UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
CURSO DE BACHARELADO EM ENGENHARIA ELÉTRICA

MARCELO UNIS BOZZETTO

GANS PARA DETECÇÃO DE ANOMALIAS EM SÉRIES TEMPORAIS: UM ESTUDO DE CASO

Porto Alegre
2022

MARCELO UNIS BOZZETTO

GANS PARA DETECÇÃO DE ANOMALIAS EM SÉRIES TEMPORAIS: UM ESTUDO DE CASO

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Engenharia Elétrica da Universidade Federal do Rio Grande do Sul como parte dos requisitos para obtenção do grau de Bacharel em Engenharia Elétrica.

Orientador: Prof. Dr. Alexandre Balbinot

Porto Alegre

2022

MARCELO UNIS BOZZETTO

GANS PARA DETECÇÃO DE ANOMALIAS EM SÉRIES TEMPORAIS: UM ESTUDO DE CASO

Trabalho de conclusão de curso de graduação apresentado ao Departamento de Engenharia Elétrica da Universidade Federal do Rio Grande do Sul como parte dos requisitos para obtenção do grau de Bacharel em Engenharia Elétrica.

Aprovado em ___ de _____ de _____

BANCA EXAMINADORA:

Prof. Dr. Alexandre Balbinot
UFRGS
Orientador

Prof. Dr. Tiago Oliveira Weber
UFRGS
Banca Examinadora

Dr. Vinicius Horn Cene
Engineer, Data Analysis Specialist,
Samsung Research Brazil
Banca Examinadora

AGRADECIMENTOS

Ao meu orientador, Alexandre Balbinot, por toda ajuda e incentivo.
À UFRGS pelo compromisso com a pesquisa e o ensino de excelência.

RESUMO

Palavras-chave: Detecção de Anomalias, Séries temporais, *Machine Learning*, *Deep Learning*, Generative Adversarial Networks;

O problema geral da detecção de anomalias se manifesta em diversos campos e se relaciona intimamente com inúmeros problemas específicos. A formulação habitual totalmente não supervisionada gera dificuldades adicionais na obtenção de representações relevantes para o problema, e restringe os métodos aplicáveis. Nesse contexto, o grande sucesso recente de soluções baseadas em GANs na modelagem de distribuições e processos arbitrários a partir de dados não supervisionados suscita grande interesse na sua aplicação ao problema de detecção de anomalias.

Com objetivo de abordar esse tema, a aplicação de soluções baseadas em GANs para detecção de anomalias no contexto não supervisionado em séries temporais foi estudada. A partir de uma revisão da literatura dos princípios gerais de GANs e detecção de anomalias, trabalhos recentes aplicando GANs à séries temporais foram compilados e apresentados. Em sequência, um método específico, TadGan (GEIGER *et al.*, 2020), foi selecionado para experimentação e estudos aprofundados sob o formato de estudo de caso. Uma implementação foi obtida e verificada, e uma metodologia para demonstrar o funcionamento e os princípios gerais do método e da aplicação de GANs às séries temporais sobre dados sintetizados a partir de funções analíticas desenvolvida e executada. Avaliou-se, em sequência, possíveis limitações do método, extraídas da literatura e propostas com base nos ensaios executados. Explorou-se a instabilidade do treinamento, e os possíveis impactos da entropia e características do processo de interesse na capacidade de detecção de anomalias. Sinais foram então sintetizados com a adição de tipos específicos de anomalias, a fim de verificar a generalidade do método quanto à natureza das anomalias, e uma coleção de sinais reais de domínios diversos compilados do conjunto *UCR Anomaly Benchmark*, de maneira a serem aplicados ao método. Por fim, alterações no método foram propostas, com maneiras alternativas de quantificar a anormalidade a partir dos modelos obtidos, e brevemente avaliadas. Os resultados obtidos permitiram a verificação e corroboração da grande aplicabilidade de GANs para detecção de anomalias em séries temporais, bem como da utilidade de experimentação com dados sintéticos analíticos para desenvolvimento de compreensão e validação de modificações. A exploração das limitações efetuadas permitiu o desenvolvimento de intuições sobre seus impactos no método, e sugeriram a possibilidade de influência de características do processo alvo na performance, e as modificações propostas apresentaram potencial de ganhos de performance, e apontaram a necessidade de estudos futuros aprofundados para a investigação posterior.

ABSTRACT

Keywords: Anomaly detection, Time series, Machine Learning, Deep learning, Generative Adversarial Networks.

The general problem of unsupervised anomaly detection in time series has applications in several different fields and is related to many specific problems. In the context of time series data, however, expert knowledge in the target application is often required in order to extract meaningful features of the process, which can be expansive and at times not possible. The field of Deep Learning provided techniques to tackle such problems with the possibility of automatic features extractions techniques, and present great potential in time series anomaly detection. The need for labeled data, however, restricts the direct application of several methods. GAN-based solutions have recently presented great performance in modeling arbitrary data distribution in unsupervised problems, showing a considerable conceptual potential in anomaly detection. In that context, with the goal of exploring the potential and applicability of GAN-based solutions for time series anomaly detection, the literature was reviewed for GAN and anomaly detection principles, and recent works specifically on GAN-based methods for time series anomaly detection summarized and presented. In sequence, a method was selected, TadGan (GEIGER *et al.*, 2020), due to the presence of the main principles of GAN application to anomaly detection and its good reported performance in public benchmarks, for detailed investigation and exploration. An implementation of the method was obtained, and verified over a partial reproduction of the original article results. A series of experiments over synthetic generated data from analytical functions were then proposed and executed in order to verify the method's principles in a controlled environment, as well as to raise intuitions of possible limitations. Limitations raised by the literature were then explored, and a new limitation, based on the influence of the signal entropy in the method performance, was informally formulated and investigated. Time series containing different types of anomalies were then synthesized, in order to verify the generality with respect to the nature of the anomalies, and data from real applications compiled from the *UCR Anomaly Benchmark*, and applied to the method. Finally, some modifications and suggestions of new scores derived from the method were presented, implemented and superficially analyzed. The results allowed to verify the great potential of the application of GAN-based techniques for unsupervised anomaly detection, as well as the benefits from exploring the method in synthetic data. The experimentation showed evidence of the explored limitations, in particular the influence of the target process entropy, and the proposed metrics showed potential of improvements and the need for further investigations.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DTFT	Discrete-time Fourier transform
FFT	Fast Fourier transform
GAN	Generative Adversarial Network
CNN	Convolutive Neural Network
RNN	Recurrent Neural Network
BRNN	Bidirectional Recurrent Neural Network
LTSM	Long Term Short Memory
MLE	Maximum Likehood Estimation (Estimação da Máxima Verossimilhança)
DTW	Dinamic Time Warping
SNR	Signal to Noise Ratio (Relação sinal-ruído)
BIGAN	Bidirectional GAN
FDD	Fault Detection and Diagnosis
PCA	Principal Component Analysis

SUMÁRIO

1 INTRODUÇÃO

O problema geral de detectar comportamentos anômalos e inconsistentes a partir de dados suscita grande interesse em diversos campos e se relaciona intimamente com inúmeros problemas específicos. Aplicações como detecção de fraudes em transações bancárias, detecção de falhas em processos, diagnósticos médicos, detecção de comportamentos novos em contextos científicos, entre inúmeros outros, são subproblemas diretos de detecção de anomalias.

No contexto de séries temporais, é somado à dificuldade de encontrar uma representação relevante para o processo de interesse, que capture as dinâmicas temporais do sinal. Uma abordagem habitual para detecção de anomalias consiste na extração de *features* que permitam a observação de desvios do comportamento esperado, e na aplicação de modelos analíticos ou aprendidos a partir de dados que possam quantificar esses desvios de normalidade (Boškoski *et al.*, 2009). Tal abordagem, entretanto, é limitada pela qualidade das *features*, que em geral dependem fortemente da existência de especialistas e conhecimento analítico no processo específico (ZHANG *et al.*, 2020), o que é extremamente custoso, pois demanda o emprego de metodologias estatísticas para validação robusta das *features* e por vezes impossível, por total ausência de conhecimento científico no processo. Além disso, a generalidade de tais soluções é limitada, uma vez que não obrigatoriamente as *features* extraídas para um problema específico serão relevantes para todas aplicações (ZHANG *et al.*, 2020).

Os avanços recentes em *Deep Learning* forneceram instrumentos para abordagem de tal problema, possibilitando a extração automática de *features* com resultados cada vez melhores e permitindo a produção de soluções mais genéricas e independentes de especialistas (ZHANG *et al.*, 2020). Os modelos clássicos de *Deep Learning* para classificação, entretanto, demandam um grande volume de dados supervisionados de todas as classes desejadas (ZHANG. *et al.*, 2020), e a apesar de diversos problemas de detecção de anomalias permitirem uma coleta de grande volume de dados, a obtenção de dados supervisionados é atravancada pela impossibilidade em amostrar arbitrariamente o comportamento anômalo, dada pela dificuldade intrínseca em gerá-las arbitrariamente, bem como de prever e coletar sinais referentes a todo conjunto de possíveis anomalias. Nesse sentido, a aplicação clássica de métodos supervisionados de *Deep Learning* é limitada pela existência de grandes conjuntos de dados para um problema específico.

Portanto, o desenvolvimento de arquiteturas que permitam a aplicação de técnicas de *Deep Learning* de extração automática de *features* e que sejam treinados de uma maneira semi-supervisionada ou não supervisionadas são de grande interesse, tem potencial de apresentar ganhos de desempenho expressivos e de possuir amplo campo prático de aplicabilidade.

Em especial, observa-se um crescente interesse em métodos baseados em GANs, pela sua aplicabilidade conceitual direta, dada pela sua capacidade de modelar a partir da dados distribuições probabilísticas arbitrárias, e pelo seus bons resultados em diversos problemas. Apesar da existência de soluções baseadas em GANs para detecção de anomalias em diversos campos de aplicação, como no processamento de imagem, dados coletados sobre o formato de séries temporais apresentam desafios específicos adicionais, em função principalmente da dificuldade em encontrar-se representações relevantes que também capturem a dependência temporal dos dados. Métodos de detecção de anomalias desenvolvidos para imagens ou outro formato em que os exemplos não apresentam dependência sequencial direta são dificilmente diretamente aplicáveis à análise de séries temporais.

Recentemente, métodos específicos foram desenvolvidos para a detecção de anomalias em séries temporais utilizando GANs, e apresentaram resultados competitivos em conjuntos de dados e *benchmark* públicos, como a TadGan (GEIGER *et al.*, 2020) e MADGAN (LI *et al.,* 2019). Entretanto, em face a falta de definições consensuais e formulações formais para o problema de detecção de anomalias, que culminam, por exemplo, na utilização de metodologias de avaliação diferentes entre trabalhos, questões sobre o escopo da aplicabilidade dessas soluções e possíveis limitações são levantadas. A avaliação discriminada para diferentes tipos de anomalias passíveis de detecção, bem como demonstrações em ambientes controlados da capacidade dos modelos envolvidos no método poderiam fornecer mais confiança e diretrizes para sua aplicação generalizada em problemas cotidianos da engenharia. Além disso, apesar do objetivo da aplicação de GANs ao problema de detecção de anomalias ser de quantificar a anormalidade, a avaliação dos métodos propostos em diversos trabalhos é realizada através do desempenho obtido em dados e *benchmarks*, que dependem não apenas da qualidade da métrica obtida, mas também dos algoritmos posteriores utilizados para decisão de sinalização da anomalia a partir dessa métrica. Tal metodologia de avaliação prejudica a capacidade de comparação direta entre os métodos, e por consequência a validação da aplicabilidade de redes complexas de *Deep Learning* como GANs para a tarefa de detecção de anomalias.

Neste trabalho uma revisão bibliográfica foi realizada para descrever o funcionamento de GANs e suas principais aplicações, bem como do campo geral de detecção de anomalias e suas particularidades para as séries temporais. Em sequência, trabalhos com uso de GANs realizados para detecção de anomalias no campo da imagem e em séries temporais foram revisados, e as principais abordagens existentes e modelos propostos sumarizados. Um método dentre os revisados, a *TadGan* (GEIGER, *et al.*, 2020) foi selecionado para estudo detalhado, pela seu desempenho competitivo em diversos problemas e presença dos elementos fundamentais da aplicação de GANs à detecção de anomalias. Uma implementação com possibilidades de realização de modificações foi obtida e inicialmente verificada reproduzindo-se resultados parciais reportados por (GEIGER *et al.,* 2020). Dessa maneira, pôde-se corroborar com os resultados reportados pelos autores originais, bem como

levantar intuições sobre o funcionamento geral do método e suas limitações.

Em sequência, uma metodologia foi proposta e executada no formato de estudo de caso para demonstrar o funcionamento geral do método, consistindo na avaliação dos modelos sob dados sintetizados a partir de funções analíticas, com o objetivo de garantir propriedades do sinal e das anomalias introduzidas presentes. Foi avaliado a capacidade do gerador de modelar processos no tempo, do *encoder* de encontrar e projetar para o espaço latente características relevantes do processo e das métricas derivadas do método em diferenciar processos distintos, e consequentemente, possibilitar a detecção de anomalias.

Explorou-se, então, limitações na aplicação geral de GANs à detecção de anomalias, oriundas da literatura especializada e de observações realizadas pelo autor. Deu-se especial atenção para a instabilidade do treinamento, originada tanto pela característica inerente do método GANs quanto pela formulação não-supervisionada do problema e às particularidades da utilização do erro de reconstrução como medida da anormalidade, e suas dependências tanto para com o treinamento como para características dos dados

O método foi então avaliado como um todo sobre um compilado de séries temporais com anomalias com causas conhecidas, bem como provenientes de aplicações diversas não presentes na avaliação original de (GEIGER *et al.,* 2020). Séries temporais foram sintetizadas com a inserção de tipos específicos de anomalias, e conjuntos de dados selecionados da base *UCR Anomaly Benchmark*, introduzido por (WU *et al.*, 2021), de modo a abranger aplicações diversas. Por fim, baseado no comportamento observado pelo experimentos, bem como as limitações levantadas, novas métricas derivadas do método para detecção de anomalias foram sugeridas e brevemente exploradas.

Uma panorama geral do trabalho pode ser visto na Figura 1, onde observa-se também a relação sequencial entre as seções, através da qual os experimentos realizados para a demonstração dos princípios do métodos são em parte derivados da análise dos resultados obtidos na reprodução do trabalho (GEIGER *et al.*, 2020), bem como a investigação das limitações do modo derivados da análise do funcionamento dos princípios do método sobre os dados sintetizados. A avaliação sobre séries reais e sintéticas é derivada em parte das limitações investigadas, e as modificações nas métricas do comportamento observado ao longo de todos experimentos.

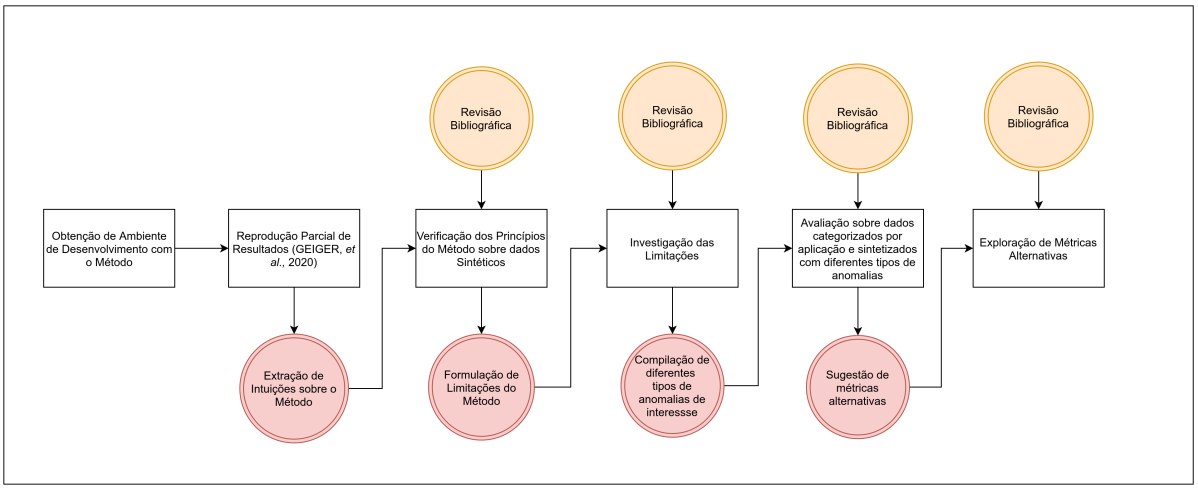

Figura 1 – Panorama geral do Trabalho.

Fonte: Autor.

2 REVISÃO BIBLIOGRÁFICA

2.1 *GENERATIVE ADVERSARIAL NETWORKS*

Nesta seção GANs são apresentadas e seus princípios básicos fundamentais explicados. Em sequência, uma formulação matemática é introduzida, assim como o procedimento típico de treinamento e principais topologias apresentados. Variações do conceito original relevantes para compreensão do funcionamento das GANs e seu potencial são listadas e brevemente discutidas.

2.1.1 Princípios Gerais

Generative Adversarial Networks é uma arquitetura para o desenvolvimento de redes generativas proposta por (GOODFELLOW *et al.*, 2014), que utiliza de um processo adversário para a obtenção do modelo gerador.

O campo de pesquisa de GANs atraiu grande interesse da comunidade acadêmica, e somente em 2018 (GUI *et al.*, 2020) estimam cerca de 11.800 publicações envolvendo GANs. Dessa maneira, uma revisão definitiva do estado da arte bem como do desenvolvimento histórico de todas aplicações e arquiteturas está além do escopo deste texto.

(GOODFELLOW, 2017) define informalmente um modelo generativo como qualquer modelo que a partir de um conjunto de dados retirados de uma mesma distribuição probabilística consegue estimar de alguma forma a distribuição. Enquanto alguns modelos resultam em uma representação explícita da função de densidade de probabilidade da distribuição, outros, como GANs, permitem apenas a geração de novos exemplos pertencentes à distribuição modelada, sendo denominados modelos de densidade probabilística implícitos (GUI, *et al.*, 2020).

As GANs podem ser formuladas matematicamente pelo princípio de Estimação da Máxima Verossimilhança (MLE), que consiste em encontrar os parâmetros para o modelo com o qual deseja-se descrever a distribuição de modo que maximizem a probabilidade do conjunto de dados de treino observados, como na expressão (1).

$$\boldsymbol{\theta}^* = \arg\max_{\theta} \prod_{i=1}^{m} p_{model}(\boldsymbol{x^i}, \boldsymbol{\theta}) \tag{1}$$

onde:

- $\boldsymbol{\theta}^*$ representa os parâmetros obtidos para o modelo em questão;
- $p_{model}(\boldsymbol{x^i}, \boldsymbol{\theta})$ representa a probabilidade dos dados x serem observados dado o modelo com os parâmetros θ.

Na Figura 2 é exemplificado o princípio da maximização da probabilidade através da regressão de uma função gaussiana para função de densidade de probabilidade de um conjunto unidimensional. Observa-se que cada amostra (representada na figura como um ponto) tenta impor uma 'elevação' da sua densidade de probabilidade.

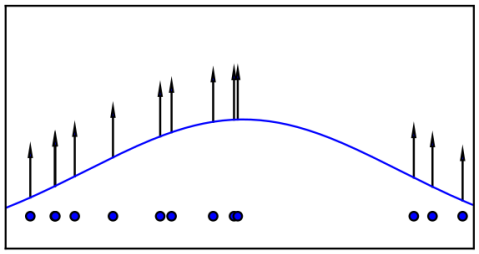

Figura 2 – Ilustração do MLE.

Fonte: GOODFELLOW, 2017

O trabalho de GOODFELLOW (2017) apresenta a taxionomia mais usada para os modelos generativos existentes baseados na formulação de maximização da probabilidade. (ver Figura 3).

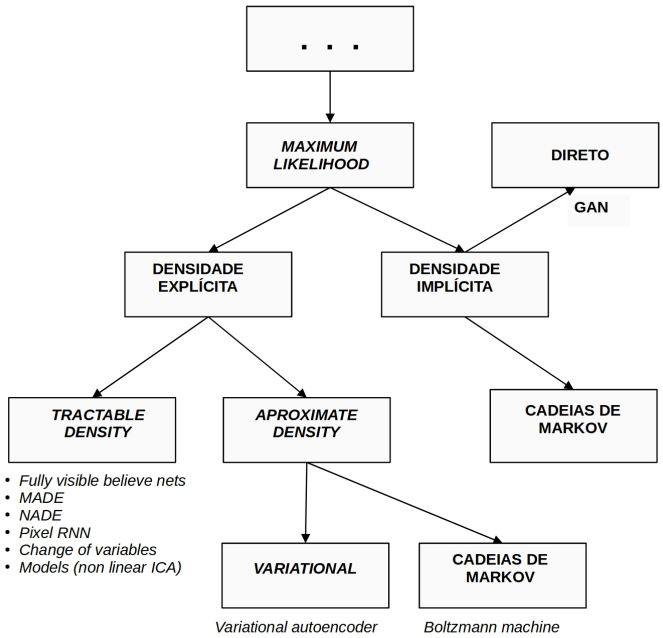

Figura 3 – Taxionomia dos Modelos Generativos.

Fonte: Adaptada livremente de GOODFELLOW, 2017

A arquitetura GAN clássica consiste em dois modelos treinados conjuntamente: uma rede **geradora**, que tem como objetivo gerar dados da distribuição de dados $\mathbf{x}$ a partir de uma entrada amostrada de uma distribuição arbitrária $\mathbf{z}$, que é tipicamente ruído,

e uma rede **discriminadora**, que classifica os exemplos entre pertencentes ao conjunto de treino ($\boldsymbol{x_{data}}$) ou sintetizados pelo gerador ($\boldsymbol{x_{gerador}}$) (GOODFELLOW *et al.*, 2014). Desse modo, pode-se associar uma função de custo ao gerador dependente da *accuracy* do discriminador, de maneira a utilizar algoritmos de otimização como gradiente descendente e ascendente. Um diagrama geral desse processo pode ser visto na Figura 4.

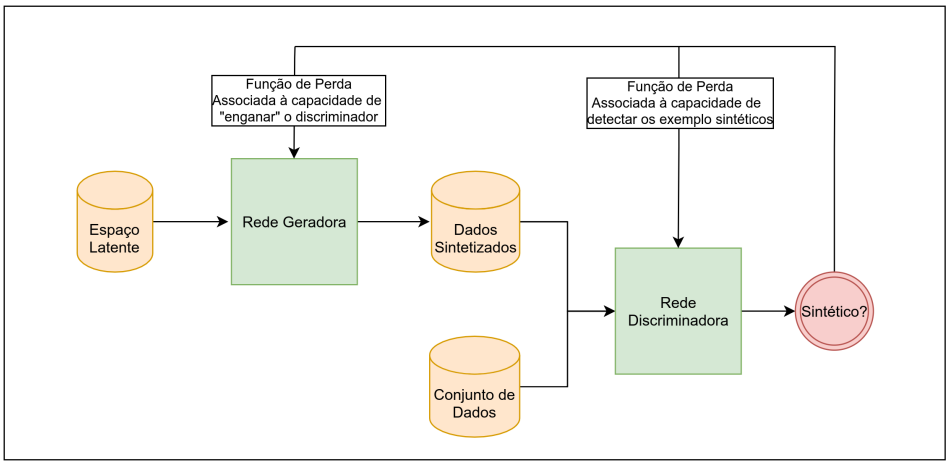

Figura 4 – Ilustração geral de uma GAN.

Fonte: Autor.

Matematicamente, uma função de custo de *cross-entropia* pode ser associada ao discriminador, dada pela equação (2):

$$J^{(D)}(\boldsymbol{\theta^D}, \boldsymbol{\theta^G}) = -\frac{1}{2}\,\mathbb{E}_{x_{data}} \log D(x) - \frac{1}{2}\,\mathbb{E}_{z_{data}} \log\left(1 - D(G(z))\right) \tag{2}$$

onde:

- $\theta^{\boldsymbol{D}}$: parâmetros do discriminador;
- $\theta^{\boldsymbol{G}}$: parâmetros do gerador;
- D(x): predições do discriminador para **x**, dentro de {0,1};
- G(z): exemplos gerados a partir de **z**;

Observa-se que a função de custo é minimizada quando a média das predições em $D(x)$ dado $\boldsymbol{x_{data}}$ se aproxima de 1 e quando $D(G(z))$ se aproxima 0, significando que o discriminador consegue distinguir entre exemplos do conjunto de dados e sintéticos. A função de perda para o gerador pode ser então simplesmente formulada como na expressão (3)

$$J^G = -J^D \tag{3}$$

Em outras palavras, o gerador é treinado para tentar "enganar" o discriminador gerando exemplos o mais próximos possíveis da distribuição **x**, enquanto o discriminador

tenta diferenciar entre exemplos gerados e originais do conjunto de dados (GOODFELLOW *et al.*, 2014) . GOODFELLOW *et al.*, (2014) ressaltam que o processo de treinamento de ambos modelos pode ser compreendido mais precisamente como um jogo, descrito formalmente pelo equilíbrio de Nash, do que como um problema de otimização tradicional, dados que as funções de custo de ambos modelos estão associados à performance dinâmica do outro modelo ao invés de um espaço de dados supervisionados estático.

De acordo com GOODFELLOW *et al.,* (2014) o treinamento pode ser representado em uma formulação de otimização *"mín máx"* como representado na Equação (4).

$$\min_G \max_D V(D,G) = \mathbb{E}_{xdata}[logD(x)] + \mathbb{E}_Z[log(1 - D(G(z)))] \tag{4}$$

Sendo $D(x)$ o discriminador, que prediz a probabilidade de um exemplo ser verdadeira em $\{0,1\}$, deseja-se escolher os parâmetros do gerador de modo a minimizar o termo $1 - D(G(z))$, implicando na predição dos dados gerados como verdadeiros pelo discriminador, e concorrentemente escolher os parâmetros do discriminador de modo a maximizar $D(x)$ e $1 - D(G(z))$, implicando na classificação dos dados sintetizados como falsos.

Desse modo, através das funções de custo definidas pode-se atualizar o gerador e o discriminador respectivamente através de gradiente descendente e ascendente. O treinamento pode ser implementado como descrito na Figura 5.

