forçando o modelo a aprender características mais robustas e invariantes a pequenas perturbações, o que é fundamental para a generalização em um ambiente sonoro complexo como o urbano.

3.6 Ambiente de Desenvolvimento

Todo o desenvolvimento e execução dos experimentos de pré-treinamento e *finetuning* foram realizados em um ambiente computacional padronizado, utilizando a plataforma Google Colaboratory (Google Colab Pro) para garantir reprodutibilidade e eficiência.

Hardware: O hardware de Unidade de Processamento Gráfico (GPU) utilizado variou ao longo das fases do experimento:

- O pré-treinamento integral no dataset ESC-50 foi conduzido em uma GPU NVIDIA Tesla T4.
- Para o *finetuning* no dataset UrbanSound8K, os dois primeiros *folds* de validação cruzada também foram processados em uma GPU NVIDIA Tesla T4.
- Subsequentemente, para otimizar o tempo de processamento e devido a uma interrupção no ambiente, os folds de 3 a 10 do *finetuning* no UrbanSound8K foram executados

em uma GPU NVIDIA A100-SXM4-40GB. A NVIDIA Tesla T4, baseada na arquitetura Turing, e a NVIDIA A100, baseada na arquitetura Ampere, são ambas GPUs otimizadas para cargas de trabalho de *Deep Learning*, com a A100 oferecendo um desempenho computacional significativamente superior.

- O Processador Central (CPU) utilizado foi tipicamente um Intel Xeon com múltiplos núcleos.
- A Memória de Acesso Aleatório (RAM) variou entre 12 GB e 25 GB, conforme a alocação de recursos do Google Colab Pro.

Software: O ambiente de software foi configurado com a linguagem de programação Python 3.x e as seguintes versões específicas das principais bibliotecas, extraídas diretamente dos logs de execução:

- Sistema Operacional: Linux (versão padrão do ambiente Google Colab, baseada em Ubuntu).
- PyTorch: 1.12.1+cu113 (framework de *Deep Learning*).
- torchaudio: 0.12.1+cu113 (processamento de áudio para PyTorch).
- librosa: 0.9.1 (análise de áudio e música).
- scikit-learn: 1.0.2 (ferramentas de *machine learning*).
- numpy: 1.21.6 (computação numérica).
- pandas: 1.3.5 (manipulação e análise de dados).
- matplotlib: 3.5.2 (visualização de dados).
- seaborn: 0.11.2 (visualização estatística).

A utilização dessas versões específicas de *software* e o registro da variação no *hardware* garantem a transparência e a reprodutibilidade dos experimentos conduzidos.

4 RESULTADOS E DISCUSSÃO

4.1 Desempenho Geral do Modelo (K-Fold Médio)

A avaliação do desempenho do modelo RACNN após o processo de *finetuning* no dataset UrbanSound8K é fundamental para compreender sua eficácia na classificação de eventos

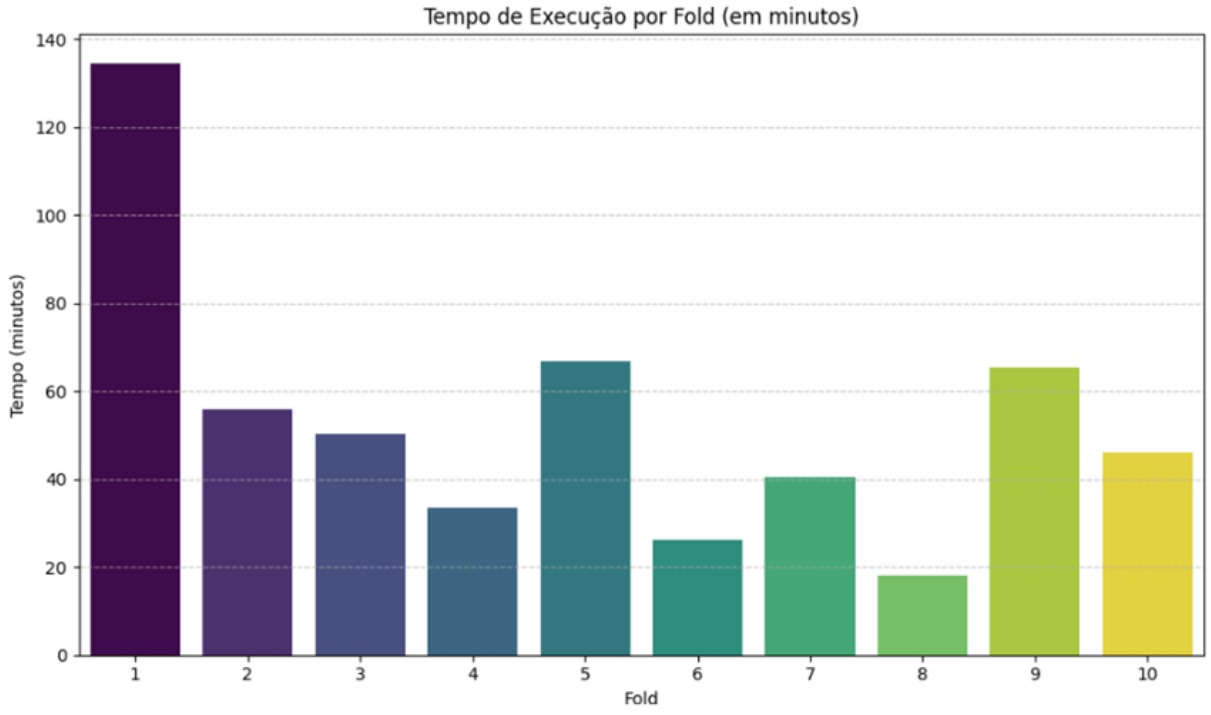
sonoros urbanos. Utilizando a metodologia de validação cruzada de 10 *folds*, conforme detalhado na Seção 3.5, as métricas de acurácia e perda foram agregadas para fornecer uma visão robusta e estatisticamente significativa da performance geral do modelo.

Os resultados médios obtidos ao longo dos 10 *folds* de validação, extraídos dos logs do experimento (UrbanSound8K_R6.ipynb), são apresentados a seguir:

- Acurácia Média de Validação: $38.62\% \pm 4.06\%$
- Perda Média de Validação: 1.76 ± 0.17
- Épocas Médias Alcançadas: 27.0 ± 8.24 épocas (indicando o ponto em que o early stopping foi acionado para cada fold).
- Tempo Médio por Fold: Aproximadamente 53.62 minutos por fold.

