UFAM

UNIVERSIDADE FEDERAL DO AMAZONAS – UFAM

INSTITUTO DE COMPUTAÇÃO – ICOMP

COORDENAÇÃO DE PÓS-GRADUAÇÃO EM INFORMÁTICA – PPGI

WILSON ARAÚJO DE OLIVEIRA NETO

MODELOS GERADORES PARA DETECÇÕES DE ANOMALIAS EM ATIVIDADES SONORAS

Manaus, AM

2023

WILSON ARAÚJO DE OLIVEIRA NETO

MODELOS GERADORES PARA DETECÇÕES DE ANOMALIAS EM ATIVIDADES SONORAS

Dissertação de mestrado submetido à Coordenação de Pós-Graduação em Informática do Instituto de Computação (ICOMP), como requisito parcial para obtenção do Título de Mestre em Informática.

Orientador: Prof. Dr. Carlos Maurício Seródio Figueiredo

Manaus, AM

2023

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Oliveira Neto, Wilson Araujo de
O48m Modelos geradores para detecções de anomalias em atividades sonoras / Wilson Araujo de Oliveira Neto . 2023
80 f.: il. color; 31 cm.

Orientador: Carlos Maurício Seródio Figueiredo
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Anomaly detection. 2. Audio. 3. GAN - Generative Adversarial Network. 4. Sound. I. Figueiredo, Carlos Maurício Seródio. II. Universidade Federal do Amazonas III. Título

Ministério da Educação
Universidade Federal do Amazonas
Coordenação do Programa de Pós-Graduação em Informática

FOLHA DE APROVAÇÃO

"MODELOS GERADORES PARA DETECÇÃO DE ANOMALIAS EM ATIVIDADES SONORAS"

WILSON ARAÚJO DE OLIVEIRA NETO

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Dr. Carlos Maurício Seródio Figueiredo - PRESIDENTE

Prof. Dr. Juan Gabriel Colonna - MEMBRO INTERNO

Profa. Dra. Elloá Barreto Guedes da Costa - MEMBRO EXTERNO

Manaus, 21 de agosto de 2023

Documento assinado eletronicamente por **Carlos Maurício Seródio Figueiredo**, **Usuário Externo**, em 22/08/2023, às 15:19, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Juan Gabriel Colonna**, **Coordenador**, em 22/08/2023, às 17:35, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Elloá Barreto Guedes da Costa**, **Usuário Externo**, em 22/08/2023, às 19:32, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufam.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código

verificador **1661245** e o código CRC **4B1ACD23**.

Avenida General Rodrigo Octávio, 6200 - Bairro Coroado I Campus Universitário Senador Arthur Virgílio Filho, Setor Norte - Telefone: (92) 3305-1181 / Ramal 1193
CEP 69080-900, Manaus/AM, coordenadorppgi@icomp.ufam.edu.br

Referência: Processo nº 23105.037720/2023-52 SEI nº 1661245

AGRADECIMENTOS

Primeiramente, desejo expressar minha gratidão aos meus pais por terem abdicado de tantas coisas para que eu pudesse manter os estudos; mesmo sem compreender exatamente o que faço, eles demonstraram interesse e me encorajaram a prosseguir. Um agradecimento especial é dirigido à minha mãe, Karla Cristina, que sempre acreditou que o futuro me reservaria coisas maravilhosas.

Às minhas filhas peludas, Sophia e Zara, por alegrarem meu dia e demonstrarem o que é o amor em sua forma mais pura e simples.

Aos meus amigos Manoel Victor, Lucas Frota, Amanda Almeida e Juliana Alencar, pelo todo o suporte emocional, companheirismo e preocupação durante esses dois anos de pesquisa. São pessoas inspiradoras e brilhantes.

À minha psicóloga, Mariana Pelizer, por me mostrar a existência de diversos caminhos e ter me ajudado a encarar vários desafios desde a graduação.

Ao Laboratório de Sistemas Inteligentes da Universidade do Estado do Amazonas, que possibilitou a utilização de sua infraestrutura para a realização de experimentos e coleta de dados.

Por fim, gostaria de agradecer aos professores com quem trabalhei ao longo deste período. À professora Elloá Guedes, pela revisão de excelência e contribuições valiosas que direcionaram para a finalização deste trabalho. Ao professor Eduardo Nakamura, por me receber em seu laboratório de pesquisa e pelo apoio. E ao meu orientador, Carlos Maurício Figueiredo, pelos anos de orientação, dedicação e conselhos que pude receber neste período.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto POSGRAD.

“Não tem progresso sem acesso.”

Marina Peralta

RESUMO

Diversos domínios de dados possibilitam a utilização de detecção de anomalias, dentre eles o áudio. Uma funcionalidade importante destes sistemas é identificar quando algo está fora da normalidade. Para isso, diversos estudos utilizando aprendizagem de máquina foram realizados. Os estado-da-arte na identificação de anomalias em imagens utilizam arquiteturas baseadas em GAN (Generative Adversarial Network), entretanto, poucos estudos demonstram a utilização destas ou outras arquiteturas geradoras no domínio de sons. Para lidar com esse problema, este trabalho propõe o desenvolvimento de um método de identificação de anomalias em atividades sonoras utilizando dados capturados através de microfones. O processo de identificação de anomalia é realizado por meio de um modelo gerador a partir de uma arquitetura de rede profunda. Testes utilizando bases de dados reais mostram que algumas alterações nas arquiteturas utilizadas para imagens podem obter resultados promissores. Validamos nossa abordagem no conjunto de dados DCASE 2021, que inclui mais de 180 horas de maquinário industrial. Avaliamos a classificação das anomalias, relatando uma média ponderada de 88,16% de AUC e 78,05% de pAUC, resultados superiores ao apresentado por *baselines*.

Palavras-chaves: Detecção de anomalias, anomalias em eventos sonoros, GANS, modelos geradores

ABSTRACT

Several data domains allow the use of anomaly detection, including audio. An important feature of these systems is to identify when something is different from ordinary. For this purpose, several studies using machine learning were performed. The state of art in anomaly detection in images uses architectures based on GAN (Generative Adversarial Network), however, few studies demonstrate the use of these or other generating architectures in the domain of sounds. To overcome this problem, this work aims to develop a method for identifying anomalies in sound activities using data captured through microphones. The anomaly identification process is carried out through a generator model from a deep network architecture. Tests using real databases show that some changes in the architectures used for images can achieve promising results. This approach has been validated using the DCASE 2021 dataset, which includes over 180 hours of audio from industrial machinery. We evaluated the classification of anomalies, reporting an weighted average of 88,16% AUC and 78,05% pAUC, results superior to those presented by baselines

Keywords: Anomaly Detection, event sound anomalous, GAN, generative neural-networks

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CNN	Rede Neural Convolucional / Convolutional Neural-Network
DNN	Rede Neural Profunda / Deep Neural-Network
AE	Rede Neural Autocodificadora / Autoencoder network
GAN	Rede Neural Geradora Adversária / Generative Adversarial Network
FFNN	Rede Neural direta / FeedForward Neural-network
DOI	Digital Object Identifier
IA	Inteligência Artificial
kNN	k-Nearest Neighbors
RF	Random Forest
FFT	Transformada rápida de Fourier / Fast Fourier Transform
RGB	Red, Green, Blue
RMSE	Raiz Quadrada do Erro-Médio / Root Mean Squared Error
SVM	Support Vector Machine
SVR	Support Vector Regression
FCN	Rede Neural Totalmente Conectada / Fully Connected Neural-network
ReLU	Unidade Linear retificada / Rectified Linear Unit
LR	Taxa de aprendizagem / Learning Rate
Beta 1	Termo de empuxo
HL	Camadas Ocultas / Hidden Layers
W Adv	Pesos de otimização adversária / Adversarial Weight
W Con	Pesos de otimização contextual / Contextual Weight
W Enc	Pesos de otimização codificadora / Encoder Weight
W Lat	Pesos de otimização latente / Latent Weight
BS	Tamanho do lote / Batch Size

AD	Aumento de dados / Data Augmentation
AUC	Área sob a curva / Area Under Curve

LISTA DE SÍMBOLOS

λ	Peso da métrica de detecção de anomalias
λ_{adv}	Pesos de otimização adversarial
λ_{con}	Pesos de otimização contextual
λ_{lat}	Pesos de otimização latente
λ_{enc}	Pesos de otimização do codificador

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Em nosso cotidiano, convivemos com diversos sons à nossa volta. São tipos de sons como vozes, músicas, tráfegos ou equipamentos e maquinários, e os utilizamos para comunicação, senso de perigo, diversão ou entendimento do ambiente que estamos. Durante décadas, pesquisas foram realizadas sobre como os seres humanos são capazes de entender os diversos sons do ambiente e como possibilitar que sistemas computacionais sejam capazes de capturar e processar esses sons apropriadamente (ZÖLZER, 2008). Tal atividade avançou significativamente com o emprego de técnicas de aprendizado de máquina, consequência também da melhoria de arquiteturas computacionais de CPUs e GPUs (RONG, 2016).

Informações extraídas de áudio têm servido para a construção de diversas aplicações relevantes, por exemplo: monitoramento urbano ou doméstico (DIMITROV et al., 2014; VACHER; SERIGNAT; CHAILLOL, 2007), diagnóstico de máquinas (DUMAN; BAYRAM; İNCE, 2013), reconhecimento de fala (Xiong et al., 2018) e detecção de cenas (Mesaros; Heittola; Virtanen, 2016). As técnicas utilizadas nessas aplicações incluem a análise de cenas acústicas para identificar lugares, situações ou atividades humanas e a análise de eventos acústicos que incluem pegadas, gritos, buzinas, trânsito dentre outros.

Cenas e eventos acústicos (atividades sonóras) são definidos precisamente por Imoto (2018). A cena acústica possui um intervalo de tempo de alguns segundos até 10 segundos, geralmente dispõe de rótulo do local que ocorreu a gravação (exemplo: ônibus, carro, parque ou casa), a situação (exemplo: durante uma reunião ou aula) e a atividade humana envolvida (exemplo: varrendo, cozinhando, andando ou conversando). O evento acústico compreende o intervalo de tempo de apenas alguns milissegundos até poucos segundos e representam um tipo específico de som, como acender um isqueiro, ventiladores ou água corrente. As cenas possuem múltiplos eventos que podem se sobrepor ao longo do tempo.

Muitas aplicações de Aprendizado de Máquina para processamento de áudio são baseadas em técnicas de aprendizagem supervisionada (VIRTANEN; PLUMBLEY; ELLIS, 2018), em que cada áudio possui um rótulo correspondente (BISHOP, 2006), cabendo aos modelos inferirem seu tipo baseada ma experiência provida pelos dados de treino. No entanto, poucos estudos são realizados na importante tarefa de detecção de anomalias em áudio, em que deseja-se identificar um evento distinto a partir de classes de eventos aprendidos durante a fase de treinamento (HABEEB et al., 2019). Nesses casos, precisamos adotar técnicas diferentes do aprendizado supervisionado, pois dados normais são comuns e os eventos de interesse(anômalos) são raros. Muitas aplicações podem compartilhar dessa situação, como monitoramento de vias públicas e auto estradas, detectar crimes

em transporte público ou monitorar violência em locais públicos como praças e parques (Foggia et al., 2016).

Infelizmente, existem poucos conjuntos de dados sonoros de larga escala para detecção de anomalias. Isso se deve, em grande parte, à raridade dos sons anômalos e à falta de definições claras, o que resulta em incertezas. Neste cenário, a demanda por inspeções automáticas em maquinários industriais tem aumentado devido à necessidade de melhorar a qualidade de sua manutenção. Atualmente, a identificação de problemas nessas máquinas depende de especialistas, o que eleva os custos de produção. (KOIZUMI et al., 2019).

A avaliação da tarefa de detecção de anomalias é majoritariamente feita utilizando a métrica AUC (Área Sob a Curva). Ela mede a capacidade do modelo em distinguir corretamente entre instâncias típicas e anômalas (FAWCETT, 2006). Quanto maior o valor da AUC, melhor é o desempenho do modelo. É uma métrica especialmente útil quando as classes estão desbalanceadas, ou seja, quando a quantidade de exemplos anômalos é muito menor em relação aos exemplos normais. Portanto, a métrica AUC desempenha um papel fundamental na avaliação e comparação de diferentes abordagens de detecção de anomalias, fornecendo uma medida objetiva e abrangente do desempenho dos modelos (KOIZUMI; KAWAGUCHI; IMOTO, 2020).

Na área de detecção de anomalias em eventos sonoros, os autocodificadores (AE) têm sido amplamente utilizados como uma abordagem eficaz. Os autocodificadores são redes neurais capazes de aprender uma representação compacta e de alta qualidade dos dados de entrada (SCHMIDHUBER, 2015). Nesse contexto, eles são projetados para reconstruir a entrada original com mínima perda de informação. Ao treinar um autocodificador em um conjunto de dados normal de eventos sonoros, ele aprende a reconstruir com precisão esses padrões normais (XU et al., 2021a).

Proposta por Goodfellow et al. (2014), as GANs (Generative Adversarial Networks) são modelos compostos por duas redes neurais, a discriminadora e a geradora, a primeira possui a tarefa de distinguir dados reais de falsos, enquanto a segunda aprende a criar dados sintéticos próximos dos reais que impedem a rede discriminadora de distingui-la como falsas. O uso desse tipo de rede se iniciou para criar imagens originais e muito próximas da realidade. Entretanto, a característica adversária de treinamento possibilitou o avanço de sua utilização em outras tarefas e, em particular, na detecção de anomalias. A ideia básica é que se um modelo é treinado para distinguir algo real de algo falso, ele pode servir para identificar se um dado é típico ou não, no último caso, uma anomalia.

Na literatura, encontramos diversos modelos GANs sendo propostos e avaliados para a detecção de anomalias, mas tendo como alvo de estudo os dados de imagens (LIU et al., 2021; AKCAY; ABARGHOUEI; BRECKON, 2019; SCHLEGL et al., 2019; AKCAY; ATAPOUR-ABARGHOUEI; BRECKON, 2018; SCHLEGL et al., 2017). Especula-se

que tais modelos sirvam para dados de áudio também, sendo um aspecto comentado por alguns desses autores. No entanto, tais trabalhos não apresentam como adequar os modelos propostos a esse novo tipo de mídia, tampouco apresentam métricas avaliativas de referência.

Diante do exposto, identificou-se a necessidade de investigação de formas de adequação de modelos GANs para aplicação de detecção de anomalias em áudio, bem como da avaliação de métricas de referência desse tipo de aplicação, sendo esses , lacunas motivadoras para realização deste trabalho. Com o intuito de realizar uma análise comparativa, este trabalho não apenas adaptou os principais modelos de detecção de anomalias da literatura, mas também os comparou com outras abordagens existentes. Para isso, forão utilizados conjuntos de dados populares na área, como o *DCASE*, *Urban Sound* e *AudioSet*, que foram amplamente utilizados em competições e, portanto, proporcionaram uma base sólida para a avaliação e comparação dos modelos.

Os resultados desse trabalho servem de base para aplicações inteligentes que colaboram para reconhecer o comportamento suspeito em contextos tais como: identificação de falha de maquinário industrial (KOIZUMI et al., 2019; PUROHIT et al., 2019), detecção de acidentes rodoviários (ROVETTA; MNASRI; MASULLI, 2020), abordagens multimodais com vídeo (KITTLER et al., 2018), identificação de falha em transmissão de sinal 5G (ZHOU et al., 2021), e etc.

1.2 HIPÓTESES

Trabalhos mais recentes mostram que modelos GANs atingiram o estado-da-arte na detecção de anomalias no domínio de imagens. Desta forma, conjectura-se que é possível aplicar a mesma abordagem para áudio, por meio de adaptações e otimizações na arquitetura e funções de perda para tornar esses modelos eficientes no novo contexto considerado.

1.3 OBJETIVOS

1.3.1 **Geral**

– Desenvolver um método eficaz e eficiente de detecção de anomalias em eventos acústicos com o aprendizado não-supervisionado, empregando arquiteturas GANs combinadas com Autocodificadores, bem como, demonstrar através da avaliação de métricas a eficácia do método proposto em conjunto de dados do mundo real.

1.3.2 **Específicos**

- Adaptar e avaliar comparativamente arquiteturas GANs combinadas com AE da literatura para o problema de detecção de anomalias em eventos acústicos, usando o paradigma de aprendizagem não-supervisionado.

- Otimizar hiper-parâmetros destas arquiteturas adaptadas para o conjunto de dados de anomalia em áudios.

- Identificar e selecionar técnicas de aumento de dados que possam ser aplicadas para melhorar a eficiência.

1.4 ORGANIZAÇÃO DO TRABALHO

Esta proposta de dissertação contém quatro capítulos posteriores, organizados como segue:

- **Capítulo 2:** Fundamentação Teórica – Apresenta o referencial teórico sobre os conceitos de detecção de anomalias, pré-processamento de áudio, redes neurais convolucionais, redes autocodificadoras e redes GANs;
- **Capítulo 3:** Trabalhos correlatos – Descreve os principais trabalhos relacionados ao contexto desta pesquisa;
- **Capítulo 4:** Solução Proposta – Apresenta os materiais e métodos utilizados no desenvolvimento da abordagem de detecção de anomalias em eventos sonoros aqui proposta;
- **Capítulo 5:** Experimentos e Resultados – Descreve os *baselines*, métricas de avaliação utilizadas e seus resultados.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção possui a finalidade de apresentar uma visão geral bibliográfica para embasamento teórico deste trabalho, a qual contempla os seguintes conceitos: detecção de anomalias, pré-processamento de áudio, redes neurais convolucionais, redes autocodificadoras e por fim redes GANs.

2.1 DETECÇÃO DE ANOMALIAS

Denomina-se anomalia um determinado elemento ou subconjunto de elementos de um conjunto de dados, quer seja da mesma natureza ou não, que é significativamente diferente do restante do conjunto (AGGARWAL, 2012). Anomalias se apresentam em qualquer tipo de representação ou domínio de dados, sejam eles estruturados como planilhas e banco de dados ou não estruturados como imagens, sons e textos (CHANDOLA; BANERJEE; KUMAR, 2009).

A Figura 2.1 ilustra a presença de anomalias em um conjunto de dados de 2 dimensões. Essa amostragem de dados possui duas regiões normais chamadas N_1 e N_2, uma vez que a maioria dos dados se encontram nestas duas regiões. Os pontos o_1, o_2 e a região O_3, que se encontram distante destas regiões principais, são classificados como anomalias.

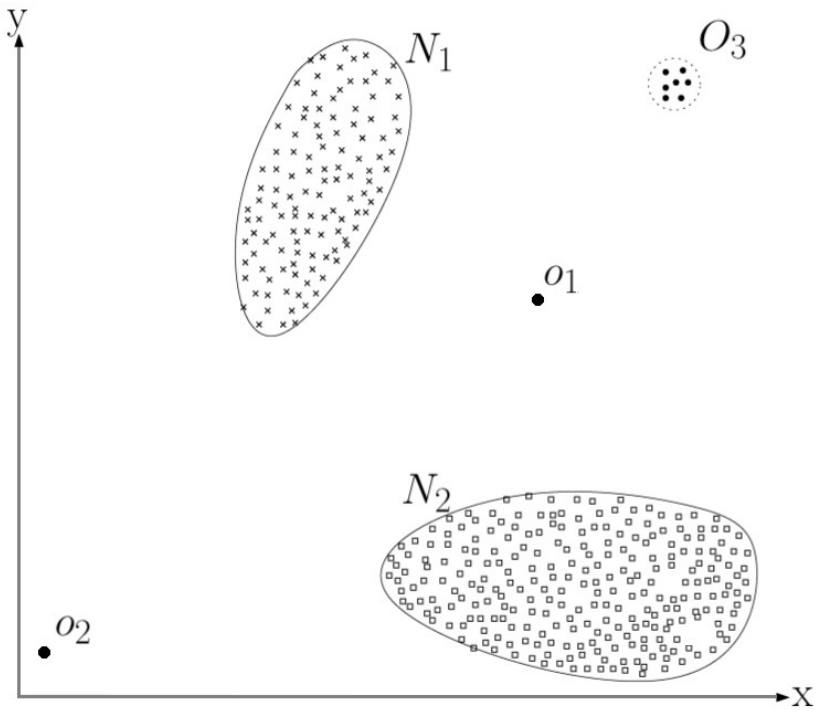


Figura 2.1 – Exemplo de anomalias em dados de 2 dimensões. Fonte (CHANDOLA; BANERJEE; KUMAR, 2009)

Um conjunto de dados pode conter anomalias por uma série de motivos, sejam eles propositais ou não. A existência de fraudes bancárias, atividade terrorista ou invasões de sistemas faz com que a detecção de anomalias se torne uma tarefa relevante. O objetivo principal de um algoritmo responsável por detectar anomalias é definir qual o limiar que distingue um dado típico ao conjunto de um dado anômalo também pertencente ao conjunto (CHANDOLA; BANERJEE; KUMAR, 2009).

A detecção de anomalias está relacionado à outras áreas como remoção de ruídos e detecção de novidade (*novelty detection*), entretanto, são tarefas distintas. Na remoção de ruídos o objetivo é remover dados não desejados para análises. A detecção de novidades possui o intuito de adicionar dados não observados anteriormente ao conjunto (Antonini et al., 2018).

A detecção de anomalias é amplamente abordada por meio da classificação binária, que envolve a distinção entre exemplos normais e anômalos. Essa abordagem se baseia na premissa de que os dados anômalos diferem significativamente dos dados normais (SURI; M; ATHITHAN, 2019). A detecção de anomalias pode ser aplicada em diversos tipos de dados, como imagens, textos ou áudios, nos quais existe uma variedade de particularidades que podem desviar-se das classes existentes. Portanto, a classificação binária desempenha um papel crucial na identificação e separação de exemplos anômalos dos normais, permitindo uma análise mais precisa e eficiente dessas ocorrências incomuns. A Figura 2.2 demonstra a utilização de um sistema de detecção de anomalias em áudio.

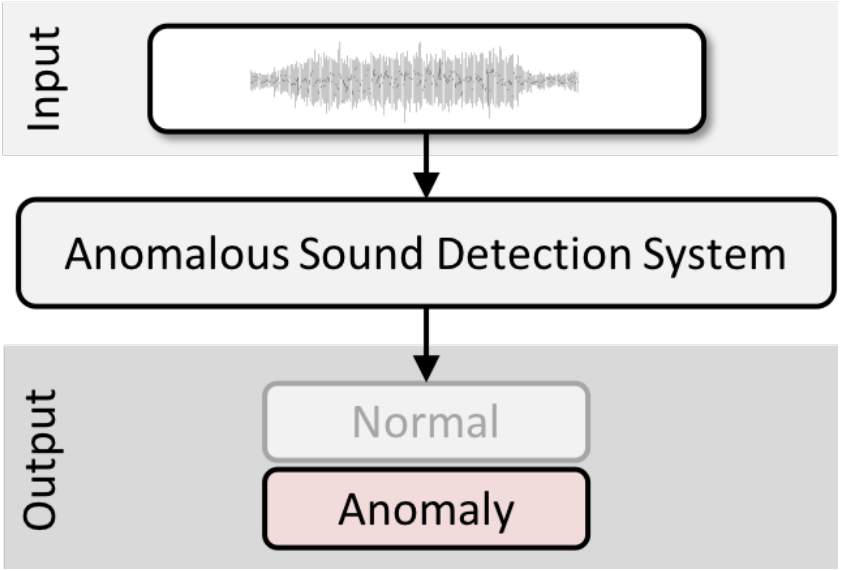


Figura 2.2 – Sistema de detecção de anomalias, Fonte: DCASE

O paradigma não supervisionado na detecção de anomalias em áudio é uma abordagem promissora que visa identificar padrões anômalos sem a necessidade de exemplos rotulados. Nesse contexto, algoritmos não supervisionados, como autocodificadores e redes generativas adversariais (GANs), são amplamente utilizados. Os autocodificadores são

capazes de aprender representações compactas dos dados de áudio típicos, e, quando apresentados a um evento sonoro anômalo, falham em reconstruí-lo de forma satisfatória. Essa diferença entre a reconstrução esperada e a realidade é explorada para identificar as anomalias. Já as GANs procuram gerar dados de áudio que se assemelhem aos normais e, assim, qualquer exemplo que não possa ser bem reproduzido pela rede geradora é considerado uma anomalia. Com esses métodos não supervisionados, é possível detectar anomalias em áudio de forma automática e sem a necessidade de rótulos, permitindo uma análise mais eficiente em cenários onde os dados anômalos podem ser escassos ou desconhecidos (SURI; M; ATHITHAN, 2019).

2.2 EXTRAÇÃO DE CARACTERÍSTICAS E PRÉ-PROCESSAMENTO DE ÁUDIO

Para os algoritmos de Aprendizagem de Máquina efetuarem treinamentos em domínios complexos, como áudio, são indicados o uso de pré-processamentos e/ou extração de características acústicas (VIRTANEN; PLUMBLEY; ELLIS, 2018). Entretanto tal pré-processamento é opcional e há variância de acordo com a atividade a ser realizada e adversidades na ocasião da captura.

