Análise de redes GANs para detecção de anomalias em atividade sonoras

Wilson A. de Oliveira Neto[1], **Carlos Maurício S. Figueiredo**[1,2]

[1] Instituto de Computação – Universidade Federal do Amazonas
Manaus – AM – Brasil

[2]Núcleo de Computação – Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brasil

wilson.oliveira@icomp.ufam.edu.br,cfigueiredo@uea.edu.br

Abstract. *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. 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 2020 dataset, which includes over 180 hours of audio from industrial machinery. We evaluated the classification of anomalies, reporting an average of 72% AUC and 69% pAUC, results superior to those presented by baselines*

Resumo. *Os trabalhos do estado-da-arte na identificação de anomalias em imagens utilizam arquiteturas baseadas em GAN (Generative Adversarial Network), entretanto, poucos estudos demonstram sua utilização no domínio de sons. 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 2020, que inclui mais de 180 horas de maquinário industrial. Avaliamos a classificação das anomalias, relatando uma média de 72% de AUC e 69% de pAUC, resultados superiores ao apresentado por baselines.*

1. Introdução

A detecção de anomalias visa identificar e diferenciar amostras anômalas das amostras típicas, essa tarefa possui extrema importância em aplicações de ambientes inteligentes e indústria 4.0, como identificação de falha de maquinário industrial [Purohit et al. 2019], detecção de acidentes rodoviários [Rovetta et al. 2020], abordagens multi-modais com vídeo [Kittler et al. 2018] e identificação de falha em transmissão de sinal 5G [Zhou et al. 2021].

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. 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 introdução de arquiteturas de modelos geradores permitiu a abordagem *Auto-Encoder* AE, onde o erro de reconstrução de áudio é utilizado para identificar e quantizar a anomalia. O objetivo das abordagens geradoras é aprender a representação dos dados através da reconstrução ou do treinamento adversário com um discriminador [Suefusa et al. 2020, Akcay et al. 2019a]. O treinamento de arquiteturas GANs proporcionam reconstruções de áudios mais fiéis, bem como a melhor capacidade de controlar o espaço latente [Creswell et al. 2017].

Na literatura, encontramos diversos modelos GANs sendo propostos e avaliados para a detecção de anomalias, entretanto, focados para imagens [Schlegl et al. 2019, Liu et al. 2021, Akcay et al. 2019b, Akcay et al. 2019a]. No entanto, tais trabalhos não apresentam como adequar os modelos propostos no domínio de áudio, tampouco apresentam métricas avaliativas de referência.

Diante do exposto, identifica-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. Nossas principais contribuições são:

- Identificamos mudanças necessárias na arquitetura dos modelos GAN's para que operem no domínio do áudio.
- Padronizamos as arquiteturas dos modelos GAN's para avaliação da métrica *Receiver Operating Characteristic* (ROC) *Curve* (AUC) e *partial-AUC* (pAUC) [Koizumi et al. 2020]
- Apresentamos os resultados das arquiteturas modificadas como forma de *baseline* para trabalhos futuros e comparações no domínio do áudio.

O restante deste artigo está organizado da seguinte forma: Na Seção 2 apresentamos os trabalhos relacionados, a Seção 3 descreve a abordagem proposta, na Seção 4 expõe os experimentos realizados, por fim, a Seção 5 detalha as conclusões do artigo.

2. Trabalhos relacionados

O problema que estamos avaliando é Detecção de Anomalias (DA) em conjunto de dados de áudio de modo não supervisionado, ou seja, sem rótulos de anomalias. Trabalhos mais recentes avançaram o desempenho de modelos geradores para a tarefa de detecção de anomalias. A abordagem típica é o treinamento de um modelo *Auto Encoder* (AE) profundo capaz de reconstruir os dados de entrada e, então, empregar o erro da reconstrução para identificar anomalias [Cheng et al. 2021]. Para aumentar a distinção entre os erros de reconstrução dos dados reais e os anômalos, células de memória [Müller et al. 2021] e discriminadores [Schlegl et al. 2019] são adicionados aos AE's. Os trabalhos que serão apresentados utilizam diferentes métricas para avaliar o *score* de anomalias.