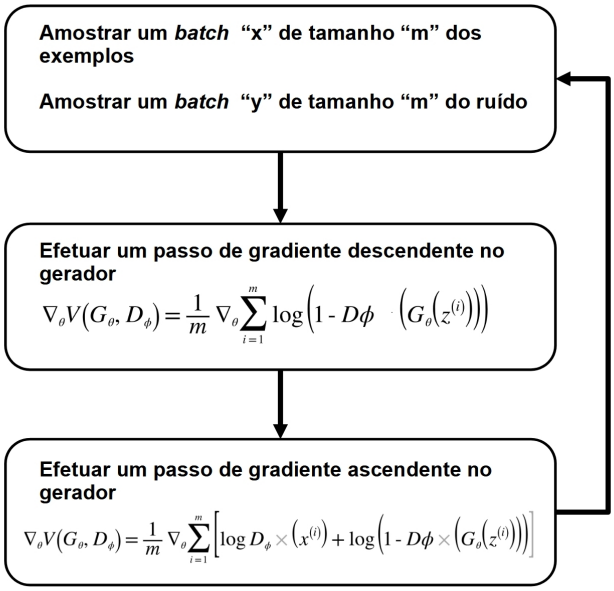

Figura 5 – Diagrama simplificado de treinamento de GANs.

Fonte: Autor.

É comum em aplicações práticas a realização de mais de uma interação sobre o discriminador para cada interação no gerador, a fim de melhorar a estabilidade do treinamento.

Apesar da existência teórica de soluções únicas para o problema de otimização de GANs, o processo de treinamento é frequentemente instável, resultando, por exemplo, no colapso do modelo generativo e dificuldades gerais na convergência de ambos modelos (GUI, *et al.*, 2020). O colapso do modelo generativo é a situação em que o gerador "memoriza"como gerar um pequeno conjunto de dados de forma realística, ao invés de verdadeiramente modelar a distribuição alvo. O estudo teórico do treinamento de GANs, bem como soluções práticas para melhor desempenho é vasto (GUI, *et al.*, 2020), mas permanecem, em geral, com os princípios descritos na Figura 5. Funções de perdas distintas e alterações na arquitetura foram amplamente estudadas e propostas para mitigar as instabilidades do treinamento foram introduzidas (GUI, *et al.*, 2020)

O comportamento do espaço **z** também é fonte de interesse e estudos. (RADFORD *et al.,* 2016) exploraram o comportamento de GANs com operações no espaço latente **z**. Na Figura 6 observa-se que ao subtrair a representação **z** responsável por gerar um homem com óculos da responsável por gerar um homem sem óculos e adicionar a de uma mulher, o gerador gerou imagens de mulheres com óculos, dando indicações não só da possibilidade de operações significativas no espaço **z** como da efetiva capacidade de GANs de capturar e mapear conceitos da distribuição alvo separadamente. Tal propriedade tem relevância para a aplicação de GANs na detecção de anomalias.

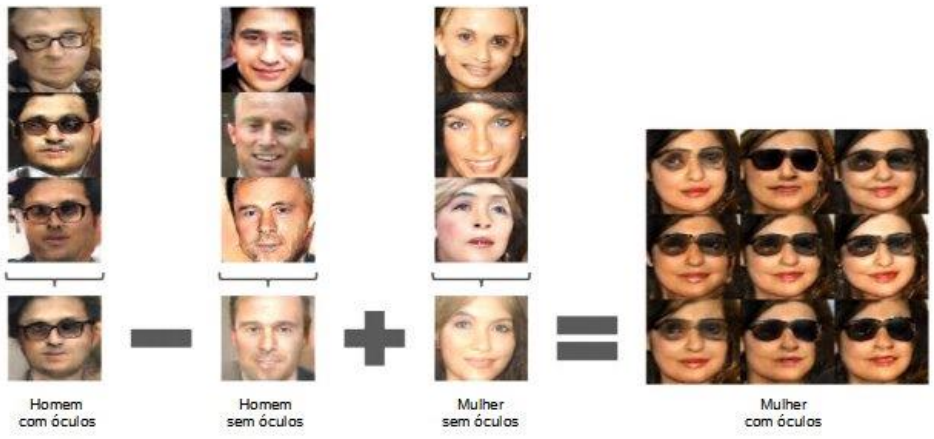

Figura 6 – Resultado de operações no espaço z na imagem gerada.

Fonte: Adaptada de RADFORD *et al.*, 2016.

2.1.2 Introdução a Arquitetura GAN e Algumas Variações Propostas

As redes Geradora e Discriminadora podem admitir diferentes arquiteturas, desde que com seus elementos constituintes diferenciáveis para possibilitar a aplicação de algorítimos de otimização de gradiente descendente/ascendente (GOODFELLOW *et al.*, 2014). Historicamente, (GOODFELLOW *et al.*, 2014) inicialmente utilizou redes neurais profundas para suas demonstrações dos princípios das GANs. (RADFORD *et al.*, 2016) consolida-

ram a utilização de redes convolucionais profundas (CNNs) com GANs para aplicações em imagens, demonstrando sua capacidade de geração de imagens com grande realismo.

No contexto de séries temporais, (HYLAND *et al.,* 2017) aplicaram RNNs para o Gerador e Discriminador e demostraram boa performance para geração de séries temporais na área médica. OUYANG *et al.*, (2018) demonstraram o uso bem sucedido de redes LSTM para sintetização de imagens a partir de textos sequênciais.

Diversas variações baseadas no conceito de GANs foram propostas. Em sequência são listadas algumas ideias relevantes ao desenvolvimento geral do campo e da compreensão da capacidade das GANs, bem como para aplicação à detecção de anomalias. As arquiteturas propostas diretamente para detecção de anomalias serão apresentadas nas seções seguintes.

InfoGAN, proposta por (CHEN *et al.*, 2016), tem como objetivo mapear o espaço de dados de forma relevante para a representação latente **z**, de maneira a garantir que ela codifique características significativas dos dados. Par tal, o espaço amostrado pelo gerador é dividido entre um componente de ruído puro **z** e uma componente denominada de *"código latente"* **c**. A infoGAN consiste em resolver o problema de otimização apresentado na (5).

$$\min_G \max_D = V(D,G) - \lambda I(C, G(z,c)) \tag{5}$$

onde I é a informação mútua entre a representação **c** e o dado gerado, que é maximizada no treino. Mecanismos adicionais são necessários para computação do termo da informação mútua, notavelmente uma rede adicional (CHEN *et al.,* 2016). CHEN *et al.,* (2016) mostraram que foi possível identificar características dos exemplos gerados diretamente na representação c obtida. A Figura 7 mostra o impacto de uma componente do espaço c resultante - c_1- nos exemplos gerados para uma infoGAN e uma GAN convencional, sobre a base de dados MNIST.

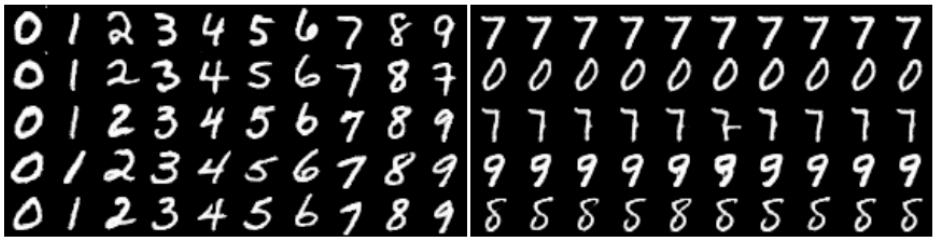

Figura 7 – Demonstração da relevância semântica da representação resultante.

Fonte: Adapatado de CHEN *et al.*, 2016.

Outra variação de GAN proposta é a *conditional GAN* (**cGANs**), com o objetivo de condicionar a geração de exemplo a alguma outra informação, como anotações sobre a imagem ou a escolha do espaço **z** como algum outro espaço de dados (GUI *et al.,* 2020). A informação condicional pode ser simplesmente concatenada à entrada z, e o discriminador

adaptado para receber concorrentemente a imagem gerada e a informação adicional (GUI, *et al.*, 2020) . O conceito tem diversas aplicações, como aumento de resolução de imagem e operações sobre faces humanas. Na Figura 8 pode-se observar a GAN *pix2pix* (ISOLA *et al.*, 2018), que colore imagens preto-e-branco.

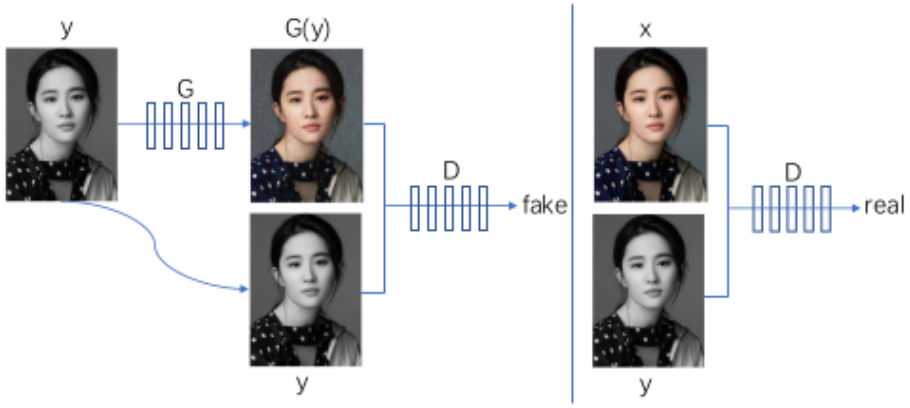

Figura 8 – Diagrama funcional da cGANs.

Fonte: GUI, *et al.,* 2020.

Nesse caso o Discriminador avalia sobre uma entrada dupla pareada, a imagem colorida e em preto-e-branco, de modo que o gerador deve gerar duplas coerentes, aprendendo a colorir a imagem e garantindo correspondência semântica entre a imagem de entrada e colorida. Nesta situação dados anotados - ou seja, imagens coloridas pareadas com não coloridas - são necessários. O problema alvo pode ser formalmente expresso pela equação (6) (GUI, *et al.*, 2020):

$$L_{cGANs}(D,G) = \mathbb{E}_{x,y}[\log(D(x,y))] + \mathbb{E}_{y}[\log(1 - D(y, G(y)))] \tag{6}$$

onde o Discriminador recebe a informação adicional, bem como o exemplo x.

Nesse contexto, ZHU *et al.,* (2017) propuseram a topologia denominada **CycleGAN** com o objetivo de possibilitar tradução imagem - imagem sem a necessidade de exemplos pareados. Ao contrario da topologia *pix-to-pix* apresentada, uma função G para mapear entre dois diferentes espaços de dados, $G : X \rightarrow Y$ pode ser obtida apenas com exemplos disponíveis de ambos espaços.

Para tal, deseja-se que um classificador não consiga diferenciar exemplos do espaço Y gerado por $G(X)$ de exemplos amostrados diretamente de Y. ZHU *et al.,* (2017) ressaltam que essa condição não é suficiente, uma vez que não implica em uma tradução significativa com correspondência semântica entre os dois espaços. Assim, é introduzido o conceito de uma função de perda pela "consistência no ciclo", garantindo que um mapeador inverso treinado $F : Y \rightarrow X$ inverta G para o exemplo original (ZHU *et al.,* 2017). Na Figura 9 pode-se ver ilustrado esse princípio.

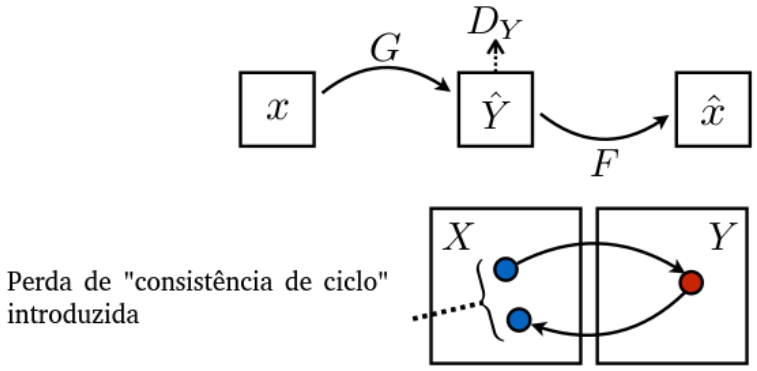

Figura 9 – Definição intuitiva da perda de consistência de ciclo.

Fonte: Adaptada de ZHU *et al.,* 2017.

Como exemplo ilustrativo (ZHU *et al.,* 2017) apresentam a tradução do modelo proposto entre uma fotografia e uma pintura de Monet (ver Figura 10).

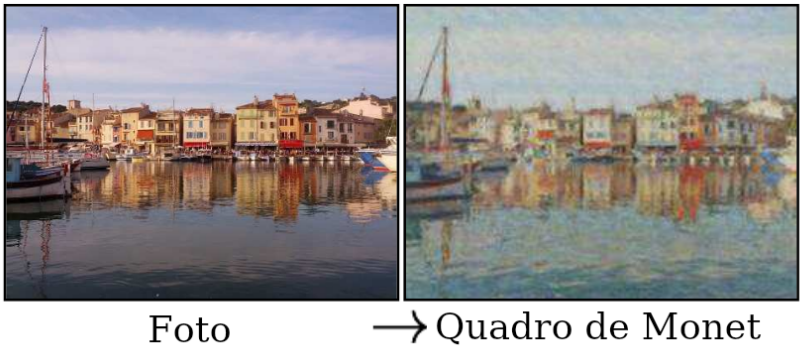

Figura 10 – Resultados obtidos na tradução de fotos para pinturas de Monet.

Fonte: Adaptada de ZHU *et al.*, 2017.

Com objetivo de melhorar a estabilidade do treinamento, diversas reformulações das funções de perda e do treinamento foram propostas. Como exemplo relevante pode-se citar a Wasserstein GAN proposta por (ARJOVSKY *et al.*, 2017), com um algoritmo alternativo para o treinamento de modo a melhorar a convergência e evitar o colapso do modelo Gerador resultante (GUI *et al.*, 2020).

Motivado pelas possibilidades originárias da existência de significado das operações realizadas no espaço $\mathbf{z}$, demonstradas por (RADFORD *et al.*, 2016), (CHEN, *et al.,* 2016) entre diversos outros autores, DONAHUE *et al.*, (2016) introduziram o conceito de *Adversarial Feature Learning*, no qual deseja-se a obtenção de uma função inversa do gerador, de modo a mapear do espaço de exemplos $\mathbf{x}$ para o espaço $\mathbf{z}$, tornando-o uma representação do espaço de exemplos. A obtenção desse mecanismo foi realizada através de uma modificação na arquitetura tradicional da GAN para o aprendizado conjunto de um *Encoder* para mapear o espaço de dados para o espaço latente. A arquitetura resultante foi

denominada de **BiGAN** (Bidirectional GAN), conforme esboço apresentado na Figura 11.

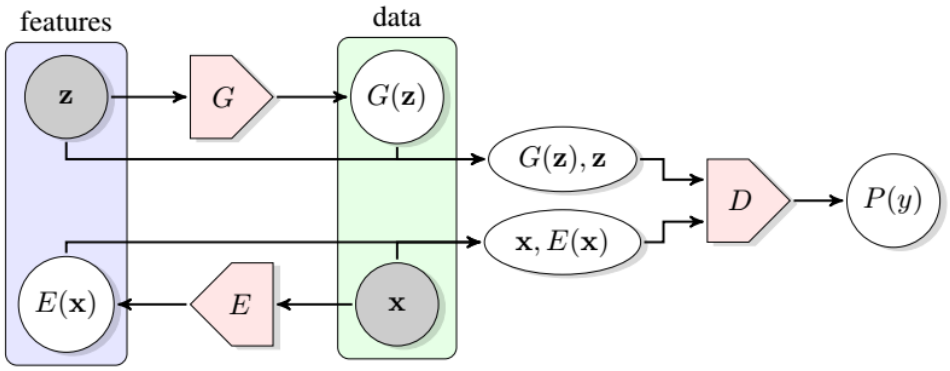

Figura 11 – Estrutura da BiGAN proposta.

Fonte: DONAHUE *et al.*, 2016

De acordo com DONAHUE *et al.*, (2016) o processo de otimização da BiGAN por ser representado pela Equação (7) :

$$\min_{G,E} \max_{D} = V(D, E, G) \tag{7}$$

onde:

$$V(D,E,G) := \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{X}}}\Big[\underbrace{\mathbb{E}_{\mathbf{z}\sim p_E(\cdot|\mathbf{x})}\left[\log D(\mathbf{x},\mathbf{z})\right]}_{\log D(\mathbf{x},E(\mathbf{x}))}\Big] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{Z}}}\Big[\underbrace{\mathbb{E}_{\mathbf{x}\sim p_G(\cdot|\mathbf{z})}\left[\log\left(1-D(\mathbf{x},\mathbf{z})\right)\right]}_{\log(1-D(G(\mathbf{z}),\mathbf{z}))}\Big]$$

Constata-se a modificação realizada no discriminador para classificar sobre duplas, analogamente ao modelo cGAN. Apesar dos resultados de (ARORA *et al.*, 2017) demonstrando as limitações teóricas de convergência do conceito de *Adversarial Feature Learning* da forma utilizada pela BiGAN, de acordo com as quais o problema de otimização formulado admite soluções que não atendem os objetivos do modelo, o conceito teve sua validade prática empiricamente demonstrada pelos resultados dos trabalhos de (ZENATI *et al.*, 2018) e (AKCAY *et al.*, 2018).

3 FUNDAMENTAÇÃO TEÓRICA

Nesse capítulo conceitos fundamentais da detecção de anomalias e suas peculiaridades na aplicação à séries temporais são apresentados. Em sequência, o método baseado em GANs selecionado para experimentação é descrito em detalhes.

O capítulo está dividido da seguinte maneira:

- A seção 3.1 revisa os fundamentos da detecção de anomalias e suas particularidades para séries temporais;
- A seção 3.2 detalha o funcionamento da TadGan para detecção de anomalias.

3.1 DETECÇÃO DE ANOMALIAS

Nesta subseção serão revisados princípios fundamentais da detecção de anomalias. O conceito de anomalia será definido e o problema formulado, abordando-se as dificuldades da sua caracterização precisa e as confusões decorrentes nesta área.

É importante ressaltar que o campo de detecção de anomalias se confunde muitas vezes com suas aplicações. *Fault Detection and Diagnosis* (FDD), por exemplo, utiliza extensivamente técnicas de detecção de anomalias, mas tem como objetivo final a detecção de defeitos relevantes ao processo alvo, e não a detecção de anomalias em seu sentido mais puro. As sutis diferenças e confusões decorrentes da atribuição ou não de valor semântico às anomalias alvos geram discordâncias na literatura. Neste trabalho, detecção de anomalias será tratada como o problema geral de detectar anomalias a partir de dados, sem compromisso semântico ou aplicado dos resultados obtidos. Nesse contexto, soluções *"model-based"*, baseadas em modelo físico analítico descritivo do processo, não constituem parte do campo de estudo, uma vez que implicam conhecimento específico sobre os processos. A possibilidade ou não de ter um estudo de detecção de anomalias genérico completamente agnóstico das aplicações também é fonte de debate. Para muitos domínios gerais de dados, como imagens e séries temporais, conjuntos de dados para detecção de anomalias concatenando vários problemas distintos foram compilados, de modo a desenvolver métodos de avaliação e ambientes de desenvolvimento o mais gerais possíveis em relação à aplicação. Críticas existem a essas abordagem, como no trabalho de (AHMED *et al.*, 2019) que ressaltaram a falta de sentido no tratamento genérico da detecção de anomalias de um ponto de vista de avaliação, e (WU *et al.*, 2021), que apontaram trivialidade em muitos exemplos presentes nesses conjuntos compilados, decorrentes do desenvolvimento não vinculado a aplicações específicas. Muitas dessas críticas parecem florescer da falta de uma taxonomia semântica para anomalias dentro dos grandes grupos de dados, de modo a discriminar a performance dos métodos resultantes de maneira

objetiva para diferentes fontes de anomalias. A tentativa de estudo da detecção de anomalias desvinculada de aplicações cria também problemas para a definição e delimitação de anomalias.

Classificações para tipos de anomalias e diferentes instâncias do problema elaboradas por diferentes autores serão comparativamente apresentadas, mostrando a falta de taxonomias consensuais e classificações relevantes para tipos de anomalias.

3.1.1 Definição e Formulação do Problema

O problema de detecção de anomalias emerge de diversos ramos com interesse em diferenciar e detectar comportamentos inesperados e inconsistentes com o comportamento esperado. Exemplos de aplicações incluem detecção de invasão em redes de comunicação (LIAO *et al.*, 2013), detecção de fraudes em diversos cenários, detecção de prevenção de falhas para monitoramento de processos (RABATEL *et al.*, 2013), auxílio à diagnóstico médicos (CHAUHAN *et al.*, 2015), aplicações gerais nas ciências, aplicações em processamento de sinais, entre diversos outros (RUFF *et al.*, 2021).

Para tipos de dados complexos e multidimensionais, a diferenciação do comportamento normal para o anormal é em geral fortemente dependente da representação escolhida (RUFF *et al.*, 2021). A possibilidade do aprendizado automático de representações relevantes motivou a aplicação de técnicas de *Deep Learning* ao ramo, com o objetivo de possibilitar soluções sem especialistas no processo específico.

Não tem-se uma definição definitiva consensual para anomalia, nem uma delimitação clara para o problema (BLÁZQUEZ-GARCÍA *et e al.*, 2021). Classicamente, (HAWKINS, 1980) define anomalia informalmente como: *"Uma observação que desvia tanto das outras observações que gera suspeita de que foi gerada por um diferente mecanismo"*

BOUGUESSA *et al.*, (2018) define uma anomalia como uma observação inconsistente com o resto do conjunto de dados.

KISHAN *et al.*, (2017) coloca como uma variação substancial da norma.

YANG *et al.*, 2021 apontam a falta de definições precisas para o problema de detecção de anomalias e as dificuldades decorrentes para desenvolvimento da área, mostrando as estreitas relações entre os problemas de *Detecção de anomalias* (Anomaly Detection), *Novelty Detection*, *Open Set Recognition*, *Out-of-Distribution Detection* e *Outlier Detection*. Os autores sugeriram ainda uma metodologia para delimitação e classificação dos problemas, colocando-os como casos particulares da detecção generalizada de *"fora da distribuição"* (*Generalized Out-of-Distribution Detection*), apontado como objetivo geral a detecção e quantificação de "desvios da distribuição"(*distribution shift*).

RUFF *et al.*, (2021) formalizaram a definição de anomalia dentro de um contexto probabilístico, colocando que dado um espaço de dados $X \subseteq R^D$, define-se o conceito de normalidade como a distribuição P^+ dentro de X. Sendo então $P^+(X)$ a função de

densidade de probabilidade de P^+, o conjunto de anomalias pode ser escrito como na equação (8).

$$A = \{\boldsymbol{x} \in \chi \mid p^+(x) \leq \tau\}, \tau \geq 0 \tag{8}$$

Sendo τ um limiar de decisão de modo que a probabilidade de A sobre P+ seja suficientemente pequena. Informalmente, então, o conjunto A é o conjunto de observações com baixa probabilidade dada a distribuição. A partir dessa definição de anomalia, o problema de detecção de anomalias torna-se um problema de estimação do conjunto de um nível de densidade de probabilidade, ou seja, de encontrar as regiões no espaço de representação com densidade de probabilidade abaixo de um liminar, de modo a caracterizar anomalias. O conjunto D formado por todos elementos que satisfazem um dado nível de densidade de probabilidade pode ser formulado como na equação (9) (CHEN *et al.*, 2016).

$$D_h \equiv D_h(\lambda) = \{x : p_h(x) = \lambda\} \tag{9}$$

Na aplicação usual de detecção de anomalia apenas dados positivos, ou seja, pertencentes ao conjunto da normalidade, estão disponíveis (GOLDSTEIN *et al.*, 2016), de modo que tipicamente a ideia principal do problema é aprender um modelo descritivo do comportamento normal, de maneira a pode-se quantificar desvios desse comportamento (RUFF *et al*, 2021).

RUFF *et al.*, (2021) apontam a existência de diversas particularidades que diferenciam e adicionam complexidade à detecção de anomalias em relação a outros problemas. A inexistência de dados anômalos em tempo de treinamento, a grande variabilidade possível dentro da classe dos dados normais, a possibilidade da não estacionaridade do comportamento da normalidade e a necessidade de diferenciar ruído de comportamentos anômalos devem ser levados em consideração no desenvolvimento de soluções.

AHMED *et al.*, (2019) abordam os problemas da falta de formalismo dos objetivos do problema, colocando em questão as abordagens genéricas agnósticas de aplicações, e como resultam em metodologias de avaliação duvidosas, baseadas em conjuntos de dados montados sem a avaliação de performance em um problema específico.

Nas próximas subseções, as diferentes configurações do problema serão apresentadas, e classificações para os tipos de anomalias e modelos utilizados nas soluções descritas.

3.1.2 Classificações do Problema

3.1.2.1 Quanto aos Dados

Em contraste com problemas clássicos de classificação de dados, a detecção de anomalias é frequentemente empregada em situações onde não existem dados supervisionados representativos das duas classes: normal e anormal, seja pela dificuldade em gerar situações anômalas arbitrariamente e de prever e coletar sinais referentes a todo conjunto de

possíveis falhas, seja pelo inerente caráter difuso do estado anômalo, como por exemplo desgastes que evoluem continuamente com o tempo. Pode-se classificar a detecção de anomalias, segundo (GOLDSTEIN *et al.*, 2016) em:

- **detecção de anomalias supervisionadas:** quando tem-se disponível um conjunto de dados com exemplos identificados relativos ao comportamento normal e anormal. Nesse caso, algoritmos tradicionais de classificação podem ser utilizados;
- **detecção de anomalias semi-supervisionada:** caso geral onde tem-se um conjunto referente ao estado normal de operação, sem a presença de anomalias ou contaminação. (RUFF *et al.*, 2021) coloca também o caso onde tem-se um conjunto de dados que misture dados supervisionados com não supervisionados;
- **detecção de anomalias não supervisionada:** caso onde não existe qualquer forma de supervisão para os dados, sendo que o conjunto disponível pode conter exemplos normais e anormais em proporções desconhecidas;

É importante ressaltar que este trabalho se concentra nos casos semi-supervisionados e não supervisionados, dada sua maior aplicabilidade e interesse prático.

3.1.2.2 Quanto ao Tipo de Anomalias

O problema de detecção de anomalias também pode ser categorizado quanto ao tipo de anomalia presente (CHANDOLA *et al.,* 2017). Definições formais não são fornecidas, de modo que a literatura por vezes apresenta interpretações ligeiramente distintas. Descrições gerais das categorias são apresentadas abaixo.

- **anomalias pontuais:** caso onde uma amostra individual consiste uma anomalia, e pode ser considerado anômalo em relação a todo o conjunto;
- **anomalias contextuais ou condicionais:** caso onde uma amostra é considerado anômalo somente dado o seu contexto. Tem grande relevância na análise de séries temporais. Como exemplo, visualizar a Figura 12.