A extração de características acústicas é uma maneira compacta de descrever um áudio, evidenciando características mais importantes e de interesse (SURI; M; ATHITHAN, 2019). Essas características podem ser de tempo ou frequência, em que cada tipo evidencia determinados atributos, como, por exemplo, amplitude do sinal, harmônica e espectrogramas. Essa extração possibilita diversas aplicações; predição de emoção (Nawaz et al., 2018), verificação automática de auto-falante (NETO; FIGUEIREDO, 2019; TODISCO et al., 2018) e classificação de gênero musical (NANNI et al., 2016; Sarkar; Saha, 2015).

2.2.1 Transformada de Fourier

A transformada de Fourier é uma técnica fundamental na extração de características e pré-processamento de áudio. Ela permite decompor um sinal de áudio em suas componentes de frequência, revelando as diferentes frequências que compõem o sinal. Ao aplicar a transformada de Fourier em um sinal de áudio, obtém-se o espectro de frequência, que mostra a intensidade das diferentes frequências presentes no sinal. Essa informação é valiosa para a análise e caracterização do áudio, permitindo identificar características como tons, harmônicos, ruídos e outros padrões sonoros. Além disso, a transformada de Fourier também é utilizada no pré-processamento de áudio, por exemplo, para remover ou atenuar frequências indesejadas, como ruídos ou interferências, ou para realizar a compressão de áudio, reduzindo a quantidade de dados necessários para representar o sinal (LATHI, 2007).

2.2.2 Transformada de Fourier de tempo curto

A transformada de Fourier de tempo curto, também conhecida como STFT (*Short-time Fourier Transform*), é uma extensão da transformada de Fourier que permite analisar a variação espectral do sinal de áudio ao longo do tempo com maior precisão. Ao contrário da transformada de Fourier convencional, que considera o sinal de áudio como um todo, a STFT divide o sinal em pequenos segmentos de tempo e, em cada segmento, calcula a transformada de Fourier (ZHAO et al., 2015). Isso permite capturar informações temporais sobre as frequências presentes no sinal. O resultado da STFT é um espectrograma, que exibe as mudanças espectrais do sinal ao longo do tempo de forma detalhada (CHACHADA; KUO, 2014). Essa técnica é amplamente utilizada em áreas como processamento de fala (Hershey et al., 2017), análise de música e reconhecimento de padrões em áudio (ESPI et al., 2015), proporcionando uma representação visual e quantitativa das características espectrais do sinal em diferentes instantes temporais.

No pré-processamento de áudio, a transformada de Fourier de tempo curto desempenha um papel importante na aplicação de técnicas de filtragem e modificação espectral. Por exemplo, é possível aplicar janelas deslizantes no sinal de áudio antes de calcular a STFT, o que permite atenuar transições bruscas no espectro e reduzir artefatos indesejados na análise. Além disso, a STFT também é utilizada em técnicas de remoção de ruído, separação de fontes sonoras e compressão de áudio. Ao dividir o sinal em segmentos menores, é possível atenuar ou remover componentes indesejadas, como ruídos ou interferências, em frequências específicas, preservando as informações relevantes do sinal. Portanto, a transformada de Fourier de tempo curto desempenha um papel crucial no pré-processamento de áudio, permitindo a manipulação e aprimoramento das características espectrais do sinal com maior precisão e adaptabilidade às características temporais presentes nos eventos sonoros (ZHAO et al., 2015). A Equação de $x[n]$ é definida como 2.1.

$$X(m,k) = \sum_{m} x[n]W[m-n]\exp(-j\frac{2\pi nk}{N}) \tag{2.1}$$

em que m é o número de amostras do tempo, k é o número de amostras da frequência, $W[n]$ é o tamanho da janela e N é o número de amostras de frequências totais. Assume-se que $W[n]$ é uma janela de tempo fixo.

2.2.3 Espectrograma

O espectrograma é uma representação visual do espectro de frequência de um sinal de áudio ao longo do tempo. Ele divide o sinal de áudio em pequenos segmentos e calcula a transformada de Fourier de cada segmento, mostrando a evolução das frequências ao longo do tempo. Essa representação permite analisar a variação espectral do sinal em diferentes momentos, sendo muito útil na análise de eventos sonoros complexos, como

fala, música e sons ambientais. O espectrograma é amplamente utilizado em tarefas como reconhecimento de fala, separação de fontes sonoras e detecção de eventos sonoros. Ao fornecer informações detalhadas sobre a distribuição espectral do áudio ao longo do tempo, a transformada de Fourier e o espectrograma desempenham um papel fundamental na extração de características e no pré-processamento de áudio (CHACHADA; KUO, 2014).

A Figura 2.3 demonstra o sinal acústico puro logo após a leitura do áudio e na Figura 2.4 o respectivo espectrograma linear. Nesse processamento utiliza-se diferentes cores para apontar a intensidade do sinal, variando do violeta ao vermelho. No espectrograma a frequência é mostrada no eixo vertical e o tempo no eixo horizontal.

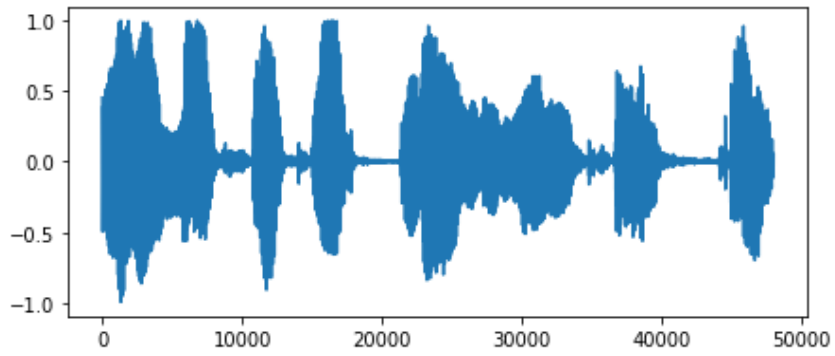


Figura 2.3 – Sinal de áudio puro. Fonte: PRÓPRIO AUTOR

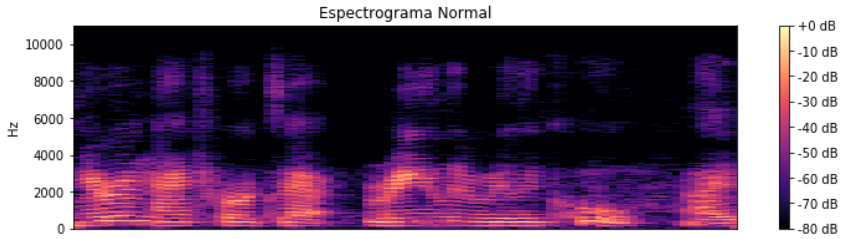


Figura 2.4 – Espectrograma do sinal do áudio. Fonte: PRÓPRIO AUTOR

2.3 AUMENTO DE DADOS

A técnica de aumento de dados pode ser utilizada em diferentes aplicações de Aprendizado de Máquina, principalmente naquelas que o número de exemplos para treinamento é baixo. Consiste em aplicar transformações e manipulações nos dados existentes, gerando novas amostras que são semelhantes às originais, porém com variações controladas (ZEIMARANI; COSTA, 2019).

O aumento de dados tem o objetivo de aumentar a diversidade do conjunto de treinamento, fornecendo ao modelo mais exemplos para aprender e, assim, melhorar sua capacidade de generalização para dados de teste. No entanto, é importante selecionar as transformações adequadas para cada tipo de dado e tarefa, evitando distorções excessivas

ou introdução de artefatos indesejados que possam comprometer a qualidade e a integridade dos dados gerados (PARK et al., 2019).

Por exemplo, no caso de dados de imagem, é importante preservar a estrutura e a semântica das imagens originais durante as transformações. Para dados de áudio, é necessário considerar a natureza temporal e espectral do som, garantindo que as manipulações preservem as características relevantes para a tarefa, como timbre e ritmo. Além disso, é fundamental ter cuidado ao aumentar dados com anomalias, pois é necessário preservar a natureza única desses exemplos durante as transformações (AGGARWAL, 2012).

2.3.1 Aumento de dados no domínio de áudio

Assim como em outros tipos de dados, o aumento de dados em áudio envolve a aplicação de transformações e manipulações nos exemplos existentes, gerando novas amostras que são semelhantes, porém com variações controladas. No contexto de áudio, as transformações podem incluir mudanças de velocidade (KO et al., 2015), ajustes de pitch (PARK et al., 2019), adição de ruído (HANNUN et al., 2014), variações de volume (KO et al., 2015) e simulações de reverberação (KANDA; TAKEDA; OBUCHI, 2013), entre outras (XIA et al., 2019).

A escolha adequada dessas transformações é fundamental para preservar as características importantes do áudio, como a qualidade do som, o timbre e o ritmo. Além disso, é importante considerar as particularidades das tarefas relacionadas ao áudio, como reconhecimento de fala, classificação de gênero musical ou detecção de eventos sonoros, para garantir que as transformações sejam relevantes para o problema em questão (PARK et al., 2019).

A Figura 2.5 apresenta o efeito do aumento de dados no sinal do áudio. Observa-se que o ganho de potência randômica no sinal é o único que altera a escala dos valores do áudio.

Para melhorar a visualização dos efeitos do aumento de dados nos conjuntos de áudio, a Figura 2.6 exibe o espectrograma dos áudios preexistente e suas modificações. A adição de ruído deforma o espectrograma, valorando de maneira randômica certas frequências.

2.4 *AUTO ENCODER - AE*

Os autocodificadores são estruturas arquiteturais empregadas com o propósito de gerar representações comprimidas do dado de entrada (SCHMIDHUBER, 2015). Esse processo é baseado em aprendizado não supervisionado, no qual a presença de rótulos ou anotações não é requerida. Essa arquitetura é composta por duas redes neurais artificiais: a rede codificadora, denotada por E, e a rede decodificadora, representada por D. A rede

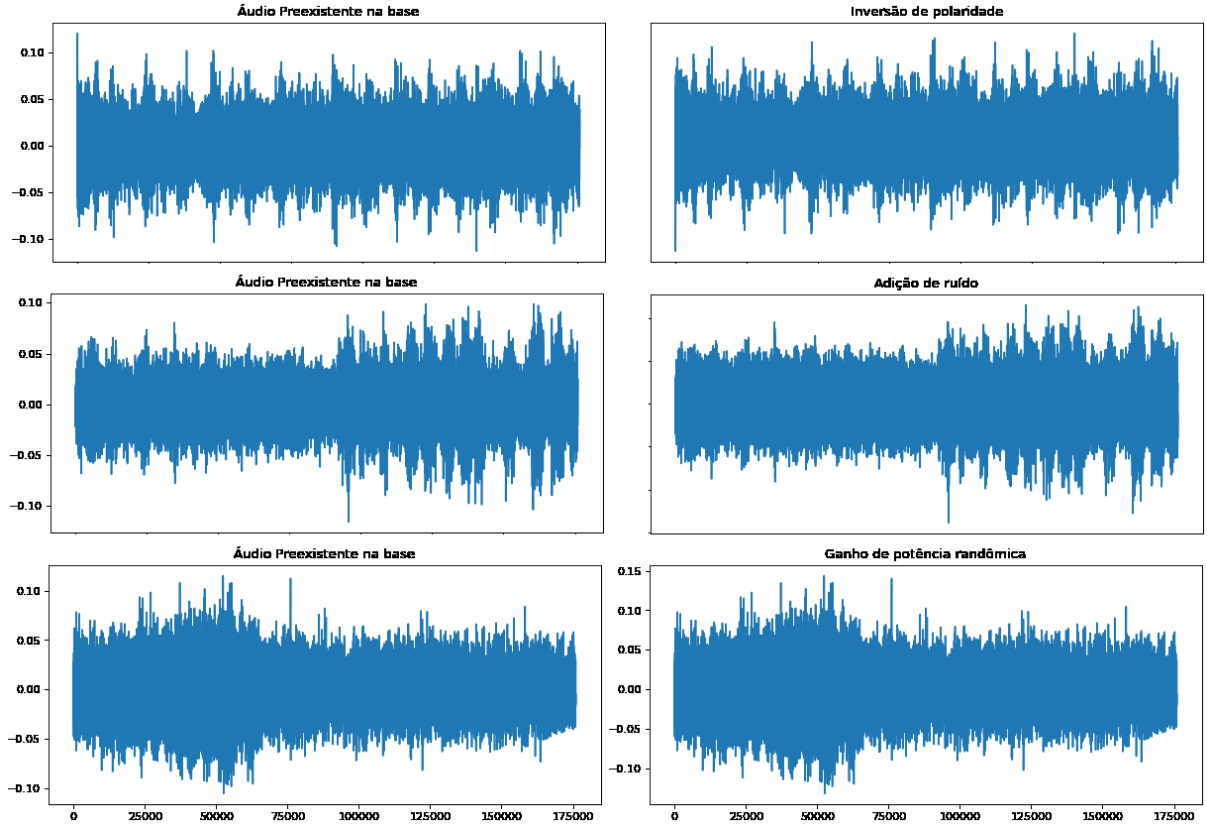


Figura 2.5 – Aumento de dados em sinais de áudio. Fonte: PRÓPRIO AUTOR

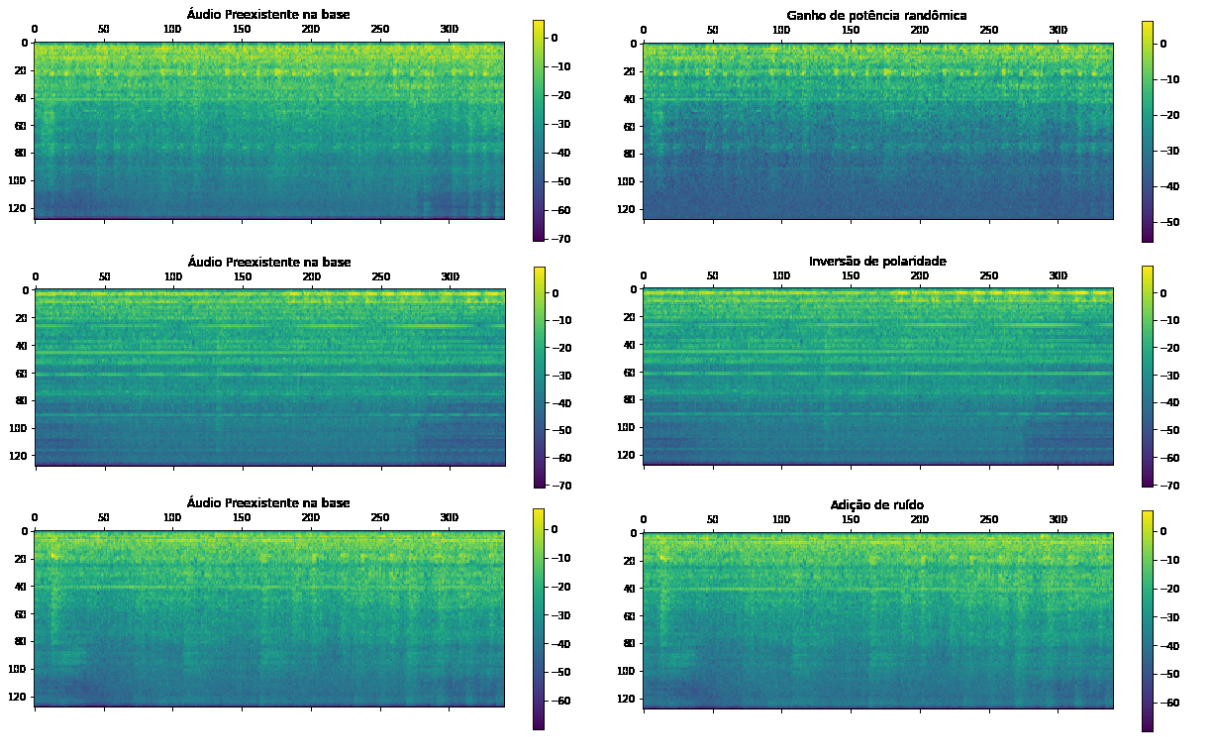


Figura 2.6 – Aumento de dados em espectrogramas do áudio. Fonte: PRÓPRIO AUTOR

codificadora E tem a função de aprender representações latentes da instância de entrada,

enquanto a rede decodificadora D utiliza esse espaço latente para reconstruir a instância original de entrada. Dessa forma, os autocodificadores visam capturar as características mais relevantes e compactas do dado, permitindo uma representação reduzida, mas informativa. A Figura 2.7 ilustra a arquitetura de uma rede AE simples que utilizou o processo de aprendizagem para representar espectrogramas de áudios.

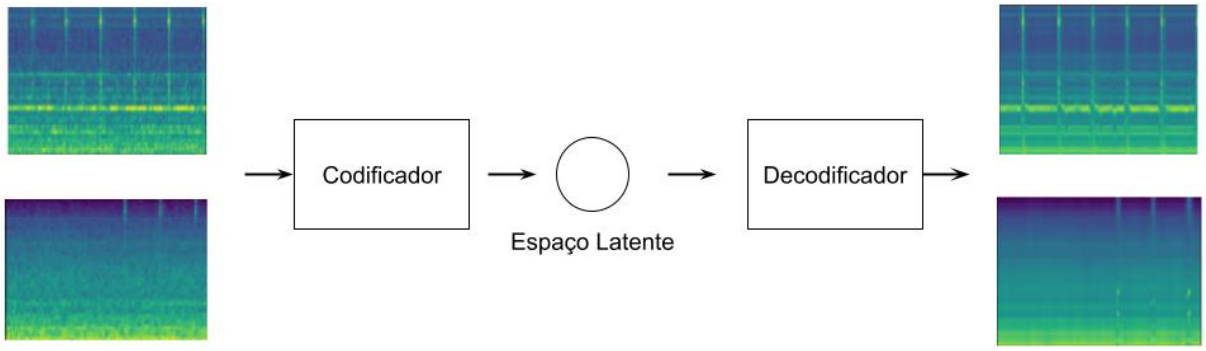


Figura 2.7 – Representação de um modelo Autocodificador simples. As imagens reais (à esquerda) e as geradas via aprendizagem (à direita). Fonte: PRÓPRIO AUTOR

Existem diversas variações de AEs por sua alta qualidade em representação de dados, alta capacidade de remoção de ruídos (ERASLAN et al., 2019; MENG et al., 2018), gerar imagens realísticas (PARMAR et al., 2021), criar novos conjuntos de dados (ISLAM et al., 2021) e também para detecção de anomalias (XU et al., 2021a; NGUYEN et al., 2021; DEEPAK; CHANDRAKALA; MOHAN, 2020).

Redes AEs são altamente eficazes em compactar informações e reconstruir os dados de treinamento, ao passo que a saída tende a ficar mais fidedigna a entrada durante o processo de treinamento visando a minimação do erro de reconstrução. Por conta de sua natureza representativa não supervisionada, esta arquitetura é amplamente utilizada em tarefas de detecção de anomalias. A utilização desta arquitetura para anomalias se dá nos seguintes passos:

1. Durante o treinamento, instâncias de treino contendo unicamente dados típicos são fornecidos à rede AE;

2. Cada camada da rede E irá aprender representações latentes cada vez mais abstratas, de modo que, a saída desta rede possuirá dimensionalidade consideravelmente menor em comparação ao dado de entrada;

3. A rede D geralmente apresenta arquitetura de camadas dispostas em ordem reversa da rede E, tal que, a medida que interpreta o espaço latente como entrada, aumenta o nível de complexidade deste dado;

4. A saída fornecida pela arquitetura AE é reconstrução da instância de entrada típica;

5. Uma instância de áudio anômalo possui características diferentes de um áudio típico. Por isso, a arquitetura AE enfrentará problemas durante a reconstrução deste tipo de entrada;

6. Por fim, é aplicado uma análise e comparação entre o áudio anômalo e sua reconstrução. Uma arquitetura bem treinada apresentará diversas divergências nesta comparação, e então, é possível identificar a anomalia através do erro de reconstrução deste áudio (XU et al., 2021a; DEEPAK; CHANDRAKALA; MOHAN, 2020).

2.4.1 U-NET

A arquitetura de rede neural U-Net é amplamente utilizada em tarefas de segmentação de imagens. Ela recebe esse nome devido à sua forma característica em U, que é composta por um caminho de contração (codificadora) e um caminho de expansão (decodificadora). O caminho de contração é responsável por capturar características de alto nível da imagem através de camadas convolucionais, reduzindo gradualmente a resolução espacial. Essas características são então transmitidas para o caminho de expansão, onde são combinadas com as informações de baixo nível para gerar uma saída segmentada de alta resolução (RONNEBERGER; FISCHER; BROX, 2015). Podemos observar a arquitetura U-Net na Figura 2.8.

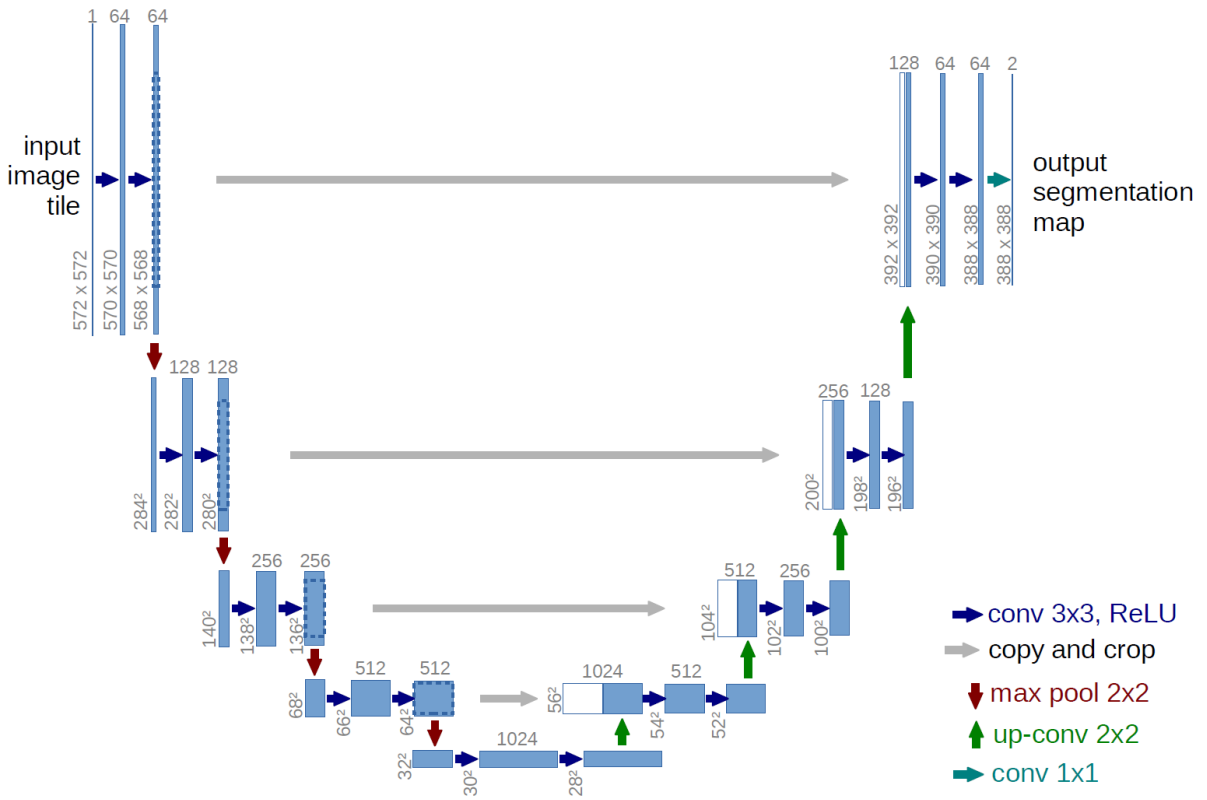


Figura 2.8 – Representação da rede U-Net. Fonte: (RONNEBERGER; FISCHER; BROX, 2015)

A U-Net é conhecida por sua capacidade de lidar com problemas de segmentação em imagens médicas, como a segmentação de órgãos ou lesões. Essa arquitetura permite que a rede capture detalhes finos e preserve a informação espacial, graças à conexão de atalho entre as camadas de contração e expansão. Essas conexões de atalho permitem que as informações de níveis mais altos sejam transmitidas diretamente para as camadas de expansão, auxiliando na reconstrução precisa da imagem segmentada (RONNEBERGER; FISCHER; BROX, 2015).

Uma das principais vantagens da arquitetura U-Net é sua flexibilidade e adaptabilidade. Ela pode ser facilmente modificada e ajustada de acordo com a natureza do problema em questão. Além disso, a U-Net é capaz de lidar com conjuntos de dados de diferentes tamanhos, pois suas camadas convolucionais são projetadas para serem independentes da resolução espacial (TRUONG et al., 2021).

2.5 *GENERATIVE ADVERSARIAL NETWORK* - GAN

A abordagem GAN foi concebida por Goodfellow et al. (2014), sua criação aprimorou a habilidade de redes neurais de gerar dados verossímeis. As GANs são um método de geração de dados realísticos, entretanto, seus resultados e abordagem de utilizar duas redes neurais em conjunto o destacaram.

As redes GANs são uma abordagem capaz gerar dados realísticos utilizando duas redes neurais distintas: um modelo Gerador G, apto a absorver a distribuição de dados e um modelo Discriminador D que estima a probabilidade de uma amostra ter ou não sido gerada por G. Essa abordagem inspirada é inspirada na Teoria dos Jogos (ou equilíbrio de Nash), onde em um jogo envolvendo dois jogadores, nenhum jogador tem a ganhar mudando sua estratégia unilateralmente. O objetivo da rede G é de gerar exemplos falsos indistinguíveis de dados reais, geralmente através de um vetor de ruídos aleatórios. O objetivo do Discriminador é determinar corretamente quando um exemplo é real.