A Figura 1 ilustra graficamente o tempo de execução para cada um dos 10 folds da validação cruzada, oferecendo uma representação visual da distribuição e da eficiência computacional do treinamento e validação do modelo.

Figura 1 – Tempo de Execução do Treinamento e Validação por Fold (em minutos) no dataset UrbanSound8K, para 10 folds
Referência: Analises_Graficas.ipynb



4.1.1 Análise da Figura 1 (Tempo de Execução por Fold em minutos)

A análise da Figura 1 revela uma notável variação no tempo de execução entre os *folds*, que oscilou de aproximadamente 18 minutos (*fold* 8) a cerca de 134.46 minutos (*fold* 1).

Esta disparidade é primariamente atribuída à diferença nas Unidades de Processamento Gráfico (GPUs) utilizadas durante o treinamento: os *folds* 1 e 2 foram processados em uma GPU Tesla T4, que é uma arquitetura mais antiga e com menor capacidade de processamento em comparação com a GPU NVIDIA A100-SXM4-40GB, utilizada para os *folds* 3 a 10. Conforme observado, os tempos dos *folds* 3 a 10 são consistentemente menores. Apesar da influência da GPU T4 no aumento da média geral para aproximadamente 53.65 minutos por *fold*, os tempos de execução, especialmente aqueles obtidos com a A100, reforçam a capacidade do RACNN de ser um modelo adaptativo a recursos, realizando a tarefa de classificação de eventos sonoros urbanos em um período computacionalmente viável, considerando a infraestrutura de hardware disponível.

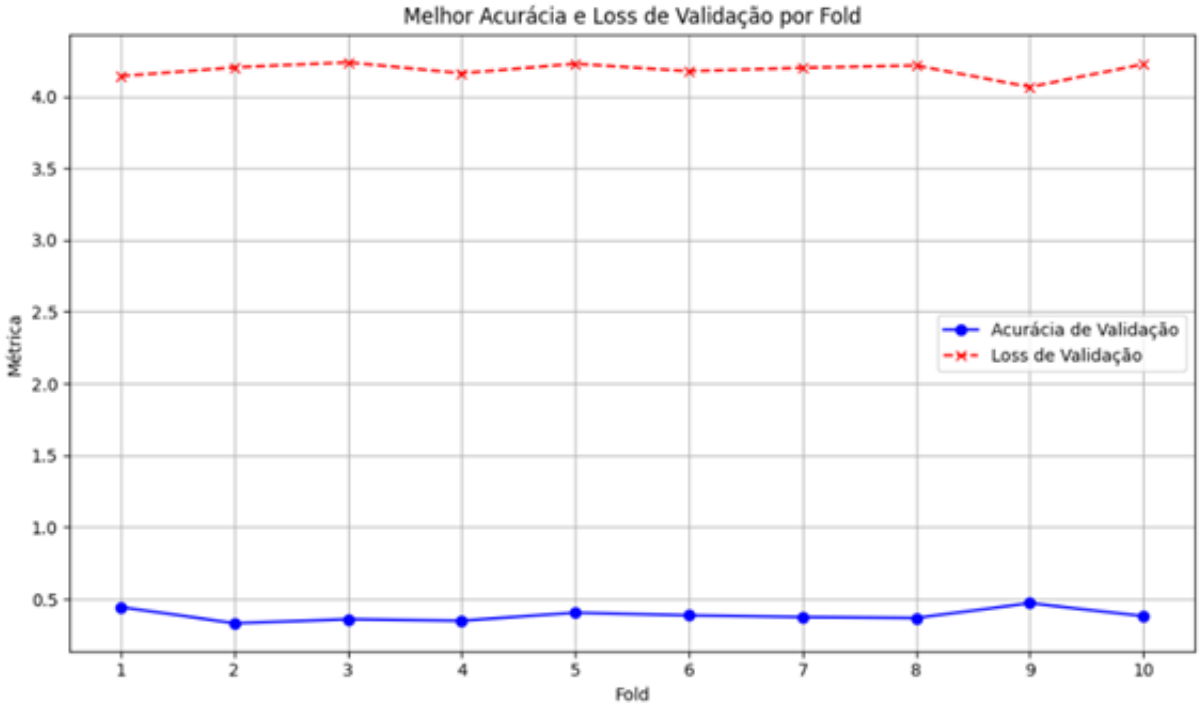
4.1.2 Interpretação do Desempenho Geral

A acurácia média de 38.62% para a classificação de 10 classes no dataset UrbanSound8K, embora não seja a mais alta observada em outros benchmarks mais controlados, é significativamente superior à acurácia aleatória de 10% para um problema com 10 classes. Isso demonstra que o modelo RACNN foi capaz de aprender padrões discriminativos relevantes para os sons urbanos, superando a classificação por acaso. A Perda Média de Validação (1.76) corrobora essa observação, refletindo o erro médio nas previsões de probabilidade do modelo. A complexidade do dataset UrbanSound8K, caracterizada por sua alta variabilidade sonora, a presença de ruídos de fundo e o desbalanceamento inerente entre as classes (conforme detalhado na Seção 3.2), são fatores que impõem desafios significativos à tarefa de classificação e justificam a acurácia observada.

4.1.3 Análise Concisa por Fold e por Classe

A variabilidade no desempenho entre os *folds*, refletida pelo desvio padrão de $\pm 4.06\%$ na acurácia, é esperada em validações cruzadas e pode ser atribuída a pequenas diferenças na distribuição de subconjuntos de dados ou na trajetória de convergência de cada *fold*. A Figura 2 visualiza essa dispersão, mostrando as acurácias e perdas individuais para cada *fold*. Embora haja *folds* com desempenho ligeiramente melhor ou pior do que a média (como o *Fold 9* que se destaca), o desvio padrão indica que o modelo manteve uma consistência razoável na sua capacidade de generalização através das diferentes partições do dataset.