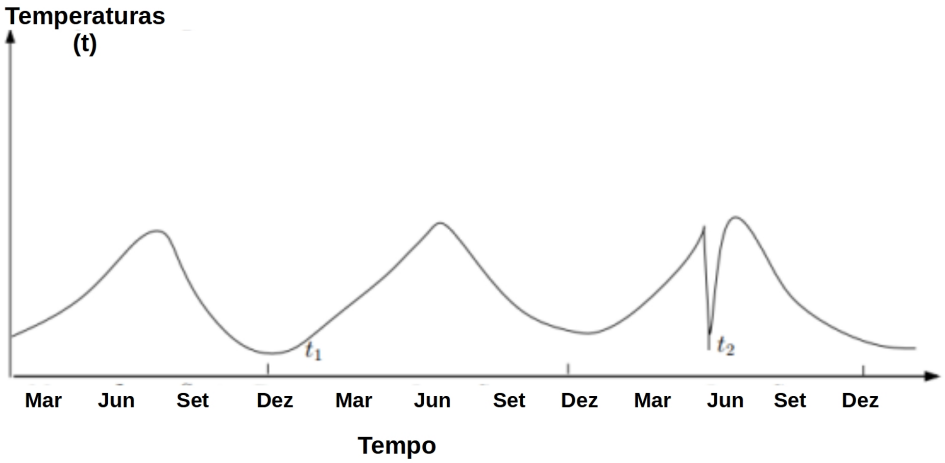

Figura 12 – Anomalias contextuais em uma série temporal.

Fonte: Adaptada de CHANDOLA *et al.*, 2017.

- **anomalias coletivas:** se a ocorrência conjunta de exemplos é anômalo, mesmo que a ocorrência individual dos exemplos não seja. (Ver a Figura 13)

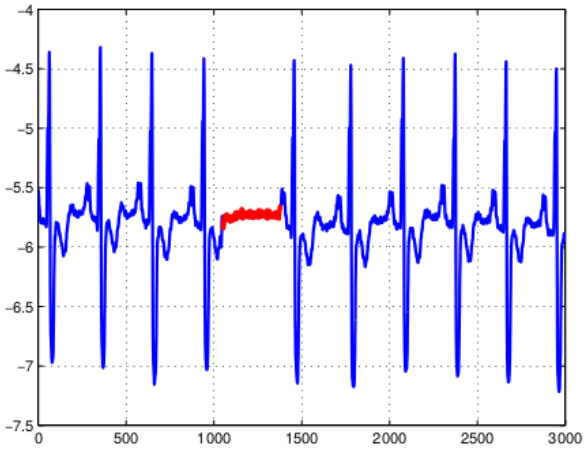

Figura 13 – Anomalias coletivas em uma série temporal.

Fonte: (CHANDOLA *et al.*, 2017.

(YANG *et al.*, 2019) introduziram duas categorias adicionais para classificação das anomalias, colocando que dada uma distribuição conjunta P(X,Y), onde X é um espaço de representação dos dados e Y um espaço das anotações dos dados, *"desvios da distribuição"* - ou seja, as anomalias - podem ser subclassificados como **desvios semânticos**, onde tem-se variação no espaço de anotações, como aparecimento de novas classes, e **desvios de covariância**, dados por variações no espaço da representação dos exemplos. Nesse contexto, o problema de detecção de anomalias tem como escopo tanto a detecção de desvios semânticos como de covariância.

(YANG *et al.*, 2019) apresentam como exemplo que dado um problema dentro de um espaço de fotos de cachorros, um desenho de um cachorro consiste em desvio em covariância, ou seja, um exemplo semanticamente consistente com a distribuição, por ser um cachorro, mas anômalo por ser um desenho. Uma foto de um gato, por outro lado, consiste em um desvio semântico, introduzindo uma nova classe. Pode-se ver essas diferenças ilustradas na Figura 14.

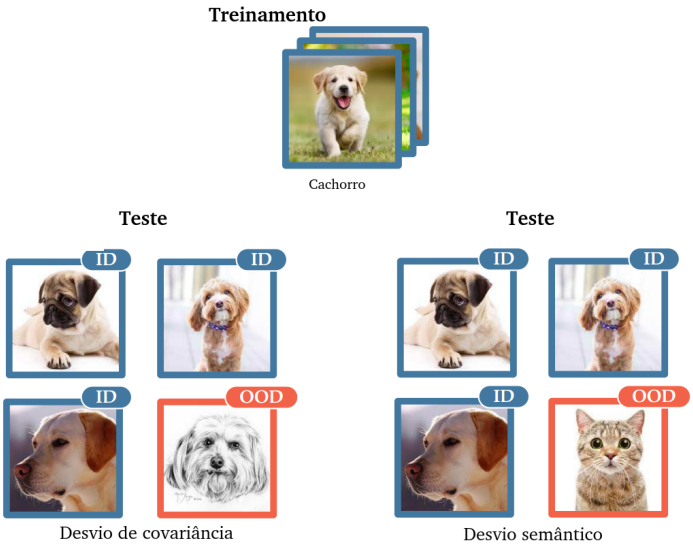

Figura 14 – Desvio de distribuição de covariância e semântico.

Fonte: Adaptado livremente de (YANG *et al.* 2019).

(RUFF *et al.*, 2021) adiciona duas categorias semelhantes, anomalias de baixo nível sensoriais, que acontecem por mudanças em propriedades de baixo nível do exemplo e anomalias de alto nível semânticas, dadas pela introdução de novas classes e grandes variações em características fundamentais .

(AHMED *et al.*, 2019) colocam a detecção de desvios semânticos como a tarefa principal da detecção de anomalias. (RUFF *et al.*, 2021) ressaltam a importância da representação do processo derivado dos dados para possibilitar a detecção de desvios semânticos, motivando ainda mais o uso de *Deep Learning* para obtenção de tais representações automaticamente a partir dos dados.

3.1.2.3 Quanto às abordagens de solução

Diversas diferentes categorizações para as classes de abordagens e soluções para detecção de anomalias estão presentes na literatura, por vezes com taxonomias contraditórias.

De forma geral, (KISHAN *et al.*, 2017) classificam os principais princípios intuitivos para desenvolvimento de algoritmos para metrificação de não normalidade como baseado em **distância** - ou seja, na avaliação explícita de alguma métrica de distância de um ponto para os outros-, baseado em agrupamento de dados, (***clustering***) - utilizando a

separação dos dados em *cluster* para identificação de pontos não pertencentes- e baseado em **modelos** - onde aprende-se um modelo descritivo do processo -. (GOLDSTEIN *et al.*, 2016) colocam mais genericamente que dada uma representação de um processo, a grande maioria das abordagens baseadas em dados para a detecção de anomalias partem da construção de um modelo descritivo do comportamento não anômalo, a partir do qual pode-se verificar a normalidade de uma amostra. As subcategorias de (KISHAN *et al.*, 2017) parecem atender essa generalização, na qual a abordagem descrita com baseada em modelos se refere a utilização de modelos paramétricos. (KISHAN *et al.*, 2017) classifica ainda a utilização de modelos paramétricos para detecção de anomalias como realizado no **espaço de parâmetros** ou no **espaço de dados**. Detecção no espaço de parâmetros consistem em avaliar a variação nos parâmetros do modelo com a inclusão de uma amostra suspeita de ser anômalo. Essa abordagem é computacionalmente custosa para modelos muito complexos, dado que diversos modelos necessitam ser aprendidos durante o processo de predição, e enfrenta problemas para conjuntos de dados muito grandes onde a influência de amostra individuais é pequena (KISHAN *et al.*, 2017). Alternativamente, a detecção no espaço dos dados consiste na utilização direta do modelo descritivo do processo obtido para quantificar a normalidade de novos exemplos.

(YANG *et al.*, 2019) e (RUFF*et al.*, 2021) apresentam uma taxonomia semelhante para as abordagens à detecção de anomalias quanto aos objetivos dos modelos utilizados, dividindo em 4 principais grupos:

1. **baseado em classificação:** modelos com objetivo de estabelecimento de uma fronteira de decisão no espaço de exemplos através de classificadores de uma classe. *One-Class SVM*, *One Class Neural Networks* são exemplos;

2. **baseado em estimação de densidade de probabilidade**: modelos com objetivo de estimar a função de densidade de probabilidade, implicitamente ou explicitamente;

3. **baseado em reconstrução**: Modelos com objetivo de aprender uma representação latente ótima para os dados não anômalos, bem como uma função para reconstruir o dado original a parir dessa representação de modo a quantificar anomalias com o erro de reconstrução;

4. **baseado em distância**: métodos que explicitamente calculam uma distância no espaço de dados para quantificação da anormalidade

(RUFF *et al.*, 2021) adicionam ainda mais uma dimensão na categorização, distinguindo entre abordagens ***Shallow* (rasas)**, nas quais a obtenção de uma representação ideal do problema não é tarefa do modelo, ou onde as representações obtidas são de baixa complexidade, e soluções ***Deep (Profundas)***, nas quais a obtenção de uma representa-

ção apropriada à complexidade do problema é parte do objetivo. Na Figura 15 pode-se ver exemplos de soluções para cada uma das classes propostas.

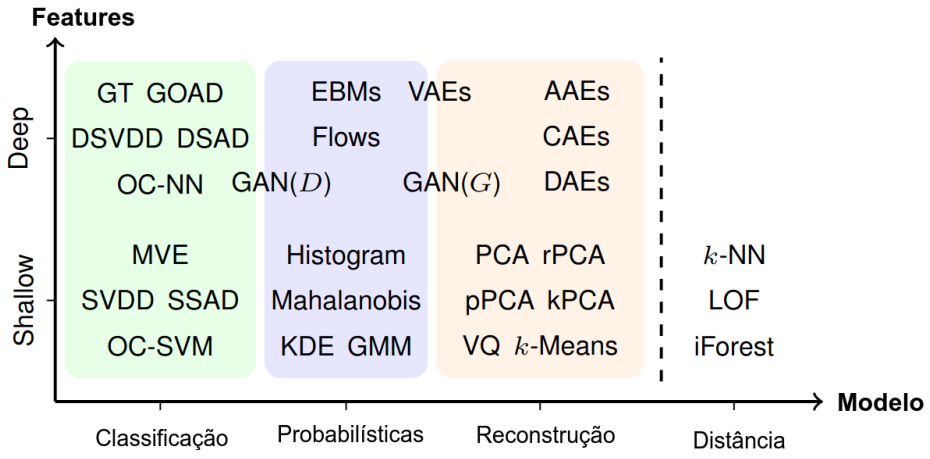

Figura 15 – Classificação das abordagens à detecção de anomalias quanto aos objetivos dos modelos

Fonte: Adaptada de (RUFF *et al.*, 2021).

Na Figura 16 pode-se ver esquematizado a natureza de cada uma das abordagens apresentadas, mostrando a função de decisão como uma fronteira no espaço dos dados. Nota-se os distintos objetivos de cada um das abordagens: para classificação uma fronteira de decisão delimitada, para as probabilísticas formas de avaliar a densidade de probabilidade do exemplo e para reconstrução a mensuração do erro de reconstrução dos exemplos.

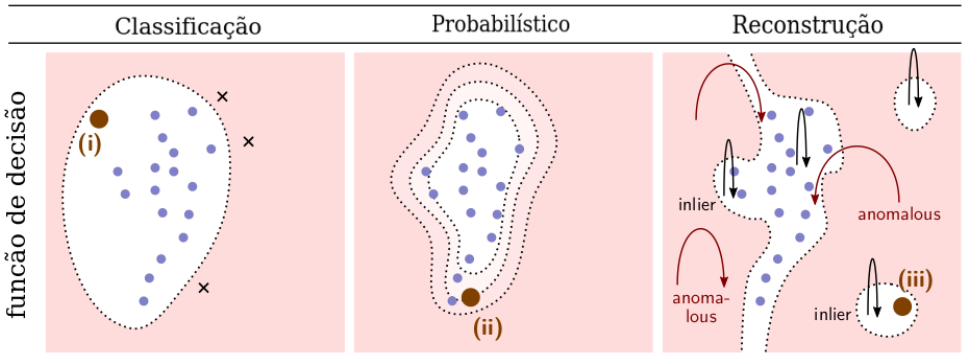

Figura 16 – Natureza das abordagens apresentadas para detecção de anomalias.

Fonte: Adaptada de (RUFF *et al.*, 2021)

3.1.3 Detecção de Anomalias em Séries Temporais

Detecção de anomalias em séries temporais difere das aplicações tradicionais, e demanda abordagens modificadas. (AGGARWAL *et al.*, 2017) ressaltam que as marcas temporais não podem ser tratadas simplesmente como uma dimensão adicional, o que

simplificaria para uma análise multidimensional, de modo que necessita-se do estabelecimento de representações e modelos especiais para a série temporal. Sendo assim, a aplicação de algoritmos para o caso de séries temporais enfrenta o problema da elaboração de representações relevantes que possam capturar o caráter sequencial dos dados, motivando ainda mais a utilização de técnicas baseadas em *Deep Learning* que permitam o aprendizado automatizado a partir dos dados (RUFF *et al.*, 2021).

(AGGARWAL *et al.*, 2017) colocam que no contexto de detecção de anomalias à séries temporais, as anomalias são ou contextuais ou coletivas.

(GEIGER *et al.*, 2020) colocam que o problema de detecção de anomalias em séries temporais consiste em encontrar subsequências anômalas de tamanho variado, e define formalmente como que dada uma série temporal $\boldsymbol{X} = (x^1, x^2, ..., x^T)$ deseja-se encontrar uma sequência $\mathbf{A}$ $(a_{\boldsymbol{seq}^1}, a_{\boldsymbol{seq}^2}, ..., a_{\boldsymbol{seq}^k})$ de exemplos anômalos, e ressalta que não tem-se nenhum conhecimento *a priori* das anomalias, nem de como segmentar o sinal temporal de modo significativo. (GEIGER *et al.*, 2020) apresentam ainda uma classificação alternativa para as abordagens dentro do contexto de séries tempos, classificando-as como:

- **baseadas em proximidade:** dada uma representação numérica do processo, define-se exemplos anômalos por medidas relacionadas com a identificação de proximidade deles, sendo por definições multivariáveis de distância ou densidade. (GEIGER *et al.*, 2020) ressaltam os problemas na captura das correlações temporais, bem como a dificuldade de formular representações significativas;

- **baseados em predição:** aprendizado de um modelo descritivo preditivo do processo, e quantificação de desvios do comportamento normal baseado através dos desvios entre o comportamento observado e predito pelo modelo. Tem como vantagem a possibilidade de utilização direta das técnicas de análise de séries temporais para modelagem e predição de comportamento;

- **baseadas em reconstrução:** aprendizado de modelos capazes de desenvolver representações latentes de menor dimensão do processo. Com a presunção do treinamento semi-supervisionado na normalidade, espera-se que a representação latente desenvolvida não deve representar efetivamente o comportamento anormal, ocasionando maiores erros de reconstrução para anomalias. A linha principal de utilização de GANs para detecção de anomalias utiliza esse princípio.

3.1.4 GANs para detecção de anomalias

Nessa subseção será apresentada uma breve revisão das aplicações recentes de GANs no problema de detecção de anomalias. Métodos gerais desenvolvidos originalmente para o campo de processamento de imagem serão revisados.

GANs e o conceito de treinamento *adversarial* mostraram-se muito bem sucedidos em modelar distribuições de dados altamente complexas para diversas aplicações, indicando sua grande aplicabilidade na detecção de anomalias, a partir da modelagem da distribuição probabilística do conjunto de dados do comportamento não-anômalo (DI MATTI *et al.*, 2021). Sendo um modelo generativo implícito, entretanto, ou seja, que não resulta na modelagem direta da distribuição de probabilidade, mas em uma função que mapeia dois espaços de dados diferentes, métodos complementares devem ser desenvolvidos para quantificação da normalidade dos dados de interesse (DI MATTI *et al.*, 2021).

DI MATTI *et al.*, (2021) colocam que todos trabalhos já realizados na aplicação de GANs diretamente para detecção de anomalias utilizam-se basicamente do conceito geral do *Adversarial Feature Learning*, apresentado por (DONAHUE *et al.*, 2016), na qual é adicionada à estrutura tradicional da GAN algum mecanismo que permita mapear exemplos do espaço de dados para o espaço $\mathbf{z}$ latente. (DONAHUE *et al.*, 2016) introduziram a obtenção desse mecanismo através de uma modificação na arquitetura tradicional de uma GAN para o aprendizado conjunto de um *encoder* para mapear o espaço de dados para o espaço latente, denominada por BiGAN. Outros métodos para obtenção desse mecanismo também formam utilizados para detecção de anomalias, como será apresentado em sequência.

O conceito geral da aplicação de GANs para detecção de anomalias parte do treinamento com apenas dados positivos, ou seja, não-anômalos, esperando-se que a GAN modele a distribuição probabilística adjacente da normalidade. Em algumas aplicações conjuntos contaminados são utilizados, com a presunção de que dada a pequena frequência relativa das amostras anormais, esse comportamento não será capturado e modelado pela GAN. Como resultado do treino obtém-se, então, um modelo que transforma do espaço de dados positivo para o espaço de *features*, e um modelo que transforma do espaço de *features* para o espaço de dados positivo. A fim de quantificar a normalidade de um novo exemplo, ou seja, sua probabilidade dada a distribuição modelada para a normalidade, utiliza-se de uma abordagem baseada em reconstrução, transformando o exemplo alvo para o espaço de *features* e em sequência novamente para o espaço de dados. Dessa forma, espera-se que a reconstrução de exemplos anômalos apresenta um maior erro, no momento em que o gerador tenta gerar sempre exemplos normais (DI MATTI *et al.* 2021). Além disso, observou-se nas aplicações à imagens que regiões anormais do exemplo apresentaram as maiores diferenças, no momento em que tiveram que ser modificadas para geração de um exemplo normal (SCHLEGL *et al.*, 2017).

O conceito de utilizar a performance de reconstrução para quantificar a normalidade é comum às aplicações de *Autoencoders* ao problema.

Diversas arquiteturas que implementam esse conceito geral foram propostas, e aplicadas à detecção de anomalias em imagens. A seguir serão apresentadas algumas das principais arquiteturas propostas.

AnoGAN, proposta por (SCHLEGL *et al.,* 2017), foi uma das primeiras arquiteturas baseadas em GANs para detecção de anomalias, e baseia-se no princípio descrito acima, através da medição do erro de reconstrução dos dados alvos. Para obtenção de uma função que mapeia o exemplo de entrada para o espaço de *features* $\mathbf{z}$, um processo interativo de minimização é realizado sobre cada exemplo alvo, através do algorítimo de *backpropagation* com uma função de perda definida como uma soma ponderada entre a *'residual loss'*, dada pelo erro de reconstrução, conforme indicado pela equação (13):

$$\mathcal{L}_R(\mathbf{z}_\gamma) = ||\mathbf{x} - G(\mathbf{z}_\gamma)||_1 \tag{10}$$

e a *'discriminator loss'*, que quantifica a semelhança entre as *features* do exemplo reconstruindo e do original através da utilização de *layers* do discriminador f, como visto na equação (11).

$$\mathcal{L}_D(\mathbf{z}_\gamma) = ||\mathbf{f}(\mathbf{x}) - \mathbf{f}(G(\mathbf{z}_\gamma))||_1 \tag{11}$$

Apesar de ter introduzido a utilização de GANs para detecção de anomalias, e de ter obtidos performance comparativamente boa em relação à outros métodos na detecção de anomalias de imagens de tecidos humanos, o processo de predição é extremamente custoso, uma vez que necessita de otimização iterativa sobre cada dado alvo (GHERBI *et al.*, 2019.).

(ZENATI *et al.*, 2018) introduziram a utilização da BiGAN para séries temporais, denominado de **EGBAD**, na qual a função operadora do espaço de dados para a representação latente é aprendida conjuntamente com o gerador, e não durante a predição para cada dado individual como na AnoGAN. Dessa forma, permitiu-se a utilização de GANS para detecção de anomalias sem a necessidade de realizar iterações de otimização durante a predição.

(AKCAY *et al.*, 2018) introduziram a **GANomaly**, inspirados pelos trabalhos anteriores. Ao invés de partir de uma distribuição arbitrária para o espaço latente, um *Autoencoder* é utilizado para encontrar uma representação $\mathbf{z}$ relevante dos exemplos, representação que é passada por dois *decoders* adicionais, formando o gerador da GAN.

(DI MATTI *et al.* 2021) realizou experimentos comparativos com as três arquiteturas descritas, e encontrou performances inferiores para GANomaly nos conjuntos de dados utilizados. Na Figura 17 pode-se ver sumarizadas as três arquiteturas apresentadas

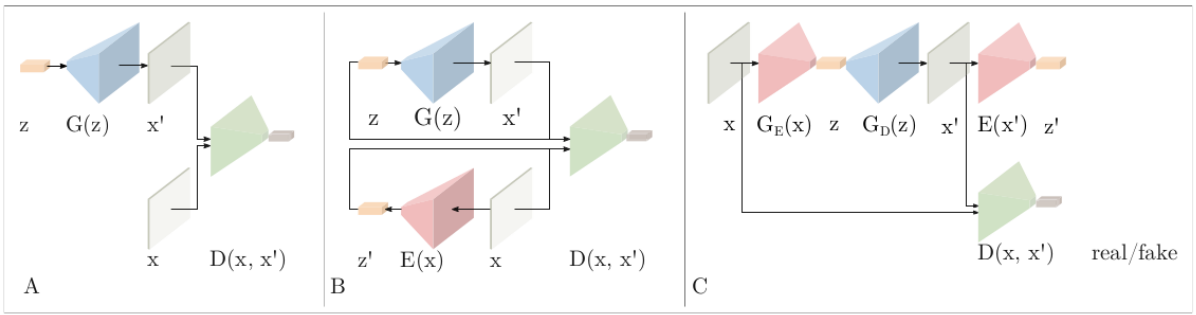

Figura 17 – Comparação entre as arquiteturas para aplicação de GAN's à detecção de anomalias, respectivamente AnoGAN, EGBAD e GANomaly.

Fonte: DI MATTI *et al.*, 2021.

3.1.5 GANs na Detecção de Anomalias em Séries Temporais

Apesar das existência de várias aplicações de GANs para detecção de anomalias envolvendo imagens, sua aplicação para detecção em séries temporais é recente e constitui uma área emergente (YIFAN *et al.,* 2021). Nessa subseção serão revisadas alguns modelos recentes baseados em GANs propostos para séries temporais, suas arquiteturas gerais e metodologias de avaliação empregadas.

Nota-se que, em geral, as abordagens desenvolvidas seguem adaptações dos princípios utilizados para imagem apresentados na secção 3.1.4, baseando-se em grande parte em modelar o comportamento não-anômalo e encontrar um mecanismo para mapear do espaço de dados $\mathbf{x}$ para a representação latente $\mathbf{z}$, utilizando principalmente o erro de reconstrução como um escore de anomalias. São também notável as diferenças metodológicas encontradas para a avaliação das soluções desenvolvidas, manifestadas em diferentes conjuntos de dados utilizados, métricas de avaliação e definições de anomalias no contexto de séries temporal.

(LI *et al.*, 2019) propuseram a **MAD-GAN**, um método não supervisionado para detecção de anomalias em séries temporais multivariável, utilizando GANs com *LSTM-RNN* como modelos de base. Um novo escore foi introduzido, *DR-score*, utilizando tanto a predição do discriminador treinado sobre a sequência $\mathbf{x}$ de entrada quanto o erro de reconstrução do gerador obtido. As diversas séries temporais são combinadas com janelas de tamanho empiricamente otimizadas. Durante o treino, o Gerador e Discriminador são treinados nos exemplos não-anômalos, com o objetivo de capturar a distribuição da normalidade. Em tempo de predição, as sequências de entrada são mapeadas para a representação latente $\mathbf{z}$, $\boldsymbol{x} \rightarrow \boldsymbol{z}$, em um procedimento análogo à (SCHLEGL *et al.*, 2017), por otimização. (LI *et al.*, 2019) ressaltam que o Discriminado obtido nesse processo tem grande *accuracy* em classificar exemplos normais, de modo que pode também identificar diferenças entre exemplos normais e anômalos. O escore final é um combinação entre a predição do discriminado e o erro de reconstrução. Na Figura 18 pode se ver um

diagrama do modelo proposto. O modelo foi avaliado usando os conjuntos de dados **SWaT** e **WADI** , com a métrica *F1*. Os autores citam que dada a criticalidade da aplicação, o *recall* apresentou maior relevância para o desenvolvimento, dado que falsos positivos são, segundo (LI *et al.*, 2019), para as aplicações de detecção de anomalias citadas pelo autor, mais aceitáveis que falsos negativos. Os autores obtiveram resultados superiores em relação aos métodos de base implementados. Na Figura 11 pode-se ver sumarizado o funcionamento da MAD-GAN. O lado esquerdo demonstra o procedimento de treinamento descrito, através do qual o Discriminador e Gerador são obtidos, enquanto a direita da figura mostra a atuação em tempo de predição, com a combinação do erro de reconstrução e escore do discriminador como métrica da anormalidade.

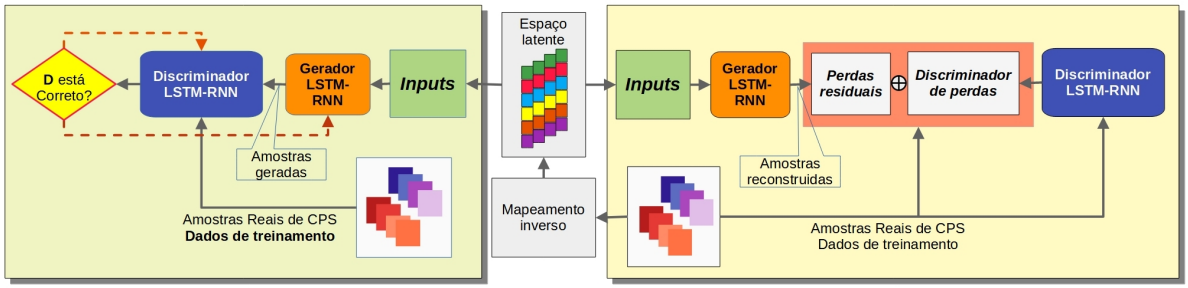

Figura 18 – Diagrama do funcionamento da MAD-GAN.

Fonte: Adaptada livremente de (LI *et al.*, 2019).