Durante o processo de treinamento, a rede geradora (G) progressivamente gera dados mais semelhantes aos dados reais. Esse treinamento ocorre de forma indireta, por meio da resposta fornecida pelas predições da rede discriminadora (D). A cada vez que a rede discriminadora classifica uma imagem gerada pela rede geradora como real, a rede geradora recebe uma resposta positiva, indicando que está produzindo imagens mais próximas das reais. Da mesma forma, a rede discriminadora recebe uma resposta positiva quando classifica corretamente uma imagem gerada pela rede geradora como falsa, o que indica que ela está corretamente identificando as imagens geradas. Esse processo iterativo de respostas permite que tanto a rede geradora quanto a rede discriminadora aprimorem suas habilidades e aprendam a gerar e distinguir, respectivamente, dados mais realistas ao longo do tempo. A Figura 2.9 demonstra um exemplo de rede GAN criando um exemplo de áudio através de um vetor de ruído.

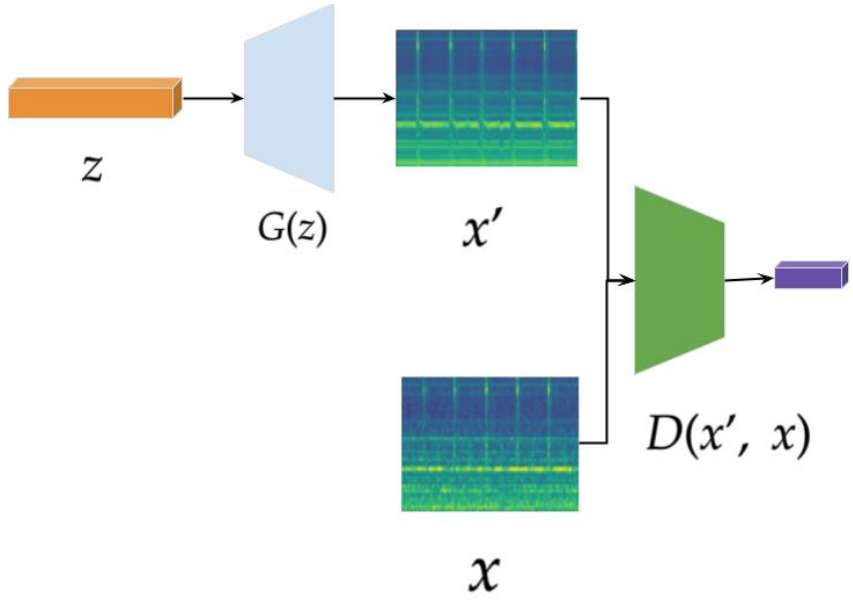


Figura 2.9 – Representação de um modelo gerador GAN simples. A imagem $x\grave{t}$ foi criada pelo Gerador G utilizando o vetor de ruído z, em seguida, o Discriminador D identifica quais os conjuntos reais e sintéticos. Fonte: PRÓPRIO AUTOR

Conforme mencionado anteriormente, o Gerador cria imagens através de uma distribuição de ruído aleatório $P_z(z)$, então representamos o mapeamento deste vetor para uma imagem específica $G(z)$, onde G é uma função diferenciável representada por uma rede neural artificial. Definimos também uma segunda rede neural artificial como $D(x)$ que produz como saída um escalar (probabilidade do dado x ser real). O objetivo do treinamento de D é maximizar a probabilidade de atribuir o rótulo correto a ambos os exemplos de treinamentos e amostras geradas por G. Simultaneamente treinamos G para minimizar $log(1 - D(G(z)))$. Portanto, D e G jogam minmax de dois jogares com função de valor $V(G, D)$, conforme mostra a Equação 2.2.

$$\min_G \max_D E_{x\sim p_{\text{data}}(x)}[\log D(x)] + E_{z\sim p_z(z)}[1 - \log D(G(z))] \tag{2.2}$$

À medida que o Gerador se torna melhor em produzir imagens realística, o Discriminador melhora a sua capacidade em distinguir exemplos artificiais de reais. As GANs possuem a analogia de um falsificador de dinheiro e a polícia, sendo o Gerador representado pelo falsificador e o Discriminador representado pela polícia. O falsificador cria e faz uso de um lote de notas falsas, se as notas falsas forem detectadas pela polícia, o falsificador sabe que precisa melhorar na falsificação. Desta maneira, uma nota detectado como falsa é gerada novamente, sem saber exatamente onde melhorar. De forma similar, a polícia precisa continuar a melhorar. Portanto, cada vez que a polícia erra em identificar as notas, novas técnicas são adicionadas ao portfólio de análise forense. Como resultado, a polícia reconhece notas falsas de alta qualidade que antes não eram identificadas. Por fim, o falsificador se aprimora a cada novo ciclo, e este ciclo se repete até que ambos não tenham

o que melhorar (GOODFELLOW et al., 2014; SOUZA, 2020).

As GANs podem ser empregadas para gerar dados sintéticos que se assemelham aos dados típicos de um determinado domínio. Ao treinar a GAN com uma grande quantidade de dados típicos, o gerador aprende a capturar a distribuição desses dados e é capaz de sintetizar amostras que seguem essa distribuição. Em seguida, durante o estágio de teste, as amostras geradas são comparadas com os dados reais observados. Caso uma amostra sintética seja considerada uma anomalia, indica-se que os dados de entrada correspondentes também são anômalos. Dessa forma, as GANs podem fornecer uma abordagem eficaz para a detecção de anomalias, permitindo identificar padrões incomuns e desvios significativos em relação aos dados típicos (SOUZA, 2020).

Além disso, as redes GANs também podem ser utilizadas para detecção de anomalias por meio do treinamento de um discriminador específico para identificar anomalias nos dados. Nesse caso, a GAN é treinada com uma combinação de dados típicos e anômalos. O discriminador é projetado para distinguir entre as amostras típicos e as amostras anômalas, enquanto o gerador é responsável por gerar dados que possam enganar o discriminador e serem classificados como típicos. O objetivo do treinamento é fazer com que o discriminador seja altamente preciso na classificação de anomalias. Durante a fase de teste, o discriminador é usado para avaliar a anomalia dos dados de entrada, classificando-os como típicos ou anômalos com base na sua resposta. Assim, as redes GANs oferecem uma abordagem inovadora e eficiente para a detecção de anomalias, combinando o poder de geração de dados sintéticos com a capacidade de discriminação entre padrões típicos e anômalos (BIAN et al., 2019).

As redes GANs se tornaram muito famosas por resultados notáveis em tarefas como gerar imagens com muita qualidade (KARRAS et al., 2020), converter imagens de desenhos em imagens reais (ISOLA et al., 2017), aplicar o estilo de uma fotografia em outro contexto diferente (XU et al., 2021b) e identificar anomalias em imagens (LIU; XU, 2020; AKCAY; ATAPOUR-ABARGHOUEI; BRECKON, 2018; BIAN et al., 2019). Todas essas tarefas são possíveis de serem executadas de maneira não supervisionada, ou seja, não precisam de nenhuma outra informação que não as próprias imagens.

3 TRABALHOS CORRELATOS

Desde discussões antigas sobre detecção de anomalias utilizando Aprendizado de Máquia (ROUSSEEUW; DRIESSEN, 1999), muitos trabalhos ampliaram essas análises (BREUNIG et al., 2000; WANG; WONG; MINER, 2004). Diante de conjunto de dados mais significativos (LIU; TING; ZHOU, 2012) e voltados para área de áudio (CONTE et al., 2012), o interesse neste nicho têm crescido.

Abordagens baseadas em Aprendizado de Máquina são capazes de generalizar suas tarefas através do uso de modelos estatísticos ou separação geométrica. Nestes métodos, um especialista cria o modelo que é ajustado de acordo com os dados reais do problema proposto. O especialista analisa os resultados do modelo de modo a ajustar os parâmetros e os limiares para futuramente detectar anomalias. Este capítulo discute trabalhos selecionados focados em detecção e classificação de áudios anômalos.

São três abordagens principais para detecção de áudios anômalos: Utilização de modelos rasos, modelos autocodificadores e modelos geradores. Fornecemos uma breve descrição dos trabalhos representativos de cada tipo nas demais seções.

3.1 ABORDAGEM TRADICIONAL DE APRENDIZADO DE MÁQUINAS

Essa abordagem baseia-se em técnicas rasas de Aprendizagem de Máquina, sejam modelos estatísticos, árvores ou separação geométrica para resolver o problema de detecção de anomalias em atividades sonoras. Essas técnicas geralmente envolvem o cálculo de estatísticas descritivas, como média, desvio padrão e valor máximo, a partir de representações de áudio, como espectrogramas ou outras transformações. Além disso, são menos complexas e computacionalmente mais eficientes do que modelos de aprendizado profundo, mas podem fornecer resultados satisfatórios em cenários de detecção de anomalias em áudio, especialmente quando há restrições de recursos computacionais ou falta de dados anotados disponíveis para treinamento de modelos mais avançados.

Song et al. (2013) apresentou o algoritmo *One-Class Conditional Random Fields* (OCCRF) capaz de identificar anomalias em conjuntos de dados de áudio e sinais de movimentação humana. O autor descreve que o objetivo do algoritmo é computar e diminuir a distribuição de probabilidade condicional, especificando uma margem mínima entre a probabilidade de cada exemplo ser classificado como típico ou anomalia. O modelo é considerado não-supervisionado por não necessitar de rótulos, pois a estratégia de aprendizado aceita que todos ou a maioria dos exemplos de treinamento são típicos. Entretanto, possui limitações de quantidades de características a serem providas como entrada ao modelo, limitando a capacidade de generalização com outros conjuntos de

dados. Além disso, o modelo é altamente sensível a ajuste de parâmetros, resultando em diversas adaptações sempre que utilizar conjuntos de dados diferentes.

Ainda nesse contexto de áudio, o trabalho de Chung et al. (2013), que teve por intento identificar sons anômalos de animais, utilizou o algoritmo *Support Vector Data Description* (SVDD) em vetores de características de áudio. Para isso, desenvolveu um sistema composto dos seguintes passos: calcular o *Mel Frequency Cepstrum Coefficients* (MFCC) do áudio, janelamento de tempo, identificação e seleção de atributos, treinamento do modelo, identificação de limiar para anomalias e alerta para usuário. O modelo é baseado em *Support Vector Machine* (SVM), que objetivam encontrar o melhor hiperplano capaz de identificar a presença de Oestrus (espécie de parasita). Contudo, o tempo de execução é extenso (cerca de 15 segundos) e crescente à medida que se adicionam novas características de áudio ao treinamento e validação.

Por fim, no trabalho de Komatsu e Kondo (2017) foi investigada a utilização de um modelo de variação temporal, na qual calcula-se a dissimilaridade entre uma cena atual e a anterior. Os conjuntos de dados de áudio são pré-processados para extrair as características MFCC. As cenas acústicas são detectadas como anomalias com base em 24 horas periódicas de dissimilaridade. O cálculo é feito através de uma função de densidade probabilística no segmento temporal, utilizando o modelo *Gaussian Mixture Model* (GMM). No entanto, limita-se em comparações muito extensas, necessitando de muito tempo de captura de novo áudio para classificação de dado típico ou anomalia. Além do mais, o conceito temporal exige sons estacionários e repetitivos, não permitindo a utilização em conjuntos de dados sem essa característica.

As técnicas tradicionais utilizadas na detecção de anomalias em áudio apresentam algumas limitações relevantes. Primeiramente, essas abordagens baseadas em características simples podem não ser capazes de capturar de forma adequada padrões complexos e sutis presentes nos dados sonoros, limitando sua sensibilidade para identificar anomalias mais sutis ou de natureza não trivial. Além disso, as técnicas rasas podem ser mais suscetíveis a ruídos e variações indesejadas nos dados, prejudicando o desempenho da detecção. Adicionalmente, a necessidade de ajuste cuidadoso de parâmetros e limiares pode ser desafiadora, demandando conhecimento prévio do domínio e das características específicas do problema.

Deste modo, a capacidade de generalização dessas técnicas pode ser limitada, especialmente para diferentes tipos de anomalias ou conjuntos de dados de áudio heterogêneos. Portanto, ao empregar técnicas rasas na detecção de anomalias em áudio, é essencial levar em consideração tais limitações.

3.2 MODELOS AUTOCODIFICADORES PARA DETECÇÃO DE ANOMALIAS

A ampla utilização de AEs destaca-se por possuir arquitetura mais simples se comparado ao de modelos generativos descritos na próxima subseção. Trabalhos recentes abordam o uso de AE em detecção de anomalias (XU et al., 2021a; NGUYEN et al., 2021; DEEPAK; CHANDRAKALA; MOHAN, 2020; WAN et al., 2019; CHENG et al., 2021).

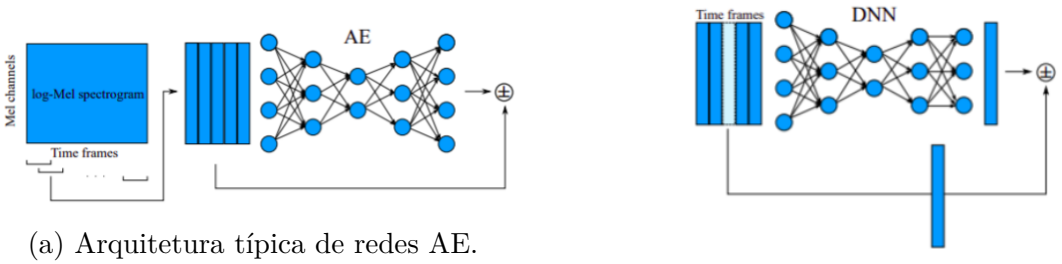
No domínio de imagens, a abordagem de Cheng et al. (2021) demonstra que transformações no dado de entrada melhoram significativamente as reconstruções do Decodificador. Assim, o espaço latente criado através de rotações das imagens aprende o significado semântico das estruturas e não somente posicional como em arquiteturas AE comuns. A avaliação da arquitetura utilizou a métrica AUC *score* nos conjuntos de dados populares como MNIST, CIFAR-10, CIFAR-100 e etc.

O trabalho de Truong et al. (2021) apresenta uma variação de uma rede neural chamada U-Net, originalmente proposta por Ronneberger, Fischer e Brox (2015) para segmentação de imagens. A arquitetura em questão possui as propriedades de uma rede AE e foi implementada utilizando somente camadas totalmente conectadas, com a vantagem de possuir ligações entre as redes Codificadora e Decodificadora. Embora a abordagem seja simples, possibilitou-lhe o uso de dados com apenas uma dimensão para o treinamento de seu modelo.

Dentro do cenário de áudio, o trabalho de Koizumi, Kawaguchi e Imoto (2020) propõe uma arquitetura simples de AE para realizar o cálculo de anomalias. Estes cálculos são feitos através do erro de reconstrução do som observado. O modelo é treinado para obter e minimizar o erro de reconstrução dos dados de treinamentos. Neste cenário, a pontuação dos sons típicos deve ser menor que os sons anormais (anômalos). Este método baseia-se na suposição mencionada na seção 2.4, que afirma a impossibilidade destas arquiteturas de reconstruir dados que não foram utilizados no treinamento, isso é, sons anômalos.

O trabalho desenvolvido por Suefusa et al. (2020) propõe a utilização de redes AE com algumas variações em sua arquitetura, chamadas *interpolation DNN* (IDNN) e *prediction DNN* (PDNN). Sua abordagem possui menos parâmetros do que a arquitetura AE, pois sua parte Decodificadora é treinada utilizando somente uma parte do dado de entrada, variando entre o centro e a borda final, com as redes IDNN e PDNN respectivamente. A Figura 3.1 ilustra a diferença entre uma rede neural AE típica e uma rede neural IDNN, representadas pela Figura 3.1a e Figura 3.1b respectivamente. O autor demonstrou sua proposta em conjuntos de dados de sons de maquinário industrial, obtendo em média 20% de AUC *score* acima do *baseline* para a arquitetura IDNN e 6% para a arquitetura PDNN.

Considerando arquiteturas AE combinadas com redes recorrentes, os autores Müller, Illium e Linnhoff-Popien (2021) propuseram melhorias no algoritmo IDNN. Durante os



(a) Arquitetura típica de redes AE.

(b) Arquitetura IDNN

Figura 3.1 – Diferenças entre arquiteturas AE e IDNN. Fonte: (SUEFUSA et al., 2020)

experimentos, verificou-se que a adição de camadas de *Long Short-Term Memory* (LSTM) em certos contextos de áudio contribuía para a melhora do desempenho, superando a arquitetura de Suefusa et al. (2020) e Koizumi, Kawaguchi e Imoto (2020). A avaliação mostrou resultados em média 9% de AUC *score* maiores que o trabalho anterior.

Embora os estudos apresentados tenham alcançado resultados relevantes e contribuído para o campo, as arquiteturas baseadas em AE possuem limitações na representação dos dados de entrada. À medida que a complexidade das instâncias de áudio aumenta, as redes AE perdem sua capacidade de representá-las adequadamente. Além disso, essas arquiteturas são altamente sensíveis a erros de entrada que diferem do conjunto de dados de treinamento, o que significa que um som típico tem maior probabilidade de ser erroneamente classificado como uma anomalia.

3.3 MODELOS GERADORES

Trabalhos mais recentes avançaram o desempenho de modelos generativos como *Generative adversarial networks* (GANs) ou *Variational Auto Encoders* (VAEs) para a tarefa de detecção de anomalias (LEE; Umar Karim Khan; KYUNG, 2021; MAN; YANG; XU, 2020; HONG; CHOE, 2020; AKCAY; ATAPOUR-ABARGHOUEI; BRECKON, 2018; ZENATI et al., 2018; SCHLEGL et al., 2017). A abordagem generativa se tornou o estado-da-arte para a tarefa de detecção anomalias em imagens, entretanto, os avanços são poucos expressivos na utilização destas arquiteturas no domínio de áudio.

O estudo realizado por Schlegl et al. (2019) propõe uma abordagem não supervisionada que combina AE e redes GAN em uma única arquitetura. A contribuição principal do trabalho está na utilização de modelos generativos para criar classes de áudio. No entanto, essa abordagem ainda não abrange todas as variações de anomalias existentes, uma vez que o foco é na diferenciação entre classes, sem necessariamente representar um domínio semântico completo.

O trabalho proposto por Zenati et al. (2018), é uma das primeiras publicações que aborda o funcionamento de uma GAN para detecção de anomalias em imagens. Isto posto,

os autores demonstram a utilização em dois conjuntos de dados famosos da literatura: **MNIST** e **KDD99**. A proposta da arquitetura consiste em aprender a representação latente de cada instância de exemplo e utilizar essa representação para otimizar as redes discriminadora e geradora. Diferentemente das redes discriminadoras convencionais, essa arquitetura utiliza uma rede capaz de avaliar tanto as instâncias de entrada (imagens) quanto suas representações latentes.

Os autores em Liu e Xu (2020) descrevem sua abordagem utilizando conjunto de dados em imagens. Sua principal contribuição é a arquitetura GAN Ortogonal para detecção de anomalias. Os conjuntos de dados validados pelo estudo foram: **KDDCUP99**, **MNIST**, **Fashion-MNIST**, **CIFAR10** e **KDDCUP99**. A combinação da GAN com o VAE, uma arquitetura ortogonal, é uma abordagem promissora. O VAE é um modelo generativo capaz de aprender a distribuição e outras características de um conjunto de dados, permitindo a geração de novos dados sem a necessidade de conhecê-los explicitamente. Essa combinação se destaca como uma estratégia eficiente na geração de dados por meio de modelos generativos.

O artigo de Akcay, Atapour-Abarghouei e Breckon (2018) descreve uma arquitetura que combina AE e GAN no contexto de imagens. O modelo generativo é construído utilizando redes AE, resultando em duas saídas: a imagem gerada e seu campo latente correspondente. Essa abordagem permite que as funções de perda mapeiem os espaços latentes das imagens, visando gerar não apenas imagens semelhantes às do conjunto de treinamento, mas também mapear os espaços latentes correspondentes. Os resultados mostram que o espaço latente da imagem gerada difere significativamente do espaço latente da imagem original, caso a imagem original não pertença aos dados normais. A contribuição do trabalho é destacada por adicionar dois codificadores à arquitetura, sendo proposto que o segundo codificador é propenso a gerar erros maiores do que apenas o decodificador.

Os autores Akcay, Abarghouei e Breckon (2019) apresentam uma abordagem utilizando redes geradoras para detecção de anomalias. Seu trabalho recente, denominado *Skip-GANomaly*, é uma evolução do trabalho mencionado anteriormente. A arquitetura proposta combina uma rede GAN com AE, mas destaca-se pelo uso da arquitetura U-Net. A U-Net, originalmente desenvolvida para segmentação de imagens médicas, substitui a rede AE anteriormente utilizada. A U-Net possui conexões residuais ou não-lineares entre suas camadas, permitindo uma recriação mais detalhada da imagem de entrada em comparação com um AE convencional, por isso o nome "Skip". Os autores também propõem novas funções de perda para essa abordagem, incluindo o cálculo de distâncias $\mathcal{L}_1$ entre a imagem real e a imagem gerada, bem como a aproximação das características latentes entre as convoluções e as previsões feitas pela rede discriminadora. Embora esse trabalho destaque as vantagens do treinamento adversarial em relação ao treinamento convencional de um AE, os autores limitam-se a descrever sua aplicação e métricas apenas

no domínio de imagens.

Ainda neste aspecto, os autores Liu et al. (2021) propuseram modificações na arquitetura *Skip-GANomaly*, incorporando mecanismos de atenção para capturar informações relevantes nas imagens. Essas modificações incluíram a adição de um módulo de atenção de bloco convolucional (CBAM) e o uso de blocos de convoluções separáveis em profundidade (DSCs), inicialmente introduzidos por Woo et al. (2018) e Chollet (2017), respectivamente. Os autores destacam que as camadas DSCs são capazes de otimizar a quantidade de parâmetros da rede em até quatro vezes.

No contexto de similaridade entre imagens, os autores em Song et al. (2021) demonstram em sua base arquitetural um modelo baseado no trabalho anterior, entretanto, sua contribuição chave é a criação de uma nova função de perda para a subrede geradora, aumentando o erro de acordo com a similaridade, brilho e outras características da imagem.

Em conclusão, a utilização de redes GANs na detecção de anomalias tem apresentado avanços significativos nos últimos anos. A combinação de GANs com redes AEs em uma única arquitetura tem se mostrado promissora, permitindo a geração de representações compactas e expressivas dos dados de entrada. Essa abordagem, representada por trabalhos como *Skip-GANomaly*, proporciona uma detecção mais eficaz de anomalias, além de uma melhor capacidade de reconstrução dos dados. Adicionalmente, a introdução de mecanismos de atenção e técnicas avançadas, como os blocos convolucionais separáveis em profundidade (DSCs), tem contribuído para melhorias adicionais na performance desses modelos. No entanto, ainda há desafios a serem superados, como a extensão dessas abordagens para outras modalidades de dados, como áudio, e a busca por métricas de avaliação robustas. O progresso nessa área tem o potencial de beneficiar diversas aplicações que envolvem detecção de anomalias, contribuindo para aprimorar a segurança e a qualidade de sistemas e processos em diversos setores.

3.4 CONSIDERAÇÕES FINAIS

Neste capítulo, foram apresentados os estudos relacionados à detecção de anomalias. Observou-se que o tema está recebendo significativa atenção da comunidade devido ao surgimento de arquiteturas profundas que têm impulsionado o avanço do estado da arte em várias áreas. Essas arquiteturas têm demonstrado resultados superiores em termos de desempenho e capacidade de detecção de anomalias, fornecendo soluções mais robustas e eficientes. Essa tendência reflete o reconhecimento da importância e das demandas crescentes por métodos avançados de detecção de anomalias, capazes de lidar com dados complexos e heterogêneos.

Embora as redes generativas sejam amplamente utilizadas na detecção de anomalias em imagens, é evidente a escassez de estudos que explorem sua aplicabilidade no domínio de

sons. Os poucos trabalhos existentes nessa área revelam que a combinação de arquiteturas codificadoras e generativas apresenta resultados promissores.

Tabela 3.1 – Comparação entre os trabalhos sobre detecção de anomalias utilizando arquiteturas profundas

Trabalhos	Arquitetura	Domínios	Paradigma
(AKCAY; ATAPOUR-ABARGHOUEI; BRECKON, 2018)	AEE+GAN	imagem	Não-supervisionado
(AKCAY; ABARGHOUEI; BRECKON, 2019)	U-Net+GAN	imagem	Não-supervisionado
(ZENATI et al., 2018)	AE+GAN	imagem	Não-supervisionado
(LIU; XU, 2020)	VAE+GAN	imagem	Não-supervisionado
(SONG et al., 2021)	U-Net+GAN	imagem	Não-supervisionado
(CHENG et al., 2021)	AE	imagem	Não-supervisionado
(LIU et al., 2021)	U-Net(DSC+CBAM)+GAN	imagem	Não-supervisionado
(SCHLEGL et al., 2019)	AE+GAN	áudio	Semi-Supervisionado
(KOIZUMI; KAWAGUCHI; IMOTO, 2020)	AE	áudio	Não-Supervisionado
(SUEFUSA et al., 2020)	AE(IDNN)	áudio	Não-supervisionado
(MüLLER; ILLIUM; LINNHOFF-POPIEN, 2021)	AE(DRINK)	áudio	Não-supervisionado
(TRUONG et al., 2021)	U-Net	áudio	Não-supervisionado
Nosso trabalho	AE+GAN	áudio	Não-supervisionado

A Tabela 3.1 apresenta os principais trabalhos na área de detecção de anomalias. Pode-se observar que a maioria dos trabalhos envolvendo arquiteturas generativas são empregadas no domínio de imagens e, por sua vez as arquiteturas autocodificadoras são empregadas em áudio.