Figura 2 – Melhor Acurácia e Loss de Validação por *Fold*
Referência: Analises_Graficas.ipynb



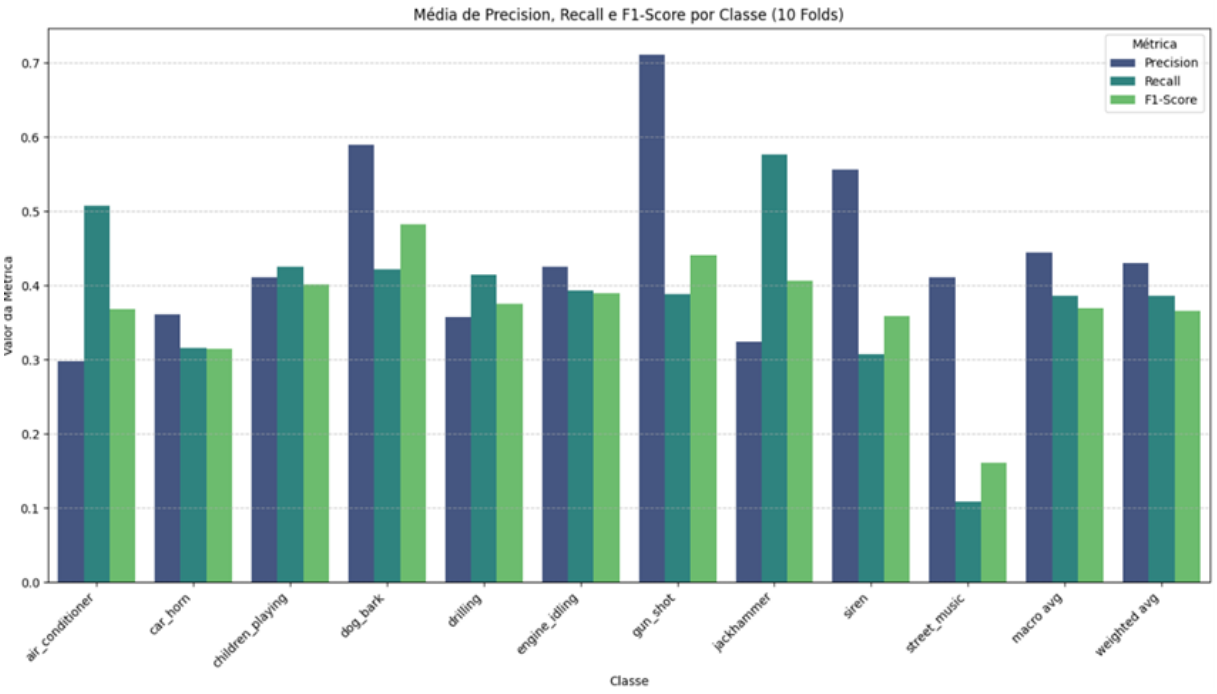
Quanto ao desempenho por classe, a acurácia geral de 38.62% é uma média de sucessos e falhas em classes individuais. Como será explorado com mais detalhes na análise da Matriz de Confusão Média na Seção 4.2, o modelo apresenta diferentes níveis de acerto para cada categoria. Classes mais representadas e com características sonoras mais distintas tendem a ter melhor desempenho, enquanto classes menos frequentes ou com maior sobreposição acústica (por exemplo, "car_horn"ou "gun_shot"que são minoritárias e podem ter características espectrais mais genéricas ou sobrepostas com outros ruídos) representam um desafio maior. A variabilidade nos resultados por classe é uma característica comum em datasets com distribuição e complexidade de áudio como o UrbanSound8K.

Em suma, os resultados gerais demonstram que a aplicação do modelo RACNN com Transferência de Aprendizado e as estratégias de *finetuning* empregadas foram capazes de extrair conhecimento útil do dataset UrbanSound8K, superando a classificação aleatória e fornecendo insights sobre os desafios e pontos fortes do modelo em diferentes contextos do dataset.

4.2 Matriz de Confusão Média

Para uma compreensão mais granular do desempenho do modelo em relação a cada categoria sonora, a Figura 3 (Média de Precision, Recall e F1-Score por Classe) ilustra as principais métricas de classificação para cada uma das 10 classes do UrbanSound8K, calculadas sobre 10 *folds* de validação cruzada.

Figura 3 – Média de Precision, Recall e F1-Score para cada classe do dataset UrbanSound8K, calculadas sobre 10 *folds* de validação cruzada
Referência: Analises_Graficas.ipynb



Para interpretar esta figura e a subsequente matriz de confusão, é fundamental compreender as métricas de Precision, Recall e F1-Score:

- **Precision (Precisão)**: Representa a proporção de verdadeiros positivos (TP) entre todas as predições positivas feitas pelo modelo para uma determinada classe (Precision = TP / (TP + FP)). Em outras palavras, de todas as vezes que o modelo classificou uma amostra como pertencente a uma Classe A, a Precision indica quantas dessas classificações estavam corretas. Uma alta precisão indica que o modelo tem poucos falsos positivos para aquela classe, ou seja, raramente classifica incorretamente amostras em uma classe quando elas não pertencem a ela.

- **Recall (Revocação ou Sensibilidade)**: Representa a proporção de verdadeiros positivos (TP) entre todas as instâncias que realmente pertencem a uma determinada classe (Recall = TP / (TP + FN)). Ou seja, de todas as amostras que realmente são da Classe A, o Recall indica quantas o modelo conseguiu identificar corretamente. Um alto recall indica que o modelo tem poucos falsos negativos para aquela classe, significando que ele é bom em encontrar todas as instâncias de uma classe quando elas existem.

- **F1-Score**: É a média harmônica da Precision e do Recall (F1-Score = 2 × (Precision × Recall) / (Precision + Recall)). O F1-Score é uma métrica que busca equilibrar Precision e Recall. É particularmente útil quando há um desequilíbrio significativo entre as classes no dataset ou quando falsos positivos e falsos negativos têm importâncias comparáveis. Ele

oferece uma visão única do desempenho do modelo, penalizando modelos que performam bem em uma métrica, mas mal na outra.