(JIANG *et al.*, 2019) propuserem um sistema para detecção de falhas em séries temporais baseados em GANs, com enfoque em problemas com grande desbalanço de classes, como a formulação semi-supervisionada de detecção de anomalias apresentada, na qual apenas exemplos normais estão disponíveis para treinamento. Foi utilizada uma abordagem análoga à de (AKCAY *et al.*, 2018), na qual um *autoencoder* treinado conjuntamente com a GAN é utilizado para mapear do espaço de dados $\mathbf{x}$ para a representação $\mathbf{z}$. O sistema apresentado não opera diretamente no sinal no tempo, entretanto, mas em *features* computadas paras as séries do conjunto de dados, selecionadas da literatura como *features* clássicas para problemas de séries temporais. O escore de anomalia atribuído da-se pela soma do erro de reconstrução do exemplo de entrada e pelo erro de reconstrução da representação latente pela estrutura do *autoencoder*, como visto na equação (12).

$$A(x) = ||x - x^{'}|| + ||z - z^{'}|| \tag{12}$$

O modelo foi empregado no conjunto de dados de falhas de rolamento de máquinas elétricas *CWRU*, disponível e descrito em (http://csegroups.case.edu/bearingdatacenter/home), bem como em dados coletados pelos autores em uma configuração análoga, obtendo-se resultados de performance relevantes. Utilizou-se como métrica principal de avaliação a **AUC** da Curva ROC, uma medida da área sobre a curva da taxa de falsos positivos e

verdadeiros positivos (FPS, TPR) com a variação do *threshold*. Na Figura 19 pode-se ver a arquitetura desenvolvida pelos autores.

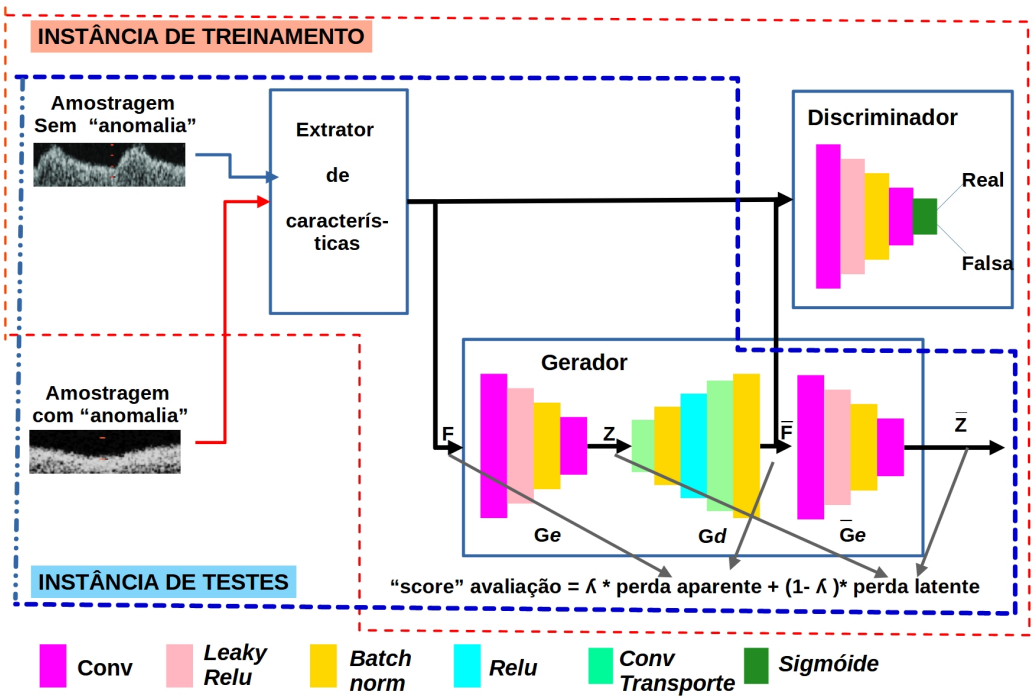

Figura 19 – Método proposto baseado em GAN para detecção de anomalias.
Fonte: Adaptada livremente de (JIANG *et al.*, 2019)

(SUN *et al.*, 2019) aplicaram um modelo baseado em GAN para detecção de falhas em veículos comerciais, provenientes de 75 sensores não especificados no trabalho instalados em 40 ônibus por 3 anos. O Discriminador teve sua arquitetura implementada com *layers* de CNN de uma dimensão, enquanto o Gerador como uma rede neural. Ambos modelos foram treinados de forma *adversarial* com os dados não-anômalos, de modo a modelar a distribuição da normalidade. A rede do Discriminador resultante foi utilizada diretamente para a predição de anomalias, sendo defendido pelos autores que após o treinamento o Discriminador consiste em um modelo capaz de distinguir dados pertencentes à classe normal dos não pertencentes. Utilizou-se como *threshold* para detecção de anomalia a predição do Discriminador dada para um dado sintético gerado pelo Gerador, sob o argumento de que quando um dado real obtivesse um escore menor que o artificial seria anormal. Métodos de base não foram implementados para comparação, e os resultados obtidos mostraram-se compatíveis com a aplicação alvo de monitoramento do estado de operação da frota de veículos.

(KHOSHNEVISAN *et al.*, 2020) propuseram a **RSM-GAN**, com objetivo de detec-

ção de anomalias em séries temporais multivariáveis, através da conversão das séries em imagens e da aplicação de uma GAN convolucional recorrente com adição de mecanismo de atenção para captura de sazonalidade. As séries temporais multivariáveis são transformadas em imagens através da matriz multicanal de correlação, (MCM), inspirada nos trabalhos de (SONG *et al.*, 2018) e (ZHANG *et al.*, 2019), com a qual para cada amostra em cada instante de tempo são extraídas janelas das diversas séries temporais, com uma defasagem definida, e a correlação entre as janelas é computada, de modo a se formar, para n variáveis, uma matriz n x n. São computadas três matrizes com tamanhos de janelas distintos, de modo a capturar características com dimensões temporais diferentes. Quatro matrizes sequenciais são finalmente definidas como a entrada do modelo, de modo a capturar as dinâmicas da evolução temporal. A quantificação da anormalidade dá-se pelo erro de reconstrução, e o mapeamento do espaço de exemplos para a representação latente analogamente ao trabalho de (AKCAY *et al.*, 2018), com o uso de uma estrutura *encoder-decoder-encoder*. A avaliação foi feita sobre um conjunto de dados proprietário referente a 36 sensores coletados de uma usina de energia juntamente com dados gerados artificialmente. O modelo apresentou performance superior em comparação com os modelos de base implementados. Na Figura 20 pode-se ver a arquitetura geral da GAN utilizada para detecção.

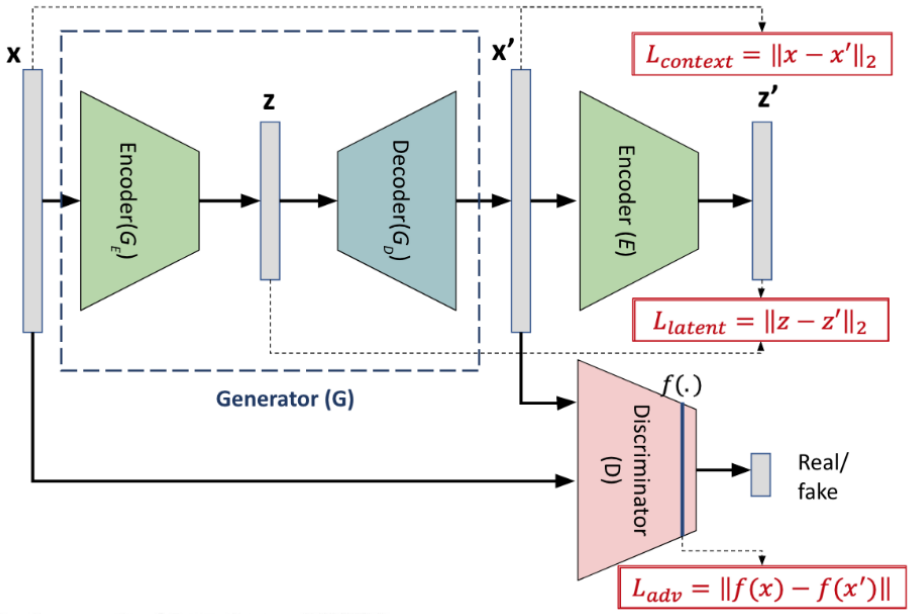

Figura 20 – Arquitetura Geral da RSM-GAN.

Fonte: KHOSHNEVISAN *et al.*, 2020

O escore de erro de reconstrução foi diretamente utilizado após o treinamento do modelo, e um *threshold* ajustado de forma não supervisionada através da divisão dos dados em validação e teste, simplesmente com a multiplicação do erro por um coeficiente definido.

(BASHAR *et al.*, 2020) propuseram a **TAnoGAN** para detecção de anomalias com GANs. A quantificação da anormalidade dá-se por uma combinação do erro de reconstrução e predição do discriminador, com pesos de cada média ajustados empiricamente. O mapeamento do espaço de dados para a representação latente dá-se através de gradiente descendente em tempo de predição, análogo ao trabalho de (SCHLEGL *et al.,* 2017). Baseado em resultados experimentais, o Gerador foi constituído de três *layers* LSTM com 32, 64 e 12 células, e o Discriminador com um *layer* LTSM com 100 unidades. As séries temporais foram divididas em janelas de tamanho 60, definidas empiricamente. O modelo foi avaliado no conjunto de dados **NAB**, e teve sua performance comparativamente avaliada em relação com MADGAN, um *Autoencoder* e uma rede LSTM baseada em reconstrução. Os resultados comparativos demonstraram o resultado superior da TAndoGAN em relação a todos modelos de base implementados sobre os dados avaliados. Utilizou-se como métrica principal de avaliação a AUC da Curva ROC, uma medida da área sobre a curva da taxa de falsos positivos e verdadeiros positivos (FPS, TPR) com a variação do *threshold*. Os autores apresentam a métrica de anormalidade dos exemplos como descrito na equação (13).

$$L = (1 - \gamma) * L_r + \gamma * L_d \tag{13}$$

onde L_r é o erro de reconstrução ponto-a-ponto e L_d o escore atribuído pelo Discriminador para o exemplo. O parâmetro γ foi otimizado empiricamente. Os autores citam como problemas a instabilidade do treinamento do modelo, notavelmente o número de Épocas utilizado, bem como a necessidade da escolha do tamanho da janela e da sua grande importância para captura das anomalias relevantes.

(GEIGER *et al.*, 2020) introduziram a **TadGAN** para detecção de anomalias não supervisionada em séries temporais, bem como uma metodologia de avaliação conjuntos de dados existentes. A TadGAN consiste em aprender conjuntamente com o gerador um mapeador inverso $\boldsymbol{x} \rightarrow \boldsymbol{z}$, denominado de *Encoder*. Isso é feito introduzindo dois discriminadores adversários: C_x, que avalia os exemplos no espaço de dados, entre os sintetizados pelo gerador $G : \boldsymbol{z} \rightarrow \boldsymbol{x}$ e os presentes no conjunto de dados original, e C_z, que avalia no espaço da representação latente, entre exemplos amostrados diretamente da distribuição e sintetizados pelo *encoder* através de $E : \boldsymbol{x} \rightarrow \boldsymbol{z}$. Ao modelo C_x é atribuído a uma perda Wasserstein, introduzida por (ARJOVSKY *et al.*, 2017) e o modelo C_z a uma perda do tipo *Cycle Consistency Losse* (ZHU, *et al.*, 2017) . Na Figura 21 pode-se ver um diagrama do funcionamento geral da TadGAN.

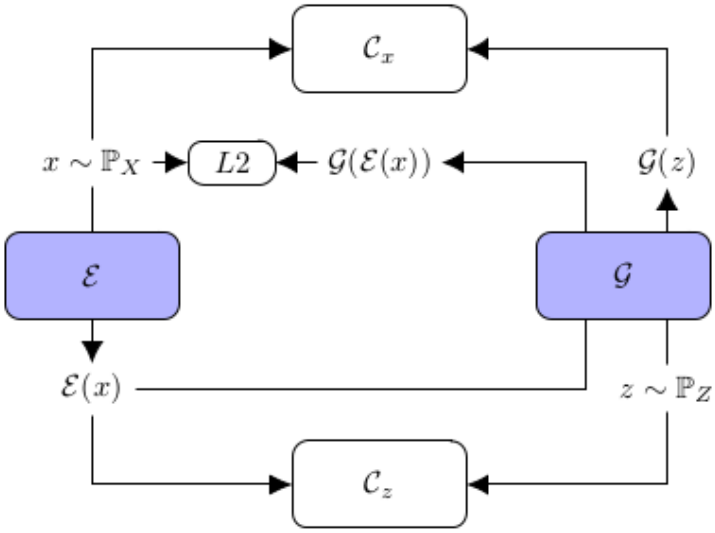

Figura 21 – Diagrama de funcionamento da TadGAN.

Fonte: GEIGER *et al.*, 2020.

Observa-se na Figura 21 os principais princípios envolvidos, onde C_x recebe dados retirados diretamente do conjuntos de dados e os sintetizados pelo gerador; C_z rebebe dados amostrados diretamente da distribuição $\mathbf{z}$ e os como sintetizados pelo *encoder*. É também ilustrada a *Cycle Consistency Losse*, gerada pela norma L2 entre o $\mathcal{G}(\mathcal{E}(x_i))$ e o exemplo original x.

Os princípios descritos resultam no problema a ser resolvido representando pela Equação (14).

$$\min_{\mathcal{E},\mathcal{G}} \max_{\mathcal{C}_x,\mathcal{C}_z} V_x(\mathcal{C}_x, \mathcal{G}) + V_z(\mathcal{C}_z, \mathcal{E}) + V_{l2}(\mathcal{G}, \mathcal{E}) \tag{14}$$

(GEIGER *et al.,* 2020) também divide as séries temporais em janelas, e assume normalidade em todo o conjunto, na formulação semi-supervisionada do problema. O escore de anomalia é dado pela combinação do erro de reconstrução e do resultado do Discriminador, análogo à (LI *et al.*, 2019). Para o erro de reconstrução foi avaliado o uso da diferença ponto-a-ponto das séries, diferença na integral e diferença após alinhamento por DTW (*Dynamic Time Warping*). Os autores relataram melhores resultados com a DTW. A combinação dos escores dá-se pela equação (15).

$$\mathcal{A}(x) = \alpha Z_{RE}(x) Z_{\mathbf{C}_x}(x) \tag{15}$$

onde Z_{RE} representa o erro de reconstrução e $Z_{\mathbf{C}_x}$ o escore do Crítico em x. O método foi avaliado em um compilado de conjuntos de dados. O funcionamento dos modelos suas arquiteturas, treinamento e avaliação serão detalhados na secção 3.2.

(NIU, Z. *et al.*, 2020) implementaram uma VAE-GAN baseado em redes LSTM, treinando conjuntamente com a GAN um *Variational Autoencoder* para mapear do espaço de

dados para a representação latente **z**. O escore de anomalia é atribuído de forma análoga à (BASHAR *et al.*, 2020), constituindo uma combinação do escore do Discriminador e do erro de reconstrução. Os autores enfatizaram o ganho de performance em tempo de predição comparativamente à abordagem da MADGAN e TAnoGAN nas quais é necessário mapear o espaço de dados para a representação latente por gradiente descendente para cada exemplo. O modelo resultante foi avaliado usando um subconjunto do *Yahoo* e *KPI*, através da métrica de F1. O *threshold* para detecção de anomalias foi determinado com anomalias do conjunto de teste, rompendo em parte com a premissa semi-supervisionada. O modelo proposto atingiu performance superior em ambos conjuntos de dados, demonstrando a validade da abordagem baseada no *Variational Autoencoder* para mapear do espaço de dados para a representação latente.

(DU *et al.,* 2021) propuseram a **FGANomaly**, com o objetivo de utilizar GANs para detecção de anomalias no contexto totalmente não supervisionado, sob o qual a assunção de um conjunto de treinamento sem anomalias não pode ser efetuada. Os autores introduziram uma função de perda dinâmica, que pondera os exemplos durante o treinamento em busca de favorecer os identificados como normais, bem como um sistema de filtragem. O modelo resultante foi testado em diversos conjuntos de dados com níveis de contaminação altos, e apresentou performance superior aos métodos de base.

(LIANG *et al.*, 2021) propuseram a **NVAE-GAN**, baseada nos avanços recentes nas topologias VAE. O modelo tem como entrada séries temporais de uma dimensão, codificadas como imagens com *Gramian Angular Field* (GAF) e *Recurrence Plots.* Um *autoencoder variacional* tem como objetivo garantir a continuidade da representação latente obtida, de modo a poder-se amostrar da representação latente com a garantia da geração de uma exemplo da distribuição de dados. O escore de anomalia é atribuído por erro de reconstrução, e as anomalias detectadas baseadas em desvios maiores que 2 desvios padrões dos escores. O modelo foi avaliado comparativamente com os conjuntos MSL e SMAP da NASA e as seleções de conjunto da NAB utilizado por TADGAN, de modo a possibilitar uma comparação direta. A métrica de avaliação utilizada foi o F1, como descrito por (GEIGER *et al.*, 2020). A média dos resultados obtidos por todos os conjuntos foi superior aos métodos de base.

3.2 TADGAN PARA DETECÇÃO DE ANOMALIAS

A detecção de anomalias aplicada a séries temporais pode ser dividida em duas etapas distintas: a obtenção de uma métrica capaz de quantificar a anormalidade de cada amostra temporal do sinal, e uma metodologia de decisão com capacidade de classificar cada amostra como normal ou anormal baseado na métrica obtida.

Nesta seção os fundamentos da aplicação da TadGan para detecção de anomalias foram apresentados. A seção está dividida da seguinte forma:

- A seção 3.2.1 detalha a arquitetura dos modelos e o treinamento da rede;
- A seção 3.2.2 detalha o procedimento de treinamento para os modelos do método;
- Por fim, a seção 3.2.3 apresenta o método de decisão utilizado por (GEIGER *et al.*, 2020) a fim de possibilitar comparação de resultados.

3.2.1 Arquitetura dos Modelos

Nessa subseção o funcionamento da TadGan de (GEIGER *et al.* 2020) será detalhado, bem como os valores utilizados na implementação. A TadGan consiste em 4 modelos treinados simultaneamente: um *encoder* do espaço de dados para a representação latente $\mathbf{z}$ ($\mathcal{E}$), um gerador que mapeia a representação latente $\mathbf{z}$ para o espaço de dados $\mathbf{x}$ ($\mathcal{G}$), um discriminador no espaço de dados ($\mathcal{C}_x$), e um discriminador no espaço de representação latente ($\mathcal{C}_z$). Uma visão geral dos modelos envolvidos e seus objetivos pode ser vista na Figura 22.

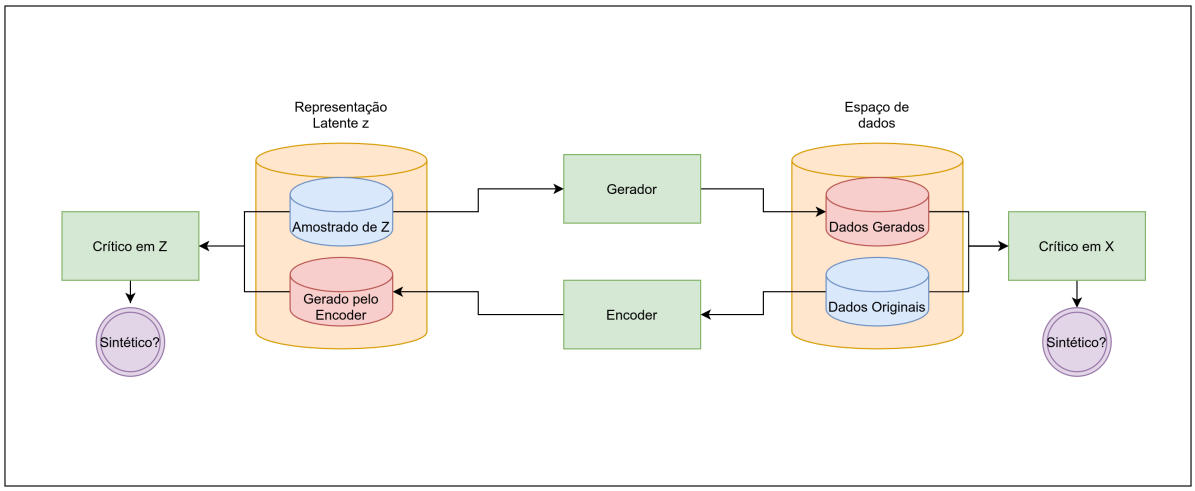

Figura 22 – Visão geral dos modelos da TadGan.

Fonte: Autor.

3.2.1.1 Crítico no Espaço de Dados

Atribui um escore as janelas de tamanho n no domínio do tempo, de modo à estimar a probabilidade da janela ser proveniente dos dados originais, e não gerada pelo Gerador. Implementada pelos autores como uma CNN de uma dimensão contendo de 4 *layers* convulsionais com 64 filtros e tamanho de *kernel* 5. A saída é gerada através de um *layer* denso de uma rede neural, para a geração de um escore. A arquitetura geral do modelo pode ser vista na Figura 23.

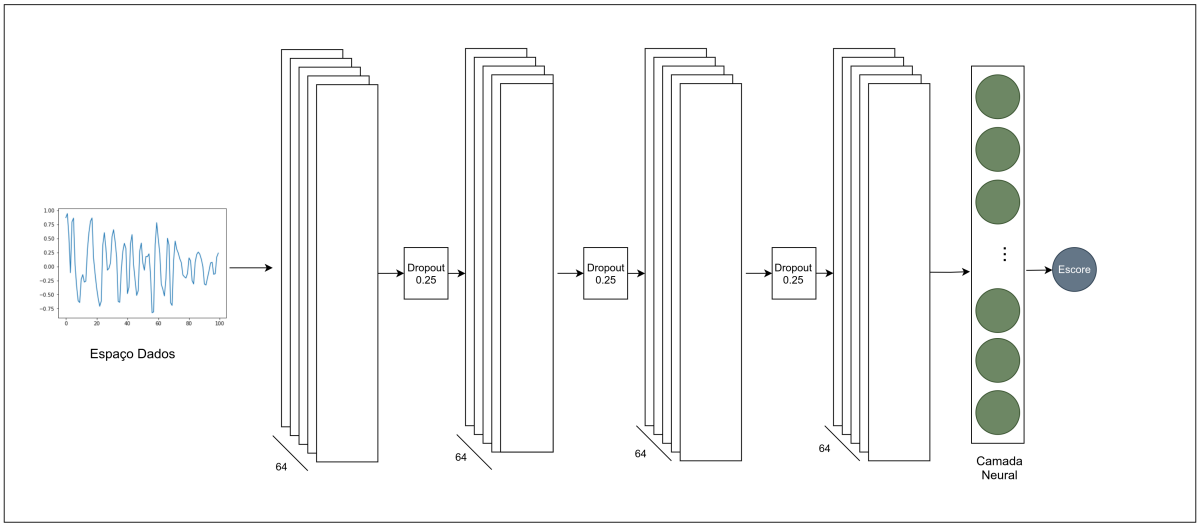

Figura 23 – Arquitetura do Modelo do Crítico no Domínio dos Dados.

Fonte: Autor.

Cada *layer* utiliza uma função de ativação LeakyReLU, com *dropout* de 25%.

3.2.1.2 Crítico no Espaço da Representação Latente

Analogamente ao crítico no espaço de dados, o crítico no Espaço da Representação Latente $\mathcal{C}_z$ atribui um escore à representação das janelas no domínio da representação latente $\mathbf{z}$, de modo a detectar se a janela foi gerada pelo *Encoder* ou foi amostrada diretamente da distribuição latente. Implementada pelos autores como uma rede neural de dois *layers* com 100 unidades cada e função de ativação logística. A arquitetura geral pode ser vista na Figura 24.

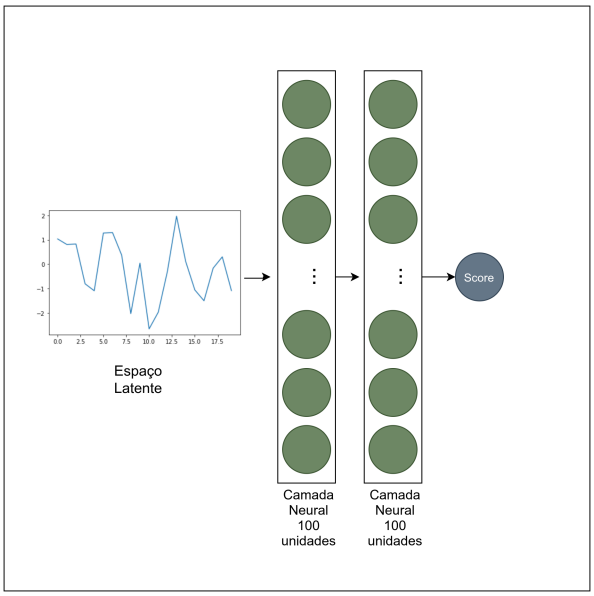

Figura 24 – Arquitetura do Modelo do Crítico em Z.

Fonte: Autor.

3.2.1.3 *Encoder*

Mapeia um exemplo do espaço de dados original para a representação latente $\mathbf{z}$, $f : \mathbb{R}^n \rightarrow \mathbb{R}^z$. Implementada como uma rede BLSTM com um *layer* de 100 unidades, e uma camada densa de uma rede neural com 20 unidades de saída. Um diagrama da arquitetura pode ser visto na Figura 25.

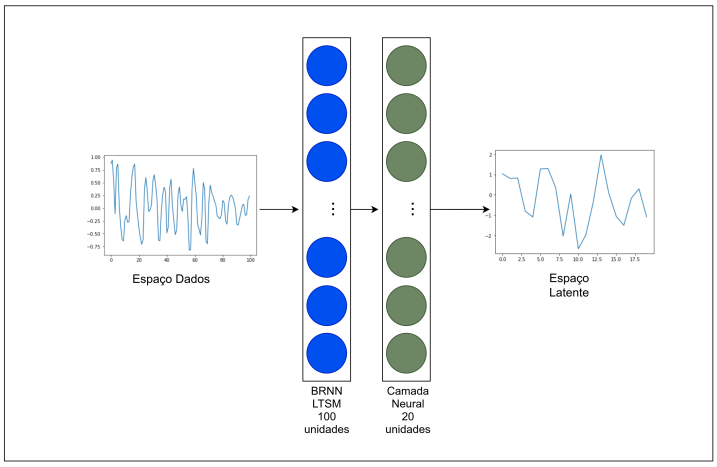

Figura 25 – Arquitetura do Modelo do Encoder.

Fonte: Autor.

3.2.1.4 Gerador

Mapeia da representação latente para o espaço de exemplos original, $f : \mathbb{R}^z \rightarrow \mathbb{R}^n$. Implementada como uma rede constituída por um *layer* neural de entrada, com 50 uni-

dades, seguido por dois *layers* BLSTM com 64 unidades, intercalador por uma camada de *upsampling* para 100. Cada ponto da saída é gerado com um neurônio que combina cada um dos 128 pontos da saída. Um diagrama geral da arquitetura pode ser visto na Figura 26.