A partir da análise desses trabalhos, constata-se que as abordagens baseadas em redes AE têm desempenhado um papel fundamental na detecção de anomalias em eventos sonoros. A simplicidade das arquiteturas AE, em comparação com outros modelos generativos, torna-as uma escolha atrativa para essa tarefa desafiadora. O uso de AE demonstrou resultados promissores, especialmente quando adaptado para o domínio de áudio, onde a escassez de dados anômalos e a complexidade das definições são fatores críticos.

Destaca-se que as redes GANs para detecção de anomalias em eventos e cenas acústicas tem sido explorada de forma menos abrangente em comparação aos avanços alcançados na detecção de anomalias em imagens. Embora os estudos nessa área ainda sejam limitados, as pesquisas existentes indicam que a aplicação de GANs no domínio acústico também pode ser promissora. A combinação de arquiteturas codificadoras e generativas tem se mostrado eficaz na criação de representações latentes e na geração de dados de áudio realistas. No entanto, é necessário um maior esforço de pesquisa para explorar e adaptar as GANs às particularidades dos eventos e cenas acústicas, a fim de obter resultados mais precisos e confiáveis na detecção de anomalias nesse contexto.

É importante destacar que a aplicação de redes GANs para detecção de anomalias requer cuidados e práticas específicas ao treinar modelos em diferentes domínios, como imagens e áudio. Enquanto as técnicas de treinamento para imagens são amplamente estudadas e estabelecidas, a adaptação dessas abordagens para o domínio de áudio apresenta desafios adicionais. Um dos principais cuidados é a escolha adequada das funções de

perda, considerando as características e a natureza dos sinais de áudio. Além disso, é necessário ajustar hiperparâmetros e arquiteturas das GANs para capturar efetivamente as particularidades do domínio acústico, como a representação temporal e a complexidade espectral. Também é fundamental garantir um conjunto de treinamento diversificado e representativo, que abranja diferentes tipos de anomalias sonoras, a fim de melhorar a capacidade de generalização do modelo. Ainda, é essencial realizar uma validação cuidadosa dos resultados, levando em consideração métricas apropriadas para avaliar a detecção de anomalias em áudio. Essas práticas e cuidados são cruciais para garantir a eficácia e confiabilidade das GANs na detecção de anomalias em áudio.

4 SOLUÇÃO PROPOSTA

Neste capítulo serão abordados os materiais e métodos utilizados no desenvolvimento da abordagem de detecção de anomalias em eventos sonoros. Inicialmente descreveremos o pré-processamento na seção 4.1, em seguida, a abordagem do método comparativo na seção 4.2 e por fim, as considerações finais na seção 4.3

4.1 PRÉ-PROCESSAMENTO

A etapa de pré-processamento pode ser descrita em 5 etapas, conforme mostrada na Figura 4.1. (1) Começando com a leitura do sinal digital de áudio, (2) aplicação da função de transformada de Fourier, (3) mapeamento da magnitude do sinal para decibéis e conversão de sinal para escala logarítmica, (4) concatenação janelada no tempo de 5 espectrogramas gerados e por fim (5) a transformação destas matrizes para vetores.

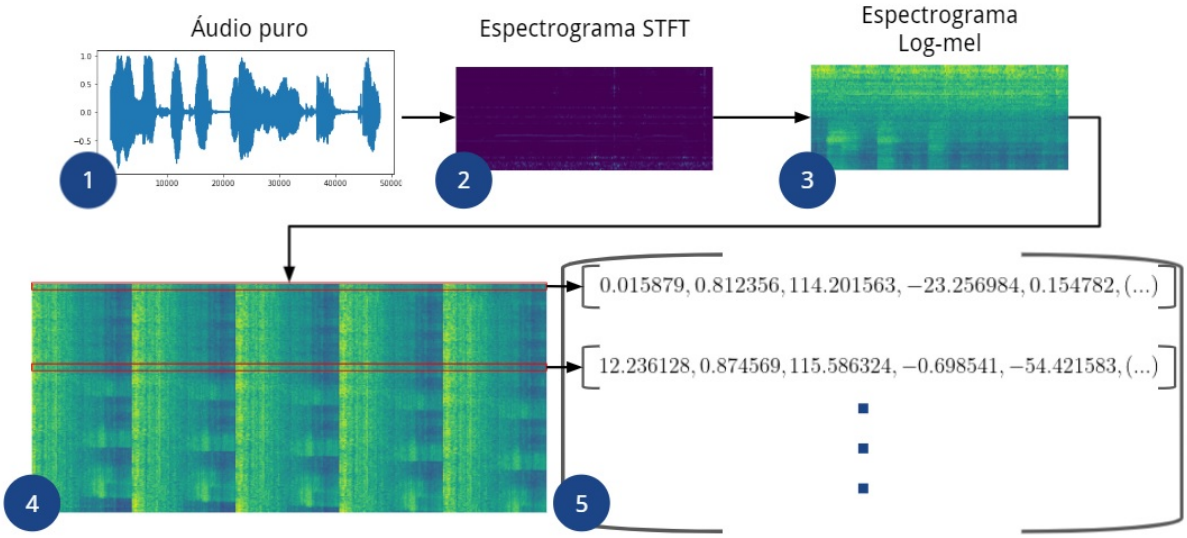


Figura 4.1 – Etapas de pré-processamento. Fonte PRÓPRIO AUTOR

Na primeira etapa o áudio é lido da base de dados, o sinal é convertido em uma matriz que denota tempo e a frequência do sinal, conforme mencionado na seção 2.2. Na segunda etapa ocorre a extração de características, onde um áudio é denotado por uma ou mais coeficientes de mel.

Na terceira etapa ocorre o mapeamento do sinal para decibéis, isto é, limita os valores do sinal somente em frequências audíveis para o ser humano. O mapeamento também aplica um limiar de frequências negativas, tal que qualquer valor abaixo de $-80dB$ se tornará $-80dB$. Esta etapa é responsável por aumentar visualmente as frequências mais baixas, tornando os sinais mais fracos equiparáveis aos mais altos. Essa técnica visa melhorar a visualização das características.

A quarta etapa compreende a concatenação e janelamento de tempo dos espectrogramas. Para que isso seja possível, o espectrograma precisa ter seu tamanho total reduzido no eixo do tempo e então, esse eixo é movido a cada novo espetrograma concatenado. De modo que cada espectrograma possua característica temporal diferente das demais. A Figura 4.2 demonstra que os espectrogramas diferem-se uns dos outros no eixo Y, à medida que novos espectrogramas são adicionados no eixo X.

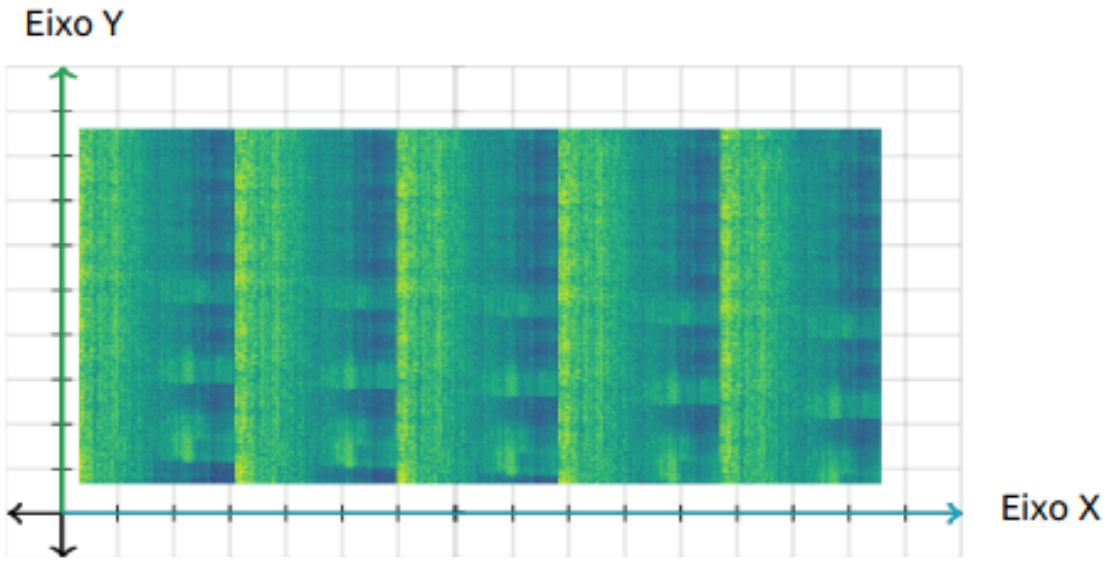


Figura 4.2 – Etapa número 4 detalhada. Fonte PRÓPRIO AUTOR

Por fim, a matriz resultante é separada em vetores, no qual cada linha corresponde a 128 características de Mel em 5 momentos diferentes do tempo. Os modelos apresentados na seção 4.2 possuem como entrada as instâncias no mesmo formato. Seguindo o formato de entrada proposto por (KOIZUMI; KAWAGUCHI; IMOTO, 2020), cada instância de áudio possui tamanho padronizado de 10 segundos repartidos em quadros de $64ms$, com janelamento de 50% entre os quadros, utilizando o algoritmo de Hop. Utilizamos 1024 pontos da transformada rápida de Fourier (FFT) e 128 Mels são utilizados para as características de cada quadro. Ao final, 5 quadros são concatenados, totalizando um vetor de dimensões $5 \times 128 = 640$ posições de entrada. Desta forma, as arquiteturas podem ser avaliadas de maneira padronizada no tocante à entrada de dados.

4.2 ABORDAGEM DO MÉTODO COMPARATIVO

Neste capítulo descrevemos 3 estruturas baseadas em aprendizagem profunda para a detecção de anomalias em eventos sonoros. Dada a revisão bibliográfica demonstrada no Capítulo 3, os trabalhos de detecção de anomalias em imagem utilizando redes GAN de modo não supervisionado superam em quantidade os trabalhos no âmbito de áudio. No

entanto, em nosso trabalho mostraremos como algumas adaptações nas arquiteturas são capazes atingir os objetivos descritos na seção 1.3.

O foco da nossa abordagem é identificar quais as menores alterações necessárias para que uma arquitetura de rede neural profunda utilizando aprendizado adversário seja capaz de identificar anomalias em eventos de áudios. Na fase de treinamento das arquiteturas, cada classe alvo descrita na subseção 5.1.2 possui um modelo correspondente, pois os modelos precisam aprender por meio dos dados as características únicas de cada classe, afim de ser possível identificar as anomalias do conjunto de teste. A Figura 4.3 exemplifica de forma resumida o comportamento esperado dos modelos. Na qual o treinamento é realizado através instâncias de sons normais e opcionalmente metadados (os detalhes dos dados serão melhor explicados na subseção 5.1.2).

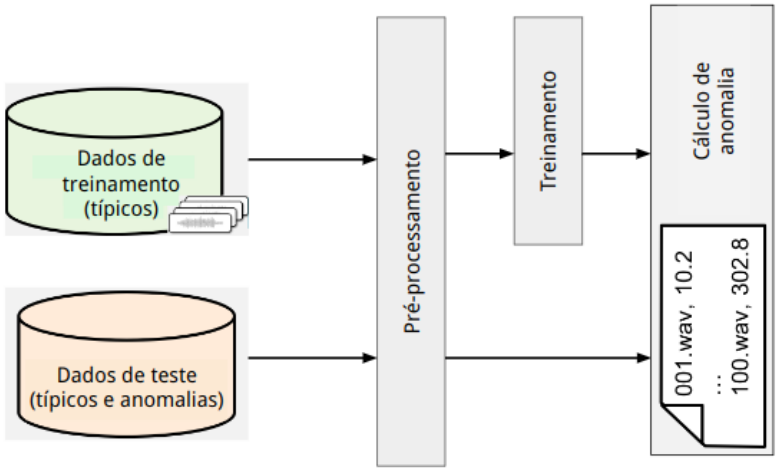


Figura 4.3 – Etapas de treinamento e teste do modelo. Fonte PRÓPRIO AUTOR

A solução proposta consiste em adaptação básica dos modelos publicados por Zenati et al. (2018), Akcay, Atapour-Abarghouei e Breckon (2018) e Akcay, Abarghouei e Breckon (2019), nas literaturas descritas no Capítulo 3 para adequação dos modelos para a detecção de anomalias sonoras. Destacamos que todas as arquiteturas utilizadas neste trabalho são abordagens de GANs, por possuírem resultados significativos em conjunto de dados de imagens. Entretanto, tais soluções apenas comentam e não demonstram sua eficácia em conjuntos de outros domínios, como áudio por exemplo.

- **Fase 1 – Pré-processamento**. Na fase inicial, os dados de treinamento e teste passam pela etapa de pré-processamento descrito na seção 4.1 e são transformados em vetores de características. Essa etapa inicial padroniza as instâncias de áudio para treinamento e validação do modelo. A padronização permite que as diferentes arquiteturas utilizem os mesmos conjuntos de dados e estabelece os requisitos para as redes GANs adaptadas realizarem a leitura dos áudios. As instâncias de treinamento e de teste foram previamente separadas, então garante-se que todas as adaptações das arquiteturas utilizem os mesmos conjuntos de dados.

- **Fase 2 – Treinamento do Modelo**. Nesta fase, treinamos as arquiteturas adaptadas utilizando somente exemplos de áudios classificados como comportamentos típicos. Após a fase 1, os dados são disponibilizados para os modelos realizarem os treinamentos, cada arquitetura possui particularidades para tal (os detalhes dos modelos serão melhores explicados na subseção 4.2.1). Entretanto, é comum a todas as arquiteturas a criação de um modelo por rótulo. Durante o treinamento as sub-redes Geradoras são treinadas para serem capazes de reproduzir os conjuntos de entrada com a maior fidelidade possível. As sub-redes Discriminadoras agem em conjunto para avaliar se os dados gerados possuem semelhanças com os originais. Essa dinâmica permite a interação entre as duas redes para um benefício mútuo, também chamado de soma zero. O resultado deste estágio são dois modelos de redes neurais, um para gerar dados próximos dos típicos e outro para classificá-los.

- **Fase 3 – Validação em Conjunto de Teste**: Por fim, o modelo Gerador recebe como entrada os áudios de teste pré-processados durante a fase 1, na qual será verificado através de algumas particularidades de cada arquitetura o grau de distância entre os dados a serem testado e sua representação gerada. Em seguida, é realizado o cálculo de média dos erros de cada áudio, este escore é chamado de cálculo de anomalia (os detalhes deste cálculo podem ser vistos na subseção 5.1.3). Este número real é utilizado posteriormente para calcular as métricas descritas com mais detalhes subseção 5.1.4.

Nossa abordagem manteve o máximo de originalidade possível das arquiteturas descritas por seus autores originais, entretanto, adequamos as arquiteturas das redes geradoras e discriminadoras para permitirem entradas de instâncias de áudio. As modificações serão descritas com mais detalhes a seguir.

4.2.1 Adaptação das Arquiteturas

As arquiteturas utilizadas na solução proposta foram construídas originalmente para o domínio de imagem, porém, com algumas adaptações é possível modificar a entrada destas arquiteturas para que seja possível o treinamento com os dados descritos na seção ??.

As adaptações destas arquiteturas foram baseadas no modelo proposto por Koizumi, Kawaguchi e Imoto (2020) que apresenta uma rede simples AE para áudio. Particularmente, a parte dos modelos GANs referentes ao Gerador e Discriminador foram adaptados para a arquitetura composta por uma entrada da rede neural totalmente conectada (FCN), três camadas ocultas FCN e uma camada de saída FCN. Cada camada oculta possui 128 neurônios e dimensão 8 no espaço latente. A função de ativação Unidade Linear retificada (ReLU) (AGARAP, 2018) e normalização em lote (*Batch Normalization*)(IOFFE;

SZEGEDY, 2015) são utilizadas em cada camada FCN, exceto na camada de saída do decodificador, conforme mostra a Figura 4.4. Esta rede foi escolhida por possuir poucos parâmetros comparado às redes originais, possibilitando a leitura das instâncias de áudios. Essas adaptações foram incorporadas em todos as arquiteturas para comparação e padronização na mesma base de dados. Visto que, desta maneira é possível avaliar as funções de otimizações e as arquiteturas em si.

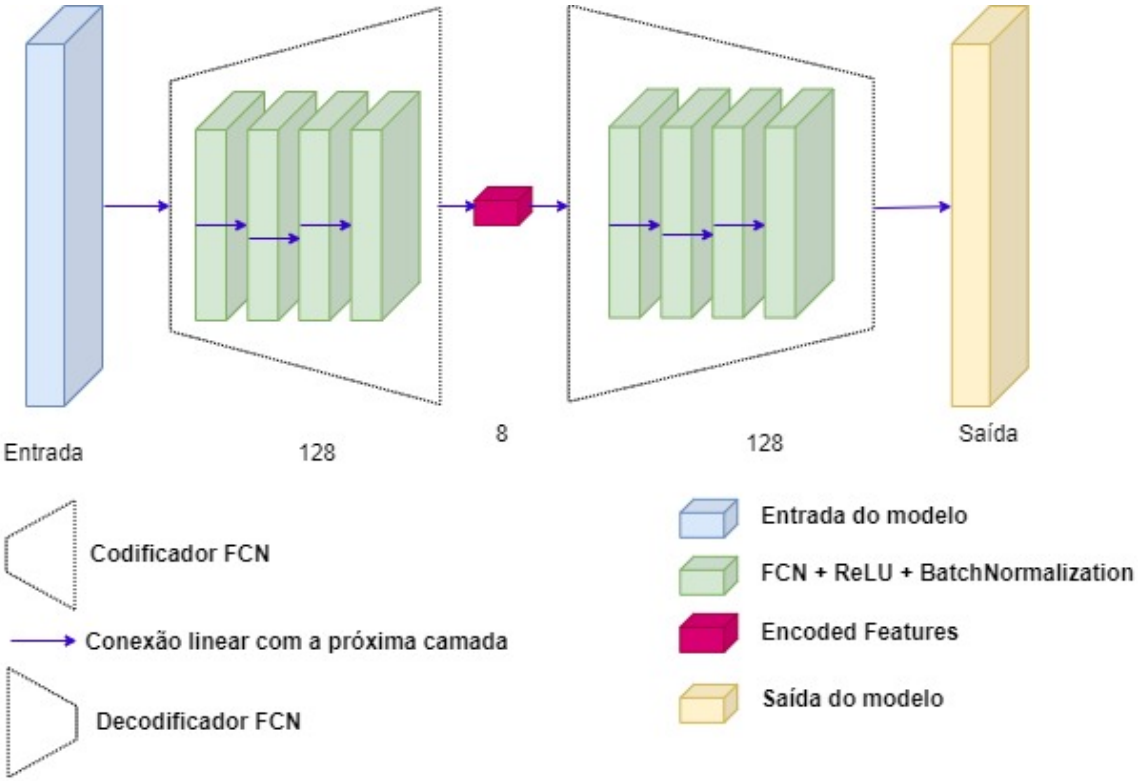


Figura 4.4 – Detalhes da arquitetura proposta por (KOIZUMI; KAWAGUCHI; IMOTO, 2020) Fonte: PRÓPRIO AUTOR

4.2.2 EGBAD – *Efficient Gan-Based Anomaly Detection*

O modelo EGBAD (ZENATI et al., 2018) descrito inicialmente na seção 3.3 é uma das arquiteturas GANs que foram adaptadas para áudio. Essa arquitetura é apresentada na Figura 4.5. Nela podemos ver 3 redes atuando em conjunto, sendo elas a rede G, E e D, respectivamente a rede Geradora, Codificadora e Discriminadora. Tais redes possuem a arquitetura adaptada conforme mencionado na subseção 4.2.1.

Particularmente, a adaptação para áudio consistiu na mudança do gerador G, Discriminador D e Codificador E conforme a descrição da subseção 4.2.1. O espelhamento das arquiteturas pode ser feito por meio da Figura 4.4, na qual o modelo G e o modelo D representam o Decodificador FCN, já o modelo E retrata o Decodificador FCN. Os demais elementos da arquitetura seguem o modelo original.

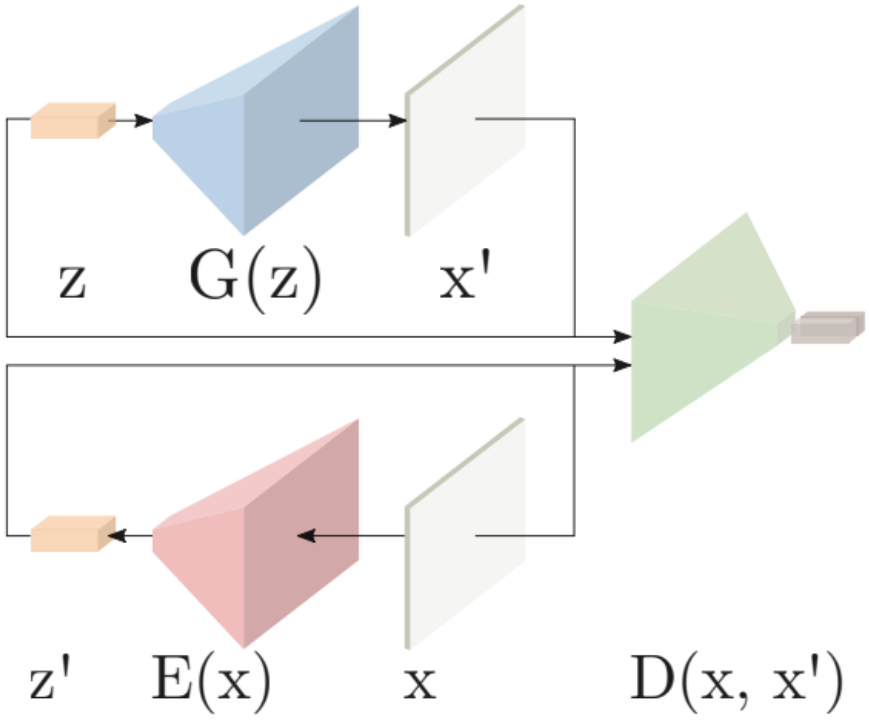


Figura 4.5 – Arquitetura da rede profunda Efficient Gan-Based Anomaly Detection. Fonte: (ZENATI et al., 2018)

A arquitetura é capaz de mapear a instância de entrada para uma representação latente z, durante o treinamento da rede G e D; Diferentemente das redes convencionais GAN, essa estratégia especifica a entrada da rede D com os dados reais, gerados e a representação latente.

A dinâmica de treinamento baseia-se na capacidade do discriminador de classificar a instância de entrada junto do vetor latente. O processo inicia com a geração de um espaço latente z', conforme a Figura 4.6. Em seguida cria-se uma instância gerada (*fake*) a partir de um espaço latente randômico z, mostrada na Figura 4.7. Por fim, a rede D recebe como entrada uma instância real ou *fake* e seu respectivo vetor latente, como observa-se na Figura 4.8.

4.2.3 GANomaly – *Semi-Supervised Anomaly Detection via Adversarial Training*

O modelo GANomaly (AKCAY; ATAPOUR-ABARGHOUEI; BRECKON, 2018) descrito inicialmente na seção 3.3 faz parte das arquiteturas GANs que foram adaptadas para áudio. A Figura 4.9 apresenta a arquitetura original do modelo. Nela podemos ver 4 redes neurais profundas atuando em conjunto. A rede Geradora G possui duas sub-redes, sendo elas: rede codificadora G_E e rede discriminadora G_D. A rede codificadora E é uma cópia da rede G_E. Por fim, temos a rede discriminadora D. A adaptação destas redes seguem em concordância com a descrição na subseção 4.2.1. Essencialmente, esta adaptação pode ser refletida através da Figura 4.4, onde constituiu-se da mudança das redes G_E

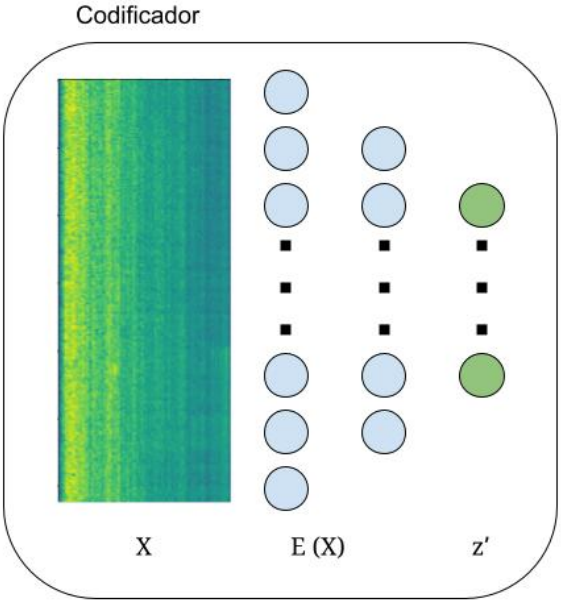


Figura 4.6 – Arquitetura Codificadora adaptada da rede profunda Efficient Gan-Based Anomaly Detection. Fonte: PRÓPRIO AUTOR

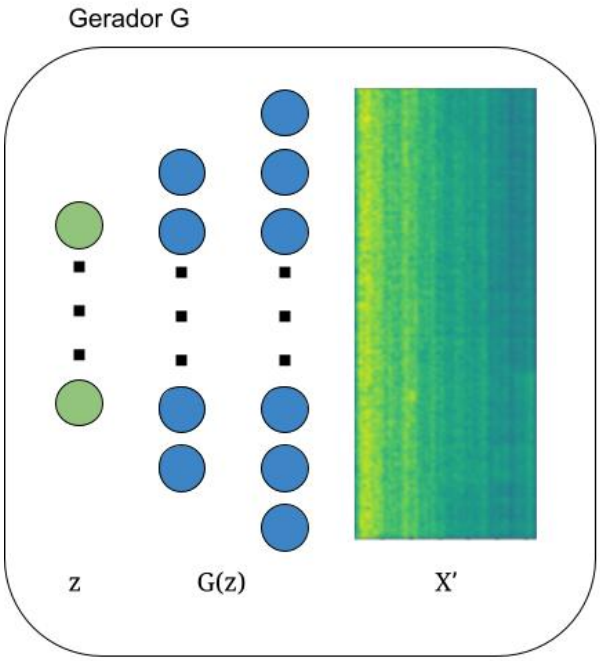


Figura 4.7 – Arquitetura Geradora adaptada da rede profunda Efficient Gan-Based Anomaly Detection. Fonte: PRÓPRIO AUTOR

e E para a rede Codificadora FCN, os modelos G_D e D foram convertidos para a rede Decodificadora FCN. Os detalhes restantes da arquitetura permaneceu conforme o original.