Nos métodos baseados em GAN, temos EGBAD [Zenati et al. 2018], a rede AE é utilizada para aprender a representação latente dos dados. Diferentemente das redes GAN's regulares, o Discriminador também considera a representação latente junto à imagem. A hipótese levantada é que os dados anômalos possuam vetores latentes que não se correlacionam com a imagem. Durante o cálculo *score* de anomalias, o autor utiliza a combinação da classificação do Discriminador e a distância entre a imagem gerada e a real. GANomaly [Akcay et al. 2019b], a arquitetura combinada AE e GAN. A sua

principal característica está na construção da rede Geradora, que utiliza uma combinação de redes AE, na qual, dispõe de duas saídas: a imagem gerada e uma representação latente desta imagem. Os autores demonstram que o espaço latente da imagem gerada possui grande diferença do espaço latente da imagem original, caso a imagem original não pertença ao dados normais. SKIP-GANomaly [Akcay et al. 2019a] apresenta a combinação entre AE e GAN. Sua rede geradora é formada por uma rede U-NET [Ronneberger et al. 2015], capaz de recriar as imagens de entrada com mais detalhes que uma rede AE padrão, pois as camadas se interligam de forma residual. Essa abordagem demonstra que o treinamento adversário possui vantagens ao treinamento de uma rede AE.

A Tabela 1 apresenta os principais trabalhos na área de detecção de anomalias. Pode-se observar que a maioria dos trabalhos envolvendo arquiteturas geradoras são empregadas no domínio de imagens e, por sua vez as arquiteturas AE são empregadas em áudio.

Trabalhos	Arquitetura	Domínios	Paradigma
[Akcay et al. 2019b]	AEE+GAN	imagem	Não-supervisionado
[Akcay et al. 2019a]	U-Net+GAN	imagem	Não-supervisionado
[Zenati et al. 2018]	AE+GAN	imagem	Não-supervisionado
[Cheng et al. 2021]	AE	imagem	Não-supervisionado
[Liu et al. 2021]	U-Net(DSC+CBAM)+GAN	imagem	Não-supervisionado
[Koizumi et al. 2020]	AE	áudio	Não-Supervisionado
[Suefusa et al. 2020]	AE(IDNN)	áudio	Não-supervisionado
[Müller et al. 2021]	AE(DRINK)	áudio	Não-supervisionado

Tabela 1. Comparação entre os trabalhos sobre detecção de anomalias utilizando arquiteturas profundas

3. Abordagem Proposta

3.1. Abordagem do método

A solução proposta consiste em adaptação básica dos modelos publicados por [Zenati et al. 2018], [Akcay et al. 2019b] e [Akcay et al. 2019a], para adequação dos modelos para a detecção de anomalias sonoras. Destacamos que todas as arquiteturas utilizadas neste trabalho são abordagens de GANs, conhecidas por seus resultados em conjuntos de dados de imagens. No entanto, não há comprovação de sua eficácia em conjuntos de outros domínios, como áudio.

- **Fase 1 – Pré-processamento**. Na fase inicial, os dados de treinamento e teste passam pela etapa de pré-processamento 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, adaptamos as arquiteturas usando apenas exemplos de áudios classificados como comportamentos típicos. Cada arquitetura possui suas particularidades, mas todas criam um modelo por rótulo. Durante o treinamento, as sub-redes Geradoras reproduzem os conjuntos de entrada com alta fidelidade, enquanto as sub-redes Discriminadoras avaliam a similaridade dos dados gerados com os originais. Essa interação entre as redes resulta em dois modelos neurais: um para gerar dados próximos aos 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, para então calcular as métricas AUC e pAUC.

3.2. Pré-processamento

A etapa de pré-processamento pode ser descrita em 5 etapas, conforme mostrada na Figura 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.

Seguindo a abordagem convencional [Koizumi et al. 2020], as entradas das adaptações das arquiteturas são as características do áudio, neste caso, o logMel espectrograma. Para calcular o espectrograma de Mel, cada áudio de 10s é dividido em quadros de 64ms, com tamanho de Hop de 32ms entre os quadros. O tamanho da janela foi ajustado para 1024-FFT e 128 Mel bins foram utilizados para obter as características de cada quadro. No total 5 logMel espectrogramas são concatenados com sobreposição de 56 quadros.