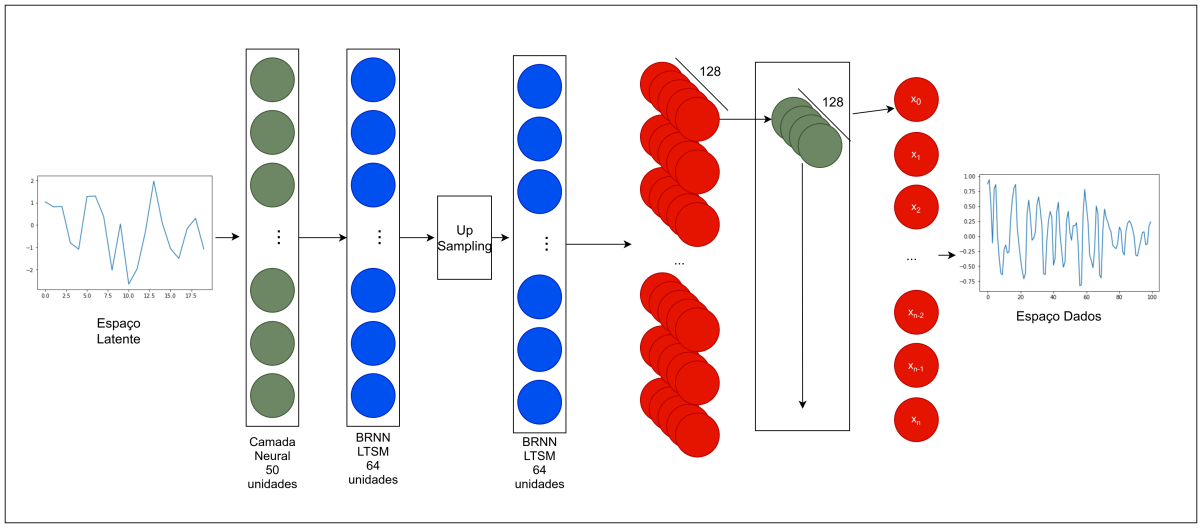

Figura 26 – Arquitetura do Modelo Gerador.

Fonte: Autor.

3.2.2 Treinamento

GEIGER *et al.,* (2020) colocam que os objetivos gerais de treinamento dos modelos são o de **modelar a distribuição das séries temporais de exemplos**, para o qual a função de custo *Wasserstein losses* é utilizada, e de **garantir a capacidade de reconstrução do método**, para o qual a função de custo *Cycle consistency losses* é utilizada.

- **Wasserstein Loss**: foi proposta por (ARJOVSKY *et al.,* 2017) como uma forma de melhorar o treinamento de GANs, em especial abordando o problema do colapso do modelo generativo, no qual o modelo gerador passa a gerar um número pequeno de exemplos diferentes, ao invés de verdadeiramente modelar a distribuição adjacente. A função de custo pode ser formulada para o par Gerador-Crítico x como na Equação (16).

$$V_x(\mathcal{C}_x, \mathcal{G}) = \mathbb{E}_{x\sim P_x}[\mathcal{C}_x(x)] - \mathbb{E}_{z\sim P_z}[\mathcal{C}_x(\mathcal{G}(z))] \tag{16}$$

  Analogamente a função de custo para o par *Encoder* - Crítico em **z** pode ser formulada como na Equação (23):

$$V_z(\mathcal{C}_z, \mathcal{E}) = \mathbb{E}_{z\sim P_z}[\mathcal{C}_z(z)] - \mathbb{E}_{x\sim P_x}[\mathcal{C}_z(\mathcal{E}(x))] \tag{17}$$

- **Cycle Consistency Loss**: foi proposta por (ZHU, *et al.*, 2017) como uma forma de garantir a consistência das redes geradores e de *encoder* para reconstrução. Para

um exemplo de entrada x_i, deseja-se que $\mathcal{G}(\mathcal{E}(x_i)) \approx x_i$, de maneira a possibilitar a reconstrução. É implementado como um termo de perda adicional para as duas redes sobre o erro de reconstrução. (GEIGER *et al.*, 2020) utilizam a norma L2, como mostrado na equação (18).

$$V_{l2}(\mathcal{G}, \mathcal{E}) = \mathbb{E}_{x \sim P_x}[||x - \mathcal{G}(\mathcal{E}(x))||_2] \tag{18}$$

Dado os dois componentes do objetivo geral, a perda de Wasserstein para a modelagem apropriada da distribuição dos dados normais pelo gerador e a perda de consistência de ciclo para garantir a reconstrução *Encoder*-Gerador, o objetivo geral do treinamento pode ser formulado como na equação (19).

$$\min_{\mathcal{E},\mathcal{G}} \max_{\mathcal{C}_x,\mathcal{C}_z} V_x(\mathcal{C}_x, \mathcal{G}) + V_z(\mathcal{C}_z, \mathcal{E}) + V_{l2}(\mathcal{G}, \mathcal{E}) \tag{19}$$

O treinamento dos quatro modelos utilizados no método é realizado simultaneamente. Para cada iteração, são realizados $n_{discriminador}$ passos de otimização sobre os discriminadores, nos quais um *batch* de exemplos reais são amostrados do espaço de dados e da distribuição da representação latente z, e exemplos artificiais gerados com o gerador e o *encoder*. A função de custo de ambos discriminadores é calculada e o gradiente descendente realizado. Em sequência, as funções de custo são computadas para o Gerador e *Encoder* no *batch* de dados amostrados, e um passo do gradiente descendente realizados. Desse modo, são realizados $n_{discriminador}$ passos do gradiente decente para os discriminadores para cada um sobre os geradores, de modo a obter-se um treinamento estável. (GEIGER *et al.*, 2020) define $n_{discriminador}$ como 5. O método de otimização escolhido foi *ADAM*.

As funções de custo para cada um dos modelos são apresentada abaixo:

- **Crítico X**: A função de custo associada ao modelo pode ser definida através do objetivo do treinamento, dada pela equação (20).

$$g_{wc_x} = \frac{1}{m}\sum_{i=0}^{m} \mathcal{C}_x(x_i) - \frac{1}{m}\sum_{i=0}^{m} \mathcal{C}_x(\mathcal{G}(z_i)) - GradientePenalty(x_i, \mathcal{G}(z_i)) \tag{20}$$

  onde GradientePenalty é a *Gradient Penalty*, descrita por GULRAJANI *et al.* (2017) como uma forma de melhorar convergência do treinamento de GANs. O primeiro termo maximiza o escore dado aos exemplos reais, enquanto o segundo minimiza o escore dos exemplos gerados artificialmente pelo gerador.

- **Crítico Z:** Analogamente ao crítico no espaço de exemplos, a função de custo foi colocada pelos autores como na equação (21).

$$g_{wc_z} = \frac{1}{m}\sum_{i=0}^{m}\mathcal{C}_z(z_i) - \frac{1}{m}\sum_{i=0}^{m}\mathcal{C}_z(\mathcal{E}(x_i)) - GradientePenalty(z_i, \mathcal{E}(x_i)) \quad (21)$$

Sendo os objetivos de cada termo análogos aos do crítico no domínio de dados, porém para a representação latente.

- **Gerador:** A função de custo também pode ser derivada do objetivo geral de treinamento, e formulada como na equação (22).

$$g_{wg} = -\frac{1}{m}\sum_{i=0}^{m}\mathcal{C}_z(\mathcal{E}(x_i)) - \frac{1}{m}\sum_{i=0}^{m}\mathcal{C}_x(\mathcal{G}(z_i)) - \frac{1}{m}\sum_{i=0}^{m}||x - \mathcal{G}(\mathcal{E}(x))||_2 \quad (22)$$

Vendo-se que para minimizar o custo deseja-se maximizar os scores dos críticos sobre os sinais gerados artificialmente e minimizar o erro de reconstrução geral do exemplo pareado.

- **Encoder:** Analogamente ao Gerador, a função de custo pode ser formulada como na equação (23).

$$g_e = -\frac{1}{m}\sum_{i=0}^{m}\mathcal{C}_z(\mathcal{E}(x_i)) - \frac{1}{m}\sum_{i=0}^{m}\mathcal{C}_x(\mathcal{G}(z_i)) - \frac{1}{m}\sum_{i=0}^{m}||x_i - \mathcal{G}(\mathcal{E}(x_i))||_2 \quad (23)$$

Com objetivos análogos ao gerador.

Dada as funções de custo apresentadas, o treinamento pode ser sumarizado no algoritmo 1

Algorítimo 1 Algoritmo para o treinamento dos modelos do método

```
for epoch = 0 ... n_epoch do
    for k = 0, ..., n_critic do
        Amostra {x_0...x_m} dos exemplos de treinamento
        Amostra {z_0...x_z} da representação latente z (ruído gaussiano)
        Calcula o custo e o gradiente para o C_x
        Calcula o custo e o gradiente para o C_z
        Realiza um passo de gradiente descendente no C_x
        Realiza um passo de gradiente descendente no C_z
    end for
    Amostra {x_0...x_m} dos exemplos de treinamento
    Amostra {z_0...x_z} da representação latente z (ruído gaussiano)
    Realiza um passo de gradiente descendente no Encoder
    Realiza um passo de gradiente descendente no Gerador
end for
```

A instabilidade do treinamento geral de GANs coloca desafios nos métodos desenvolvidos, e é um campo de estudo. Nas seções seguintes serão propostos experimentos para

explorar a instabilidade do treinamento, bem como a sensibilidade do modelo para com o número de épocas escolhidas

3.2.3 Detecção de Anomalias

Para a aplicação da TadGan para detecção de anomalias, as séries temporais alvo são pré-processadas e divididas em janelas de tamanho constante T, geradas com um passo n. Todos resultados reportados por (GEIGER, A. *et al.,* 2020) foram para $T = 100$ e $n = 1$. O processo geral de obtenção das janelas a partir da série temporal alvo pode ser visto na Figura 27.

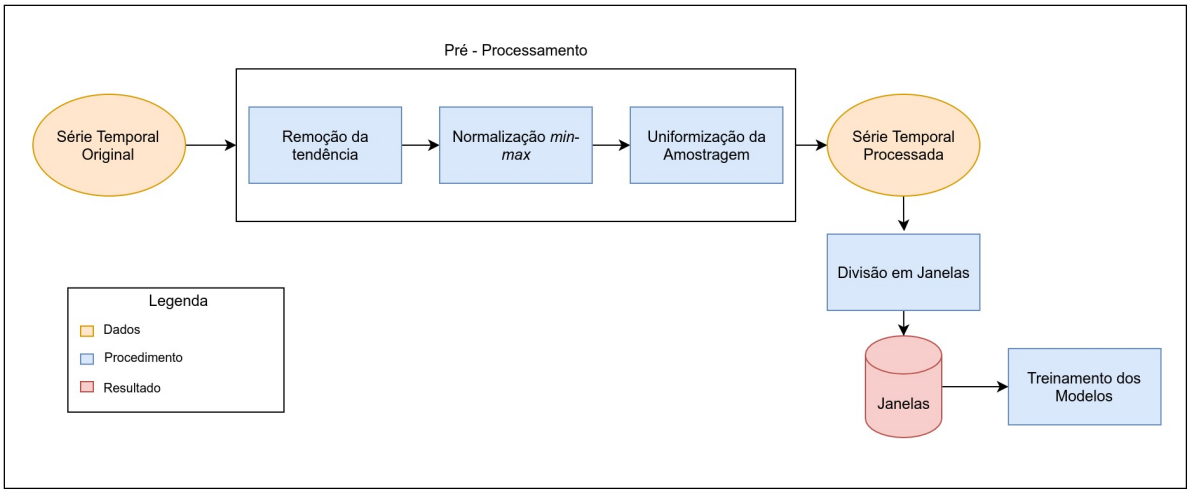

Figura 27 – Pré-Processamento do Método.

Fonte: Autor.

O processo de remoção da tendência é realizado através da interpolação de uma reta por mínimos quadráticos e da subtração do valor de cada ponto para o da reta interpolada. Em sequência, a série é normalizada como *min-max*, e a amostragem uniformizada com a remoção de amostras/interpolação por média.

As redes descritas são treinadas sobre as janelas obtidas, de modo a obter-se o *Encoder*-Gerador e o crítico no domínio de dados.

Duas métricas são utilizadas para a detecção de anomalias - o escore do crítico em X sobre as janelas originais e o erro de reconstrução.

Em tempo de predição, as janelas são codificadas e reconstruídas pelo gerador. A partir das janelas reconstruídas, o sinal total é reconstruído pela aplicação da mediana sobre todos pontos temporais repetidos, visto que cada ponto está presente em T janelas (dado o passo de avanço das janelas de 1).

Com o sinal total reconstruído, para o erro de reconstrução, os autores experimentaram com três métodos distintos, a área sobre o erro (para janelas de tamanho 10 sobre o sinal reconstruído), erro ponto-a-ponto quadrático e DTW (Dynamic Time Warping)(também

para janelas de tamanho 10 sobre o sinal reconstruído), concluindo que o último apresenta resultados marginalmente superiores.

O crítico em X é diretamente aplicado às janelas originais, e os valores de cada amostra calculados de forma análoga ao escore, com a mediano das n janelas que a amostra esta presente.

Um panorama geral da aplicação dos modelos obtidos para a obtenção dos escores descritos pode ser visto na Figura 28.

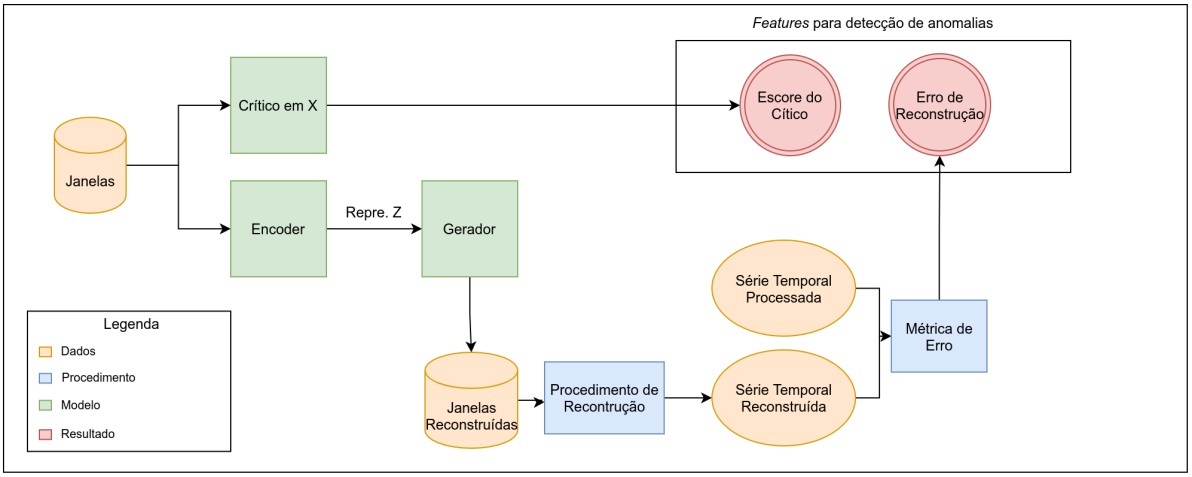

Figura 28 – Obtenção das Métricas de Anormalidade.

Fonte: Autor.

Com a obtenção de um espaço de *features* com capacidade de detecção das anomalias presentes no sinal no tempo, os autores de (GEIGER *et al* 2020) propõem uma metodologia para a detecção de anomalias na série. Os dois valores obtidos -escore do crítico e erro de reconstrução- são combinados em uma única métrica da anormalidade do ponto. Para tal, dois principais métodos foram sugeridos. como apresentado nas equações (24) e (25).

$$\mathcal{A}(x) = \alpha Z_{RE}(x) + (1 - \alpha) Z_{\mathbf{C}_x}(x) \tag{24}$$

e

$$\mathcal{A}(x) = \alpha Z_{RE}(x) Z_{\mathbf{C}_x}(x) \tag{25}$$

com $\alpha = 0.5$ e 1 respectivamente para os dois casos. Nos ensaios realizados sobre os dados de avaliação, os autores reportam resultados marginalmente superiores com a segunda abordagem.

Sob escore obtido é aplicado um simples algoritmo de *threshold* local, no qual uma janela de tamanho $\frac{T}{3}$ é aplicada sobre a sequência resultante, é um *threshold* local da janela definido como 4 desvios padrões. A janela avança com um passo definido como $\frac{T}{30}$, onde T é o tamanho da janela original utilizado para o treinamento do modelo. Em sequência,

um algoritmo inspirado por (HUNDMAN *et al.,* 2018) é aplicado para a redução de falsos positivos, no qual para cada janela, anomalias com o escore de anomalia menor que 10% do escore máximo dentro da janela são descartadas.

4 METODOLOGIA

Neste capítulo a metodologia empregada no trabalho é descrita e detalhada. É importante ressaltar que ao longo do texto, o termo *"dados normais"* é utilizado para referir-se aos dados do processo não anômalo, e não tem relação com a distribuição gaussiana normal. Dados normais são também por vezes denominados de *positivos*, como frequentemente é encontrado na literatura.

Inicialmente a implementação utilizada da TadGan foi detalhada, bem como o procedimento de verificação em relação aos resultados do trabalho original de (GEIGER *et al.*, 2020), efetuada através da reprodução parcial dos resultados reportados no trabalho original, de forma a corroborar com a validade da implementação utilizada.

Em sequência, uma metodologia para verificar o funcionamento geral do método foi desenvolvida e executada, consistindo na avaliação dos modelos presentes no método sob dados sintetizados a partir de funções analíticas, com o objetivo de garantir propriedades do sinal e das anomalias introduzidas presentes. Avaliou-se nessa etapa a capacidade do método de produzir um modelo gerador capaz de efetivamente modelar o processo dos dados não anômalos, a capacidade do *Encoder* de projetar na representação latente características relevantes do sinal, a fim de possibilitar a reconstrução, e a relevância das métricas derivadas do método utilizadas para detecção de anomalias -o escore do crítico sobre o dado original e o erro de reconstrução - na geração de uma espaço de *features* com possibilidade de discriminar entre dois diferentes processos no tempo, e consequentemente, de efetivamente detectar anomalias. O método desenvolvido para a verificação das propriedades listadas teve como objetivo não a demonstração formal de um ponto de vista matemático e metodológico dessas propriedades, mas a possibilidade de corroborar através de exemplos didáticos com os princípios fundamentais do método. Dessa maneira, espera-se não só adquirir uma compreensão profunda sobre seu funcionamento e eventuais limitações, como também aumentar a confiança da sua aplicação em situações práticas da engenharia e ciências.

Explorou-se, então, limitações na aplicação geral de GANs à detecção de anomalias, oriundas tanto da literatura especializada quanto de observações realizadas pelo autor. Deu-se especial atenção para a instabilidade do treinamento, que pode ser potencialmente originada tanto pelas características do treinamento das GANs quanto pela formulação não-supervisionada do problema e às particularidades da utilização do erro de reconstrução como medida da anormalidade, e suas dependências tanto para com o treinamento como para características dos dados.

O método foi então avaliado de forma geral utilizando-se um compilado de séries temporais com anomalias com causas conhecidas, bem como provenientes de aplicações diversas não presentes na avaliação original de GEIGER *et al.*, 2020. Séries temporais

foram sintetizadas com a inserção de tipos específicos de anomalias, e conjuntos de dados selecionados da base *UCR Anomaly Benchmark*, introduzida por (WU *et al.*, 2021) de modo a abranger aplicações diversas. Por fim, baseado no comportamento observado pelo experimentos, bem como as limitações levantadas, novas métricas derivadas do método para detecção de anomalias foram sugeridas e brevemente exploradas.

Uma panorama geral do trabalho pode ser visto na Figura 29, onde observa-se também a relação sequencial entre as seções, através da qual os experimentos realizados para a demonstração dos princípios do métodos são em parte derivados da análise dos resultados obtidos na reprodução do trabalho de (GEIGER *et al.*, 2020), bem como a investigação das limitações derivadas da análise do funcionamento dos princípios do método sobre os dados sintetizados. A avaliação sobre séries experimentais e sintéticas é derivada em parte das limitações investigadas, e as modificações nas métricas do comportamento observado ao longo de todos experimentos.

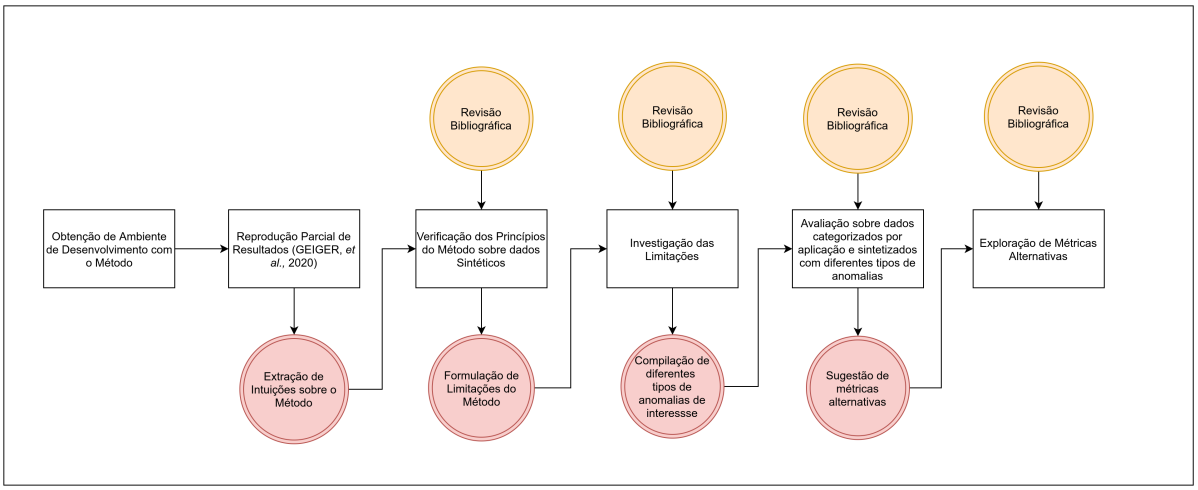

Figura 29 – Panorama geral da Metodologia Aplicada.

Fonte: Autor.

Este capítulo está divido da seguinte maneira:

- a seção 4.1 descreve a implementação utilizada da TadGan e sua verificação em relação ao trabalho de (GEIGER *et al.*, 2020), bem a metodologia empregada para a reprodução parcial dos resultados.

- a seção 4.2 descreve a metodologia proposta para verificação do princípios fundamentais do método e dos modelos envolvidos, englobando a geração dos dados sintéticos utilizados e os métodos de análise empregados.

- a seção 4.3 descreve a metodologia proposta para a exploração de algumas limitações da TadGan e da aplicação geral de GANs para detecção de anomalias em séries temporais.

- a seção 4.4 descreve os dados sintetizados com tipos de anomalias diversos e compilados de diversos subproblemas e a metodologia de avaliação empregada.
- por fim, a secção 4.5 descreve as métricas alternativas propostas para quantificação da anormalidade.

4.1 IMPLEMENTAÇÃO UTILIZADA DO MÉTODO

(GEIGER *et al.*, 2020) disponibilizou uma implementação colaborativa de código aberto do método através de um *Framework* voltado para detecção de anomalias em séries temporais, *Orion*, contando com a implementação dos autores para a *TadGan*. A implementação disponibilizada foi inspecionada, de modo a a compreender o código inicialmente disponibilizado assim como a GAN proposta. Um ambiente de desenvolvimento foi configurado de modo a possibilitar a execução de alterações na implementação, bem como a execução na GPU NVIDIA GeForce MX150 , devido ao alto custo computacional no treinamento e execução dos modelos. Em sequência, modificações foram efetuadas para possibilitar a inspeção das curvas de treinamento e das características da representação latente $\mathbf{z}$ obtida. A implementação utilizada foi realizada em Python3 com a biblioteca de inteligência de máquina TensorFlow 1.14, através da API keras. Os métodos de pré-processamento e pós-processamento descritos na seção 3.2.3 foram implementados com a biblioteca de computação numérica Numpy. O erro de reconstrução por DTW foi utilizado através da biblioteca *pymetrics*. O desenvolvimento deu-se através do ambiente *Jupyter Notebook,* pelas facilidades fornecidas no campo de análise de dados.

4.1.1 Verificação da Implementação

A fim de garantir a aplicação correta do método em relação ao trabalho de (GEIGER *et al.*, 2020), bem como de corroborar com a implementação utilizada, resultados parciais do trabalho de (GEIGER *et al.*, 2020) foram reproduzidos, com a seleção de um subconjunto dos sinais utilizados na avaliação do trabalho referido. A seleção de um subconjunto deu-se pelo alto custo computacional na reprodução integral dos resultados, e na não necessidade da reprodução integral para apenas verificação da implementação utilizada. As etapas de pré-processamento e pós-processamento descritos na seção 3.2.3 foram também reproduzidas, dado que os resultados fornecidos por (GEIGER *et al.,* 2020) foram computados sobre o método como um todo. Nas secções seguintes, a base de dados utilizada é descrita, e o procedimento de avaliação empregado apresentado.

4.1.1.1 Descrição dos Dados Utilizados

(GEIGER *et al.*, 2020) utilizou 5 diferentes conjuntos de dados para a avaliação, cada um contendo diversos sinais. Dentre os conjuntos utilizados, sinais do conjunto NAB

(*Nubenta Anomaly Bechmark*) foram selecionados para a validação da implementação. O conjunto de dados NAB foi introduzido por (LAVEIN, A. *et al.* 2015), juntamente com uma metodologia de avaliação. Os dados disponíveis são divididos em séries temporais, contendo originalmente 58 diferentes sinais, provenientes de diversas aplicações, cada uma com 1000 a 22000 amostras, totalizando cerca de 365550 amostras por todos os subconjuntos de dados. Os dados disponíveis foram agrupados pelos autores nos seguintes subconjuntos:

- **Artificial sem anomalias**: cinco séries geradas artificialmente não contendo anomalias;
- **Artificial com anomalia**: seis séries geradas artificialmente contendo anomalias;
- **Real AWS Cloud Watch**: dezessete séries coletadas pelo serviço AmazonCloudwatch, contendo medições de utilização de CPU, Entrada de Bytes, Leitura de Bytes;
- **Real causa conhecida**: sete sinais provenientes de processos com anomalias cuja causa é conhecida;
- **Real tráfico**: sete séries com dados reais de tráfego coletados pelo departamento de transporte de Minnesota. Comporta métricas como ocupação, velocidade e tempo de viagem para veículos do transporte público;
- **Real Tweets**: dez séries temporais contendo a contagem de menções no twitter de dez diferentes empresas;
- **Real AdExchange**: cinco séries com métricas de custo para serviços de propaganda na internet, com CPC (custo por click) e CPM (custo por mil impressões).