A dinâmica de treinamento baseia-se no pressuposto de que o espaço latente gerado pela rede Ez' acumula erros mais expressivos que os gerados pela rede GD x'. Pode-se observar na Figura 4.10 a interação entre as sub-redes pertencentes ao modelo Gerador e de que modo a função de otimização converge para que este modelo crie características de

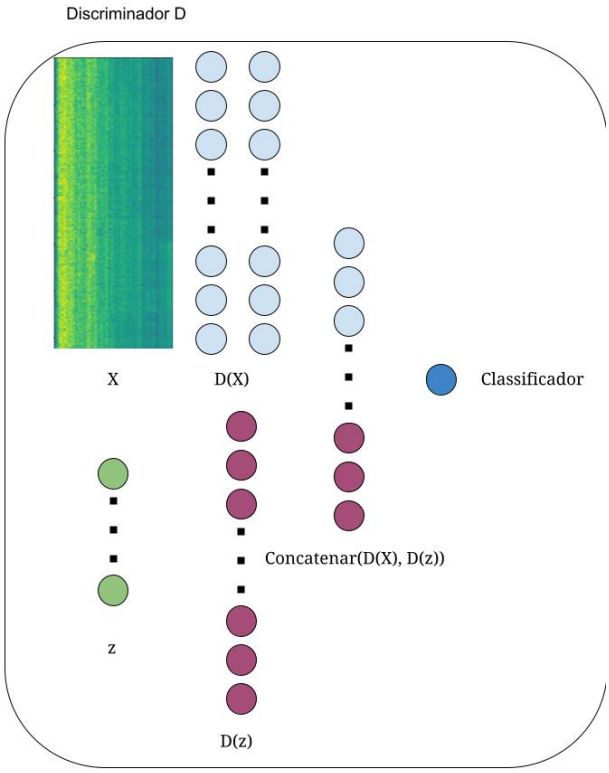


Figura 4.8 – Arquitetura Discriminadora adaptada da rede profunda Efficient Gan-Based Anomaly Detection. Fonte: PRÓPRIO AUTOR

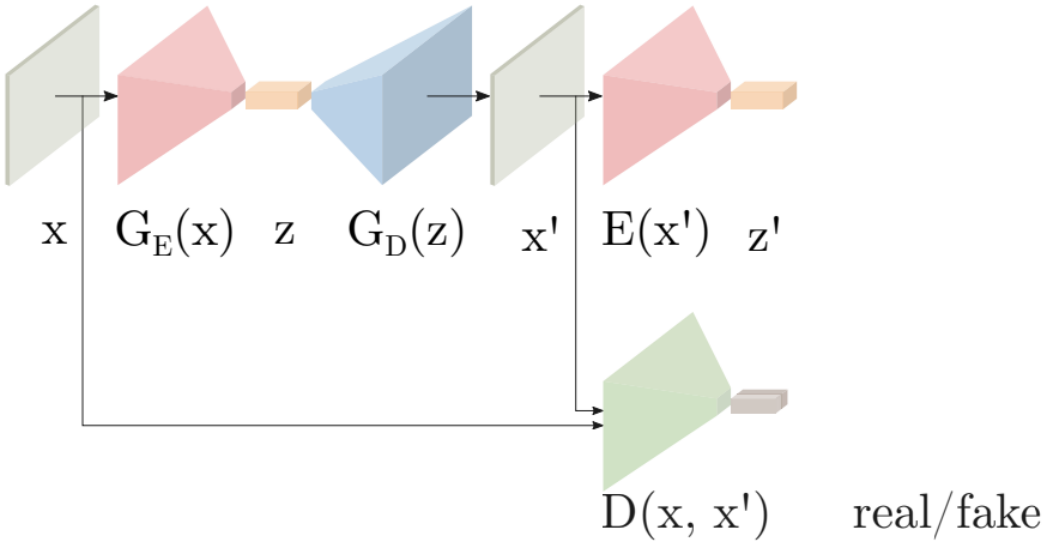


Figura 4.9 – Arquitetura da rede profunda Semi-Supervised Anomaly Detection via Adversarial Training. Fonte: (AKCAY; ATAPOUR-ABARGHOUEI; BRECKON, 2018)

áudio similares às instâncias de treino. Já a Figura 4.11 apresenta o modelo Discriminador e sua função de otimização.

Descrevendo o processo de treinamento desta rede de forma detalhada, temos:

- **Rede Geradora** – A instância de treino x é codificada para um campo latente z e

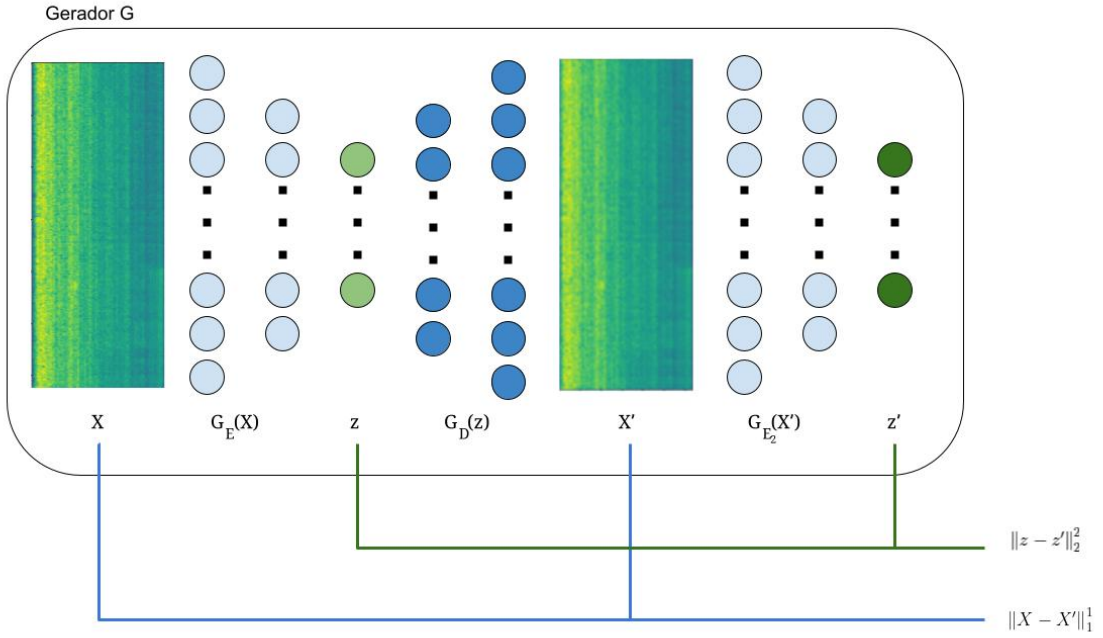


Figura 4.10 – Arquitetura da rede Geradora adaptada GANomaly. Fonte: PRÓPRIO AUTOR

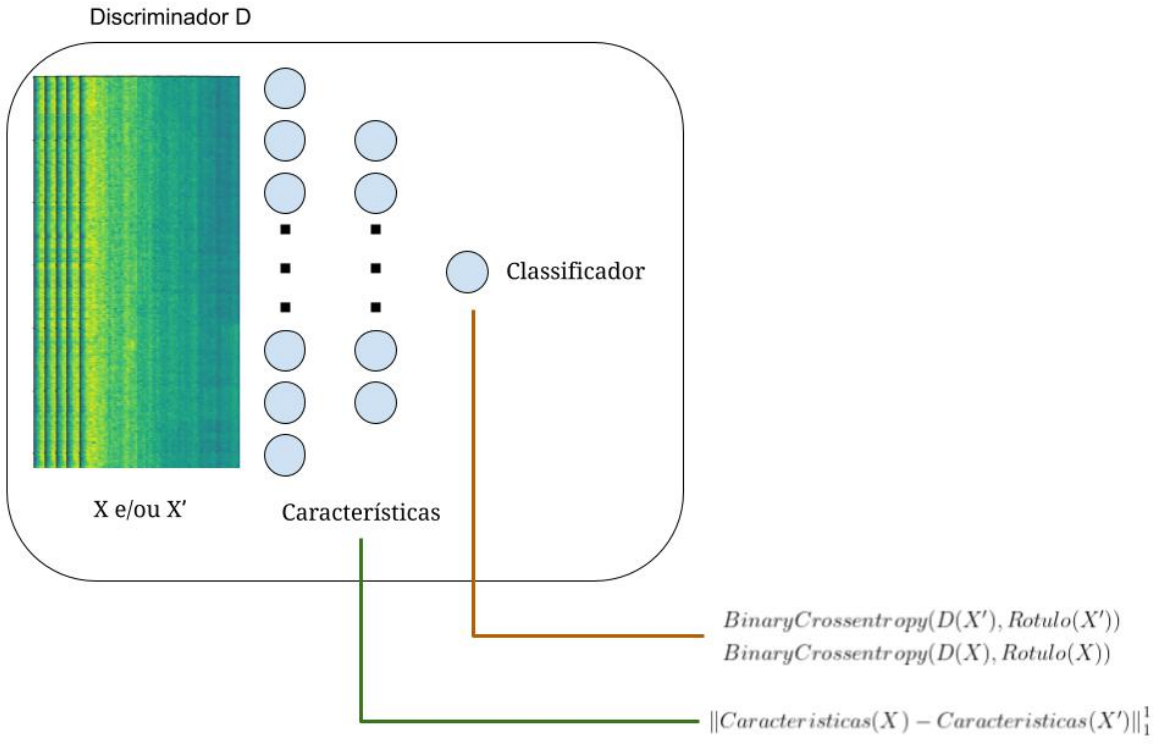


Figura 4.11 – Arquitetura da rede Discriminadora adaptada GANomaly. Fonte: PRÓPRIO AUTOR

decodificada, gerando assim, uma versão artificial da instância de treino chamada x'. O passo seguinte é a codificação desta instância artificial que chamaremos de z'. Desta maneira, pode-se otimizar a reconstrução das características do áudio, bem como sua representação latente. O modelo Gerador converge para criar características de áudio cada vez mais realistas.

- **Rede Discriminadora** – Classifica as instâncias de treino (dados típicos) e as instâncias artificiais geradas no passo anterior. Esta classificação ocorre de maneira sintética, pois os rótulos de dados típicos e artificiais são sempre considerados como 0 e 1 respectivamente.

A hipótese demonstrada pelos autores afirma que o espaço latente gerado pela rede E z' acumula erros maiores que os gerados pela rede GD x'. Para que seja possível a reconstrução das instâncias de entrada, são necessárias 3 funções de otimizações nas redes G e D, chamadas *Adversarial Loss*, *Contextual Loss* e *Encoder Loss*.

Adversarial Loss. Para maximizar a capacidade de reconstrução dos espectrogramas y durante a fase de treino, utilizaremos a função de perda Adversária, proposta por Goodfellow et al. (2014). A Equação 4.1 garante que a rede G reconstrua os dados de maneira mais realística possível, enquanto a rede D distingue corretamente entre espectrograma real ou gerado (falso).

$$\mathcal{L}_{adv} = E_{x \sim p_x}[\log D(y)] + E_{x \sim p_x}[1 - \log D(\hat{y})] \tag{4.1}$$

Contextual Loss. Calcula a distancia L1 entre o dado real x e o dado gerado ($\hat{x} = G(x)$). É repensável por calcular a diferença entre um espectrograma real e o gerado pela rede G. A função de otimização $\mathcal{L}_{con}$ é defina pela Equação 5.2:

$$\mathcal{L}_{con} = \|x - G(x)\|_1^1 \tag{4.2}$$

Encoder Loss. As funções de otimização anteriores produzem dados $\hat{y}$ próximos dos originais y, adicionalmente é empregada a função de perda $\mathcal{L}_{lat}$ que é capaz de produzir representações contextuais dos exemplos. Para reconstruir suas representações, transformando-os em representações latentes, tal que ($z = GE(y)$) e ($\hat{z} = GE(\hat{y})$). A função de otimização $\mathcal{L}_{enc}$ é definido pela Equação 5.3:

$$\mathcal{L}_{enc} = \|z - \hat{z}\|_2 \tag{4.3}$$

Para finalizar, a função total de otimização $\mathcal{L}_{gen}$ do modelo G é dada pela Equação 4.4:

$$\mathcal{L}_{gen} = \lambda_{adv} \times \mathcal{L}_{adv} + \lambda_{con} \times \mathcal{L}_{con} + \lambda_{enc} \times \mathcal{L}_{enc} \tag{4.4}$$

Os valores de λ_{adv}, λ_{con} e λ_{enc} são pesos atribuídos aos otimizadores arbitrariamente e serão apresentados na subseção 5.3.1.

4.2.4 SGANomaly – *Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection*

A arquitetura Skip-GANomaly (AKCAY; ABARGHOUEI; BRECKON, 2019) apresentado inicialmente na seção 3.3 constitui as arquiteturas GANs que receberam adaptações para áudio. A arquitetura em questão utiliza blocos interconectados no modelo G dispostos em formato de U, isto é, o modelo G é um Autocodificador cujo as ordem das camadas do codificador são ligadas à ordem inversa camadas do decodificador, de tal maneira que são passadas informações da primeira rede para a segunda sem seguir o fluxo direto.

A Figura 4.12 demonstra a arquitetura geral do modelo, formada por um modelo gerador (G) e um descriminador (D), respectivamente. A rede G é formada por um Autocodificador simétrico em formato de gravata borboleta, utilizando uma sub-rede Codificadora (G_E) e outra Decodificadora (G_D). As adaptações destas redes seguem conforme descrito na subseção 4.2.1, entretanto, algumas alterações pontuais são feitas na rede G, para respeitar as conexões originais entre camadas.

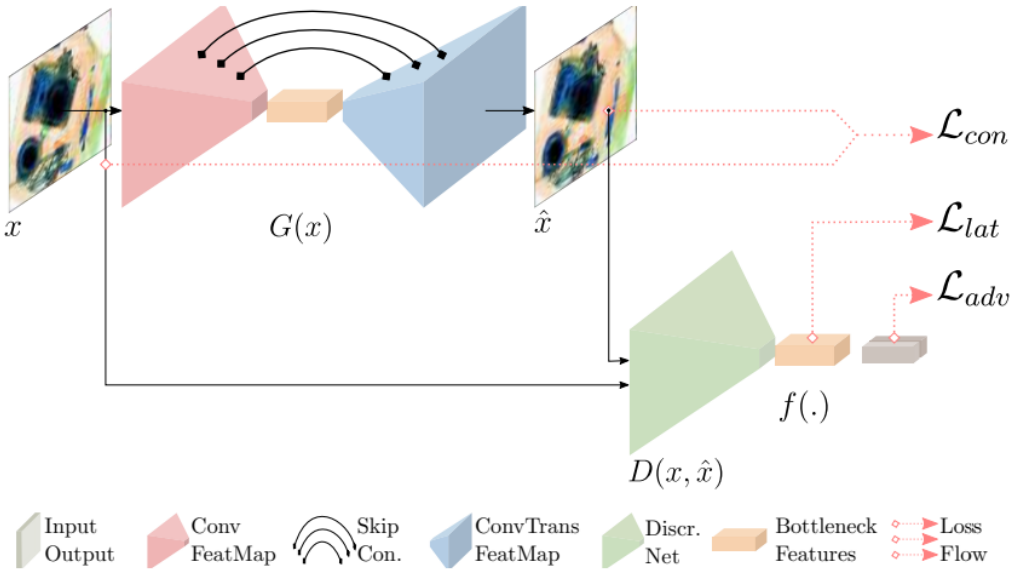


Figura 4.12 – Arquitetura da rede profunda Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection. Fonte: (AKCAY; ABARGHOUEI; BRECKON, 2019)

Fundamentalmente a adaptação desta arquitetura pode ser observada na Figura 4.13. Modificou-se a rede G_E e rede G_D para redes Codificadoras FCN e Decodificadoras

FCN respectivamente, o modelo D é substituído pela rede Codificadora FCN. Por fim, são adicionada conexões e concatenações entre a rede G_E e G_D, de modo que haja vantagem substancial na transferência de informações entre as camadas, preservando informações locais e globais, além de resultar em reconstruções mais próximas do dado original.

A dinâmica de treinamento feita através do uso apenas de características de áudios típicos. Estima-se que o modelo é capaz de reconstruir as instâncias típicas e seus vetores latentes, entretanto que falhe durante a reconstrução de instâncias consideradas anômalas.

Apresentamos o processo de treinamento de modo detalhado:

- **Rede Geradora** – A rede G_E captura e aprende a distribuição dos dados de entrada x (somente típicos) e os mapeia para representações latentes z, tal que $G_E : x \rightarrow z$, onde $x \in R^{w \times h}$ e $z \in R^d$.
- **Rede Discriminadora** – Classifica as instâncias recebidas. Neste contexto, é realizada a classificação das imagens reais (x) e das imagens geradas no passo anterior (x'). A pesar ser um classificador, esta rede também é utilizada como extratora de características capaz de aproximar as representações latentes entre um áudio de entrada e um áudio reconstruído. Esta classificação ocorre de maneira sintética, pois os rótulos de dados típicos e artificiais são sempre considerados como 0 e 1 respectivamente.

Os autores demonstram que a arquitetura falha em reconstruir instâncias anômalas, já que as reconstruções se limitam aos típicos. Para estes dados anômalos, são esperados valores de erro maiores na reconstrução da saída x' ou na representação do espaço latente z'. A validação é feita através de 3 funções de otimização (*Adversarial Loss*, *Contextual Loss* e *Latent Loss*). As funções de otimização *Adversarial Loss* e *Contextual Loss* podem ser observadas na Equação 4.1 e Equação 5.2 respectivamente.

Latent Loss. Em adição aos objetivos das funções de otimização anteriores, os autores propõe a reconstrução do espaço latente para as instâncias de entrada x e as instâncias geradas $\hat{x}$ com maior similaridade possível. Isso garante que a rede neural é capaz de reproduzir além das representações de entrada, bem como representações de características destas entradas. Conforme pode ser observado na Figura 4.12, é utilizado a camada final do discriminador D, capaz de extrair as características de x e $\hat{x}$ tal que $z = f(x)$ e $\hat{z} = f(\hat{x})$. Essa função de otimização de aproximação das características latentes da entrada pode ser descrita conforme a Equação 5.3.

$$\mathcal{L}_{lat} = E_{x \sim p_x} \| f(x) - f(\hat{x}) \|_2 \tag{4.5}$$

Por fim, a função total de otimização $\mathcal{L}_{sganomaly}$ é dada pela Equação 4.6

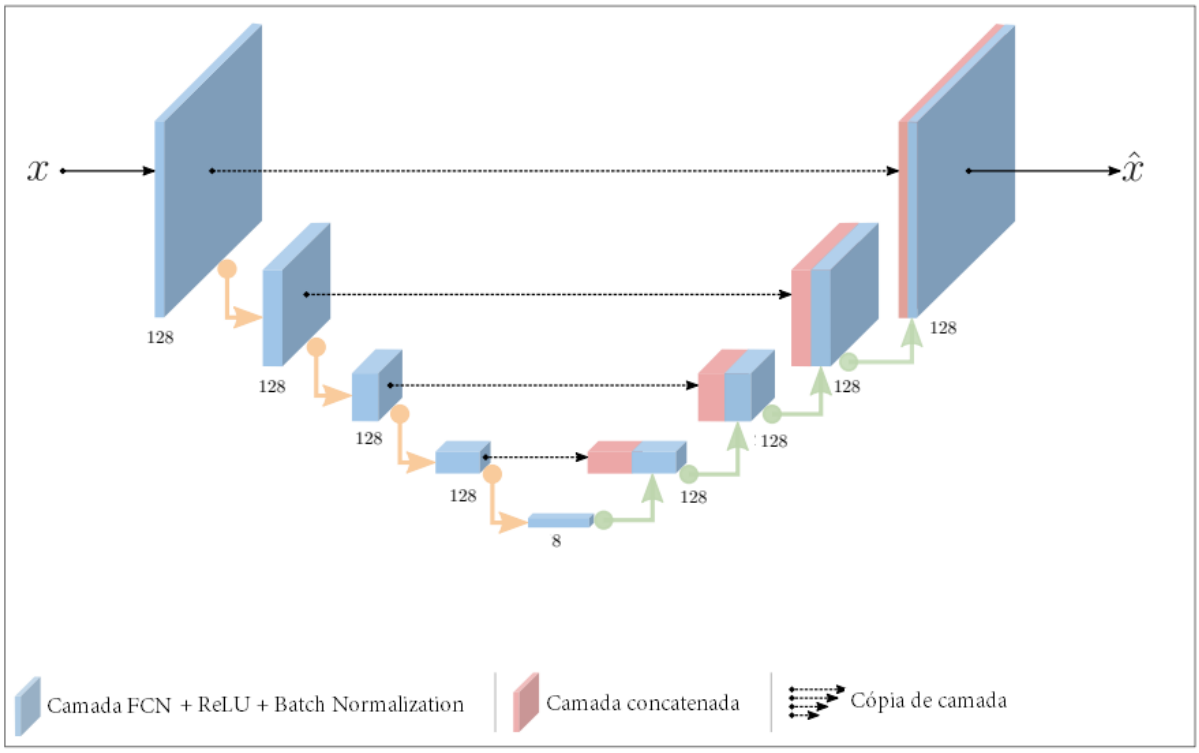


Figura 4.13 – Arquitetura adaptada da rede geradora Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection. Fonte: PRÓPRIO AUTOR

$$\mathcal{L}_{sganomaly} = \lambda_{adv} \times \mathcal{L}_{adv} + \lambda_{con} \times \mathcal{L}_{con} + \lambda_{lat} \times \mathcal{L}_{lat} \tag{4.6}$$

Os valores de λ_{adv}, λ_{con} e λ_{lat} são pesos atribuídos aos otimizadores arbitrariamente e serão apresentados na subseção 5.3.1.

4.3 CONSIDERAÇÕES FINAIS

No presente estudo, abordamos a detecção de anomalias em eventos sonoros, uma tarefa desafiadora que envolve a adaptação de arquiteturas de aprendizado profundo originalmente concebidas para o domínio de imagens. Durante todo o capítulo, exploramos as adaptações realizadas nessas arquiteturas. Inicialmente, destacamos a importância do pré-processamento de dados de áudio, padronizando instâncias para um formato uniforme que serviria como entrada para as redes GAN adaptadas. Isso possibilitou uma avaliação justa e padronizada das arquiteturas nos mesmos conjuntos de dados, permitindo a análise das otimizações e o desempenho geral das redes.

A adaptação das arquiteturas de GAN para o contexto de áudio foi influenciada pelo modelo proposto por Koizumi, Kawaguchi e Imoto (2020), que apresentou uma abordagem de rede neural simples para processamento de áudio. As modificações incorporadas nas arquiteturas dos Geradores e Discriminadores garantiram que essas redes fossem capazes

de lidar com dados de áudio, uma mudança fundamental para a detecção de anomalias em eventos sonoros. Essas adaptações, embora preservassem a essência das arquiteturas originais, abriram novas possibilidades para a aplicação desses modelos em um domínio até então pouco explorado.

Três arquiteturas específicas foram adaptadas e examinadas em detalhes neste estudo: EGBAD, GANomaly e SGANomaly. Cada uma dessas arquiteturas foi ajustada para a detecção de anomalias em áudio, incluindo mudanças nos Geradores, Discriminadores e Codificadores. A dinâmica de treinamento de cada arquitetura foi explicada, destacando como essas redes eram capazes de aprender a representação latente dos dados e reconstruir instâncias de áudio de maneira eficaz. As hipóteses subjacentes a esses modelos e sua capacidade de diferenciar instâncias típicas de anômalas foram discutidas com base nas funções de otimização específicas empregadas em cada arquitetura.