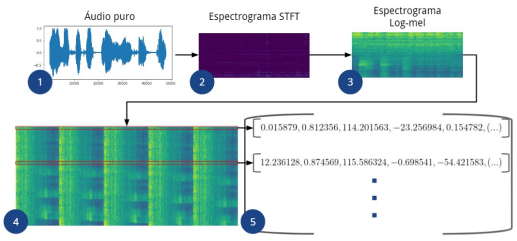


Figura 1. Etapas de pré-processamento.

3.3. Adaptação das arquiteturas

As adaptações destas arquiteturas foram baseadas no modelo proposto por [Koizumi et al. 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) e normalização em lote (*Batch Normalization*) são

utilizadas em cada camada FCN, exceto na camada de saída do decodificador, conforme mostra a Figura 2. A escolha da rede foi baseada em sua baixa quantidade de parâmetros, facilitando a leitura dos áudios. As adaptações foram aplicadas em todas as arquiteturas para padronização na mesma base de dados, permitindo a avaliação das funções de otimização e das arquiteturas.

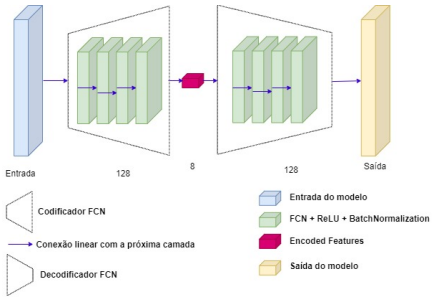


Figura 2. Detalhes da arquitetura proposta por [Koizumi et al. 2020]

3.3.1. EGBAD – *Efficient Gan-Based Anomaly Detection*

A adaptação para áudio consistiu na mudança do gerador G, Discriminador D e Codificador E. O modelo G e o modelo D representam o Decodificador FCN, já o modelo E retrata o Codificador FCN. Os demais elementos da arquitetura seguem o modelo original.

3.3.2. GANomaly – *Semi-Supervised Anomaly Detection via Adversarial Training*.

A adaptação pode ser vista através da Figura 2, onde constituiu-se da mudança das redes G_E 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 permaneceram 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'. 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 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.
- **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.

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

A adaptação desta arquitetura pode ser observada na Figura 3. Modificou-se a rede G_E e rede G_D para redes Codificadoras FCN e Decodificadoras FCN respectivamente, o mo-

delo D é substituído pela rede Codificadora FCN. Por fim, são adicionadas 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 é detalhada a seguir:

- **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.
- **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 similar a arquitetura anterior.

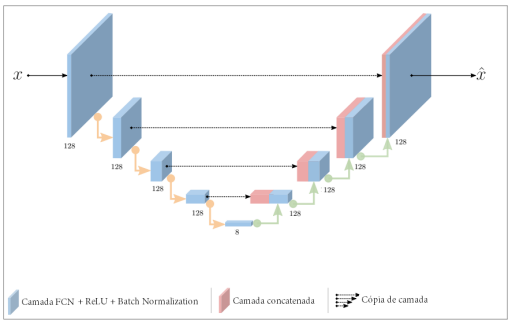


Figura 3. Arquitetura adaptada da rede geradora Skip-GANomaly.

4. Experimentos

4.1. Modelo e Conjunto de dados

Adaptamos as arquiteturas de aprendizado não supervisionado EGBAD [Zenati et al. 2018], GANomaly [Akcay et al. 2019b] e Skip-GANomaly [Akcay et al. 2019a]. Para validar as arquiteturas utilizamos os conjuntos de dados de áudio no mundo real ToyADMOS [Koizumi et al. 2019] e MIMII [Purohit et al. 2019]. Os áudios são de seis tipos de maquinários industriais, chamados de: *Toy-car (ToyADMOS)*, *Toy-conveyor (ToyADMOS)*, *Valve (MIMII Dataset)*, *Pump (MIMII Dataset)*, *Fan (MIMII Dataset)* e *Slide rail (MIMII Dataset)*. Os dados são divididos entre típicos e anômalos, com aproximadamente 10 segundos cada. Entretanto, durante a fase de treinamento somente temos acesso aos áudios de funcionamento típico. Os dados encontram-se agrupados por identificadores (ID.), classes e propósito de utilização.