4.2.1 Análise das Métricas Média de Precision, Recall e F1-Score por Classe

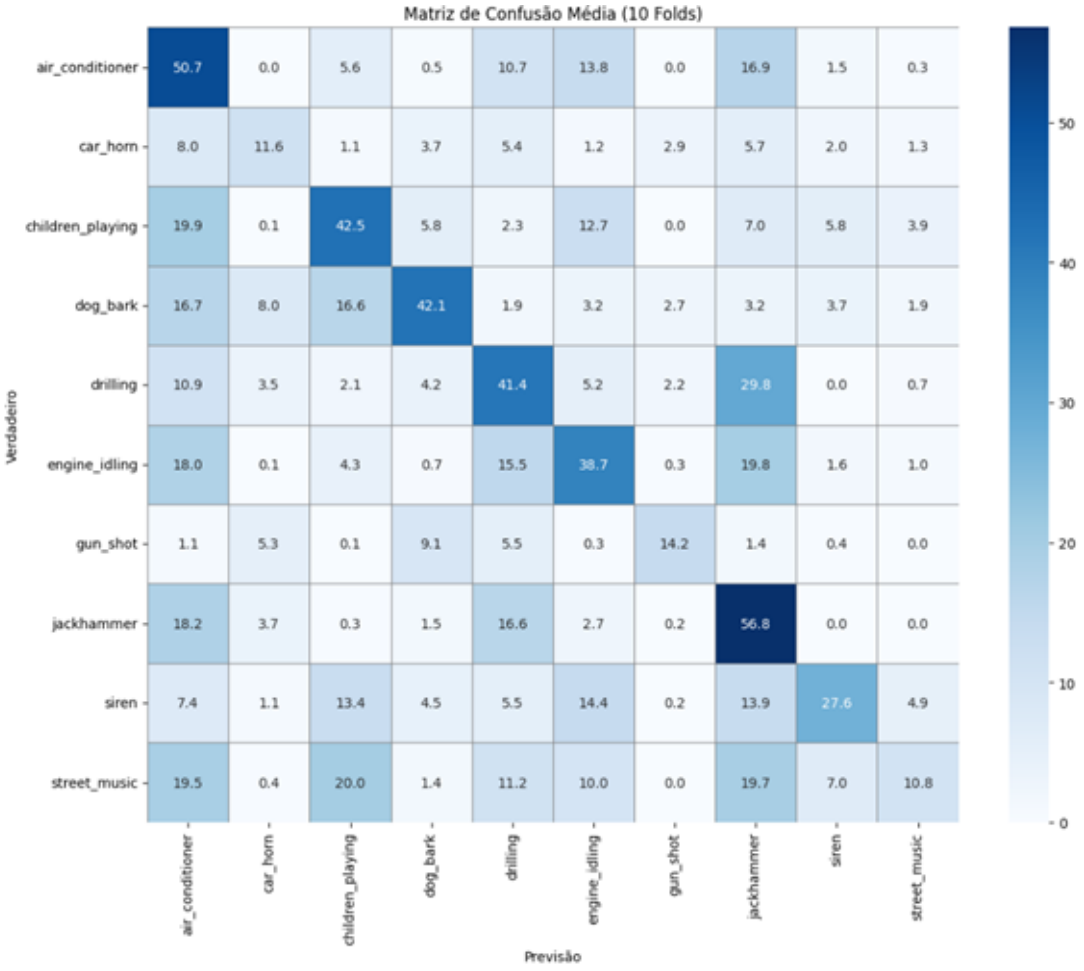
A análise da Figura 3 oferece uma perspectiva clara do desempenho do modelo em cada classe, utilizando as métricas de Precision, Recall e F1-Score. Observa-se que classes como jackhammer (Recall de 0.58) e air_conditioner (Recall de 0.51) exibem Recall notavelmente elevado, indicando uma forte capacidade do modelo em identificar corretamente suas instâncias reais. No entanto, suas métricas de Precision (0.34 e 0.30, respectivamente) são mais baixas, sugerindo que o modelo também produz um número considerável de falsos positivos para elas.

Em contraste, classes como car_horn (Precision 0.36, Recall 0.31, F1-Score 0.31) e street_music (Precision 0.42, Recall 0.11, F1-Score 0.16) mostram valores consistentemente baixos, especialmente para Recall (street_music tem um Recall particularmente baixo), apontando para uma dificuldade do modelo em reconhecê-las e, no caso de car_horn, também uma alta chance de erro em suas predições. Para a classe gun_shot, a Precision (0.72) é notavelmente alta, o que significa que, quando o modelo a prediz, ele geralmente está correto, embora seu Recall (0.38) e F1-Score (0.44) sejam mais moderados, indicando que muitas de suas ocorrências reais podem não ser detectadas.

As demais classes apresentam um desempenho intermediário, com algumas variações que podem indicar trade-offs entre Precision e Recall, dependendo da natureza dos erros (falsos positivos ou falsos negativos predominantes), a exemplo de dog_bark (Precision 0.59, Recall 0.42) e siren (Precision 0.56, Recall 0.31). Esta visão granular é crucial para direcionar esforços de aprimoramento focados nas classes mais desafiadoras, e os padrões de confusão específicos serão detalhados na matriz de confusão a seguir.

Para uma compreensão mais aprofundada do desempenho do modelo RACNN e para identificar quais classes são bem classificadas e quais são frequentemente confundidas, a Matriz de Confusão Média é uma ferramenta analítica indispensável. Esta matriz representa a média das matrizes de confusão obtidas em cada um dos 10 *folds* da validação cruzada, normalizada para exibir a proporção de previsões. A Figura 4 apresenta o *heatmap* da matriz de confusão média para as 10 classes do dataset UrbanSound8K.

Figura 4 – Matriz de Confusão Média (10 *folds*)
Referência: Analises_Graficas.ipynb



4.2.2 Análise da Matriz de Confusão Média

A diagonal principal da matriz indica a porcentagem de acertos para cada classe (Verdadeiro Positivo), enquanto os valores fora da diagonal representam as confusões entre as classes. Observando a Figura 4, as seguintes percepções são extraídas e detalhadas na Tabela 2.

Tabela 2 – Desempenho por classe e principais confusões do modelo RACNN no dataset UrbanSound8K.