Em conclusão, este estudo proporcionou uma contribuição significativa para a detecção de anomalias em eventos sonoros por meio da adaptação de arquiteturas de aprendizado profundo.

5 EXPERIMENTOS E DISCUSSÕES

Este Capítulo descreve os experimentos realizados e os resultados obtidos. Para tanto, inicia-se com a descrição do protocolo experimental na seção 5.1. Na seção 5.2 explicamos as definições dos *baselines*, na seção 5.3 são apresentados os resultados obtidos e por fim, seção 5.4 sintetiza as considerações finais.

5.1 PROTOCOLO EXPERIMENTAL

Este seção apresenta uma visão geral de como os experimentos foram planejados, organizados e executados para analisar a capacidade de detectar anomalias em atividades sonoras. Para isso, apresentamos as configurações utilizadas nos experimentos na subseção 5.1.1, a visão geral da coleta de dados subseção 5.1.2, a métrica de detecção de anomalias na subseção 5.1.3, bem como as métricas de comparação para avaliar as arquiteturas na subseção 5.1.4 e por fim, a definição dos trabalhos de *baseline* na subseção 5.2

5.1.1 Configuração dos Experimentos

Todos os experimentos foram realizados em um servidor Intel Core i7-9700, 3.00GHz, 32GiB com sistema operacional Ubuntu 18.04.5 64bits. Este servidor possui uma placa GeForce RTX 2080 utilizando o CUDA 11.0. O modelo foi implementado usando a linguagem Python 3.9 e bibliotecas públicas de aprendizagem de máquina como Tensorflow 2.6.1 e Scikit-learn 1.0.1.

5.1.2 Coleta dos dados

Os conjuntos de dados mais populares utilizados pelos trabalhos descritos no Capítulo 3 são: *Audioset* por Gemmeke et al. (2017), *Industrial Sound* por Grollmisch et al. (2019), *ToyADMOS* por Koizumi et al. (2019) e *MIMII Dataset* por Purohit et al. (2019). Escolhemos para o estudo experimental apenas o *ToyADMOS* e *MIMII Dataset*, pois são utilizados pelos trabalhos de *baseline* 5.2. Esses conjuntos descrevem o funcionamento operacional de 6 máquinas industriais e brinquedos, isto é, inclui-se também o ruído do ambiente. Os dados são divididos entre típicos e anômalos, com aproximadamente 10 segundos cada. Entretanto, durante a fase de treinamento somente temos acesso aos dados de funcionamento típico, assim, o modelo precisa aprender a representação do funcionamento para na fase de testes identificar as anomalias. O seguintes tipos de máquinas e brinquedos utilizados nessa tarefa:

- Toy-car (ToyADMOS)

- Toy-conveyor (ToyADMOS)
- Valve (MIMII Dataset)
- Pump (MIMII Dataset)
- Fan (MIMII Dataset)
- Slide rail (MIMII Dataset)

Este conjunto de dados é amplamente utilizado em diversos trabalhos e competições para identificação de eventos anômalos: (AGRAWAL; MAURYA, 2020; DANILUK et al., 2020; GIRI et al., 2020; AHMED; OTHMAN; SALEM, 2021; LIU et al., 2022).

Os dados estão organizados em grupos identificados por identificadores (ID.), classes e propósito de utilização. Os identificadores são utilizados para representar diferentes tipos de equipamentos, os quais possuem distintos ruídos de fundo e foram gravados em momentos temporais diversos. A Figura 5.1 ilustra a estrutura do conjunto de dados, destacando a separação entre os dados de treinamento, que consistem exclusivamente em exemplos de funcionamento típico, e os dados de teste, que contêm exemplos anômalos.

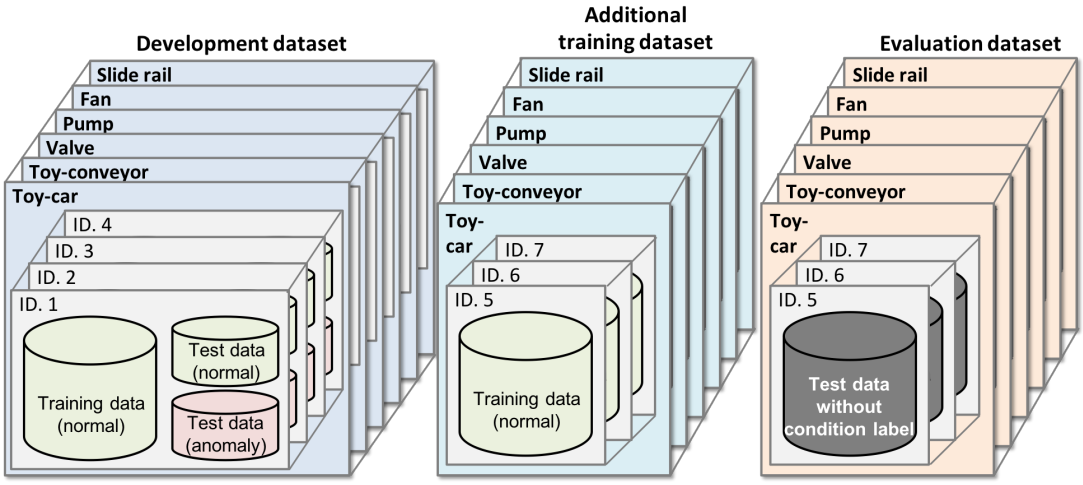


Figura 5.1 – Estrutura de organização da base dados. Fonte: (KOIZUMI; KAWAGUCHI; IMOTO, 2020)

Na Figura 5.2 são apresentados dois exemplos aleatórios de cada uma das seis classes. Observa-se que, mesmo pertencendo à uma mesma classe, os dados apresentam diferenças visuais devido às condições variáveis de ambiente e tempo durante a gravação dos espectrogramas. O objetivo deste trabalho é identificar anomalias de funcionamento dentro de uma mesma classe, a despeito da variação intra-classe.

Os dados apresentados na Tabela 5.1 mostram a distribuição das classes entre os conjuntos de treinamento e teste. Cada classe representa um tipo específico de objeto ou evento a ser detectado. Observa-se que as quantidades de exemplos de treinamento

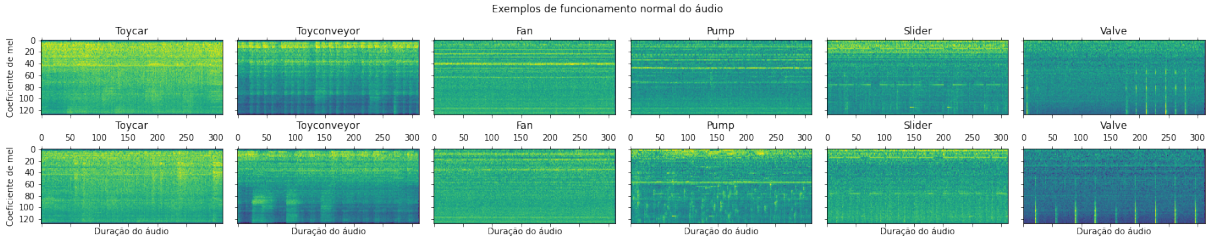


Figura 5.2 – Exemplo de áudios durante o funcionamento típico das máquinas. Fonte: PRÓPRIO AUTOR

e teste variam para cada classe. A classe *ToyCar* possui 4000 exemplos no conjunto de treinamento e 2459 exemplos no conjunto de teste. A classe *ToyConveyor* tem 3000 exemplos de treinamento e 3509 exemplos de teste. As demais classes, *fan*, *pump*, *slider* e *valve*, também apresentam variações na quantidade de exemplos entre os conjuntos de treinamento e teste. Estas variações são importantes para que se possa obter a métrica da média ponderada utilizada na avaliação geral do modelo.

Tabela 5.1 – Quantidade dos conjuntos de dados separados em treino e teste

Classes	Treino	Teste
ToyCar	4000	2459
ToyConveyor	3000	3509
fan	3675	1875
pump	3349	856
slider	2804	1290
valve	3291	879

5.1.3 Métricas de Detecção de Anomalias

No escopo deste trabalho, para que um dado seja classificado como anomalia, o modelo considerado não deve ser capaz de reconstruí-lo de maneira satisfatória, de modo que o resultado gerado possua diversos erros ao se comparar com a entrada. Diante disso, é necessário descrever um método para identificar e padronizar o cálculo de anomalias.

Neste contexto, utilizamos no conjunto de teste o método descrito por Akcay, Atapour-Abarghouei e Breckon (2018), que determina o **cálculo de anomalia** como uma representação numérica quantitativa para identificar os áudios anômalos. Desta maneira, a Equação 5.1 é descrita como:

$$A(x) = \lambda C(x) + (1 - \lambda)L(x), \tag{5.1}$$

na qual o **cálculo contextual** $C(x)$ e o **cálculo latente** $L(x)$ são definidos pelas Equação *contextual loss* $\mathcal{L}_{con}$ na Equação 5.2 e *latent loss* $\mathcal{L}_{lat}$ na Equação 5.3 respectivamente. A variável x representa um espectrograma do conjunto de testes, enquanto λ descreve

um parâmetro de peso que controla a importância das funções, por fim as variáveis z e $\hat{z}$ são respectivamente as representações dos espaços latente da entrada original e do dado gerado pelo Gerador. O modelo produz valores altos de **cálculo de anomalia** $A(x)$ para instância anômalas, enquanto baixos valores para instâncias típicas. A escala dos valores pode variar de acordo com a classe e a arquitetura, entretanto, valores baixos estão próximos de 0 e 1 e valores altos estão acima de 1.

$$\mathcal{L}_{con} = \|x - G(x)\|_1^1 \tag{5.2}$$

$$\mathcal{L}_{lat} = \|z - \hat{z}\|_2 \tag{5.3}$$

O **cálculo de anomalia** descrito na Equação 5.1 é um vetor que contém os escores conjunto de teste, seus valores são variados e sem escalas determinadas. Para isso, os autores Akcay, Atapour-Abarghouei e Breckon (2018) também descrevem um método de padronizar as anomalias na escala $[0, 1]$, logo, o **cálculo de anomalia final** $\hat{A}$ de um instância de entrada x é conforme a Equação 5.4:

$$\hat{A} = \frac{A(x) - min(A)}{max(A) - min(A)} \tag{5.4}$$

5.1.4 Métricas de comparação

A Figura 5.3 representa uma matriz de confusão binária. Nela é possível entender a viabilidade do modelo, bem como sua performance no experimento. A divisão de classes é feita entre a classificação produzida pelo modelo de aprendizagem e classe verdadeira. Em nosso contexto podemos traduzir as duas classes para: anômala e não-anômala. A matriz de confusão torna-se estaticamente confiável através de várias separações aleatórias entre os conjuntos de dados de treino e teste.

		Classe Predita	
		Positivo	Negativo
Classe Verdadeira	Positivo	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 5.3 – Matriz de Confusão Binária. Fonte (DUARTE, 2021)

A seguir há alguns termos básicos demonstrados na matriz de confusão:

- **Verdadeiro Positivo - VP** - É a quantidade de dados anômalos que o modelo conseguiu prever corretamente.
- **Falso Positivo - FP** - É a quantidade de dados identificados como anômalos, mas que são não-anômalos.
- **Falso Negativo - FN** - É a quantidade de dados considerados não-anômalos, mas que são anômalos.
- **Verdadeiro Negativo - VN** - É a quantidade de dados não-anômalos previstos corretamente pelo modelo.

A partir dos termos descritos anteriormente, é possível avaliar e comparar o modelo de aprendizagem através de várias métricas. Portanto, neste trabalho serão utilizadas as métricas de: *Area Under the ROC Curve* e *Partial Area Under the ROC Curve*, calculada a partir de:

- ***True Positive Rate* - TPR** – Também conhecido como *recall*, essa taxa mede a porcentagem da identificação correta feita pelo modelo de aprendizagem dentre todas possíveis anomalias. Pode ser calculada a partir da Equação 5.5:

$$FAR = \frac{VP}{VP + FN} \tag{5.5}$$

- ***False Positive Rate* - FPR** – Também conhecida como *fall-out* Essa taxa mede a porcentagem de dados identificados como anômalos dentre todos os não-anômalos. Pode ser calculada a partir da seguinte Equação 5.6:

$$FRR = \frac{FP}{FP + VN} \tag{5.6}$$

- ***ROC Curve*** – Representação gráfica de uma curva probabilística que varia através de um limiar, ou seja, um gráfico que analisa a performance de uma classificação binária feita por um modelo de aprendizagem dado um limiar para determinar se a classe prevista é anômala ou não-anômala. A curva ROC utiliza os valores de TPR contra FPR em diferentes limiares de classificação (FAWCETT, 2006).
- ***Area Under the ROC Curve* - AUC** – Avalia a separabilidade entre as classes e a probabilidade do modelo de aprendizagem possuir maior confiança na classe prevista (FAWCETT, 2006).
- ***Partial Area Under the ROC Curve* - pAUC** – É o cálculo da proporção da curva ROC sobre um intervalo de interesse. Lida principalmente com o desbalanceamento de classes, dando mais ênfase em **TPR**. Em nosso trabalho, a métrica

será calculada sobre a baixa quantidade de FPR. Nesta tarefa, utilizaremos p=0.1 conforme definida por (KOIZUMI; KAWAGUCHI; IMOTO, 2020).

Essas métricas estão relacionadas com a matriz de confusão que é uma ferramenta importante na coleta dessas informações e na avaliação de modelos de aprendizagem de máquina. A utilização de duas métricas para avaliar cada classe se dá7 pelo fato de entender se o modelo produz alerta falsos com frequência, logo, pouco confiável.

5.2 DEFINIÇÃO DE *BASELINES*

Comparamos a nossa abordagem com *baselines* no estado-da-arte baseados em GANs no domínio de imagens para detectar eventos de áudio anômalos. Nossa comparação focou em abordagens GANs por possuírem resultados significativamente melhores em conjunto de dados de imagens, também apresentamos uma abordagem de AE fornecida pelos autores dos conjuntos de dados em questão. Além disso, é importante observar que nenhum dos esforços anteriores baseados em GAN possuem comparação direta usando esse tipo de conjunto de dados, ou seja, não demonstram suas soluções aplicadas em domínios diferentes do proposto. Portanto, apresentaremos as soluções escolhidas no domínio de áudio. Os *baselines* escolhidos foram:

- **DCASE Baseline**: O *baseline* proposto por Koizumi, Kawaguchi e Imoto (2020), conforme descrito na seção 3.2 apresenta o limite inferior para nossa comparação. Esta margem nos guia para identificar a validade das alterações descritas na subseção 4.2.1 realizadas nos modelos apresentados na seção 4.2.
- **GMADE**: *Unsupervised Anomalous Sound Detection Using Self-Supervised Classification And Group Masked AutoEncoder For Density Estimation* proposto por Giri et al. (2020), é composta por um *ensamble* de redes capazes de identificar anomalias em áudio. O trabalho propõe a utilização da rede MADE (*Masked AutoEncoder for Density Estimation*), uma arquitetura baseada em auto-codificadores capaz de remover conexões de pesos entre camadas e se tornar um estimador de densidade. Os autores propõe o método de treinamento semi-supervisionado, no qual utilizam-se os dados de áudio como a identificação (ID) como rótulo. Esta técnica de treinamento depende de meta-dados e isto torna a solução específica para o conjunto de dados mencionado na subseção 5.1.2. Portanto, embora o método seja usado na avaliação comparativa, serve de base como limite superior, assumindo a existência de informações adicionais não disponíveis normalmente em problemas de detecção de anomalia (KOIZUMI; KAWAGUCHI; IMOTO, 2020).

Cada implementação seguiu rigorosamente os parâmetros e configuração referente a arquitetura fornecida, com apenas a alteração da rede Geradora e Discriminadora. Em

relação aos hiperparâmetros como o número de épocas, treinamo-os todos sob as mesmas condições, conforme apresentado a seguir.

5.3 RESULTADOS

Nesta seção serão apresentados os resultados e as considerações obtidos dos experimentos realizados pelo método proposto. Os experimentos foram agrupados em cinco tópicos utilizando o conjunto de dados *ToyADMOS* e *MIMII Dataset*. Os tópicos visam atender os objetivos geral e específicos descritos na seção 1.3:

1. Descrição dos experimentos na subseção 5.3.1.
2. Avaliação de arquiteturas adaptadas e otimizadas da subseção 5.3.2 até a subseção 5.3.4.
3. Discussão resultados *baseline* da subseção 5.3.5 até a subseção 5.3.6.
4. Resultados gerais na subseção 5.3.7.
5. Avaliação de técnicas de aumento de dados nestas adaptações na subseção 5.3.8.

O primeiro tópico aborda sobre a experimentação de diferentes funções de otimização no domínio de sons, isto inclui: avaliação de otimizadores e avaliação de cálculo de otimização, inclui-se também a descrição dos hiper-parâmetros utilizados para obtenção das arquiteturas otimizadas. O segundo tópico apresenta a avaliação das arquiteturas adaptadas os melhores resultados obtidos. O terceiro tópico discute os resultados dos limites inferiores e superiores. O quarto tópico aborda os resultados gerais entre diferentes arquiteturas. Por fim, o quinto tópico apresenta um comparativo das melhores arquiteturas adaptadas e a utilização da técnica de aumento de dados.

5.3.1 Experimentos

Os experimentos tiveram como objetivo verificar a eficácia de adaptações nos modelos GANs mencionados na seção 4.2. As arquiteturas foram avaliadas utilizando os dados descritos na etapa de coleta de dados segundo a subseção 5.1.2. Desta maneira, cada arquitetura possui um modelo único por classe, ou seja, as avaliações são feitas utilizando os dados de teste de apenas uma classe por modelo. De modo que possa garantir o aprendizado do funcionamento típico de cada maquinário industrial.

Cada classe possui particularidades, bem como os modelos que realizam a aprendizagem. Por isso, é esperado que uma arquitetura tenha resultados variados em diferentes tipos de áudio. Vale ressaltar que as adaptações propostas na subseção 4.2.1 utilizam somente as instâncias de áudio, sem a adição de metadados para o treinamento, visto que,

este trabalho tem como objetivo demonstrar a eficiência das adaptações em conjuntos de dados do mundo real.

A escolha adequada dos hiperparâmetros pode impactar significativamente a capacidade de generalização e o desempenho do modelo. Geralmente, os hiperparâmetros são ajustados por meio de tentativa e erro ou por métodos mais avançados, como busca exaustiva, busca em grade ou otimização *bayesiana*, a fim de encontrar a configuração que maximize o desempenho do modelo para uma determinada tarefa (RAGAB et al., 2021).

Os modelos adaptados foram instanciados com os melhores parâmetros obtidos pela metodologia de busca em grade ou *grid search*. A particularidade das classes demandam hiper-parâmetros diferentes para obter melhores resultados. Considerando os seguintes espaços de busca de hiper-parâmetros, temos a Tabela 5.2:

Tabela 5.2 – Tabela de hiper-parâmetros utilizados nestes experimentos

Parâmetros	**Intervalos**
Número de épocas	10, 20, 30 e 90
Taxa de aprendizagem	0.0002, 0.001 e 0.005
Dimensão do espaço latente	8, 16 e 100
Valores de regularização (*dropout*)	0, 0.25 e 0.5
Iniciadores de pesos	*Glorot Uniform* e *Random Normal*
Termo de impulso	0.3 e 0.5
Número de camadas ocultas	3 e 4
λ_{adv}	1 e 50
λ_{con}	1 e 50
λ_{enc}	1 e 50
λ_{lat}	1 e 50
Tamanhos de batch	8, 16, 32, 128, 512 e 1024

Também avaliamos diferentes otimizadores. Cada otimizador possui seus prós e contras. Alguns fatores considerados na seleção de um otimizador incluem a convergência do treinamento, a velocidade de convergência, a capacidade de lidar com diferentes escalas de gradiente e a eficiência computacional. A seleção cuidadosa do otimizador apropriado pode levar a um treinamento mais rápido e a melhores resultados de desempenho para o modelo (SUTSKEVER et al., 2013). Neste trabalho, foram avaliados os seguintes otimizadores: Adamax (KINGMA; BA, 2014), Adam (REDDI; KALE; KUMAR, 2018) e AdamW (LOSHCHILOV; HUTTER, 2019). Combinados com as variações de valores λ descritas na subseção 5.1.4: 0, 0.5 e 1. A Tabela 5.3 mostra as funções de otimização e os valores λ experimentados.

Através da seleção dos melhores parâmetros, alguns se tornaram constantes e por isso não farão parte das demais tabelas de hiper-prâmetros de cada arquitetura. São eles: Taxa de aprendizagem (L.R.) igual a 0.0002, dimensão do espaço latente igual a 8, função de otimização Adam, valor λ igual 1.0, iniciador de peso *Glorot Uniform*, valor de

Tabela 5.3 – Relacionamento entre funções de otimização e valores λ

Funções de Otimização	Valores λ
Adamax	0, 0.5 e 1.0
Adam	0, 0.5 e 1.0
AdamW	0, 0.5 e 1.0

regularização (*dropout*) igual a 0 e termo de impulso (*Beta 1*) igual a 0.5.

5.3.2 Avaliação da arquitetura adaptada GANomaly

A arquitetura adaptada GANomaly (AKCAY; ATAPOUR-ABARGHOUEI; BRECKON, 2018) obteve bons resultados após a busca exaustiva de parâmetros, conforme apresentada na Figura 5.4. O gráfico de barras contém o desempenho desta arquitetura para todas as classes. Podemos observar os resultados acima de 90% em AUC nas classes *Toycar* e *fan*, enquanto todas as classes restantes possuem métricas de AUC acima ou iguais a 75%. Isso indica que o modelo foi capaz de identificar e distinguir corretamente os dados anômalos dos típicos.

A métrica pAUC também mostrou-se satisfatória, visto que a maioria das classes obteve resultados acima de 75%. Com relação aos rótulos *pump* e *valve*, nota-se os resultados mais baixo. Embora estejam acima de 50%, os dados apontam que a rede neural não foi capaz de generalizar os casos onde o ruído anômalo se concentra em frequências próximas demais dos sons típicos.

A Tabela 5.4 apresenta os melhores parâmetros utilizados para cada classe de áudio, nela podemos observar pouca variação comparada as demais arquiteturas adaptadas. Entende-se que esses hiper-parâmetros estão correlacionados ao modo de treinamento da rede, pois certas classes obtêm melhores performances ao analisar poucos exemplos de lote.

Tabela 5.4 – Melhores hiper-parâmetros definidos para cada classe na arquitetura adaptada GANomaly

Classes	**Épocas**	**H.L.**	λ_{adv}	λ_{con}	λ_{enc}	**B.S.**
Toycar	90	4	1	50	1	512
ToyConveyor	30	4	1	50	1	512
fan	10	3	1	50	1	16
pump	10	3	50	1	50	1024
slider	10	3	1	50	1	16
valve	10	3	1	50	1	16

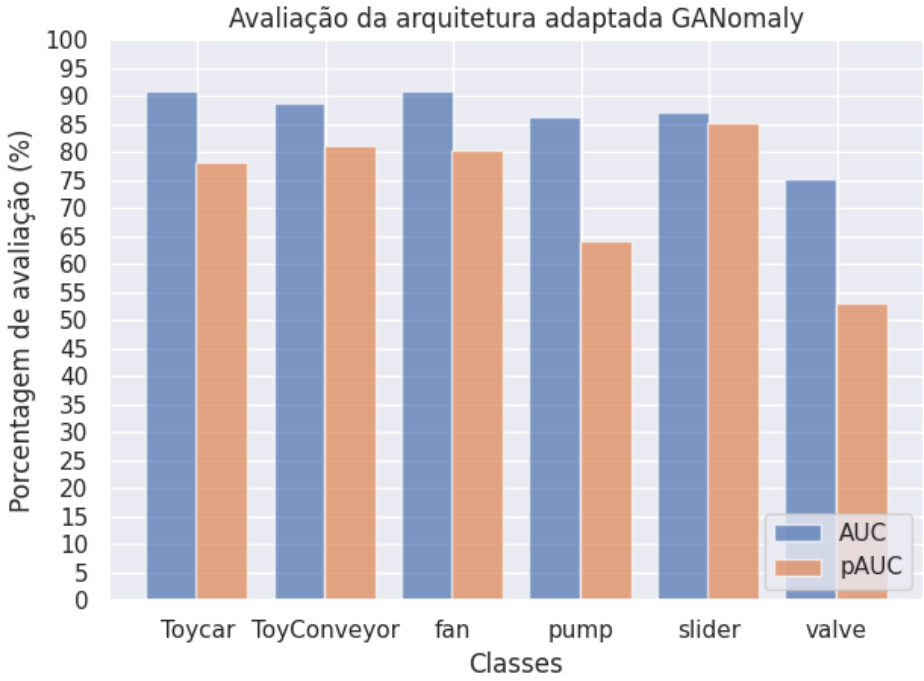


Figura 5.4 – Avaliação da arquitetura adaptada GANomaly (AKCAY; ATAPOUR-ABARGHOUEI; BRECKON, 2018). Fonte: PRÓPRIO AUTOR

5.3.3 Avaliação da arquitetura adaptada SGANomaly

A Figura 5.5 demonstra um gráfico de barras contendo os resultados da arquitetura adaptada Skip-GANomaly (AKCAY; ABARGHOUEI; BRECKON, 2019) em todas as classes. Esta arquitetura apresenta resultados acima de 75% e 64% na métricas AUC e pAUC respectivamente, para todas as classes. Destaca-se o rótulo *slider* e *fan* como o maior resultado em AUC e pAUC na devida ordem. Pode-se afirmar que a rede neural aprendeu a reconhecer corretamente o áudios anômalos e típicos.