4.2. *Baseline*

Para a comparação, priorizamos abordagens baseadas em GANs, conhecidas por resultados superiores em conjuntos de dados de imagens. No entanto, não foram encontradas comparações diretas dessas abordagens em conjuntos de dados de áudio. Assim, apresentaremos as soluções selecionadas para o domínio de áudio. Os *baselines* escolhidos foram:

- **DCASE *Baseline* 2020** [Koizumi et al. 2020]: Método de aprendizagem não supervisionada. Representa o limite inferior para nossa comparação. Esta margem nos guia para identificar a validade das adaptações das redes GAN's.
- **GMADE** [Giri et al. 2020]: Utiliza método de aprendizagem semi-supervisionado por utilizar metadados do conjunto treino para classificar os áudios. Portanto, embora o método seja usado na avaliação comparativa, ele serve de base como limite superior, assumindo a existência de informações adicionais não disponíveis normalmente em problemas de detecção de anomalia [Koizumi et al. 2020].

4.3. Resultado dos experimentos

Cada arquitetura possui um modelo único para cada classe, sendo as avaliações feitas apenas com dados de teste dessa classe. Isso garante o aprendizado do funcionamento típico de cada maquinário industrial. Tanto as classes quanto os modelos têm particularidades que afetam a aprendizagem. Portanto, espera-se que uma arquitetura apresente resultados variados em diferentes tipos de áudio. É importante destacar que as adaptações das arquiteturas usam apenas instâncias de áudio, sem adicionar metadados para o treinamento. O objetivo deste trabalho é demonstrar a eficiência das adaptações em conjuntos de dados do mundo real.

A avaliação da arquitetura adaptada GANomaly é demonstrada na Figura 4a que apresenta um gráfico de barras contendo os resultados da arquitetura adaptada GANomaly [Akcay et al. 2019b] em todas as classes. Podemos observar que esta arquitetura obteve desempenho acima de 85% em AUC nas classes *Toycar* e *ToyConveyor*, enquanto todas as classes restantes possuem resultados acima de 70%. Isso indica que o modelo foi capaz de identificar e distinguir corretamente os dados anômalos dos típicos. A métrica pAUC também se mostrou satisfatória, visto que a maioria das classes obteve resultados acima de 60%. Com relação aos rótulos *fan* 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.

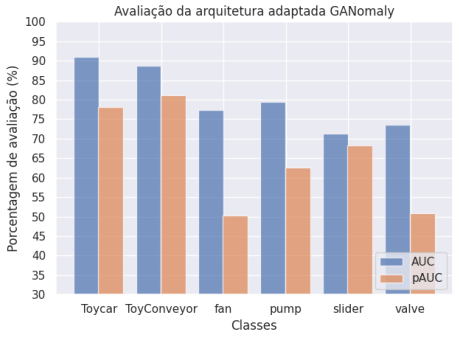
Já a avaliação arquitetura adaptada SGANomaly apresentada na Figura 4b demonstra um gráfico de barras contendo os resultados da arquitetura adaptada Skip-GANomaly [Akcay et al. 2019a] em todas as classes. Esta arquitetura apresenta resultados acima de 65% e 50% na métricas AUC e pAUC respectivamente, para todas as classes. Destaca-se o rótulo *ToyConveyor* e *pump* como o maior resultado em AUC e pAUC na devida ordem.

Por fim, a avaliação da arquitetura adaptada EGBAD exibida no gráfico de barras da Figura 4c, apresenta o resultado da arquitetura adaptada EGBAD [Zenati et al. 2018] em todos os rótulos. A arquitetura obteve sucesso na identificação somente na classe *valve* com as métricas AUC e pAUC de 69% e 57% respectivamente. Nota-se que o desempenho nas demais classes não foi satisfatório e em classes como a *ToyConveyor*, *pump* e *fan* obteve resultados abaixo do limiar de 50%, ou seja, assegura-se que o modelo não foi capaz de discernir entre áudios normais e anômalos.