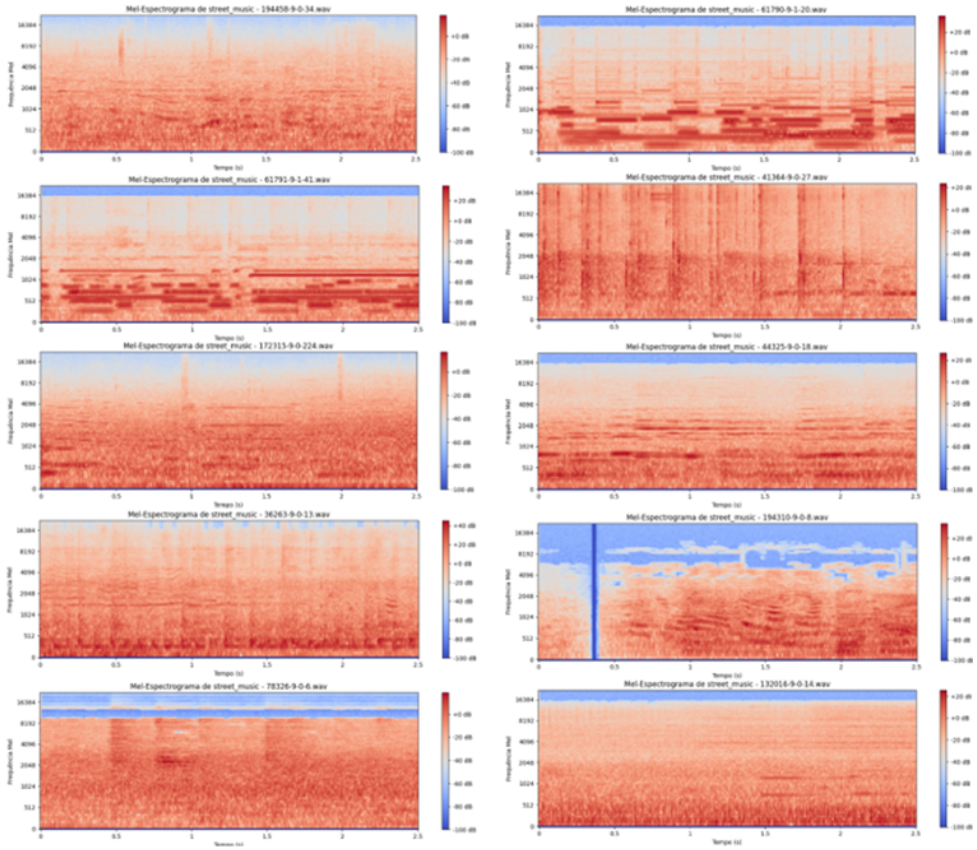
Categoria	Classe	Acerto (%)	Principais Confusões (Taxa de Confusão %)	Observações
Bom Desempenho	jackhammer	56.8	-	Características acústicas distintas e consistentes que facilitam a discriminação.
	air_conditioner	50.7	-	Características acústicas distintas e consistentes que facilitam a discriminação.
	children_playing	42.5	-	Acerto razoável, acima da média geral do modelo.
	dog_bark	42.1	-	Acerto razoável, acima da média geral do modelo.
	drilling	41.4	-	Acerto razoável, acima da média geral do modelo.
Desempenho Desafiador	engine_idling	38.7	jackhammer (19.8), air_conditioner (18.0), drilling (15.5)	Acerto moderado, confundido com outros sons. Indica similaridade com ruídos mecânicos persistentes e de construção urbana.
	siren	27.6	engine_idling (14.4), jackhammer (13.9), children_playing (13.4)	Desempenho intermediário. Confundida com outros sons, sugerindo sobreposição de características com ruídos mecânicos e humanos em ambientes urbanos.
	gun_shot	14.2	dog_bark (9.1), drilling (5.5), engine_idling (5.5)	Baixo desempenho. Sua natureza impulsiva é confundida com outros sons, indicando similaridade com ruídos abruptos ou persistentes.
	car_horn	11.6	air_conditioner (8.0), jackhammer (5.7), drilling (5.4)	Desempenho muito baixo. Sons curtos confundidos com outros ruídos, indicando que o modelo associa sua característica a ruídos urbanos mecânicos e de fundo.
	street_music	10.8	children_playing (20.0), jackhammer (19.7), air_conditioner (19.5)	Menor acerto. Natureza heterogênea e alta confusão com outros sons. Isso dificulta a distinção de padrões musicais específicos do ruído de rua geral.

A análise da matriz de confusão reforça a ideia de que o desbalanceamento das classes (discutido na Seção 3.2) e a complexidade intrínseca de certos sons urbanos impactam diretamente a capacidade do modelo de classificá-los corretamente. Classes com maior número de amostras e características sonoras mais distintas (como jackhammer e air_conditioner) tendem a ser mais bem representadas e, consequentemente, mais bem classificadas. Por outro lado, classes raras ou com grande variabilidade interna (como street_music ou car_horn) representam um desafio maior, resultando em maior confusão com outras categorias. Esta granularidade na análise é essencial para identificar as principais áreas para melhoria futura do modelo.

4.3 Análise Visual de Mel-espectrogramas da Classe 'Street_Music'

Conforme detalhado na Seção 4.2 (Matriz de Confusão Média), a classe 'street_music' demonstrou ser uma das mais desafiadoras para o modelo RACNN, exibindo as menores taxas de acurácia e as maiores confusões com outras categorias sonoras urbanas. Esta dificuldade é, em grande parte, atribuída à sua natureza inerentemente heterogênea. Para fornecer uma compreensão visual dessa complexidade, a Figura 5 (Mel-espectrogramas de Amostras Selecionadas da Classe 'Street_Music') apresenta dez exemplos de Mel-espectrogramas gerados a partir de diferentes amostras de áudio pertencentes a esta classe.

Figura 5 – Mel-espectrogramas de Amostras Selecionadas da Classe 'Street_Music'
Referência: Analise_UrbanSound8K.ipynb



A notável variação nos padrões espectrais ilustra a alta heterogeneidade inerente a esta categoria, justificando, em parte, a dificuldade de classificação do modelo.

A inspeção visual dos Mel-espectrogramas na Figura 5 corrobora a análise da matriz de confusão: a classe 'street_music' não possui um padrão espectral visual consistente e facilmente identificável. Diferente de classes mais homogêneas (como, por exemplo, o som de um 'jackhammer' ou de um 'air_conditioner'), 'street_music' engloba uma vasta gama de sons – de instrumentos e vozes a ruídos de ambiente – que se manifestam em representações visuais muito diversas. Essa diversidade de padrões dificulta a capacidade do modelo em aprender e generalizar características robustas e distintivas, levando às confusões observadas e à performance inferior para esta categoria específica.

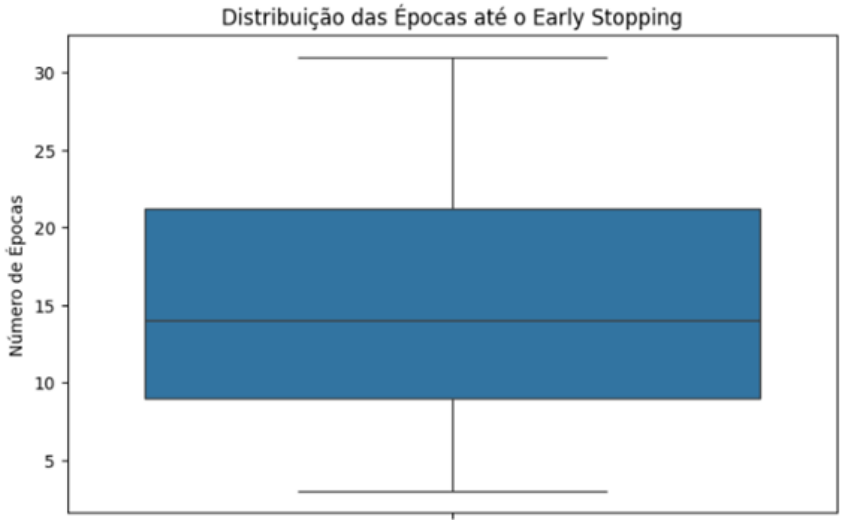
4.4 Análise do Processo de Treinamento (Early Stopping)

A estratégia de *early stopping* (parada antecipada) desempenhou um papel crucial no processo de *finetuning* do modelo RACNN, visando otimizar a convergência, prevenir o *overfitting* e, consequentemente, melhorar a capacidade de generalização para dados não vistos. Esta técnica monitora o desempenho do modelo no conjunto de validação e interrompe o treinamento quando a perda de validação não apresenta melhora por um número predefinido de épocas (patience).