A Tabela 5.5 exibe os melhores parâmetros escolhidos para cada classe de áudio. Nota-se que esta arquitetura possui diversas variações de parâmetros para cada classe. Por conta de sua estrutura residual, a arquitetura demanda de mudanças específicas no tamanho do lote e nos pesos atribuídos às funções de otimização.

Tabela 5.5 – Melhores hiper-parâmetros definidos para cada classe na arquitetura adaptada SGANomaly

Classes	**Épocas**	**H.L.**	λ_{adv}	λ_{con}	λ_{lat}	**B.S.**
Toycar	20	3	1	1	50	128
ToyConveyor	30	4	1	50	1	512
fan	10	3	1	50	1	16
pump	10	3	50	1	50	1024
slider	10	3	50	50	50	512
valve	30	4	1	50	1	512

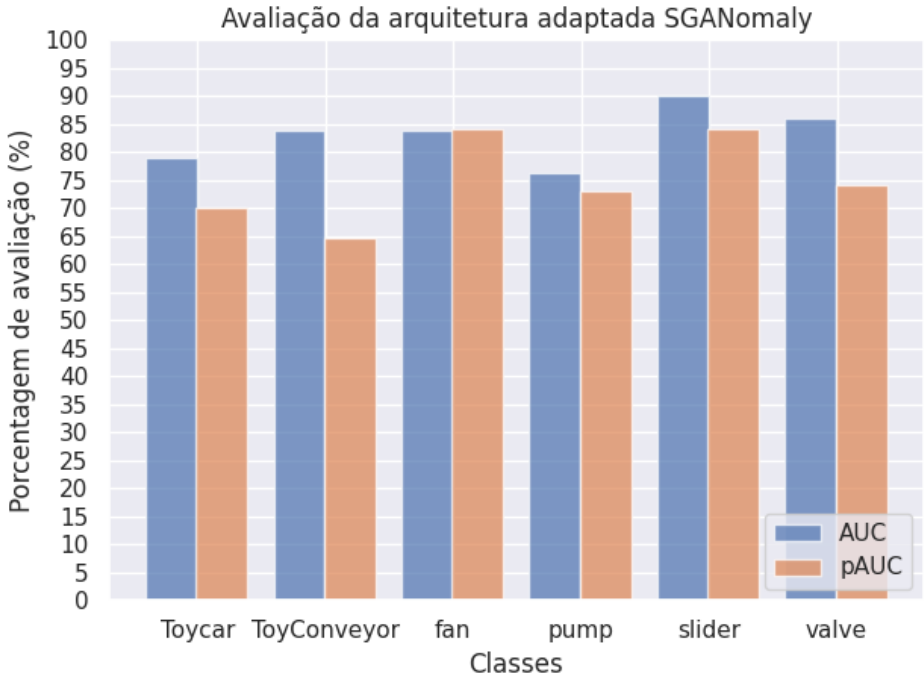


Figura 5.5 – Avaliação da arquitetura Skip-GANomaly (AKCAY; ABARGHOUEI; BRECKON, 2019) Fonte: PRÓPRIO AUTOR

5.3.4 Avaliação da arquitetura adaptada EGBAD

O gráfico de barras na Figura 5.6 apresenta os resultados da arquitetura adaptada EGBAD (ZENATI et al., 2018) para todos os rótulos. A arquitetura obteve sucesso na identificação apenas na classe *valve* com as métricas AUC e pAUC de 69% e 57%, respectivamente. Observa-se que o desempenho nas demais classes não foi satisfatório, com resultados abaixo do limiar de 50% em classes como *ToyConveyor*, *pump* e *fan*. Isso indica que o modelo não foi capaz de distinguir entre áudios normais e anômalos nessas classes.

Esta arquitetura obteve os piores resultados por conta dos vetores latentes aleatórios também serem considerados na avaliação do modelo Discriminador, assim, o vetor latente do espectrograma tende a sempre possuir mais erros nesta comparação direta. Influenciando o cálculo de anomalias descrito na subseção 5.1.3. Por possuir resultados inferiores ao *baseline*, esta arquitetura não foi eleita para fazer mudanças de hiper-parâmetros.

5.3.5 GMADE

A Figura 5.7 apresenta os resultados publicados pelos autores Giri et al. (2020), na qual apresenta os maiores resultados da competição publicada por Koizumi, Kawaguchi e Imoto (2020). Esta arquitetura possui resultados relevantes para a tarefa de detecção de anomalia, já que 3 rótulos apresentam métricas de AUC e pAUC acima de 90% e 78% de modo respectivo. Observa-se também que todas as métricas estão pelo menos em 65%, que sugere a alta habilidade desta arquitetura em identificar as anomalias.

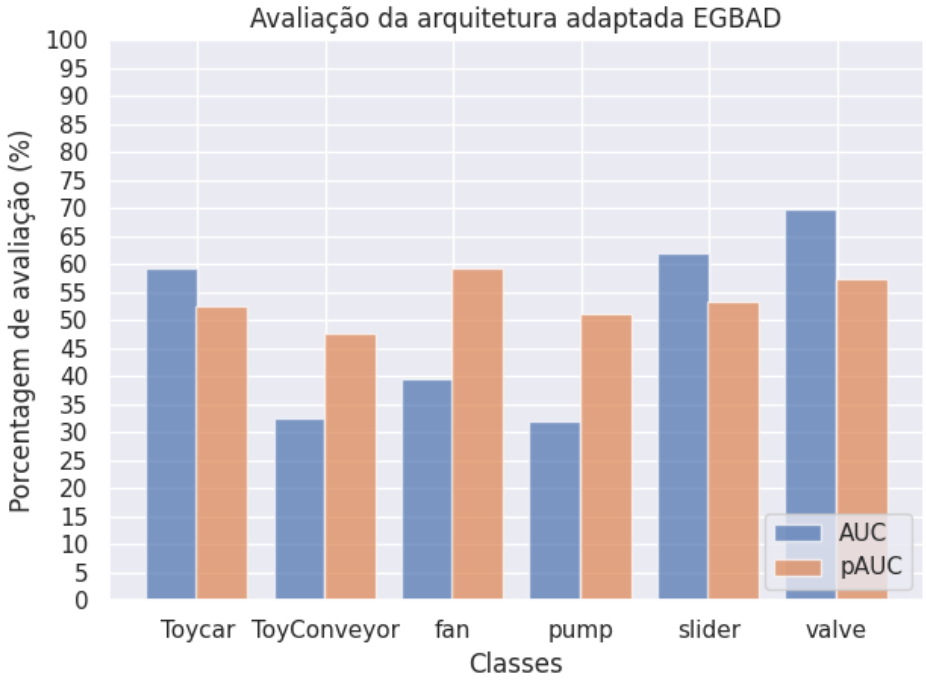


Figura 5.6 – Avaliação da arquitetura adaptada EGBAD (ZENATI et al., 2018) Fonte: PRÓPRIO AUTOR

É importante ressaltar que esta abordagem não possui disponibilidade de código e apresenta uma solução composta por diversos modelos como um comitê, ou seja, diversos modelos são combinados a fim de apresentar o melhor resultado. Isso o torna um modelo custoso computacionalmente e dependente e de várias soluções que não são intrinsecamente baseadas em GANs. Além disso, esta arquitetura utiliza metadados como rótulos, tornando a tarefa não-supervisionada em semi-supervisionada. Os resultados apresentados foram extraídos do trabalho proposto por Giri et al. (2020).

5.3.6 *Baseline*

Conforme mencionado na seção 5.2, esta abordagem é considerada o limite inferior, na qual servirá para identificar se os modelos de treinamento adversários adaptados são capazes de distinguir as anomalias dos áudios típicos. A Figura 5.8 ilustra um gráfico de barras do apresentado por Koizumi, Kawaguchi e Imoto (2020). Nela podemos observar inicialmente que todas as métricas estão acima de 50%, ou seja, ouve um aprendizado consistente. Dentre os rótulos, destaca-se o *slider* e *Toycar* pois possuem os valores mais altos da métrica AUC e pAUC sendo elas 80% e 67% para o primeiro rótulo e 84% e 66% para o segundo.

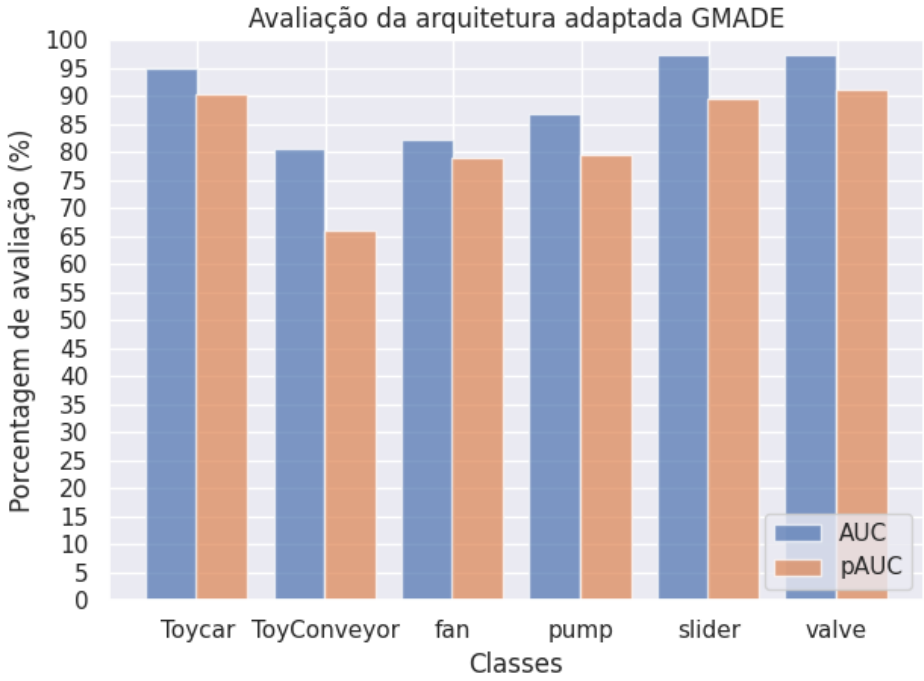


Figura 5.7 – Avaliação da arquitetura GMADE (AKCAY; ABARGHOUEI; BRECKON, 2019) Fonte: PRÓPRIO AUTOR

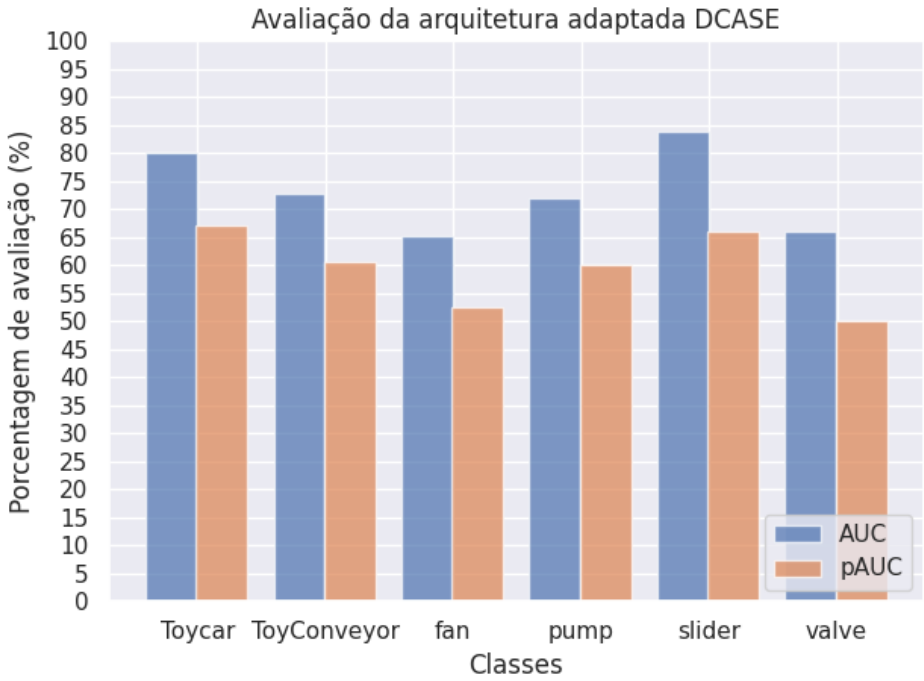


Figura 5.8 – Avaliação da arquitetura *baseline* (KOIZUMI; KAWAGUCHI; IMOTO, 2020) Fonte: PRÓPRIO AUTOR

5.3.7 Resultados gerais

As médias aritméticas dos experimentos realizados podem ser observados na Figura 5.9. Nela, são apresentados as médias e desvio padrão para todos os modelos, considerando

todas as classes presente no experimento. Com isso, podemos comparar os modelos de forma geral. O modelo DCASE é o guia para identificar a performance dos demais modelos, pois para que a solução seja considerada eficiente, deve obter dados superiores através da técnica de treinamento adversária. O modelo DCASE possui o resultado de $73,32\%$ e $59,42\%$ nas métricas AUC e pAUC, nesta ordem. Adicionalmente exibe $6,86\%$ e $6,34\%$ de desvio padrão em AUC e pAUC respectivamente. Conforme observa-se na Tabela 5.6

Tabela 5.6 – Média aritmética dos experimentos realizados

Arquiteturas	Métricas			
	AUC	Desvio-Padrão AUC	pAUC	Desvio-Padrão pAUC
DCASE	73,32%	6,86%	59,41%	6,34%
GMADE	**89,94%**	6,93%	**82,60%**	8,98%
GANomaly	86,52%	5,42%	75,54%	11,30%
SGANomaly	86,89%	4,45%	74,86%	7,09%
EGBAD	49,17%	15,10%	53,58%	3,86%

Nota-se que através que o modelo GMADE obteve maior resultado na métrica AUC e pAUC que as demais arquiteturas, mesmo com os melhores hiper-parâmetros da arquitetura adaptada GANomaly e SGANomaly. Entretanto, a média aritmética não leva em consideração a quantidade de elementos no conjunto de dados de validação. O modelo GMADE possui resultado de $89,94\%$ de AUC e $82,60\%$ de pAUC, com os desvios-padrão de $6,93\%$ e $8,89\%$ de AUC e pAUC na devida ordem.

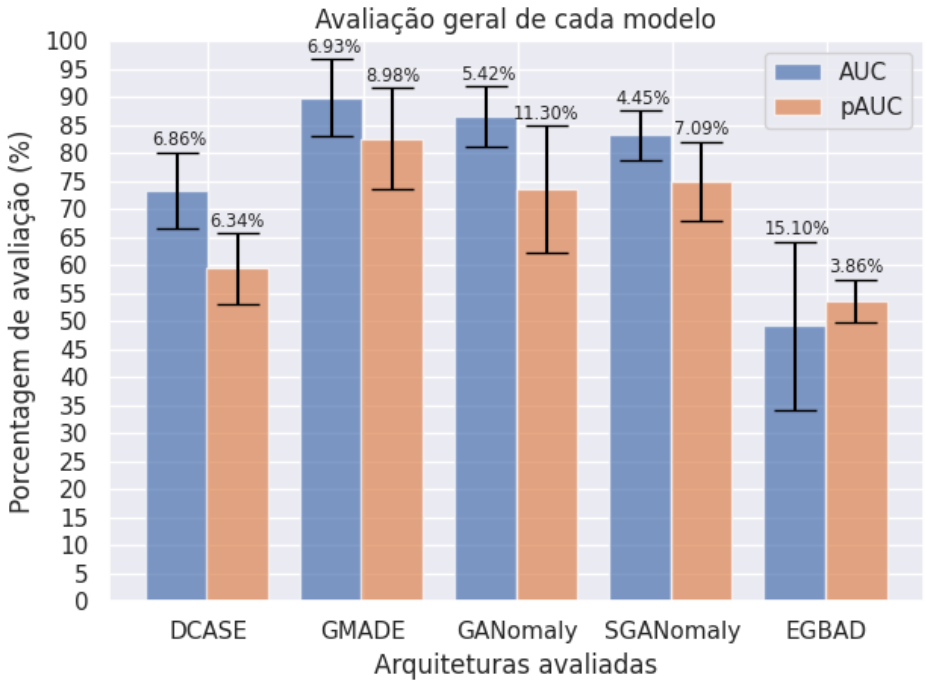


Figura 5.9 – Avaliação da média aritmética para cada modelo. Fonte: PRÓPRIO AUTOR

A Figura 5.10 demonstra a média ponderada utilizando a quantidade de dados no conjunto de validação. Esse dado é importante, pois resultados mais altos em conjuntos de dados bastante populosos possuem maiores confianças estatísticas. Através desta avaliação, observa-se que a arquitetura adaptada GANomaly possui resultados superiores ao modelo GMADE na métrica AUC, demonstrando que a quantidade de dados é um fator importante de avaliação. Verifica-se também que o desvio padrão da métrica AUC nos modelos GANomaly e SGANomaly são menores que o modelo GMADE, respectivamente $4,19\%$ e $3,73\%$.

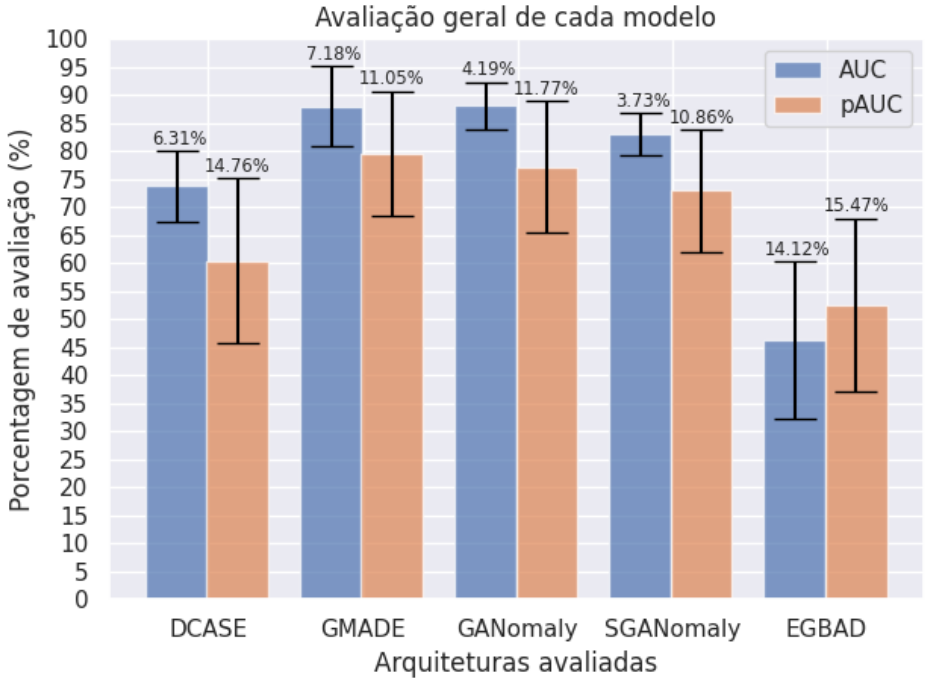


Figura 5.10 – Avaliação da média ponderada para cada modelo. Fonte: PRÓPRIO AUTOR

A Tabela 5.7 apresenta as médias ponderadas das métricas AUC e pAUC, bem como seus respectivos desvios-padrão para todas as arquiteturas avaliadas. Podemos observar que os modelos GANomaly e SGANomaly obtiveram resultados da métrica AUC acima do *baseline* oferecido pelo DCASE. Entretanto, o modelo GANomaly apresentou os melhores resultados gerais. Isso se deve ao fato de que a comparação entre os vetores latentes são melhores em reconstruir o contexto de um áudio.

Destacamos que os modelos GANomaly, SGANomaly e EGBAD são arquiteturas adaptadas para áudio, já que sua concepção foi feita para imagens. Embora as arquiteturas possuam algumas alterações, suas funções de otimização e estrutura foram mantidas originais, conforme mencionado na seção 4.2.

Na Tabela 5.8 demonstra-se que as arquiteturas adaptadas GANomaly e SGANomaly possuem resultados superando o limite superior nas classes *ToyConveyor*, *fan* e *pump* nas métricas AUC e pAUC.

Tabela 5.7 – Média ponderada dos experimentos realizados

Arquiteturas	Métricas			
	AUC	Desvio-Padrão AUC	pAUC	Desvio-Padrão pAUC
DCASE	73,81%	6,31%	60,47%	14,76%
GMADE	88,02%	7,18%	**79,63%**	11,05%
GANomaly	**88,16%**	4,19%	78,05%	11,77%
SGANomaly	86,11%	3,73%	72,64%	10,86%
EGBAD	46,23%	14,12%	52,55%	15,47%

Tabela 5.8 – Comparativos de baseline

Trabalhos	ToyCar		ToyConveyor		fan		pump		slider		valve	
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
DCASE	80,09%	67,22%	72,68%	60,65%	65,15%	52,59%	72,00%	60,00%	84,00%	66,00%	66,00%	50,00%
GMADE	**95,04%**	**90,39%**	80,67%	65,90%	82,33%	78,97%	86,94%	**79,60%**	**97,28%**	**89,54%**	**97,38%**	**91,21%**
GANomaly	91,00%	78,10%	**88,80%**	**81,20%**	**91,00%**	80,00%	86,19%	75,81%	87,16%	85,11%	75,00%	53,00%
SGANomaly	87,41%	67,42%	83,90%	64,60%	84,00%	**84,00%**	**89,52%**	71,04%	90,49%	88,09%	86,00%	74,00%
GMADE	59,40%	52,60%	32,40%	47,70%	39,50%	59,40%	31,90%	51,10%	62,00%	53,40%	69,80%	57,30%

5.3.8 Experimentos adicionais com aumento de dados

Avaliamos três diferentes técnicas de aumento de dados voltados para domínio de áudio. São eles: Inversão de polaridade, Ganho de potência randômica e Adição de ruído (PARK et al., 2019), respectivamente com os intervalos de: $(-1, 1)$, $(1.0, 1.5)$ e $(0.1, 0.5)$.

A Tabela 5.9 apresenta o comparativo entre a otimização dos modelos adaptados e a adição de aumento de dados. Pode-se observar que as classes *ToyConveyor*, *fan*, *pump* e *slider* apresentam resultados bem abaixo dos modelos otimizados. Isso demonstra que são rótulos muito sensíveis e que a inclusão de aumento de dados pode aumentar o espaço amostral do comportamento típico, levando o modelo a avaliar instâncias anômalas como típicas. Entretanto, as classes *ToyCar* e *Valve* obtiveram respectivamente valores de pAUC e AUC maiores.

Tabela 5.9 – Comparativo dos modelos com a inclusão do aumento de dados no treinamento

Trabalhos	ToyCar		ToyConveyor		fan		pump		slider		valve	
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
GANomaly(A.D.)	90,6%	**78,80%**	85,90%	73,60%	75,20%	55,90%	73,70%	56,00%	86,60%	79,20%	**81,80%**	50,90%
GANomaly	**91,00%**	78,10%	**88,80%**	**81,20%**	**91,00%**	**80,00%**	**86,19%**	**75,81%**	**87,16%**	**85,11%**	75,00%	**53,00%**
SGANomaly(A.D.)	84,95%	**67,80%**	78,66%	62,11%	71,92%	56,89%	74,78%	56,81%	73,90%	56,97%	82,90%	56,65%
SGANomaly	**87,41%**	67,42%	**83,90%**	**64,60%**	**84,00%**	**84,00%**	**89,52%**	**71,04%**	**90,49%**	**88,09%**	**86,00%**	**74,00%**

Além disso, as atividades sonoras geralmente possuem características intrínsecas que são sensíveis a essas distorções, como padrões de frequência e variações temporais. É importante buscar um equilíbrio entre o aumento da quantidade de dados e a preservação das características relevantes das atividades sonoras. Para isso, experimentamos os valores mais baixos e mais altos do intervalos.

5.4 CONSIDERAÇÕES FINAIS

No decorrer deste estudo, foram investigadas diversas arquiteturas de detecção de anomalias em eventos acústicos. Os resultados obtidos demonstram que as abordagens baseadas em redes GANs têm mostrado um grande potencial para lidar com a detecção de anomalias em áudio. A utilização de arquiteturas como GANomaly, SGANomaly e EGBAD permitiu a identificação de anomalias em diferentes contextos de maquinários industriais. Essas arquiteturas apresentaram desempenho promissor, alcançando métricas de avaliação significativas, como AUC e pAUC. No entanto, é importante ressaltar que ainda há espaço para aprimoramentos, especialmente no que diz respeito à detecção de anomalias em classes específicas, como *ToyConveyor*, *fan* e *pump*.

Durante o processo de experimentação, foi possível observar a importância dos hiperparâmetros na obtenção de resultados satisfatórios. A escolha adequada dos hiperparâmetros, como taxa de aprendizado, tamanho do *batch* e número de épocas, desempenha um papel crucial no desempenho das arquiteturas de detecção de anomalias.

Além disso, a utilização de técnicas de aumento de dados, como ganho de potencia e adição de ruído, mostraram-se benéficas para melhorar a capacidade de generalização dos modelos nas classes *ToyCar* e *Valve*. Portanto, é essencial realizar uma análise criteriosa dos hiperparâmetros e explorar estratégias de aumento de dados para otimizar o desempenho das arquiteturas propostas.