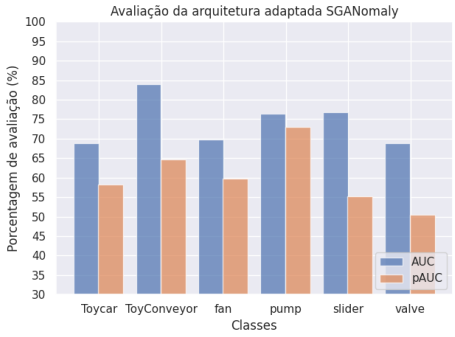
4.3.1. Resultados Gerais

O resultado geral dos experimentos realizados podem ser observados na Figura 4d. Nela, são apresentados as médias e desvios-padrão para todos os modelos, considerando todas as classes presentes 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 0.66 e 0.53 nas métricas AUC e pAUC respectivamente.

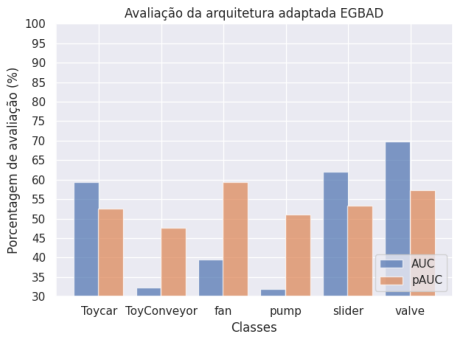
Destacamos que os modelos GANomaly, SGANomaly e EGBAD são arquiteturas adaptadas para áudio, já que sua concepção foi feita para imagens. Embora as arquiteturas possuam algumas alterações, suas funções de otimização e estrutura foram mantidas originais. A avaliação geral demonstra que o modelo GMADE possui os melhores resultados nas métricas AUC e pAUC, apresentando média de 0.8 na AUC e 0.72 na pAUC. Seguido pelo modelo GANomaly que apresenta média de 0.72 na AUC e 0.69 na pAUC. O modelo SGANomaly apresenta 0.66 na AUC e 0.54 na pAUC.



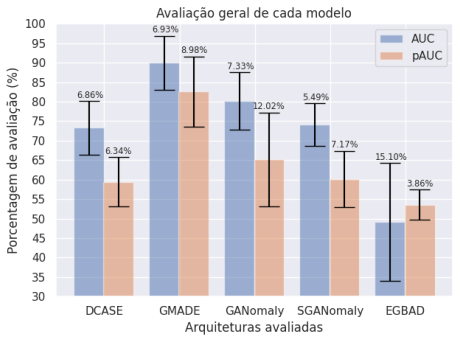
(a) Avaliação da arquitetura adaptada GANomaly [Akcay et al. 2019b].



(b) Avaliação da arquitetura adaptada Skip-GANomaly [Akcay et al. 2019a].



(c) Avaliação da arquitetura adaptada EGBAD [Zenati et al. 2018].



(d) Avaliação geral para cada modelo. A imagem apresenta as médias e o desvio padrão para cada classe.

Figura 4. Gráficos de avaliação de arquiteturas.

A Tabela 2 mostra os resultados dos modelos mencionados anteriormente. Podemos observar que os modelos GANomaly e SGANomaly obtiveram resultados da métrica

AUC próximos e acima do *baseline* oferecido pelo DCASE. Entretanto, o modelo GANomaly apresentou os melhores resultados comparativamente. Isso se deve ao fato de que a comparação entre os vetores latentes são melhores em reconstruir o contexto de um áudio.