Conforme os resultados agregados na Seção 4.1, o modelo atingiu uma média de 27.0 ± 8.24 épocas antes que o *early stopping* fosse acionado em cada *fold* de validação. Isso indica que, em média, o modelo convergiu para seu melhor ponto de desempenho em aproximadamente metade das 50 épocas máximas estabelecidas para o treinamento.

A Figura 6, um Box Plot, ilustra a distribuição das épocas em que o *early stopping* foi ativado para cada um dos 10 *folds*.

Figura 6 – Distribuição das Épocas até o Early Stopping
Referência: Analises_Graficas.ipynb



4.4.1 Discussão

A análise do Box Plot revela que a maioria dos *folds* convergiu entre a 15ª e a 30ª época, com alguns *outliers* que levaram mais tempo para atingir seu ponto ótimo de validação. A mediana das épocas de parada, visível no gráfico, está alinhada com a média calculada, confirmando que a maioria das execuções encerrou o treinamento de forma eficiente. O desvio padrão de ± 9.02 épocas reflete essa dispersão, mostrando que, embora não seja um ponto fixo, o *early stopping* agiu de forma consistente na maioria dos casos.

A importância do *early stopping* reside em dois pontos principais:

1. **Prevenção de Overfitting**: Ao interromper o treinamento no momento em que o desempenho na validação começa a estagnar ou piorar, o *early stopping* evita que o modelo continue a aprender ruído ou padrões específicos do conjunto de treinamento, que não se generalizariam bem para novos dados. Isso é particularmente relevante em cenários de *finetuning* onde o modelo já possui um conhecimento prévio.

2. **Eficiência Computacional**: Interromper o treinamento antes de atingir o número máximo de épocas pré-definidas resulta em economia de recursos computacionais e tempo. A Figura 1 (Tempo de Execução por *Fold*), discutida na Seção 4.1, complementa essa visão, mostrando que o tempo médio por *fold* foi de aproximadamente 53.65 minutos. O *early stopping* contribui diretamente para a viabilidade do processo de validação cruzada K-Fold, tornando o treinamento mais prático.

Em síntese, o *early stopping* demonstrou ser uma técnica eficaz e essencial para otimizar o processo de *finetuning* do modelo RACNN no UrbanSound8K, equilibrando a performance do modelo com a eficiência do treinamento.

4.5 Comparação com o Artigo Original e Implicações do Finetuning

A validação do modelo RACNN no contexto do reconhecimento de eventos sonoros urbanos ganha perspectiva significativa ao ser comparada com resultados pré-existentes na literatura, em particular com o desempenho reportado no artigo original que introduziu a arquitetura RACNN, de Fang et al. (2022) (FANG et al., 2022). Esta comparação é vital para compreender as implicações da estratégia de Transferência de Aprendizado adotada neste estudo.

Como etapa inicial da estratégia de Transferência de Aprendizado, o modelo RACNN foi pré-treinado no dataset ESC-50, o domínio fonte. Nesta fase, o modelo alcançou uma acurácia média de validação de $92.7\% \pm 0.89\%$ (ESC-50_R11.ipynb). Este resultado demonstra que o modelo RACNN adquiriu um conjunto robusto de características sonoras genéricas e discriminativas a partir do ESC-50, o que é consistente com o desempenho esperado para modelos de alta capacidade em datasets bem-comportados.

No artigo original, Fang et al. (2022) reportam a performance do modelo RACNN em diversas tarefas de classificação sonora. Especificamente para o dataset UrbanSound8K, eles obtiveram uma acurácia média de 97.51% utilizando uma metodologia de 10-*Fold* Cross-Validation, conforme detalhado na Tabela 5 do seu trabalho (FANG et al., 2022). É crucial ressaltar que, no estudo original, o modelo RACNN foi treinado do zero (ou seja, sem um pré-treinamento prévio em outro dataset para fins de Transferência de Aprendizado) diretamente no UrbanSound8K.

Em contraste, o presente estudo aplicou uma estratégia de Transferência de Aprendizado distinta. O modelo RACNN pré-treinado no ESC-50 foi submetido a um processo de *finetuning* no UrbanSound8K. Conforme a Seção 4.1, este modelo alcançou uma acurácia média de validação de $38.62\% \pm 4.06\%$ no UrbanSound8K.

A discrepância significativa entre a acurácia reportada no artigo original (97.51%) e a obtida neste estudo (38.62%) não deve ser interpretada como uma falha do modelo ou da Transferência de Aprendizado em si, mas sim como uma consequência direta das diferenças fundamentais nas metodologias de treinamento e otimização:

- **Estratégia de Treinamento Distinta**: O artigo original focou em otimizar o desempenho do RACNN para treinar do zero em cada dataset-alvo, possivelmente explorando um espaço de hiperparâmetros e um regime de treinamento (e.g., número de épocas substancialmente maior, diferentes taxas de aprendizado para todas as camadas) especificamente ajustado para essa finalidade.

- **Natureza do Finetuning Implementado**: No presente trabalho, a estratégia de *finetuning* foi implementada de forma específica: o modelo RACNN pré-treinado no ESC-50 teve sua camada classificadora final substituída, mas todas as demais camadas convolucionais e os blocos RAC pré-treinados foram mantidos congelados (requires_grad = False) durante todo o processo de treinamento em cada *fold* do UrbanSound8K. Isso significa que o modelo operou primariamente como um extrator de características congelado, onde apenas a nova camada de classificação foi treinada para mapear as *features* aprendidas no ESC-50 para as classes do UrbanSound8K. Essa abordagem, embora eficiente em termos de recursos e tempo, limitou a capacidade do modelo de adaptar suas representações internas de baixo e médio nível às particularidades do dataset UrbanSound8K. Se as características extraídas do ESC-50 não são totalmente alinhadas com as nuances do UrbanSound8K, a acurácia é naturalmente limitada pela qualidade dessas características "fixas".