O procedimento adotado pelos autores para a obtenção dos intervalos anômalos nas séries é referenciado em (LAVEIN, A. *et al.* 2015), consistindo na inspeção visual por humanos e marcação das regiões visualmente anômalas. Alguns aspectos desse procedimento devem ser ressaltados, pelo impacto direto na avaliação obtida:

- nos primeiros 15% do sinal não são marcadas anomalias, e é utilizado para estabelecer o comportamento normal;
- se um padrão de comportamento anômalo se mantém por mais de 10% da duração do sinal, ele não é mais computado como anômalo após esse momento;
- a partir de anomalias pontuais identificadas, são criadas janelas ao redor do ponto de tamanho $\frac{0.1 TamanhoSinal}{NumeroAnomalias}$.

A criação de janelas de tamanho fixo marcadas como anômalas ao redor das amostras efetivamente identificadas como anômalas prejudica a avaliação da performance amostra-a-amostra, pois exemplos marcados como anômalos podem não conter de fato anomalias. A possibilidade de um comportamento anormal tornar-se normal, ou seja, a não estacionariedade da anormalidade é inconsistente com as premissas utilizadas para a aplicação da TadGan e deve ser considerado na avaliação da performance.

O subconjunto da *AdExchange* e *Artificial com anomalia* foram selecionado para a validação. Uma breve descrição dos sinais presentes pode ser vista na Tabela 1.

Tabela 1 – Sinais Utilizados

Nome do Sinal	N. Pontos	N. Pontos Anômalos	N. Janelas Anômalas
exchange-2_ cpc_ results	1624	162	1
exchange-2_ cpm_ results	1624	161	2
exchange-3_ cpc_ results	1538	154	3
exchange-3_ cpm_ results	1624	158	1
exchange-4_ cpm_ results	1643	162	4
art_ daily_ flatmiddle	4032	402	1
art_ daily_ jumpsdown	4032	402	1
art_ daily_ jumpsup	4032	402	1
art_ daily_ nojump	4032	402	1
art_ increase_ spike_ density	4032	402	1
art_ load_ balancer_ spikes	4032	402	1

Um exemplo de um dos sinais utilizados pode ser visto na Figura 30.

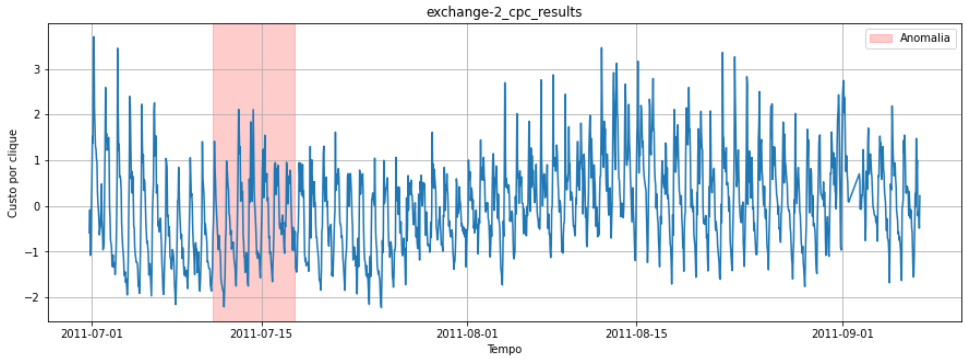

Figura 30 – Série temporal referente ao sinal *exchange-2_ cpc_ results*.

Fonte: Autor.

Observa-se a presença de uma janela anômala, formadas por anomalias coletivas, de acordo com a classificação de (CHANDOLA *et al.*, 2017).

4.1.1.2 Método de Avaliação da Performance

A fim de possibilitar a comparação direta com os resultados de (GEIGER *et al.*, 2020), o mesmo procedimento de avaliação foi replicado. Esse procedimento apresenta

peculiaridades, com a utilização de uma versão modificada do escore de F1 para o caso de detecção de anomalias em séries temporais, que segue as regras apresentadas em sequência:

- se uma janela detectada possui intersecção com uma janela anotada como anômala, um Verdadeiro Positivo é registrado;
- se uma janela detectada não possui intersecção com nenhuma janela anotada, um Falso Positivo é registrado;
- se uma janela anotada não possui interseção com nenhuma janela detectada, um Falso Negativo é detectado.

É importante ressaltar que esse método é alvo de críticas da literatura especializada, como por (KIM *et al.*, 2019), que colocam a possibilidade de superestimação da performance do método. Além disso, dado o baixo número de janelas anômalas nos dados, grande variância de performance é observada, visto que a não detecção de apenas uma janela representa uma grande variação percentual.

Baseado nos valores obtidos de verdadeiros positivos e falsos positivos através do procedimento anteriormente descrito, o F1 é calculado através da maneira tradicional, como representado pelas Equações (26), (27), (28).

$$Precision = \frac{Verdadeiros\ Positivos}{Total\ Detectados\ como\ Positivos} = \frac{TP}{TP + FP} \tag{26}$$

$$Recall = \frac{Verdadeiros\ Positivos}{Total\ Positivos} = \frac{TP}{TP + FN} \tag{27}$$

e o F1 é então calculado como a média harmônica entre a precisão e o *recall*

$$F1 = 2\frac{Precision \cdot Recall}{Precision + Recall} \tag{28}$$

Este procedimento se relaciona com algumas abordagem utilizadas em aplicações de detecção em imagem, mas para fins de clareza será referido na continuidade do texto como **F1 (GEIGER *et al.*, 2020)** enquanto o F1 tradicional realizado sobre cada amostra como **F1 amostra-a-amostra**.

4.1.1.3 Análise dos Resultados

O método foi treinado por 100 épocas, com os hiper-parâmetros descritos em 3.2. As curvas de custo dos modelos foram inspecionadas, bem como exemplos de janelas reconstruídas a fim de garantir a convergência do treinamento. Em sequência, o procedimento descrito em 3.2.3 foi aplicado sobre os intervalos de anomalias detectados, e o F1 (GEIGER *et al.*, 2020) computado. Os resultados obtidos foram numericamente comparados com os reportados por (GEIGER *et al.*, 2020), e eventuais divergências investigadas.

Em sequência, com objetivo de adquirir intuições sobre o funcionamento do método, os elementos individuais constituintes do escore de anomalia, - a predição do crítico em X e do erro de reconstrução - também foram analisados. Seu comportamento foi observado no tempo, de modo a investigar sua capacidade de detecção de anomalias.

4.2 VERIFICAÇÃO DOS PRINCÍPIOS DO MÉTODO

Nesta subseção a metodologia desenvolvida para visualizar e verificar o funcionamento do método é descrita. Ensaios foram propostos de forma a avaliar diferentes aspectos do método, e corroborar com seus princípios, bem como identificar limitações e particularidades. Esta seção está dividida da seguinte forma:

- A seção 4.2.1 descreve a metodologia desenvolvida para verificação do modelo gerador sobre dados sintéticos analíticos, bem como da inspeção do espaço formado pelas *features* resultantes na sua capacidade de separar diferentes processos.
- A seção 4.2.2 descreve a metodologia desenvolvida para a investigação das capacidades do *encoder* de projetar determinadas características no espaço latente.

Os ensaios foram realizados com repetição, de modo que pode-se também inferir sobre a estabilidade e reprodutibilidade do treinamento do método.

4.2.1 Análise Geral do Método

Com a proposta de demonstrar o funcionamento geral da TadGan, com enfoque na capacidade do modelo gerador de efetivamente capturar a distribuição dos exemplos não anômalos e de gerar sinais coerentes com os dados de entrada, séries temporais sob o formato de janelas de 100 amostras foram geradas a partir de funções arbitrárias com propriedades conhecidas e utilizadas para o treinamento dos modelos. Com a utilização de funções analíticas, com as quais sabe-se ser possível de um ponto de vista teórico a obtenção de um projeção em um espaço de menor dimensão e reconstrução sem perda , espera-se poder validar a capacidade de modelagem de processos no tempo pelo método. Dessa forma, poderá ter-se mais confiança na aplicação do método em situações com dados experimentais.

A utilização de dados gerados artificialmente para avaliação de métodos de detecção de anomalias é vasta, podendo-se colocar como exemplos os trabalhos de (LEE *et al.,* 2021) e o conjunto de avaliação NAB.

Diferentemente da aplicação habitual da TadGan, no qual uma série temporal é dividida em janelas para análise, sintetizou-se diretamente as janelas de tamanho 100, dado o interesse em verificar os princípios dos modelos do método, e não avaliar sua performance sob algum tipo de dado. As perguntas que deseja-se investigar são: *"Dadas janelas de treinamento provenientes de um mesmo processo A no tempo":*

1. O modelo gerador consegue modelar o processo e gerar novas janelas consistentes com esse processo?

2. A rede *Encoder-Gerador* consegue reconstruir com baixo erro janelas proveniente do processo A?

3. As métricas utilizadas para detecção de anomalias são capazes de diferenciar janelas do processo A de um outro processo arbitrário B?

Com a utilização de funções analíticas, com as quais sabe-se ser teoricamente possível a reconstrução sem perda com uma projeção num espaço de 20 dimensões e a diferenciação entre os processos, pode-se analisar a performance do método de forma mais robusta, sem a necessidade de considerar a possível influência nos resultados de fenômenos desconhecidos eventualmente presentes em dados experimentais.

Duas funções foram definidas, uma para o processo A, com presença majoritária no conjunto formado, de modo a simular o comportamento não anômalo, e uma para o processo B, presente em baixa proporção, de modo a emular um comportamento anômalo. A implementação das funções deu-se em Python, através da biblioteca de computação numérica Numpy.

Com os conjuntos de dados sintetizados obtidos os modelos foram treinados. Exemplos de reconstrução foram analisados, bem como métricas do erro total, com objetivo de verificar a capacidade de reconstrução do método. O procedimento de análise realizado está detalhado nas subsecções seguintes. Na Figura 31 pode-se ver um panorama geral da metodologia desenvolvida nessa subseção. Com os modelos treinados sobre dados sintéticos, avaliou-se a performance de reconstrução do método, sua variabilidade em diferentes execuções sobre os mesmos dados e a capacidade das *features* resultantes de diferenciar diferentes processos, e consequentemente, de detectar anomalias.

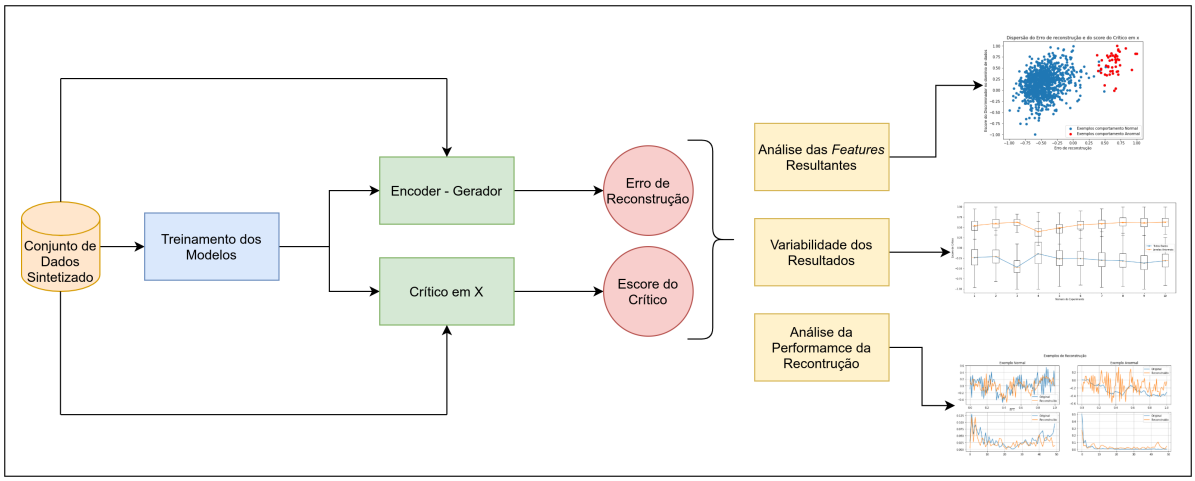

Figura 31 – Visão geral da metodologia para verificação dos princípios fundamentais do método.

Fonte: Autor.

4.2.1.1 Dados Sintetizados

A metodologia geral adotada para a geração dos conjuntos de dados sintéticos pode ser vista na Figura 32 e será detalhada nas secções seguintes. A implementação das funções responsáveis pela geração dos dados sintéticos foi realizada em Python, com o uso de números pseudo-aleatórios para geração do ruído.

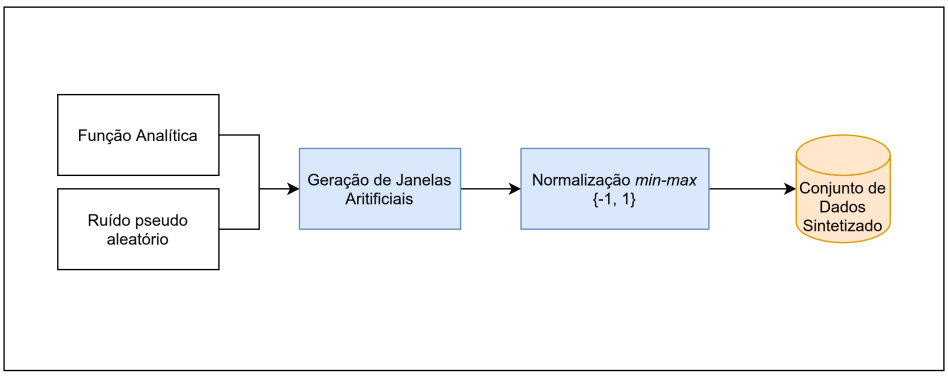

Figura 32 – Procedimento Geral de Sintetização das Janelas.

Fonte: Autor.

Nas subsecções seguintes são detalhados os dois tipos de sinais sintetizados, justificando-se a escolha dos parâmetros.

4.2.1.1.1 Senoides Atenuadas

Dado o objetivo de avaliar a capacidade de modelagem do gerador, escolheu-se funções que apresentassem moderada complexidade teórica e que contemplassem distintos

comportamentos matemáticos e famílias de funções. Para a geração do comportamento positivo, utilizou-se uma função constituída de duas senoides de frequências diferentes com ruído na frequência e na amplitude, atenuadas por um envelope exponencial, como pode se visto na equação (29):

$$D_{normal} = e^{-x}(\sin(2\pi f_1(x + \mathcal{N}(\mu_1, \sigma_1)) + \sin(2\pi f_2(x + \mathcal{N}(\mu_1, \sigma_2))) + \mathcal{N}(\mu_2, \sigma_3)) \quad (29)$$

onde os valores foram definidas arbitrariamente como $f_1 = 10Hz$, $f_2 = 20Hz$, $\mu_1 = 0$, $\sigma_1 = 0.01$, $\sigma_2 = 0.005$ e $\sigma_3 = 0.25$.
Os exemplos anormais foram gerado de acordo com a equação (30), sem a presença da atenuação exponencial e formado apenas por uma senoide.

$$D_{Anormal} = \sin(2\pi f_3 x + \mathcal{N}(\mu_2, \sigma_2)) + \mathcal{N}(\mu_2, \sigma_2) \quad (30)$$

com:

$$f_3 = 40Hz$$

Sendo uma função majoritariamente determinística, é garantido de um ponto de vista matemático a existência de uma representação na dimensão latente que permita a reconstrução sem erro. As janelas foram normalizadas com normalização *min-max* em {-1, 1}. O desvio padrão médio das janelas do comportamento anormal foi igualado ao do comportamento normal.

A diferença entre os dois processos das funções definidas é dado pela não presença do decaimento exponencial, da diferente frequência fundamental e da não presença da adição da segunda senoide. É importante ressaltar que o objetivo do ensaio não depende da real relevância prática desse tipo de anomalias, mas apenas busca a validação da capacidade do método de modelar e reconstruir com baixo erro funções no tempo, bem como de verificar a capacidade das métricas envolvidas de formarem um espaço que apresente separabilidade entre dois processos distintos arbitrários. A não obtenção de um espaço separável não demonstra a não usabilidade do método, apenas indica restrições quando ao tipo de anomalia passível de detecção.

Gerou-se um conjunto de dados contendo 1000 exemplos, dos quais 50 anômalos, mantendo a proporção de anomalias no conjunto de treinamento próxima a grande maioria dos sinais da NAB de $\sim 5\%$, de modo a obter uma situação condizente em termos de contaminação do conjunto.

4.2.1.1.2 Processo ARMA

Gerou-se também janelas artificiais provenientes de um processo ARMA(1,1), reproduzindo-se a metologia aplicada por (LEE *et al.*, 2021) na avaliação realizado de GANs para detecção de anomalias. A utilização de um processo ARMA é justificada

pela sua consolidada capacidade de modelar séries temporais provenientes de aplicações práticas, de modo que a constatação da capacidade do método de modelagem desse tipo de processo corrobora sua aplicabilidade geral.

Para um processos ARMA(1,1), como visto na Equação(31).

$$ARMA(1,1) := x_t = c + \phi_1 x_{t-1} + \gamma_1 \epsilon_{t-1} \tag{31}$$

Replicou-se a metodologia de (LEE *et al.,* 2021), sendo as janelas do comportamento não-anômalo geradas com c = 0, e janelas de anomalias com c = 1. Analogamente ao realizado para os sinais senoidais sintetizados, 1000 janelas foram sintetizadas, 950 para normalidade e 50 anômalas.

4.2.1.2 Avaliação

O conjunto de dados sintetizado obtido foi inspecionado, analisando-se seu comportamento no domínio da frequência e o histograma de todos valores gerados, bem como exemplos de janelas. Em sequência, a TadGan foi treinado por 300 épocas sobre os dados gerados, e a convergência dos objetivos pelos modelos foi verificada através do gráfico das funções de custo. Uma vez verificada a convergência, o procedimento foi repetido 10 vezes sobre os dados sintetizados, de modo a analisar a performance com repetições de forma mais robusta. Em cada iteração, calculou-se o erro de reconstrução quadrático e por DTW de todas as janelas, bem como escore do crítico em X. Com os dados coletados, examinou-se o erro de reconstrução médio sobre as janelas e inspecionou-se visualmente exemplos de reconstrução. A manutenção das propriedades gerais do conjunto também foram verificadas, através do comportamento em frequência médio e do histograma. As três métricas extraídas foram exibidas graficamente, de modo a verificar-se a dispersão das janelas não-anômalas e anômalas. A variabilidade dos resultados nas repetições também foi inspecionada. Os valores médios para cada métrica foram calculados para as janelas normais e anormais, e a capacidade de separação entre os dois processos verificada comparativamente pela contagem das janelas com o valor da métrica maior que 3 desvios padrões em relação à todo conjunto.

4.2.2 Capacidade do *Encoder*

Com o objetivo de verificar a capacidade do *Encoder* de extrair e projetar na representação latente características relevantes do processo modelado, um conjunto de dados foi gerado artificialmente com inserção de variação dentro da classe não anômala através da variação de parâmetros das funções geradoras. A situação motivadora é entender a capacidade do método de generalizar um processo geral que apresente variância sobre determinados parâmetros nos dados disponíveis. Por exemplo, um instrumento musical pode produzir uma determinada forma de onda característica, porém com diferentes frequências

fundamentais. Espera-se que um conjunto de dados definidor do comportamento *positivo* para esse instrumento apresente séries com diferentes frequências fundamentais, porém todas com as propriedades comuns do tipo de som gerado pelo instrumento. De modo a diferenciar da onda característica de outro instrumento, é necessário que o método consiga reconstruir a série temporal para diferentes frequências.

O objetivo do experimento busca abordar a seguinte questão:

Dado um processo A definido por $f(\theta)$, onde θ são parâmetros de f, $\theta \in R^n$, e sendo o método apresentado à um conjunto de dados formado por $\{ f(\theta_1), f(\theta_2)...f(\theta_n) \}$, é o *Encoder* é capaz de encontrar uma representação para θ e projeta-lo na representação latente?

Como o conjunto de treinamento apresenta variação sobre o parâmetro, o menor erro de reconstrução será dado se essa informação for codificada no espaço latente de forma a ser reconstruído pelo gerador.

Dada a importância do comportamento em frequência na análise de sinais temporais e em diversas aplicações, em especial a frequência fundamental, avaliou-se como exemplo a capacidade do sistema *Encoder-Gerador* de reconstruir sinais com diferentes frequências fundamentais. Para isso, utilizou-se de funções análogas às funções descritas na seção anterior, porém sem a presença da segunda componente senoidal, com a diferença que as frequência das senoides geradas não são mais constantes, mas amostradas de uma distribuição uniforme em um intervalo definido. Espera-se que dada a não mais constância das frequências fundamentais, a rede Geradora não mais poderá "*memorizar*"a frequência a ser gerada, cabendo ao *Encoder* efetivamente extrair e projetar essa informação no espaço latente. A função utilizada para sintetizar o processo positivo pode ser visto na Equação (32):

$$D_{normal} = e^{-x}(\sin(2\pi f_1(x + \mathcal{N}(\mu_1, \sigma_1)) + \mathcal{N}(\mu_2, \sigma_3)) \tag{32}$$

e para a anormalidade como na Equação (33):

$$D_{Anormal} = \sin(2\pi f_3 x + \mathcal{N}(\mu_2, \sigma_2))) + \mathcal{N}(\mu_2, \sigma_2) \tag{33}$$

As frequências foram definidas como:

$$f_1 \sim \mathcal{U}(0, 20)$$

$$f_3 := 40$$

onde $\mathcal{U}$ denota uma distribuição uniforme sobre os reais no intervalo.

O conjunto de dados sintetizados foi inspecionada de maneira análoga às seções anteriores.

A fim de quantificar a reconstrução coerente da frequência, extraiu-se a frequência fundamental do sinal reconstruído e do original, através do Algoritmo 2.

Algorítimo 2 Determinação da Frequência Dominante

1: $C, Freq \leftarrow DFT(Janela)$
2: $Mag \leftarrow \sqrt{Re(C)^2 + Im(C)^2}$
3: $Freq.\,Dominante \leftarrow Freq[\arg\max(Mag)]$

Foi também computada a centroide espectral das janelas originais e reconstruídas, através da Equação (34):

$$Centroide\,Espectral = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \tag{34}$$

onde $f(n)$ representa a amplitude no intervalo n, e $x(n)$ o valor da frequência.

Com base nesses valores, verificou-se a correlação entre as frequências dominantes e centroides espectral das janelas reconstruídas e originais. Dessa forma, pôde-se intuir sobre a capacidade do método de capturar a variância nos parâmetros do processo definidor do comportamento positivo, e de efetivamente extrai-las e projetá-las na dimensão latente.

O método foi treinado por 200 épocas sobre os dados gerados. Um diagrama geral da metodologia adotada pode ser visto na Figura 33.

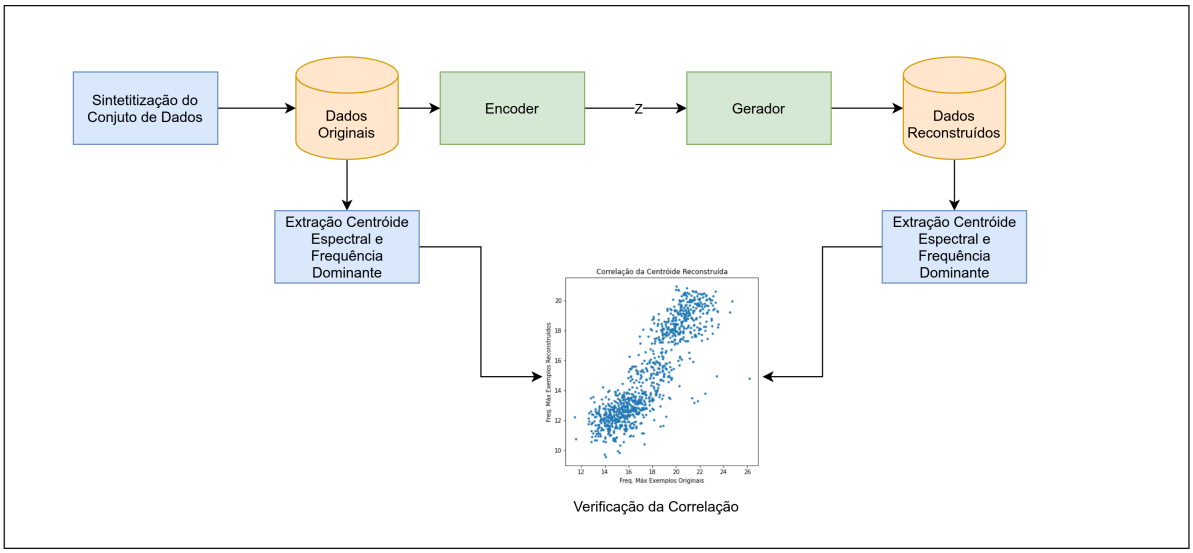

Figura 33 – Metodologia para verificação da capacidade do Encoder.

Fonte: Autor.

4.3 LIMITAÇÕES DO MÉTODO

Nesta subseção a metodologia aplicada para as investigações sobre algumas limitações do método é detalhada. As limitações selecionadas para estudo foram extraídas da literatura e também propostas a partir da análise dos resultados preliminares. Duas principais possíveis limitações foram abordadas: a investigação da estabilidade do treinamento do modelo e a dependência da utilidade do erro de reconstrução como métrica de anormalidade com características do sinal, em especial a entropia.

Essa seção está dividida da seguinte forma:

- na subseção 4.3.1 uma possível limitação do uso do erro de reconstrução é apresentada, e um procedimento experimental descrito;
- na subseção 4.3.2 duas diferentes perspectivas são abordadas quanto a instabilidades do treinamento: a variabilidade de performance entre execuções sobre os mesmos dados, e a possibilidade da influência do número de épocas de treinamento na performance do método em conjuntos de dados contaminados.

4.3.1 Entropia do Sinal

A principal métrica utilizada na aplicação de GANs à detecção de anomalias é o erro de reconstrução. Como previamente apresentado e discutido, na grande maioria dos métodos presentes na literatura o sinal de interesse é projetado através de diferentes metodologias para um espaço de menor dimensão, e um modelo treinado sobre exemplos positivos o reconstrói, com a presunção de que, como o modelo generativo foi treinado sobre a distribuição dos dados positivos apenas, será gerado um exemplo pertencente àquela distribuição. Dessa maneira, anomalias que espera-se não pertencer a distribuição positiva devem possuir um maior erro de reconstrução. Além disso, assume-se que o método de projeção no espaço latente menor dimensional aprendeu uma representação relevante apenas para a distribuição positiva, de modo que os exemplos anômalos não tem suas características apropriadamente representadas no espaço latente.