Trabalhos	ToyCar		ToyConveyor		fan		pump		slider		valve	
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
DCASE	0.8009	0.6722	0.7268	0.6065	0.6515	0.5259	0.72	0.60	0.84	0.66	0.66	0.50
GMADE	**0.9504**	**0.9039**	0.8067	0.6590	**0.8233**	**0.7897**	**0.8694**	**0.7960**	**0.9728**	**0.8954**	**0.9738**	**0.9121**
GANomaly	0.9100	0.7810	**0.888**	**0.812**	0.773	0.503	0.794	0.626	0.713	0.682	0.736	0.508
SGANomaly	0.6890	0.5820	0.839	0.646	0.698	0.598	0.764	0.73	0.767	0.552	0.689	0.504
EGBAD	0.5940	0.5260	0.324	0.477	0.395	0.594	0.319	0.511	0.62	0.534	0.698	0.573

Tabela 2. Comparativos de baseline

5. Conclusão

Este trabalho comprovou a eficácia dos modelos baseados em GAN para a detecção de anomalias em áudios. Os modelos GANomaly e SGAnomaly apresentaram resultados superiores ao *baseline*. Adaptamos três modelos GANs de detecção de anomalias em imagens para o domínio de áudio, utilizando o conjunto de dados do desafio DCASE como referência popular de sons.

O estudo apresentado mostra referências 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.

Trabalhos futuros podem incluir: (1) Utilizar técnicas de aumento de dados ou adicionar ruídos para melhorar as métricas estabelecidas. (2) Identificar funções de otimização relevantes e explorar variações ou combinações de arquiteturas baseadas em treinamento adversário. (3) Investigar variações nas características do áudio, como utilizar diferentes números de coeficientes de Mel e avaliar seu impacto.

Referências

Akcay, S., Atapour Abarghouei, A., and Breckon, T. (2019a). Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. pages 1–8.

Akcay, S., Atapour-Abarghouei, A., and Breckon, T. P. (2019b). GANomaly: Semi-supervised Anomaly Detection via Adversarial Training. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11363 LNCS, pages 622–637. Springer Verlag.

Cheng, Z., Zhu, E., Wang, S., Zhang, P., and Li, W. (2021). Unsupervised outlier detection via transformation invariant autoencoder. *IEEE Access*, 9:43991–44002.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2017). Generative adversarial networks: An overview. *CoRR*, abs/1710.07035.

Giri, R., Tenneti, S. V., Helwani, K., Cheng, F., Isik, U., and Krishnaswamy, A. (2020). Unsupervised anomalous sound detection using self-supervised classification and group masked autoencoder for density estimation. Technical report, DCASE2020 Challenge.

Kittler, J., Kaloskampis, I., Zor, C., Xu, Y., Hicks, Y., and Wang, W. (2018). 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)*, pages 6563–6567.

Koizumi, Y., Kawaguchi, Y., and Imoto, K. (2020). Description and discussion on dcase2020 challenge task2: unsupervised anomalous sound detection for machine condition monitoring. Technical report, DCASE2020 Challenge.

Koizumi, Y., Saito, S., Uematsu, H., Harada, N., and Imoto, K. (2019). 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)*, pages 308–312.

Liu, G., Lan, S., Zhang, T., Huang, W., and Wang, W. (2021). Sagan: Skip-attention gan for anomaly detection. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2468–2472.

Müller, R., Illium, S., and Linnhoff-Popien, C. (2021). Deep recurrent interpolation networks for anomalous sound detection. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Purohit, H., Tanabe, R., Ichige, T., Endo, T., Nikaido, Y., Suefusa, K., and Kawaguchi, Y. (2019). 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)*, pages 209–213.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Rovetta, S., Mnasri, Z., and Masulli, F. (2020). Detection of hazardous road events from audio streams: An ensemble outlier detection approach. In *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, pages 1–6.

Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., and Schmidt-Erfurth, U. (2019). f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30 – 44.

Suefusa, K., Nishida, T., Purohit, H., Tanabe, R., Endo, T., and Kawaguchi, Y. (2020). Anomalous sound detection based on interpolation deep neural network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 271–275.

Zenati, H., Foo, C. S., Lecouat, B., Manek, G., and Chandrasekhar, V. R. (2018). Efficient gan-based anomaly detection.

Zhou, X., Xiong, J., Zhang, X., Liu, X., and Wei, J. (2021). A radio anomaly detection algorithm based on modified generative adversarial network. *IEEE Wireless Communications Letters*, 10(7):1552–1556.