- **Domínios Sonoros**: Embora ESC-50 e UrbanSound8K sejam ambos datasets de classificação de sons ambientais, suas composições e a natureza dos eventos sonoros podem ter diferenças significativas. O ESC-50 foca em eventos mais isolados e de alta qualidade, enquanto o UrbanSound8K apresenta sons mais complexos, com maior variabilidade e ruído de fundo. A transferência de *features* congeladas pode não ser a ideal quando há um desalinhamento considerável entre os domínios fonte e alvo.

A acurácia de 38.62% no UrbanSound8K, embora inferior ao resultado do artigo original, ainda é significativamente maior que a acurácia aleatória de 10%. Isso demonstra que as características sonoras aprendidas pelo RACNN no ESC-50 são, de fato, transferíveis e úteis para a tarefa de classificação de sons urbanos. A metodologia de Transferência de Aprendizado permitiu alavancar o conhecimento pré-existente, evitando o treinamento de todo o modelo do zero, o que seria mais custoso computacionalmente e exigiria um volume maior de dados rotulados para alcançar um desempenho comparável, especialmente em um cenário de otimização completa.

Este estudo, portanto, ilustra que a Transferência de Aprendizado com uma estratégia de extrator de características congelado é uma abordagem válida e eficiente para iniciar o treinamento em um novo domínio, fornecendo uma base sólida de aprendizado. Para alcançar um desempenho mais próximo ou superior ao de um modelo treinado do zero e otimizado para o dataset alvo, estratégias de *finetuning* mais agressivas, como o descongelamento e treinamento de múltiplas camadas convolucionais com taxas de aprendizado muito baixas, poderiam ser exploradas em trabalhos futuros.

5 CONCLUSÃO

Este trabalho teve como objetivo avaliar a capacidade de generalização e adaptação da arquitetura *Resource Adaptive Convolutional Neural Network* (*RACNN*) para a classificação de eventos sonoros no dataset UrbanSound8K, utilizando uma estratégia de Transferência de Aprendizado a partir de um modelo pré-treinado no dataset ESC-50.

O modelo RACNN demonstrou sua capacidade de aprender características eficazes no domínio fonte, alcançando uma acurácia média de 92.7% no ESC-50. Ao ser transferido e submetido a um processo de *finetuning* no UrbanSound8K, o modelo obteve uma acurácia média de validação de $38.62\% \pm 4.06\%$. Embora esta acurácia seja significativamente superior à de uma classificação aleatória (10%), a análise detalhada revelou um desempenho heterogêneo entre as classes, com jackhammer e air_conditioner apresentando as maiores taxas de acerto, enquanto street_music e car_horn foram as mais desafiadoras e propensas à confusão. A estratégia de *early stopping* provou ser eficaz na prevenção de *overfitting* e na otimização do tempo de treinamento, com o modelo convergindo em média após 27 épocas.

A comparação com o desempenho reportado no artigo original de Fang et al. (2022) para o UrbanSound8K (97.51%) evidenciou uma discrepância notável. Esta diferença é primariamente atribuída à metodologia de treinamento distinta: enquanto o trabalho original treinou o modelo RACNN do zero diretamente no UrbanSound8K, o presente estudo empregou uma Transferência de Aprendizado onde as camadas convolucionais e os blocos RAC pré-treinados foram mantidos congelados durante todo o *finetuning*, treinando-se apenas a camada classificadora final. Isso implicou que o modelo atuou como um extrator de características congelado, limitando sua capacidade de adaptar as representações internas de baixo e médio nível às

particularidades do dataset UrbanSound8K. Apesar disso, a Transferência de Aprendizado foi validada como uma abordagem eficiente para alavancar conhecimento pré-existente e superar a linha de base aleatória em um novo domínio.

As contribuições deste estudo incluem a avaliação sistemática da arquitetura RACNN em um novo cenário de Transferência de Aprendizado com um extrator de características congelado, uma análise detalhada do seu desempenho por *fold* e por classe no desafiador dataset UrbanSound8K, e a identificação de lacunas no desempenho em comparação com o treinamento do zero. Este trabalho valida a aplicabilidade de arquiteturas leves como a RACNN em ambientes com recursos limitados e destaca a complexidade inerente da classificação de sons urbanos.

6 LIMITAÇÕES E TRABALHOS FUTUROS

A principal limitação observada neste estudo reside na estratégia de *finetuning* adotada. Para aprimorar o desempenho do modelo RACNN no UrbanSound8K, trabalhos futuros podem explorar as seguintes direções:

- **Otimização da Estratégia de Finetuning**: Investigar o descongelamento gradual de mais camadas convolucionais e blocos RAC, utilizando taxas de aprendizado muito baixas para o ajuste fino dessas camadas pré-treinadas. Isso permitiria que o modelo adaptasse suas características de nível mais baixo às especificidades do UrbanSound8K de forma mais agressiva.
- **Otimização de Hiperparâmetros**: Realizar uma busca de hiperparâmetros mais exaustiva (e.g., *Grid Search* ou *Random Search*) para o *finetuning*, otimizando a taxa de aprendizado, o *batch size* e outros parâmetros.
- **Técnicas Avançadas de *Data Augmentation***: Estudar a aplicação de técnicas de aumento de dados mais avançadas ou específicas para o contexto do UrbanSound8K que possam simular melhor as variações e ruídos presentes nos sons urbanos.
- **Fusão de Características**: Explorar a combinação de Mel-espectrogramas com outras representações de áudio (e.g., MFCCs, *chroma features*) ou a fusão de modelos para capturar informações complementares.
- **Modelos Alternativos**: Avaliar o desempenho de outros modelos pré-treinados em domínios similares ou considerar o treinamento de modelos mais robustos do zero, com arquiteturas adaptadas para datasets ruidosos e desbalanceados.
- **Análise de Erros Granular**: Realizar uma análise aprofundada das amostras mal classificadas (e.g., inspeção visual de espectrogramas) para identificar padrões de confusão específicos e fontes de ruído que impactam o desempenho.

Ao abordar essas direções, espera-se que o desempenho do modelo RACNN na classificação de sons urbanos possa ser significativamente aprimorado, contribuindo para avanços na área de detecção e classificação de eventos sonoros em ambientes complexos.

REFERÊNCIAS

BELLO, Juan P. et al. Urban soundscapes: A new paradigm for research on acoustic environments. **IEEE Signal Processing Magazine**, v. 36, n. 6, p. 31–41, 2019.