Limitações já foram apontadas para a utilização do erro de reconstrução como métrica da anormalidade. GEIGER *et al.*, (2020) levantam nas suas considerações finais a possibilidade de influência da complexidade do modelo gerador na utilidade do erro de reconstrução, dado que modelos muito complexos podem também conseguir modelar o comportamento anormal na formulação totalmente não-supervisionada do problema, na qual assume-se o conjunto disponível como contaminado com exemplos anômalos.

BATTIKH *et al.*, (2021) exploraram o problema da modelagem do comportamento anormal no contexto de *Autoencoders* para detecção de anomalias, levantando também a hipótese da influência do tipo de dados, na situação onde exemplos anômalos e positivos

pertencem a domínios muito próximos, levando a representação obtida para o comportamento positivo apresentar relevância para as anomalias, resultando no possível colapso da métrica.

Nesta secção uma limitação alternativa para o uso do erro de reconstrução é apresentada e investigada no contexto de utilização de GANs para detecção de anomalias. Dado uma distribuição arbitrária, a taxa de compressão sem perda máxima está vinculada à entropia, como descrito pela teoria da informação. A máxima taxa de compressão com distorção - alvo da *Rate–distortion theory* - também é vinculado à entropia da distribuição do sinal. Dessa forma, o erro de reconstrução obtido sobre uma distribuição de dados depende não apenas da capacidade do modelo de modelar o processo, mas também da sua natureza. Aplicada à detecção de anomalias, a diferença do erro de reconstrução entre dois processos pode depender não apenas de alguma métrica de dessemelhança efetiva correlacionado com a anormalidade, mas também da relação de entropia entre os processos estocásticos geradores. Esse fato apresenta interesse em aplicações onde o comportamento esperado apresenta alto ruído, por exemplo, que é estreitamente vinculado à entropia, e onde sinais determinísticos podem ser considerados anômalos. No melhor dos casos, pode-se esperar que a métrica não apresenta usabilidade, dado que o erro de reconstrução será alto e semelhante tanto para o processo majoritário esperado quanto para anomalias não modeladas pelo método. Em conjuntos de dados contaminados, entretanto, o treinamento dos modelos podem ser levados a convergir para a modelagem dos exemplos anômalos, dado que o menor erro de reconstrução passa a ser obtido na possibilidade de modelagem dos poucos exemplos de baixo entropia. A fim de investigar esse problema foi proposta uma metodologia sem formalismo matemático, mas que possibilita levantar hipóteses e intuições sobre o tema

Dois sinais provenientes de funções distintas foram selecionadas. O método foi treinado por 300 épocas, e a capacidade do erro de reconstrução e escore do crítico de separar os dois diferentes processos verificadas, validando-se as funções escolhidas. Para o conjunto de dados formado, mesma proporção de janelas anômalas e não-anômalas, de 950/50, foi mantida.

Uma vez validada as funções, ruído gaussiano branco foi adicionado às janelas positivas, com diferentes desvios padrões, e o SNR (*Signal to Noise Ratio*), relação sinal ruído em uma tradução livre, foi calculado para cada caso. O SNR está, em geral, relacionado com a entropia do sinal, de forma inversa. Dessa forma, espera-se poder treinar diferentes modelos sobre conjuntos de dados com níveis de entropia variados, e analisar os impactos na capacidade do erro de reconstrução de discriminar os diferentes processos

Para o comportamento positivo não anômalo, as janelas foram sintetizadas a partir de uma senoide de frequência fixa, e ruído gaussiano aditivo foi somado, como na Equação (35):

$$D_{Normal} = \sin(2\pi f_1 x) + \mathcal{N}(0, \sigma) \tag{35}$$

com $f1 = 5Hz$ e σ variado no experimento.

Para o comportamento anormal foi sintetizada uma onda quadrada de frequência 10 Hz, pela baixa entropia relativa, determinismo teórico e comportamento suficientemente diferente para ser desejável a detecção em relação à um seno de frequência distinta. Um panorama geral da geração do conjunto de dados pode se vista na Figura 34 .

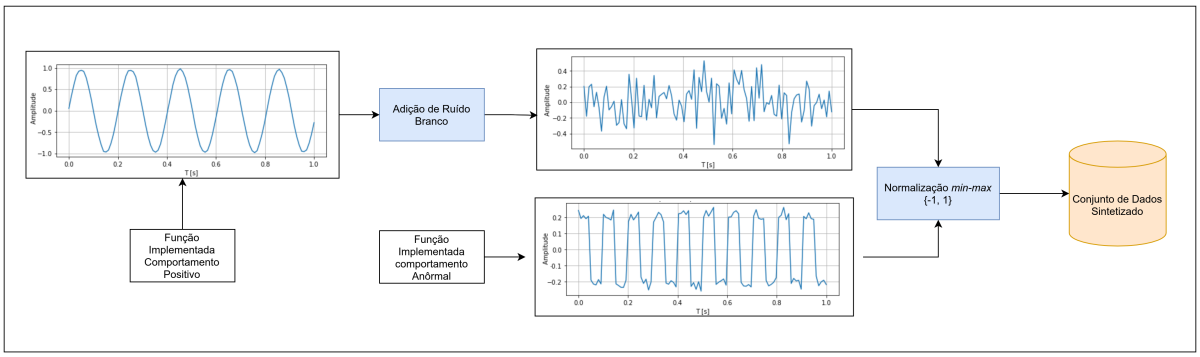

Figura 34 – Geração dos Dados par teste da performance do erro de reconstrução com a entropia.

Fonte: Autor.

Seis diferentes valores de desvio padrão σ foram selecionados, e o SNR do comportamento positivo calculado para cada caso. Na Tabela 2 pode-se ver sumarizados as propriedades dos conjuntos gerados.

Tabela 2 – Dados Gerados.

Desvio Padrão	**SNR (positivo)**	**SNR (Anormal)**	**N . Janelas**
0.1	19.58 dB	23.13 dB	1000 (950/50)
0.5	3.48 dB	23.09 dB	1000 (950/50)
1.0	-3.46 dB	23.00 dB	1000 (950/50)
5.0	-19.54 dB	23.20 dB	1000 (950/50)
10.0	-26.51 dB	23.08 dB	1000 (950/50)
100.0	-49.49 dB	22.92 dB	1000 (950/50)

Com a obtenção dos conjuntos de dados sintéticos, a TadGan foi treinada com 300 épocas sobre cada conjunto. O erro quadrático médio e o escore do crítico foram calculados sobre todas as janelas, analogamente ao procedimento realizado em 4.2.1.1. A dispersão dos dados foi também inspecionada no espaço de *features* resultante, através do qual buscou-se também analisar a influência da entropia no escore gerado pelo crítico. A partir da observação dos resultados, obteve-se indicações sobre a possibilidade de influência da entropia no sinal na usabilidade do erro de reconstrução.

Por fim, uma métrica correlacionada com a entropia do sinal foi selecionada, e os erros de reconstrução de cada janela foram normalizados para essa estimativa da entropia da janela, de modo a verificar-se a correlação entre o colapso da métrica e a entropia do sinal. Dessa forma, a métrica de erro passa a ser medida em $\frac{erro}{entropia}$. Assim, considerando-se a suposição que o erro de reconstrução deve aumentar com a entropia do sinal, a razão $\frac{erro}{entropia}$ deve apresentar menor dependência para o aumento da entropia . Escolheu-se a *fuzzy entropy*, implementada na biblioteca de código aberto *EntropyHub*, pela seu bom desempenho reportada no trabalho de (Weiting Chen, *et al.,* 2007) na mensuração da entropia de séries temporais complexas.

4.3.2 Estabilidade do Treinamento

A literatura apresenta diversos relatos de instabilidades gerais do treinamento de métodos baseados em GANs para detecção de anomalias, como os elencados abaixo: DU *et al.,* (2021) ressaltam as implicações da utilização de GANs na formulação totalmente não supervisionada (conjunto de dados passível de contaminação com anomalias), como utilizado por (GEIGER *et al.,* 2020), e na possibilidade de modelagem do processo anormal devido a contaminação do conjunto de treinamento.

BASHAR *et al.*, (2020) relatou dependência de métodos baseados em GANs para detecção de anomalias para com o número de épocas de treinamento, possivelmente relacionado também com a utilização dos conjuntos contaminado e da modelagem também do comportamento anormal. Também reportou instabilidades gerais no treinamento, supostamente derivadas da instabilidade conhecida dos modelos GANs.

Nessa seção a estabilidade do treinamento do modelo é explorada sobre duas diferentes perspectivas:

1. dependência com o número de épocas: proveniente da contaminação do conjunto de dados e da subsequente possibilidade de modelagem do comportamento anômalo a partir de um determinado número de iterações;

2. variabilidade entre execuções: a da variabilidade dos resultados obtidos em diferentes execuções sobre os mesmos dados, possivelmente originárias da instabilidade do treinamento de GANs.

4.3.2.1 Quanto ao Número de Épocas

Com o objetivo de avaliar a possibilidade de influência do número de épocas de treinamento na performance do método causada pelo treinamento dos modelos sobre conjuntos de dados contaminados, ou seja, com a presença de anomalias, um experimento foi elaborado.

O mesmo conjunto de dados sintético descrito na secção 4.2.1.1.1 foi utilizado, e os modelos treinados por 1000 épocas. A cada 2 épocas, o erro de reconstrução e o escore do crítico foram computados para todas janelas. Ao final do treinamento, avaliou-se a evolução do erro médio para os exemplo anormais, bem como a diferença entre as centroides geométricas das janelas positivas e anormais no espaço de *features* formado pelo escore do crítico e o erro de reconstrução, que espera-se estar relacionado com capacidade de diferenciar os dois processos. Essa distância foi calculada de acordo com a Equação (36):

$$D_{media} = \sqrt{(ER_{positivos} - ER_{anomalos})^2 + (Crit_{positivos} - Crit_{anomalo})^2} \tag{36}$$

onde $ER_{positivos}$ é o erro de reconstrução médio para as janelas não anômalas, $ER_{anomalos}$ o erro de reconstrução para a janelas anômalas e $Crit_{positivos}$ e $Crit_{anomalo}$ o escore do crítico para as janelas não anômalas e anômalas, respectivamente.

Além disso, mediu-se a dispersão dos janelas em torno dessa centroide, através do desvio padrão para cada um dos processos. Os resultados foram analisados graficamente em busca de indicações da dependência para com o número de épocas.

4.3.2.2 Variabilidade do Treinamento

Com objetivo de investigar a influência da instabilidade do treinamento geral de GANs na aplicação do método, utilizou-se um sinal real proveniente do conjunto de dados NAB, *exchange-2_ cpm_ results*, e a performance obtida em 10 diferentes execuções comparada. Os resultados foram avaliados em busca de sinais da existência de variação significativa da performance entre execuções através da comparação do escore F1 do protocolo (GEIGER *et al.,* 2020) obtido, bem como do F1 tradicional amostra-a-amostra.

4.4 AVALIAÇÃO DISCRIMINADA POR PROBLEMA

Com o objetivo de avaliar a TadGan em séries com anomalias de causa conhecida, levantando assim indicações da performance para diferentes tipos de anomalias e possíveis limitações, assim como em sinais reais de aplicações diversas não presentes na avaliação do trabalho original de (GEIGER *et al.,* 2020), séries foram sintetizadas com a introdução de tipos específicos de anomalias, e conjuntos de dados selecionados do *UCR Anomaly Benchmark* para a avaliação do método. Os resultados foram apresentados através do F1 (protocolo GEIGER *et al.,* 2020) e F1 amostra-a-amostra.

4.4.1 Anomalias Sintetizadas

Um processo de base foi definido, e 10 tipos diferentes de anomalias sintetizadas artificialmente e inseridas no sinal. Para o processo de base, definiu-se uma função senoidal como visto na equação (37):

$$S = \sin(2\pi f_1(x + \mathcal{N}(\mu_1, \sigma_1)) + \mathcal{N}(\mu_2, \sigma_3) \tag{37}$$

Foram geradas séries com 1000 amostras, com $f_1 := 10$. As anomalias propostas são:

- **Anomalia na energia:** 50 amostras foram multiplicadas por 2 ($x_n(anorm.) := 2x_n$);
- **Anomalia na frequência fundamental**: 50 amostras foram gerada com $f_1(anorm.) := 1.5f_1$;
- **Anomalia na fase:** 50 amostras sequenciais foram invertidas $x_n(anorm.) := -x_n$;
- **Anomalia exemplos faltantes:** 50 amostras foram igualadas a zero ($x_n(anorm.) := 0$);
- **Anomalia no processo**: 50 amostras foram geradas por uma onda quadrada de mesma frequência;
- **Anomalia por adição ruído**: foi adicionado ruído gaussiano a 50 amostras ($x_n(anorm.) := x_n + \mathcal{N}(\mu, \sigma)$);
- **Anomalia por remoção de ruido:** 50 amostras foram geradas sem ruído;
- **Anomalia por mudança na média:** à 50 amostras foi feito $x_n(anorm.) := x_n + 3$;
- **Anomalia por mudança na tendência:** à 50 amostras sequenciais foi somadas uma reta ($x_n(anorm.) := x_n + a(x_n - n) + b$);
- **Anomalia pontual amplitude:** a uma amostra n foi feito $x_n(anorm.) := x_n + 3$.

As séries geradas foram inspecionadas no domínio do tempo e da frequência para verificação do comprimento dos objetivos estabelecidos. A TadGan foi então treinada com 200 épocas sobre as séries sintetizadas. A dispersão do erro de reconstrução e escore do crítico foram analisados graficamente para cada série, a fim de verificar a possibilidade de detecção. Exemplos de reconstrução e do comportamento das métricas no tempo também foram examinados. Em sequência, um detector ingênuo foi implementado, considerando anômalo todo exemplo cuja amplitude do escore de anomalia detectado fosse maior que 3 desvios padrões, como $Amplitude\,Escore\,Anomalia > 3\sigma$. Dessa forma pode-se computar os valores do F1, e avaliar quantitativamente a performance nos diferentes sinais. Os pontos detectados foram ainda concatenados em janelas, seguindo a regra de que pontos à menos de n amostras de outros pontos são concatenados na mesma janela. n foi escolhido como 100.

4.4.2 Diferentes Problemas

Em sequência, alguns conjuntos de dados foram compilados de campos de aplicações diversos para a avaliação da performance, provenientes do conjunto *UCR Anomaly Benchmark*, introduzido por (WU *et al.*, 2021) com o objetivo de abordar problemas apontados com os conjuntos de avaliação disponíveis, como o NAB. A avaliação se justifica para verificar a performance do método em dados obtidos com outra metodologia de anotação e entediamento de anomalia, corroborando assim com a generalidade da abordagem. Além disso, as aplicações selecionadas não estão presentes no conjunto de avaliação original de (GEIGER *et al.,* 2020), servindo também para fornecer intuições sobre performance e aplicabilidade em problemas alternativos.

Realizou-se a avaliação sobre 7 diferentes sinais, provenientes de 6 áreas distintas. Uma breve discrição de cada sinal é fornecida em sequência:

- 032_UCR_Anomaly_DISTORTEDInternalBleeding4_1000_4675_5033: pressão arterial medida de um porco sangrando;
- 162_UCR_Anomaly_WalkingAceleration5_2700_5920_5979: sinal de passos coletados de um indivíduo através de um acelerômetro;
- 145_UCR_Anomaly_Lab2Cmac011215EPG1_5000_17210_17260: EPG gravado de um inseto;
- 153_UCR_Anomaly_PowerDemand2_14000_23357_23717: variação da demanda de energia elétrica na Itália;
- 161_UCR_Anomaly_WalkingAceleration1_1500_2764_2995: sinal de passos coletados de um indivíduo através de um acelerômetro;

- 131_UCR_Anomaly_GP711MarkerLFM5z5_5000_8612_8716: sinal de um indivíduo do *GaitPhase Database*. Caminhada em 1.1 m/s;
- 117_UCR_Anomaly_CIMIS44AirTemperature5_4000_4852_4900: Temperatura do ar registrada em um estação meteorológica.

A TadGan foi treinada por 200 épocas sobre cada série, os resultados avaliados graficamento e o F1 obtido com a aplicação do classificador ingênuo sobre o escore combinado. Um procedimento adicional de pré-processamento foi adotado, no qual realizou-se o *downsampling* de algumas séries, de modo a permitir que as janelas de 100 exemplos capturassem características relevantes do sinal.

4.5 PROPOSTAS DE MODIFICAÇÕES

Baseado nos experimentos conduzidos, propostas de modificações na aplicação do método foram efetuadas, implementadas e, neste trabalho, superficialmente testadas. Deu-se atenção para a possibilidade de obtenção de métricas alternativas derivadas do método que pudessem auxiliar na mensuração da anormalidade, bem como da metodologia para combinação desses valores em um escore final da anomalia do exemplo. Além do erro de reconstrução e escore do critico sobre as janelas originais, foram propostas três métricas alternativas:

- **Escore do Critico Diferencial**: computa-se a diferença entre o escore do crítico da janela original e da janela reconstruída. Espera-se dessa forma obter-se uma métrica correlacionada com o erro de reconstrução, mas aproveitando-se a capacidade do crítico de identificar pontos chave do processo alvo. É calculada como na equação (38).

$$Err_{C_x} = \frac{|\mathcal{C}_x(\mathcal{G}(\mathcal{E}(X))) - \mathcal{C}_x(X)|}{\mathcal{C}_x(X)} \tag{38}$$

- **Variância do ponto**: desvio padrão medido no valor de um ponto reconstruído entre as n janelas que ele está presente, divido sobre a amplitude original do ponto. Espera-se que a reconstrução de um ponto em uma janela anômala apresente maior variância, visto que por ser um comportamento inconsistente, será reconstruído de diferentes maneira dependendo do contexto da janela formada. É calculado como na equação (39):

$$VarRec_t = \frac{\sigma(\{x_t^{(1)}...x_t^{(n)}\})}{|x_{original}|} \tag{39}$$

  Observa-se a possibilidade de problemas de instabilidade númerica pela divisão com a amplitude original do ponto

- **Distância no espaço Z**: distância euclidiana entre a representação **z** do exemplo original e do reconstruído. Espera-se também obter uma métrica correlacionada com o erro de reconstrução, porém aproveitando-se as características extraídas pelo *Encoder*. Em outras palavras, pode-se avaliar o erro de reconstrução de características de difícil comparação diretamente nos valores no tempo. Essa é a métrica que aparente apresentar maior potencial, por ser conceitualmente uma medição direta de semelhança entre as janelas na representação do processo encontrada no treinamento. Pode ser calculada como na equação (40);

$$D_z = ||\mathcal{E}(\mathcal{G}(\mathcal{E}(X))) - \mathcal{E}(X)||_2 \tag{40}$$

  As componentes da dimensão z são normalizadas para desvio padrão unitário e média zero antes do cálculo da distância.

Para encontrar a melhor combinação entre as características apresentadas para a métrica final de anormalidade, utilizou-se o método PCA, onde encontra-se a direção do espaço de variáveis de maior variância, e todas dimensões são projetadas nessa direção. Dessa maneira, se espera encontrar para cada processo a combinação dos escores que discriminem mais os exemplos divergentes. Possíveis instabilidades podem emergir da utilização do PCA para combinar as diferentemente métricas obtidas, e deve ser futuramente estudadas.

As métricas foram implementadas, e avaliadas comparativamente em relação às métricas originais sobre os dados descritos na seção 4.4.2, através do F1. Observou-se também o comportamento do escore final através de gráficos.

5 RESULTADOS E DISCUSSÕES

5.1 VERIFICAÇÃO DA IMPLEMENTAÇÃO

Os dados selecionados descritos na secção 4.1.1 foram aplicados à TadGan e a performance foi avaliada de acordo com o protocolo descrito. Observou-se um alto custo computacional na reprodução dos resultados, resultando em altos tempos de treinamento na configuração de hardware utilizada.

Na Tabela 6 pode-se ver os resultados obtidos em comparação com os reportados por GEIGER *et al.,* (2020).

Tabela 3 – Comparação Entre os Resultados Obtidos para o F1 (GEIGER *et al.,* 2020).

Nome Sinal	**Recall**	**Precision**	**F1**	**F1(Geiger et al)**
exchange-2_ cpc_ results	N.D.	N.D.	N.D.	N.D.
exchange-2_ cpm_ results	1.00	0.667	0.800	0.667
exchange-3_ cpc_ results	1.00	0.750	0.857	1.000
exchange-3_ cpm_ results	1.00	0.500	0.667	0.667
exchange-4_ cpm_ results	1.00	0.800	0.889	0.889
art_ daily_ flatmiddle	1.00	1.000	1.000	1.00
art_ daily_ jumpsdown	1.00	1.000	1.000	1.000
art_ daily_ jumpsup	1.00	1.000	1.000	1.000
art_ daily_ nojump	N.D.	N.D.	N.D.	N.D.
art_ increase_ spike_ density	0.750	0.750	0.750	0.889
art_ load_ balancer_ spikes	0.750	0.750	0.750	0.889

onde N.D. denota que nenhuma anomalia foi detectada, e portanto o F1 não está definido.

Observa-se a proximidade entre os resultados reportados no trabalho original de autores e os obtidos com a implementação utilizada. As diferenças observadas são ilustrativas da instabilidade do treinamento, que gera modelos diferentes sobre mesmos dados, e também de eventuais diferenças nas implementações dos métodos utilizados de pré-pŕocessamento e pós-processamento dos escores obtidos. É importante ressaltar também que devido ao protocolo utilizado para a avaliação, a não detecção de uma janela apenas, ou detecção de uma falso positivo, gera grande variação numérica na performance resultante.

A fim de ilustração, o sinal *exchange-2_ cpm_ results* foi inspecionado. Na Figura 35 pode-se ver o sinal original e reconstruído, bem como o erro de reconstrução para cada amostra e o escore do crítico . No gráfico inferior é mostrado também o escore combinado resultante e as anomalias detectadas a partir desse escore e do procedimento relatado por (GEIGER *et al.,* 2020). O eixo x das métricas não possui unidade, e representa apenas o valor obtido para aquela métrica. No eixo y foi apresentado diretamente o *index* do vetor

da série temporal, sem qualquer correspondência com alguma unidade de tempo, visto a irrelevância para a esta análise.

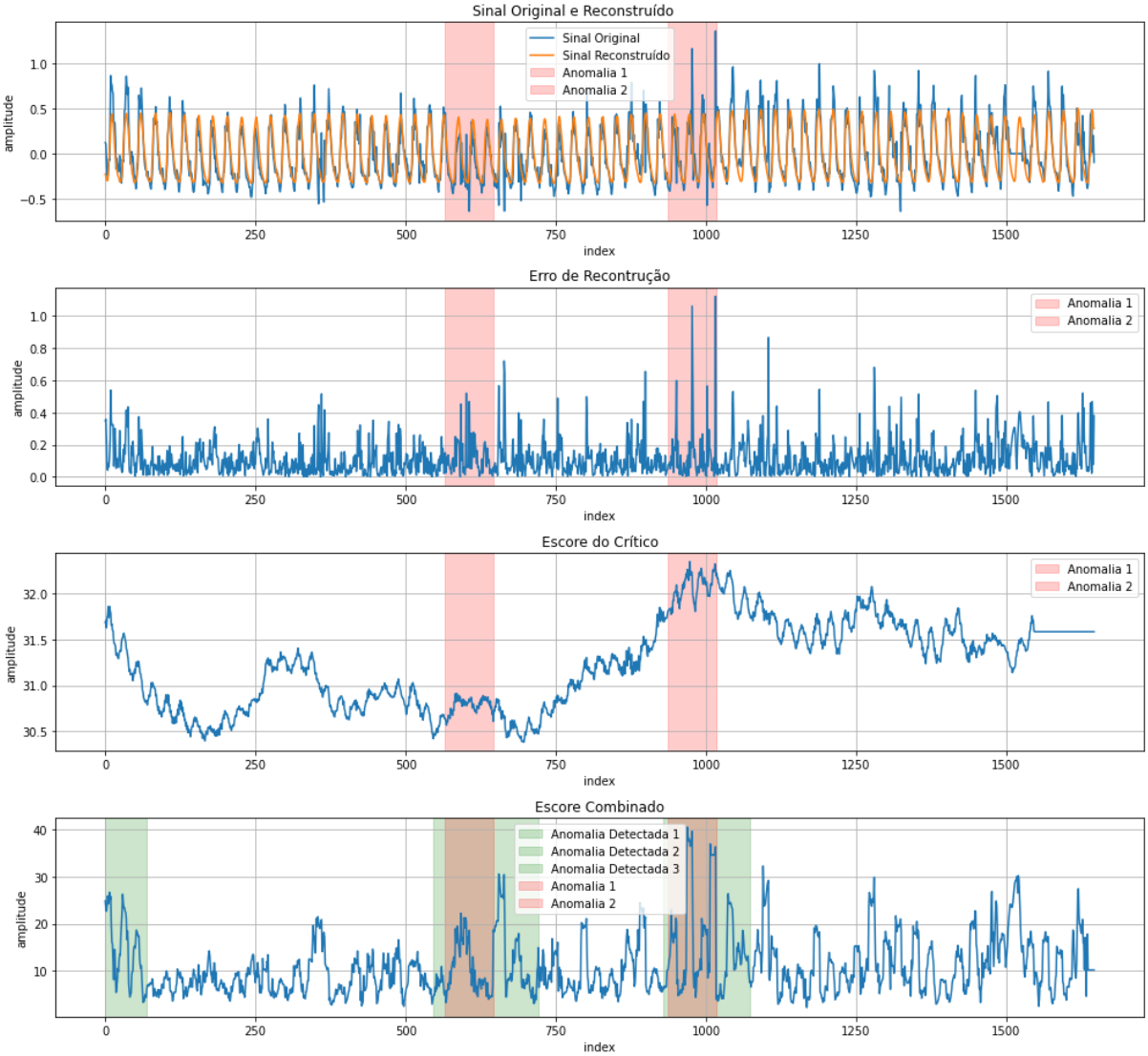

Figura 35 – Sinal exchange-2_cpm_results.

Fonte: Autor.

Na Figura 36 pode-se ver dois exemplos de reconstrução de uma janela normal e anormal, corroborando com a capacidade de modelagem do método. Observa-se a consistência da janela reconstruída tanto no domínio do tempo quanto da frequência.