Considerando os resultados obtidos, é importante ressaltar que a detecção de anomalias em eventos acústicos ainda apresenta desafios a serem enfrentados. Embora as arquiteturas de detecção de anomalias baseadas em GANs tenham mostrado resultados promissores, é necessário um maior aprofundamento nas estratégias de treinamento e avaliação. Além disso, é fundamental explorar abordagens complementares, como o uso de técnicas de pré-processamento de áudio e a combinação de múltiplas arquiteturas, a fim de aprimorar ainda mais a detecção de anomalias. Essas considerações podem abrir caminho para futuras pesquisas e contribuições na área de detecção de anomalias em áudio, visando aprimorar a eficiência e a precisão dos modelos desenvolvidos.

Por tanto, nestes experimentos, mostrou-se que modelos desenvolvidos para identificar anomalias em imagens podem ser utilizados também em áudio com algumas adaptações, considerando conjuntos de dados muito populares como DCASE. Por tanto, a utilização de GANs apresenta resultados superiores ao *baseline* oferecido pela competição DCASE (KOIZUMI; KAWAGUCHI; IMOTO, 2020) e ao modelo GMADE (GIRI et al., 2020).

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho evidenciou a viabilidade da utilização de modelos baseados em GAN para detecção de anomalias em áudios. Particularmente, foram adaptados 3 modelos GANs para detecção de anomalias de imagens para áudio, afim de destacar sua eficácia em eventos acústicos. Esse modelos foram avaliados no conjunto de dados populares de sons como o DCASE (KOIZUMI; KAWAGUCHI; IMOTO, 2020).

Nos resultados apresentados, os modelos adaptados GANomaly e SGAnomaly se destacaram por obter médias superiores ao *baseline*. Em especial o modelo GANomaly adaptado obteve resultados maiores que o limite superior GMADE proposto por Giri et al. (2020) na métrica AUC. Este limite superior é uma arquitetura de viabilidade limitada, pela utilização de meta-dados na classificação de anomalias, situação esta que não é viável em aplicações práticas.

O estudo apresentado mostra referencias de desempenho destes modelos GANs no contexto de áudio, diferentemente dos trabalhos originais e outros da literatura. Através da adaptação dos modelos, foi possível obter referências de métricas que servirão de base para novos avanços.

6.1 CONTRIBUIÇÕES

As principais contribuições deste trabalho são:

1. Adaptação e parametrização otimizada de arquiteturas GANs para detecção de anomalias em eventos sonoros.
    a) A adaptação das arquiteturas GAN se deu por meio de uma padronização das redes geradoras e discriminadoras, utilizando como base a rede autocodificadora proposta por Koizumi, Kawaguchi e Imoto (2020).
    b) As parametrizações otimizadas seguiram a metodologia de busca exaustiva avaliando diversos hiper-parâmetros e otimizadores.
2. Artigo aceito na conferência SBCUP - Simpósio Brasileiro de Computação Ubíqua e Pervasiva 2023

6.2 TRABALHOS FUTUROS

Como trabalhos futuros ou variações deste trabalho podem ser elencados:

- Explorar a aplicação de redes convolucionais 2D em espectrogramas com o intuito de aprimorar a detecção de anomalias em áudios, sua adaptação para o domínio de espectrogramas pode proporcionar uma representação mais robusta e discriminativa dos padrões acústicos.
- Investigação sobre o uso de transferência de aprendizado, de modo que a reconstrução deixe de ser das características do áudio, como o coeficientes de Mel para que a reconstrução de um *embedding* que é um vetor de características descrita em um espaço vetorial.
- Identificar variações de características do áudio. Atualmente utilizamos os mesmos pré-processamento descrito pelo *baseline*, pretende-se utilizar números diferentes de coeficientes de Mel e avaliar seus impactos.

REFERÊNCIAS

AGARAP, A. F. Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375, 2018. Citado na página 41.

AGGARWAL, C. C. An introduction to outlier analysis. In: Outlier Analysis. Springer New York, 2012. p. 1–40. Disponível em: <https://doi.org/10.1007/978-1-4614-6396-2_1>. Citado 2 vezes nas páginas 18 e 23.

AGRAWAL, V. K.; MAURYA, S. S. Unsupervised Detection of Anomalous Sounds for Machine Condition Monitoring. [S.l.], 2020. Citado na página 53.

AHMED, Y. A.; OTHMAN, H.; SALEM, M. A.-M. Comparative study of different activation functions for anomalous sound detection. In: 2021 International Conference on Microelectronics (ICM). [S.l.: s.n.], 2021. p. 207–211. Citado na página 53.

AKCAY, S.; ABARGHOUEI, A. A.; BRECKON, T. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In: . [S.l.: s.n.], 2019. p. 1–8. Citado 9 vezes nas páginas 7, 15, 34, 36, 40, 48, 61, 62 e 64.

AKCAY, S.; ATAPOUR-ABARGHOUEI, A.; BRECKON, T. P. Ganomaly: Semi-supervised anomaly detection via adversarial training. In: SPRINGER. Asian Conference on Computer Vision. [S.l.], 2018. p. 622–637. Citado 14 vezes nas páginas 6, 7, 15, 29, 33, 34, 36, 40, 43, 45, 54, 55, 60 e 61.

Antonini, M. et al. Smart audio sensors in the internet of things edge for anomaly detection. IEEE Access, v. 6, p. 67594–67610, 2018. ISSN 2169-3536. Citado na página 19.

BIAN, J. et al. A Novel and Efficient CVAE-GAN-Based Approach With Informative Manifold for Semi-Supervised Anomaly Detection. v. 7, p. 88903–88916, 2019. Disponível em: <https://ieeexplore.ieee.org/document/8727990>. Citado na página 29.

BISHOP, C. M. Pattern Recognition and Machine Learning (Information Science and Statistics). Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387310738. Citado na página 14.

BREUNIG, M. M. et al. Lof: Identifying density-based local outliers. SIGMOD Rec., Association for Computing Machinery, New York, NY, USA, v. 29, n. 2, p. 93–104, may 2000. ISSN 0163-5808. Disponível em: <https://doi.org/10.1145/335191.335388>. Citado na página 30.

CHACHADA, S.; KUO, C.-C. J. Environmental sound recognition: a survey. APSIPA Transactions on Signal and Information Processing, Cambridge University Press, v. 3, p. e14, 2014. Citado 2 vezes nas páginas 21 e 22.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. ACM Comput. Surv., Association for Computing Machinery, New York, NY, USA, v. 41, n. 3, jul 2009. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/1541880.1541882>. Citado 3 vezes nas páginas 6, 18 e 19.

CHENG, Z. et al. Unsupervised outlier detection via transformation invariant autoencoder. IEEE Access, v. 9, p. 43991–44002, 2021. Citado 2 vezes nas páginas 32 e 36.

CHOLLET, F. Xception: Deep learning with depthwise separable convolutions. In: . [S.l.: s.n.], 2017. p. 1800–1807. Citado na página 35.

CHUNG, Y. et al. Automatic detection of cow's oestrus in audio surveillance system. Asian-Australasian Journal of Animal Sciences, Asian Australasian Association of Animal Production Societies, v. 26, n. 7, p. 1030–1037, jul. 2013. Disponível em: <https://doi.org/10.5713/ajas.2012.12628>. Citado na página 31.

CONTE, D. et al. An ensemble of rejecting classifiers for anomaly detection of audio events. In: IEEE. 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance. [S.l.], 2012. p. 76–81. Citado na página 30.

DANILUK, P. et al. ENSEMBLE OF AUTO-ENCODER BASED SYSTEMS FOR ANOMALY DETECTION. [S.l.], 2020. Citado na página 53.

DEEPAK, K.; CHANDRAKALA, S.; MOHAN, C. K. Residual spatiotemporal autoencoder for unsupervised video anomaly detection. Signal, Image and Video Processing, Springer Science and Business Media LLC, v. 15, n. 1, p. 215–222, jul. 2020. Disponível em: <https://doi.org/10.1007/s11760-020-01740-1>. Citado 3 vezes nas páginas 25, 26 e 32.

DIMITROV, S. et al. Analyzing sounds of home environment for device recognition. v. 8850, 01 2014. Citado na página 14.

DUARTE, N. C. V. Previsão de geração eólica baseada na classificação do tipo de clima em parque eólico. 2021. Citado 2 vezes nas páginas 7 e 55.

DUMAN, T. B.; BAYRAM, B.; İNCE, G. Acoustic anomaly detection using convolutional autoencoders in industrial processes. In: ÁLVAREZ, F. M. et al. (Ed.). 14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019). Cham: Springer International Publishing, 2013. p. 432–442. ISBN 978-3-030-20055-8. Citado na página 14.

ERASLAN, G. et al. Single-cell RNA-seq denoising using a deep count autoencoder. Nature Communications, Springer Science and Business Media LLC, v. 10, n. 1, jan. 2019. Disponível em: <https://doi.org/10.1038/s41467-018-07931-2>. Citado na página 25.

ESPI, M. et al. Exploiting spectro-temporal locality in deep learning based acoustic event detection. EURASIP Journal on Audio, Speech, and Music Processing, v. 2015, n. 1, p. 26, Sep 2015. ISSN 1687-4722. Disponível em: <https://doi.org/10.1186/s13636-015-0069-2>. Citado na página 21.

FAWCETT, T. An introduction to roc analysis. Pattern Recognition Letters, v. 27, n. 8, p. 861–874, 2006. ISSN 0167-8655. ROC Analysis in Pattern Recognition. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016786550500303X>. Citado 2 vezes nas páginas 15 e 56.

Foggia, P. et al. Audio surveillance of roads: A system for detecting anomalous sounds. IEEE Transactions on Intelligent Transportation Systems, v. 17, n. 1, p. 279–288, Jan 2016. ISSN 1558-0016. Citado na página 15.

Gemmeke, J. F. et al. Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.: s.n.], 2017. p. 776–780. ISSN 2379-190X. Citado na página 52.

GIRI, R. et al. Unsupervised Anomalous Sound Detection Using Self-Supervised Classification and Group Masked Autoencoder for Density Estimation. [S.l.], 2020. Citado 6 vezes nas páginas 53, 57, 62, 63, 68 e 69.

GOODFELLOW, I. et al. Generative adversarial nets. In: GHAHRAMANI, Z. et al. (Ed.). Advances in Neural Information Processing Systems. Curran Associates, Inc., 2014. v. 27. Disponível em: <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>. Citado 4 vezes nas páginas 15, 27, 29 e 47.

GROLLMISCH, S. et al. Sounding industry: Challenges and datasets for industrial sound analysis. In: 2019 27th European Signal Processing Conference (EUSIPCO). [S.l.: s.n.], 2019. p. 1–5. Citado na página 52.

HABEEB, R. A. A. et al. Real-time big data processing for anomaly detection: A survey. International Journal of Information Management, v. 45, p. 289 – 307, 2019. ISSN 0268-4012. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0268401218301658>. Citado na página 14.

HANNUN, A. et al. Deepspeech: Scaling up end-to-end speech recognition. 12 2014. Citado na página 23.

Hershey, S. et al. Cnn architectures for large-scale audio classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.: s.n.], 2017. p. 131–135. ISSN 2379-190X. Citado na página 21.

HONG, E.; CHOE, Y. Latent Feature Decentralization Loss for One-Class Anomaly Detection. IEEE Access, v. 8, p. 165658–165669, 2020. ISSN 2169-3536. Disponível em: <https://ieeexplore.ieee.org/document/9187777>. Citado na página 33.

IMOTO, K. Introduction to acoustic event and scene analysis. Acoustical Science and Technology, v. 39, 05 2018. Citado na página 14.

IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: BACH, F.; BLEI, D. (Ed.). Proceedings of the 32nd International Conference on Machine Learning. Lille, France: PMLR, 2015. (Proceedings of Machine Learning Research, v. 37), p. 448–456. Disponível em: <https://proceedings.mlr.press/v37/ioffe15.html>. Citado na página 42.

ISLAM, Z. et al. Crash data augmentation using variational autoencoder. Accident Analysis Prevention, v. 151, p. 105950, 2021. ISSN 0001-4575. Disponível em: <https://www.sciencedirect.com/science/article/pii/S000145752031770X>. Citado na página 25.

ISOLA, P. et al. Image-to-image translation with conditional adversarial networks. In: . [S.l.: s.n.], 2017. p. 5967–5976. Citado na página 29.

KANDA, N.; TAKEDA, R.; OBUCHI, Y. Elastic spectral distortion for low resource speech recognition with deep neural networks. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. [S.l.: s.n.], 2013. p. 309–314. Citado na página 23.

KARRAS, T. et al. Analyzing and Improving the Image Quality of StyleGAN. 2020. Citado na página 29.

KINGMA, D.; BA, J. Adam: A method for stochastic optimization. International Conference on Learning Representations, 12 2014. Citado na página 59.

KITTLER, J. et al. Intelligent signal processing mechanisms for nuanced anomaly detection in action audio-visual data streams. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.: s.n.], 2018. p. 6563–6567. Citado na página 16.

KO, T. et al. Audio augmentation for speech recognition. In: Proc. Interspeech 2015. [S.l.: s.n.], 2015. p. 3586–3589. Citado na página 23.

KOIZUMI, Y.; KAWAGUCHI, Y.; IMOTO, K. Description and discussion on DCASE2020 challenge task2: unsupervised anomalous sound detection for machine condition monitoring. [S.l.], 2020. Citado 17 vezes nas páginas 6, 7, 15, 32, 33, 36, 39, 41, 42, 50, 53, 57, 62, 63, 64, 68 e 69.

KOIZUMI, Y. et al. ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection. In: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). [s.n.], 2019. p. 308–312. Disponível em: <https://ieeexplore.ieee.org/document/8937164>. Citado 3 vezes nas páginas 15, 16 e 52.

Komatsu, T.; Kondo, R. Detection of anomaly acoustic scenes based on a temporal dissimilarity model. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.: s.n.], 2017. p. 376–380. ISSN 2379-190X. Citado na página 31.

LATHI, B. Sinais e Sistemas Lineares - 2.ed. Bookman, 2007. ISBN 9788560031139. Disponível em: <https://books.google.com.br/books?id=ySxoo2TVeeYC>. Citado na página 20.

LEE, J.; Umar Karim Khan, M.; KYUNG, C.-M. Hybrid Discriminator With Correlative Autoencoder for Anomaly Detection. IEEE Access, v. 9, p. 49098–49109, 2021. ISSN 2169-3536. Disponível em: <https://ieeexplore.ieee.org/document/9258885>. Citado na página 33.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation-based anomaly detection. ACM Trans. Knowl. Discov. Data, Association for Computing Machinery, New York, NY, USA, v. 6, n. 1, mar 2012. ISSN 1556-4681. Disponível em: <https://doi.org/10.1145/2133360.2133363>. Citado na página 30.

LIU, G. et al. Sagan: Skip-attention gan for anomaly detection. In: 2021 IEEE International Conference on Image Processing (ICIP). [S.l.: s.n.], 2021. p. 2468–2472. Citado 3 vezes nas páginas 15, 35 e 36.

LIU, S.; XU, L. An Integrated Model Based on O-GAN and Density Estimation for Anomaly Detection. IEEE Access, v. 8, p. 204471–204482, 2020. ISSN 2169-3536. Disponível em: <https://ieeexplore.ieee.org/document/9256316>. Citado 3 vezes nas páginas 29, 34 e 36.

LIU, Y. et al. Anomalous sound detection using spectral-temporal information fusion. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.: s.n.], 2022. p. 816–820. Citado na página 53.

LOSHCHILOV, I.; HUTTER, F. Decoupled weight decay regularization. In: International Conference on Learning Representations. [s.n.], 2019. Disponível em: <https://openreview.net/forum?id=Bkg6RiCqY7>. Citado na página 59.

MAN, R.; YANG, P.; XU, B. Classification of Breast Cancer Histopathological Images Using Discriminative Patches Screened by Generative Adversarial Networks. IEEE Access, v. 8, p. 155362–155377, 2020. ISSN 2169-3536. Disponível em: <https://ieeexplore.ieee.org/document/9177091>. Citado na página 33.

MENG, Z. et al. An enhancement denoising autoencoder for rolling bearing fault diagnosis. Measurement, v. 130, p. 448–454, 2018. ISSN 0263-2241. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0263224118307474>. Citado na página 25.

Mesaros, A.; Heittola, T.; Virtanen, T. Tut database for acoustic scene classification and sound event detection. In: 2016 24th European Signal Processing Conference (EUSIPCO). [S.l.: s.n.], 2016. p. 1128–1132. ISSN 2076-1465. Citado na página 14.

MüLLER, R.; ILLIUM, S.; LINNHOFF-POPIEN, C. Deep recurrent interpolation networks for anomalous sound detection. In: 2021 International Joint Conference on Neural Networks (IJCNN). [S.l.: s.n.], 2021. p. 1–7. Citado 2 vezes nas páginas 32 e 36.

NANNI, L. et al. Combining visual and acoustic features for music genre classification. Expert Systems with Applications, v. 45, p. 108 – 117, 2016. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417415006326>. Citado na página 20.

Nawaz, R. et al. Acoustic feature extraction from music songs to predict emotions using neural networks. In: 2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS). [S.l.: s.n.], 2018. p. 166–170. ISSN null. Citado na página 20.

NETO, W.; FIGUEIREDO, C. Detecção de tentativa de invasão por dados sintéticos em aplicações de biometria por voz. In: Anais do XI Simposio Brasileiro de Computacao Ubiqua e Pervasiva. Porto Alegre, RS, Brasil: SBC, 2019. ISSN 2595-6183. Disponível em: <https://sol.sbc.org.br/index.php/sbcup/article/view/6587>. Citado na página 20.

NGUYEN, H. et al. Forecasting and anomaly detection approaches using lstm and lstm autoencoder techniques with the applications in supply chain management. International Journal of Information Management, v. 57, p. 102282, 2021. ISSN 0268-4012. Disponível em: <https://www.sciencedirect.com/science/article/pii/S026840122031481X>. Citado 2 vezes nas páginas 25 e 32.

PARK, D. S. et al. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In: Proc. Interspeech 2019. [S.l.: s.n.], 2019. p. 2613–2617. Citado 2 vezes nas páginas 23 e 67.

PARMAR, G. et al. Dual contradistinctive generative autoencoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2021. p. 823–832. Citado na página 25.

PUROHIT, H. et al. MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019). [s.n.], 2019. p. 209–213. Disponível em: <http://dcase.community/documents/workshop2019/proceedings/DCASE2019Workshop_Purohit_21.pdf>. Citado 2 vezes nas páginas 16 e 52.

RAGAB, M. G. et al. An ensemble one dimensional convolutional neural network with bayesian optimization for environmental sound classification. Applied Sciences, v. 11, n. 10, 2021. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/11/10/4660>. Citado na página 59.

REDDI, S. J.; KALE, S.; KUMAR, S. On the convergence of adam and beyond. In: International Conference on Learning Representations. [s.n.], 2018. Disponível em: <https://openreview.net/forum?id=ryQu7f-RZ>. Citado na página 59.

RONG, F. Audio classification method based on machine learning. In: 2016 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS). [S.l.: s.n.], 2016. p. 81–84. Citado na página 14.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. CoRR, abs/1505.04597, 2015. Disponível em: <http://arxiv.org/abs/1505.04597>. Citado 4 vezes nas páginas 6, 26, 27 e 32.

ROUSSEEUW, P. J.; DRIESSEN, K. van. A fast algorithm for the minimum covariance determinant estimator. Technometrics, [Taylor Francis, Ltd., American Statistical Association, American Society for Quality], v. 41, n. 3, p. 212–223, 1999. ISSN 00401706. Disponível em: <http://www.jstor.org/stable/1270566>. Citado na página 30.

ROVETTA, S.; MNASRI, Z.; MASULLI, F. Detection of hazardous road events from audio streams: An ensemble outlier detection approach. In: 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS). [S.l.: s.n.], 2020. p. 1–6. Citado na página 16.

Sarkar, R.; Saha, S. K. Music genre classification using emd and pitch based feature. In: 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR). [S.l.: s.n.], 2015. p. 1–6. ISSN null. Citado na página 20.

SCHLEGL, T. et al. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. 2017. Citado 2 vezes nas páginas 15 e 33.

SCHLEGL, T. et al. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Medical Image Analysis, v. 54, p. 30 – 44, 2019. ISSN 1361-8415. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1361841518302640>. Citado 3 vezes nas páginas 15, 33 e 36.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. Neural Networks, v. 61, p. 85–117, 2015. ISSN 0893-6080. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0893608014002135>. Citado 2 vezes nas páginas 15 e 23.

SONG, S. et al. A Mura Detection Model Based on Unsupervised Adversarial Learning. IEEE Access, v. 9, p. 49920–49928, 2021. ISSN 2169-3536. Disponível em: <https://ieeexplore.ieee.org/document/9389548>. Citado 2 vezes nas páginas 35 e 36.

SONG, Y. et al. One-Class Conditional Random Fields for Sequential Anomaly Detection. 2013. Disponível em: <https://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/6579/6951>. Citado na página 30.

SOUZA, E. F. de. Mmi-gan : Multi medical imaging translation using generative adversarial network. 2020. Citado na página 29.

SUEFUSA, K. et al. Anomalous sound detection based on interpolation deep neural network. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.: s.n.], 2020. p. 271–275. Citado 4 vezes nas páginas 6, 32, 33 e 36.

SURI, N. N. R. R.; M, N. M.; ATHITHAN, G. Outlier Detection: Techniques and Applications. Springer International Publishing, 2019. Disponível em: <https://doi.org/10.1007/978-3-030-05127-3>. Citado 2 vezes nas páginas 19 e 20.

SUTSKEVER, I. et al. On the importance of initialization and momentum in deep learning. In: DASGUPTA, S.; MCALLESTER, D. (Ed.). Proceedings of the 30th International Conference on Machine Learning. Atlanta, Georgia, USA: PMLR, 2013. (Proceedings of Machine Learning Research, 3), p. 1139–1147. Disponível em: <https://proceedings.mlr.press/v28/sutskever13.html>. Citado na página 59.

TODISCO, M. et al. Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion. In: Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association Hyderabad, India: ISCA, 2018. Disponível em: <https://hal.inria.fr/hal-01889934>. Citado na página 20.

TRUONG, H. V. et al. Unsupervised detection of anomalous sound for machine condition monitoring using fully connected u-net. In: . [S.l.: s.n.], 2021. Citado 3 vezes nas páginas 27, 32 e 36.

VACHER, M.; SERIGNAT, J.-F.; CHAILLOL, S. Sound classification in a smart room environment: an approach using gmm and hmm methods. 05 2007. Citado na página 14.

VIRTANEN, T.; PLUMBLEY, M. D.; ELLIS, D. (Ed.). Computational Analysis of Sound Scenes and Events. Springer International Publishing, 2018. Disponível em: <https://doi.org/10.1007/978-3-319-63450-0>. Citado 2 vezes nas páginas 14 e 20.

WAN, F. et al. Outlier detection for monitoring data using stacked autoencoder. IEEE Access, v. 7, p. 173827–173837, 2019. Citado na página 32.

WANG, Y.; WONG, J.; MINER, A. Anomaly intrusion detection using one class svm. In: Proceedings from the Fifth Annual IEEE SMC Information Assurance Workshop, 2004. [S.l.: s.n.], 2004. p. 358–364. Citado na página 30.

WOO, S. et al. CBAM: convolutional block attention module. CoRR, abs/1807.06521, 2018. Disponível em: <http://arxiv.org/abs/1807.06521>. Citado na página 35.

XIA, M. et al. Generalized data augmentation for low-resource translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019. p. 5786–5796. Disponível em: <https://aclanthology.org/P19-1579>. Citado na página 23.

Xiong, W. et al. The microsoft 2017 conversational speech recognition system. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.: s.n.], 2018. p. 5934–5938. ISSN 2379-190X. Citado na página 14.

XU, W. et al. Improving performance of autoencoder-based network anomaly detection on nsl-kdd dataset. IEEE Access, v. 9, p. 140136–140146, 2021. Citado 4 vezes nas páginas 15, 25, 26 e 32.

XU, W. et al. DRB-GAN: A Dynamic ResBlock Generative Adversarial Network for Artistic Style Transfer. 2021. Citado na página 29.

ZEIMARANI, C. F. F. C. F. B.; COSTA, M. G. F. Breast Tumor Classification in Ultrasound Images using Deep Convolutional Neural Network. In: Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal do Amazonas. [S.l.: s.n.], 2019. p. 27–28. Citado na página 22.

ZENATI, H. et al. Efficient GAN-Based Anomaly Detection. 2018. Citado 9 vezes nas páginas 6, 7, 33, 36, 40, 42, 43, 62 e 63.

ZHAO, Y. et al. Frequency detection algorithm for frequency diversity signal based on stft. In: 2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC). [S.l.: s.n.], 2015. p. 790–793. Citado na página 21.

ZHOU, X. et al. A radio anomaly detection algorithm based on modified generative adversarial network. IEEE Wireless Communications Letters, v. 10, n. 7, p. 1552–1556, 2021. Citado na página 16.

ZÖLZER, U. Digital Audio Signal Processing. Wiley, 2008. ISBN 9780470680025. Disponível em: <https://books.google.com.br/books?id=fyV86ge9wqUC>. Citado na página 14.