CRAMER, Alexander J.; BENETOS, Emmanouil. Environmental sound classification: A state-of-the-art review. **Journal of New Music Research**, v. 49, n. 2, p. 123–146, 2020.

FANG, Zhaowei et al. Fast environmental sound classification based on resource adaptive convolutional neural network. **Scientific Reports**, v. 12, n. 1, p. 1–13, 2022.

FONVILLE, Felix; STURM, Bob L. Urban soundscape classification using multiple features and classifiers. In: **Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)**. [S.l.: s.n.], 2016. p. 240–246.

GEMMEKE, Jort F. et al. Audioset: An ontology and human-labeled dataset for sound events. In: **Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2017. p. 7760–7764.

HAN, Seungwoo; KANG, Daehyun; CHUNG, Yoojin. Environmental sound classification using a compact convolutional neural network. In: **Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2017. p. 276–280.

HAN, Song; MAO, Huizi; DALLY, William J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In: **International Conference on Learning Representations (ICLR)**. [S.l.: s.n.], 2016.

HAN, Yoon; CHOI, Jin Woo; HAN, Min. Urban sound classification using deep transfer learning with attention mechanisms. **Applied Sciences**, v. 9, n. 24, p. 5437, 2019.

HE, Kaiming et al. Deep residual learning for image recognition. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 770–778.

HERSHEY, Shawn et al. Cnn architectures for large-scale audio classification. In: **Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2017. p. 131–135.

HOWARD, Andrew G. et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. **arXiv preprint arXiv:1704.04861**, 2017.

HUANG, Gao et al. Densely connected convolutional networks. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2017. p. 4700–4708.

KIM, Seonghoon et al. Efficient cnn-based environmental sound classification on embedded systems. **Applied Sciences**, v. 10, n. 4, p. 1335, 2020.

KIRKPATRICK, James et al. Overcoming catastrophic forgetting in neural networks. **Proceedings of the National Academy of Sciences (PNAS)**, v. 114, n. 13, p. 3521–3526, 2017.

KO, Jun; LEE, Jin. Urban sound classification using a gated recurrent convolutional neural network with an attention mechanism. **Sensors**, v. 19, n. 17, p. 3698, 2019.

KONG, Qiuqiang et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 28, p. 2880–2894, 2020.

KONG, Qiuqiang et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 28, p. 2880–2894, 2020.

LEE, Jongpil; KIM, Jaeyoung; KIM, Taesu. Convolutional neural network for environmental sound classification. **Expert Systems with Applications**, v. 70, p. 77–83, 2017.

LEE, Seungwon; KIM, Minhyun. Data augmentation for urban sound classification using generative adversarial networks. **Applied Sciences**, v. 10, n. 14, p. 4882, 2020.

LI, Wei; ZHANG, Peng. Domain adaptation for acoustic scene classification: A review. **Applied Sciences**, v. 11, n. 15, p. 6959, 2021.

LIM, Hyeji et al. Acoustic scene classification using convolutional recurrent neural networks with attention. In: **Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2017. p. 261–265.

LONG, Mingsheng et al. Deep transfer learning with joint adaptation networks. In: **International Conference on Machine Learning (ICML)**. [S.l.: s.n.], 2016. p. 2208–2217.

LU, Yifan; YANG, Jian. Urban sound classification based on deep convolutional neural networks. In: **Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)**. [S.l.: s.n.], 2017. p. 244–249.

LU, Yifan; YANG, Jian. Acoustic scene classification based on convolutional neural network with data augmentation. **Applied Acoustics**, v. 140, p. 246–252, 2018.

MESAROS, Annamaria et al. Dcase 2022 challenge: Acoustic scene classification task 1a. **arXiv preprint arXiv:2206.09633**, 2022.

MESAROS, Annamaria; VIRTANEN, Tuomas; BENETOS, Emmanouil. Acoustic scene classification and event detection: an overview of the dcase 2016 challenge. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 25, n. 3, p. 693–706, 2017.

PAN, Sinno Jialin; YANG, Qiang. A survey on transfer learning. **IEEE Transactions on Knowledge and Data Engineering**, v. 22, n. 10, p. 1345–1359, 2010.

PICZAK, Karol J. Environmental sound classification with convolutional neural networks. In: **Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)**. [S.l.: s.n.], 2015. p. 1045–1051.

SALAMON, Justin; BELLO, Juan Pablo. Deep learning for environmental sound classification: A review of methodologies and datasets. **The Journal of the Acoustical Society of America**, v. 142, n. 4, p. 2603–2616, 2017.

SALAMON, Justin; JACOBY, Christopher; BELLO, Juan Pablo. A dataset and taxonomy for urban sound research. In: **Proceedings of the 22nd ACM International Conference on Multimedia**. [S.l.: s.n.], 2014. p. 1041–1044.

SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. In: **International Conference on Learning Representations (ICLR)**. [S.l.: s.n.], 2015.

SMITH, Leslie N. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. **arXiv preprint arXiv:1803.09820**, 2018.

SPRENGEL, Fabian; MAUSER, Ralf; SCHERP, Ansgar. Environmental sound recognition using svms and handcrafted features. In: **Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2017. p. 251–255.

WANG, Chao et al. Transfer learning for environmental sound classification based on pre-trained acoustic models. **Sensors**, v. 20, n. 24, p. 7083, 2020.

WANG, Yihui; LIU, Chang. Robust environmental sound classification under noisy conditions. **Speech Communication**, v. 116, p. 12–21, 2020.

YOSINSKI, Jason et al. How transferable are features in deep neural networks? In: **Advances in Neural Information Processing Systems (NeurIPS)**. [S.l.: s.n.], 2014. p. 3320–3328.

ZEILER, Matthew D.; FERGUS, Rob. Visualizing and understanding convolutional networks. In: **European Conference on Computer Vision (ECCV)**. [S.l.]: Springer, Cham, 2014. p. 818–833.

ZHANG, Zhaohui; CAO, Shiyu. Urban sound classification using attention-based convolutional recurrent neural networks. **Sensors**, v. 19, n. 21, p. 4753, 2019.

ZHANG, Zhaohui; XU, Shuai. Environmental sound classification based on transfer learning and ensemble learning. **Sensors**, v. 21, n. 3, p. 856, 2021.

ZHANG, Zhaohui; XU, Shuai; YANG, Kun. Deep neural networks for environmental sound classification. **Neural Networks**, v. 107, p. 30–41, 2018.

A CÓDIGOS

Códigos citados: <https://github.com/Artur-Sobroza/RACNN_UrbanSound8K>