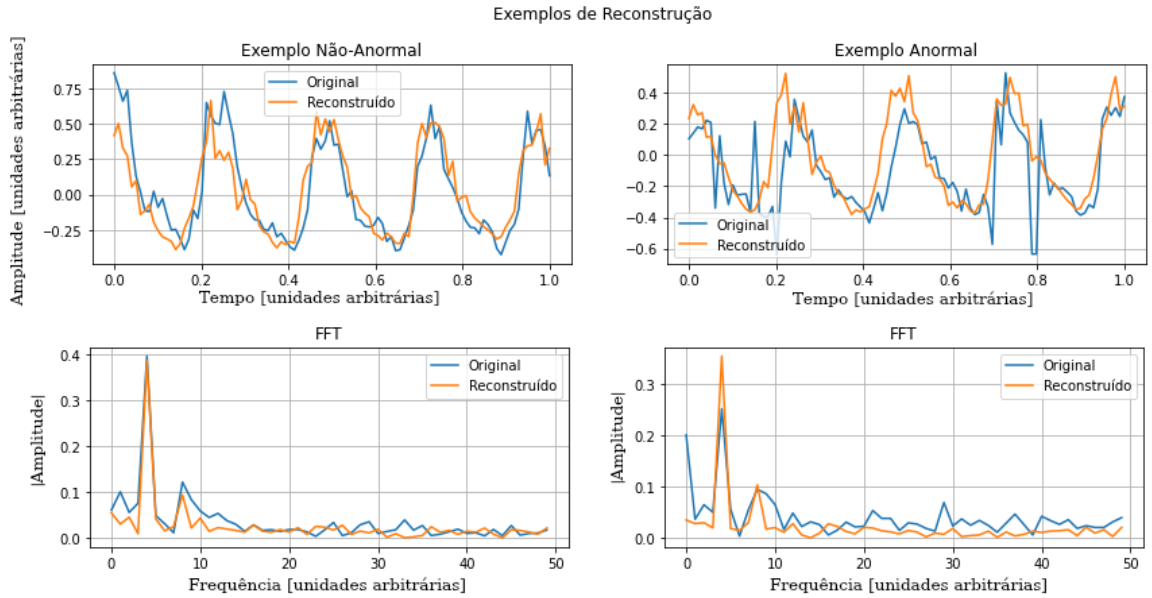

Figura 36 – Exemplos de Janelas Reconstruídas.

Fonte: Autor.

O método falhou em detectar anomalias no sinal *art_daily_nojump*. Vê-se o sinal na figura 37, onde a anomalia introduzida consiste na ausência do comportamento periódico de elevação da amplitude como observado nas amostras anteriores.

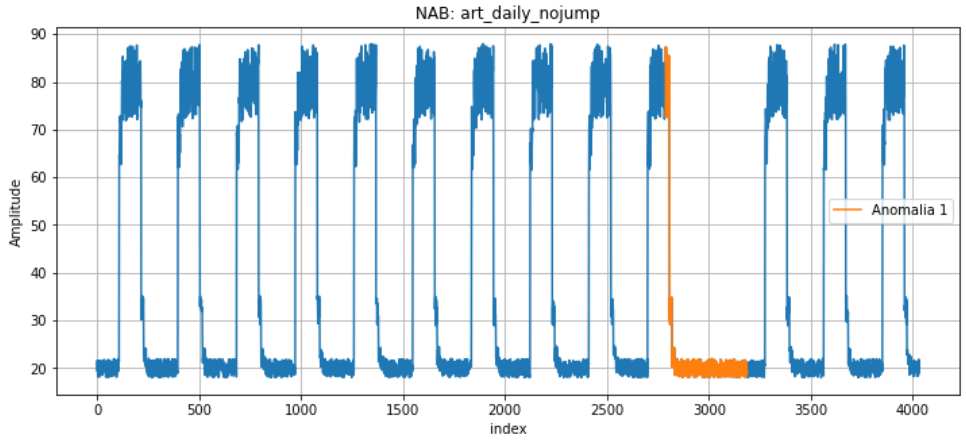

Figura 37 – Sinal art_daily_nojump.

Fonte: Autor

Esse comportamento periódico, entretanto, não é observável dentro do tamanho definido da janela, de 100 amostras. Esse fato é facilmente verificável através da medição do período, que é de 144 amostras. Dessa forma, o gerador não consegue modelar esse comportamento periódico como parte do comportamento não-anômalo. Além disso, as amostras observadas na região são condizentes com as observadas anteriormente, resultando na formação de janelas não anômalas e em erros de reconstrução e escore do critico

semelhantes com o do resto do sinal. O resultado de ambos escores pode ser visto na Figura 38.

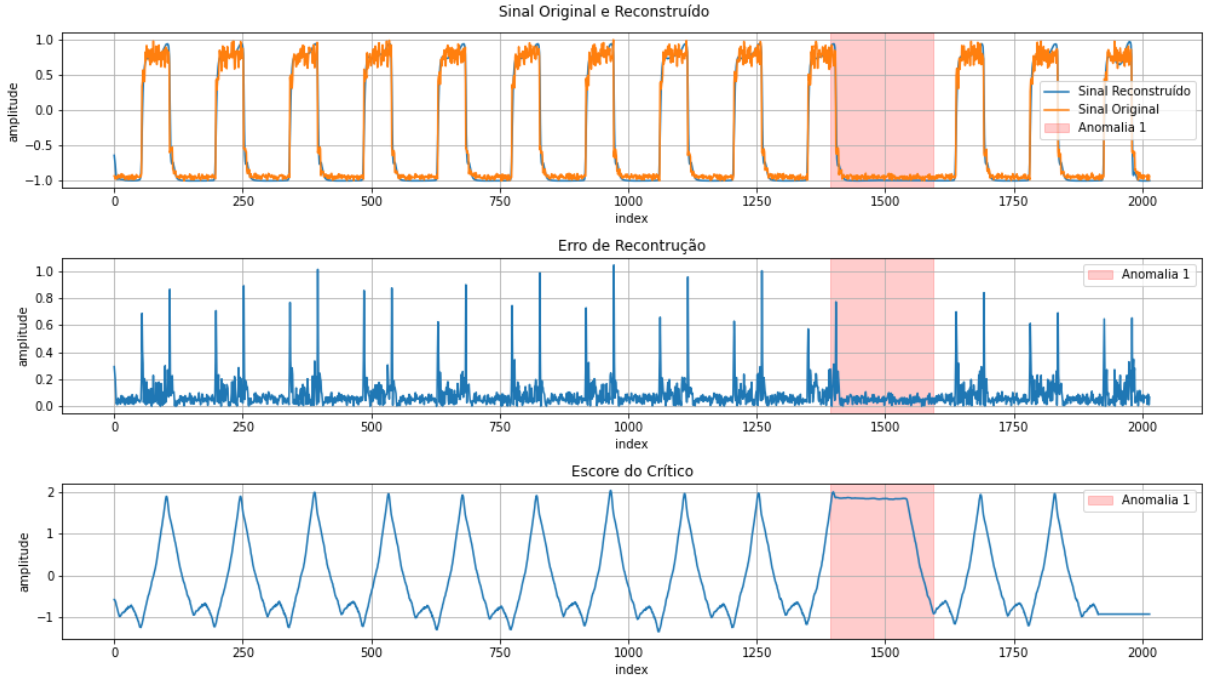

Figura 38 – Sinal art_daily_nojump.

Fonte: Autor.

A característica da dependência da escala de detecção e padrões para com o tamanho da janela representa uma limitação significativa do método, e deve ser considerado na sua aplicação à novos problemas.

5.2 VERIFICAÇÃO DOS PRINCÍPIOS DO MÉTODO

5.2.1 Análise Geral do Método

Os resultados obtidos para os dois tipos de dados sintetizados são apresentados e discutidos nas secção seguintes.

5.2.1.1 Senoides Atenuadas

As funções definidas foram implementadas, e as janelas geradas inspecionadas. Na Figura 39 pode-se ver dois exemplos de janelas sintetizadas, do comportamento normal (direita) e anormal (esquerda), bem como o seu espectro em frequência (FFT). No terceiro e quarto gráficos são mostrados o histograma total das amostras, para o comportamento normal e anormal, bem como a média das FFT, corroborando com a obtenção de um conjunto de acordo com a expressão analítica apresentada. Uma "frequência de amostragem"foi arbitrariamente definida de modo a possibilitar a análise em unidades reais de tempo e frequência.

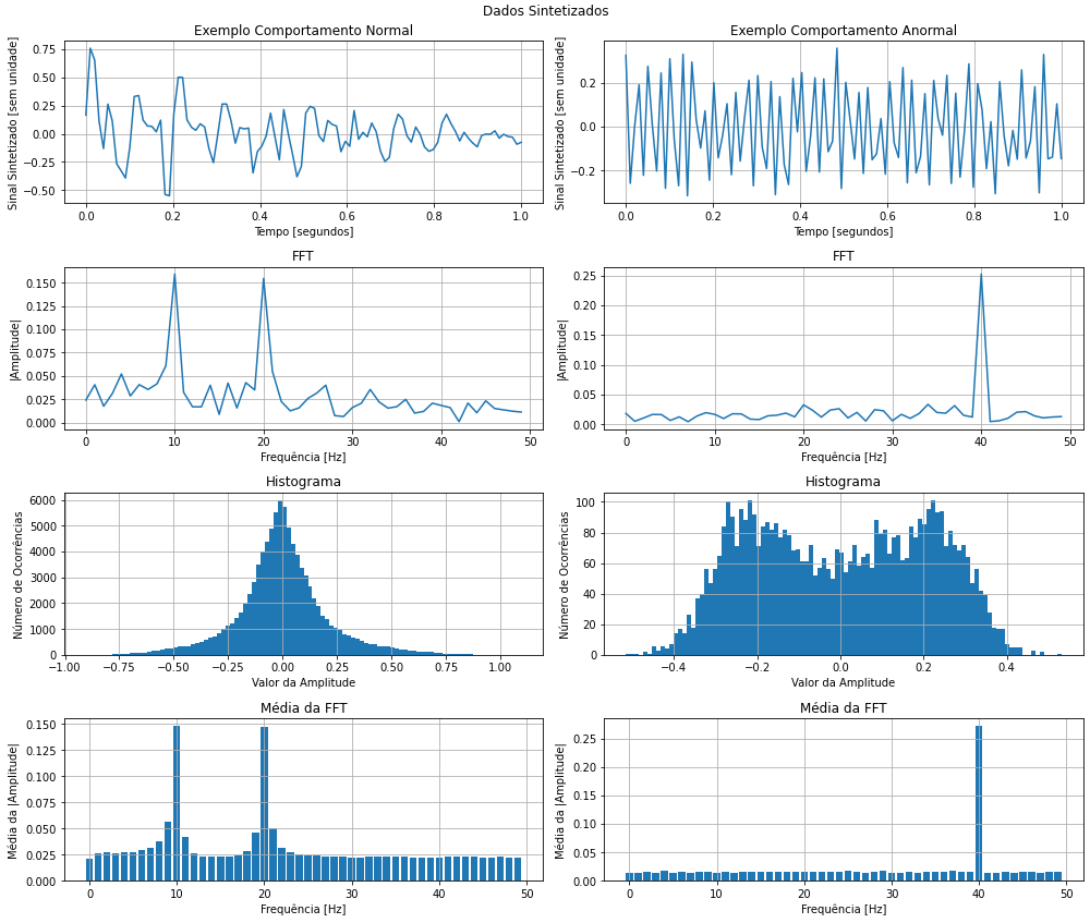

Figura 39 – Janelas Sintetizadas para Verificação do Método.

Fonte: Autor.

O modelo foi treinado por 300 épocas, e as curvas de custo inspecionadas para validação da convergência dos objetivos. As funções de custo para cada um dos modelos treinados podem ser vistas na Figura 40, na qual a primeira coluna mostra as componentes individuais do custo, e a segunda o valor total utilizado para a realização do gradiente descendente.

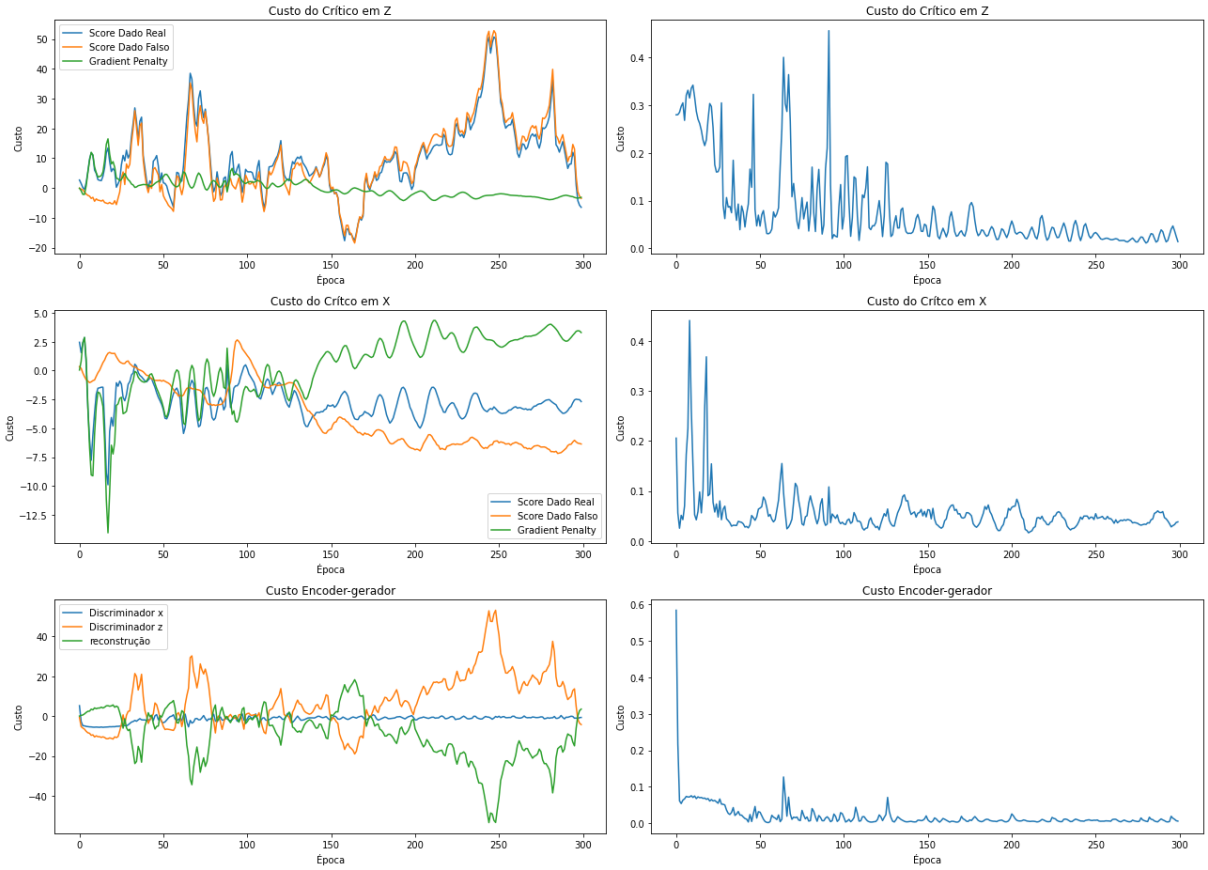

Figura 40 – Funções de custo do treinamento.

Fonte: Autor.

Observa-se a convergência geral dos objetivos, dada pela diminuição para com as épocas de treinamentos das componentes individuais do custo. É também ressaltado o caráter *adversarial* do treinamento, onde com a diminuição do erro de reconstrução, por exemplo, observa-se um aumento dos erros dos discriminados, visto que os modelos geradores estão gerando amostras mais "realistas".

Em sequência, explorou-se exemplos de sinais reconstruídos pelo modelo, verificando visualmente a coerência tanto no domínio do tempo quanto da frequência. Os resultados observados corroboram com a capacidade do método de capturar o processo gerador, dada a consistência do sinal reconstruído. Observou-se também o comportamento da reconstrução de exemplos anormais, com os quais o resultado do gerador foi um de sinal próximo da função definidora do comportamento positivo, corroborando com a premissa

da captura do processo majoritário positivo e da utilidade do erro de reconstrução como métrica de anormalidade.

O resultado para o conjunto de dados reconstruído para um conjunto de modelos treinado por 300 épocas pode ser visto na Figura 41.

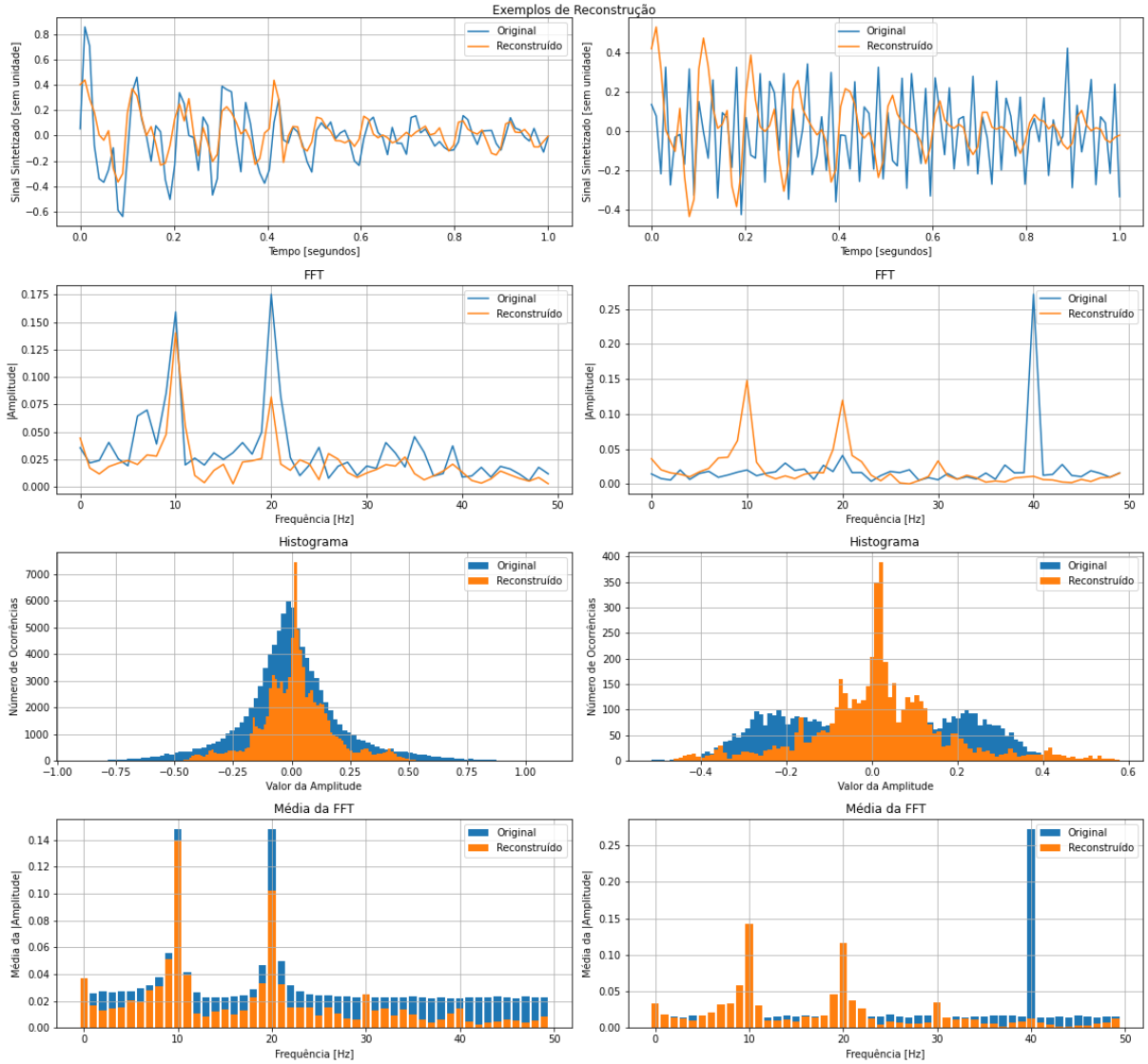

Figura 41 – Visão geral das janelas reconstruídas no experimento para as senóides.

Fonte: Autor.

Os quatro gráficos superiores apresentam um exemplo de uma janela reconstruída do processo definido como não-anômalo e anômalo, respectivamente, no domínio do tempo e da frequência Nos quatro gráficos inferiores vê-se o histograma original e reconstruído, bem como as médias de cada *bin* em frequência sobre todas janelas. Observa-se para os exemplos não anômalos a manutenção das características gerais da distribuição, bem como o comportamento em frequência. Por outro lado, para os exemplos anormais, vê-se que os exemplos reconstruídos aproximam-se do processo definido para a normalidade, corro-

borando com o conceito do método no qual o gerador modela somente o comportamento positivo, e sempre gera exemplos relativos a esse processo.

Uma vez validada visualmente a convergência geral do treinamento e dos objetivos, o ensaio foi repetido 10 vezes, de forma a possibilitar afirmações mais robustas e de verificar a estabilidade do treinamento. Foram compilados os valores do erro de reconstrução e escore do crítico para cada janela ao fim das 200 épocas de treinamento. Na Tabela 6 vê-se sumarizados os valores médios obtidos sobre todos o conjunto e somente sobre os exemplos anormais.

Observa-se a diferença entre as médias dos valores, que implica na possibilidade de diferenciação entre os processos com a utilização da métrica. Na última coluna é colocada a porcentagem dos exemplos anômalos maiores que 3 desvios padrões em relação à média sobre todos exemplos do conjunto, como uma maneira de quantificar a separabilidade entre os processos fornecidas pela métrica em questão.

Tabela 4 – Resultados Sumarizados Obtidos.

	Janelas Positivas		**Janelas Anômalas**		**% Janelas Anormais $>3\sigma$**
	Média	**Desvio Padrão**	**Média**	**Desvio Padrão**	
Escore do Crítico	0.147	0.109	-0.683	0.0870	0 %
Erro Reconstrução (Quadrático)	0.191	0.02	0.250	0.01	47,00%
Erro Reconstrução (DTW)	1.415	0.108	1.903	0.034	36,00%

Os resultados individuais de cada ensaio também foram apresentados graficamente. Nas Figuras 42, 43 e 44, vê-se, respectivamente, os resultados obtidos para cada um dos 10 ensaios para o erro de reconstrução quadrático, por DTW e do escore do crítico (normalizado por min-max em {-1.1}).

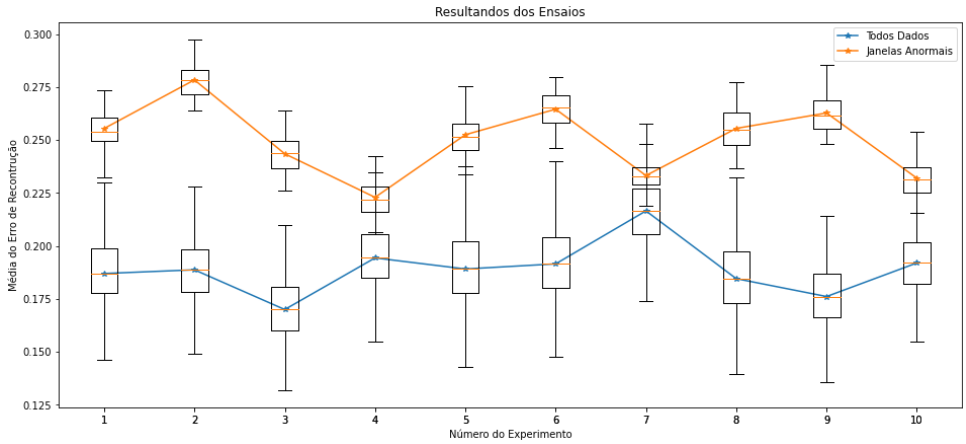

Figura 42 – Erro de Reconstrução quadrático em cada execução.

Fonte: Autor.

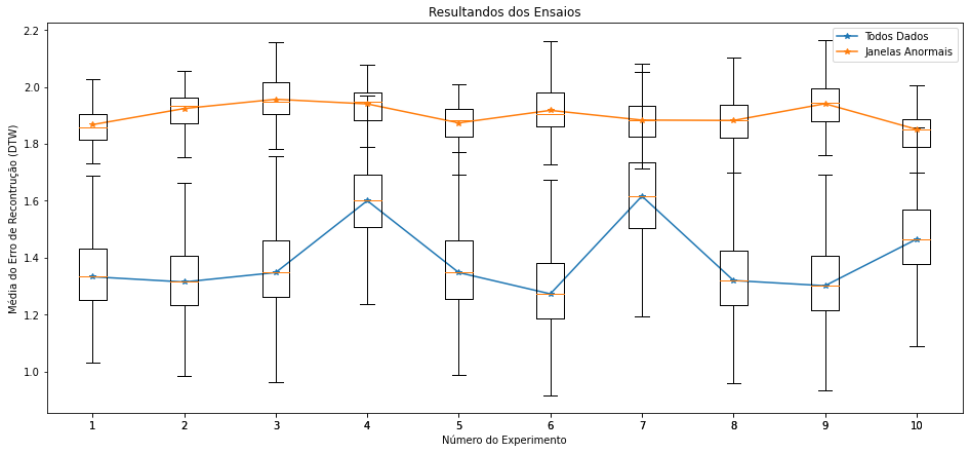

Figura 43 – Erro de Reconstrução DTW em cada execução.

Fonte: Autor.

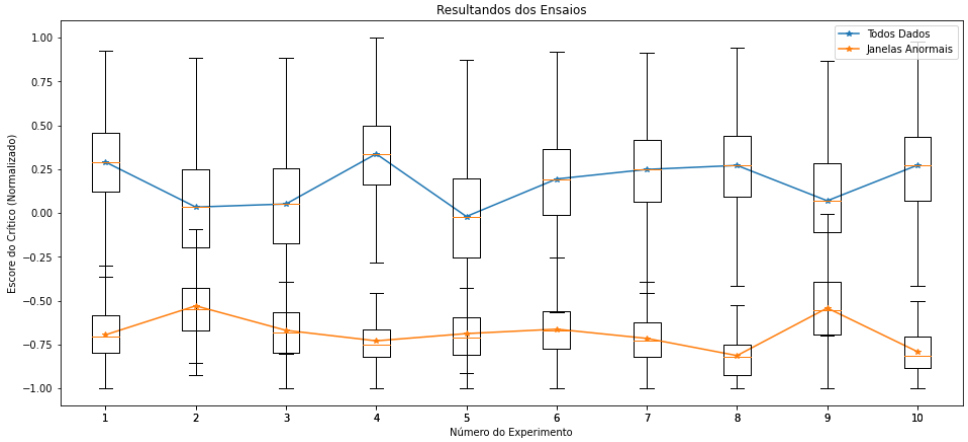

Figura 44 – Escore do Crítico em cada execução.

Fonte: Autor.

Observa-se a capacidade de separação geral entre as janelas positivas e anômalas com as métricas, bem como a existência de variação na performance em diferentes execuções.

Por fim, exemplos do gráfico da dispersão do erro de reconstrução e o escore do crítico podem ser visto na Figura 45 para seis dos ensaios realizados. Valida-se a capacidade de separação do espaço entre janelas dos dois processos, corroborando com sua aplicação para detecção de anomalias.

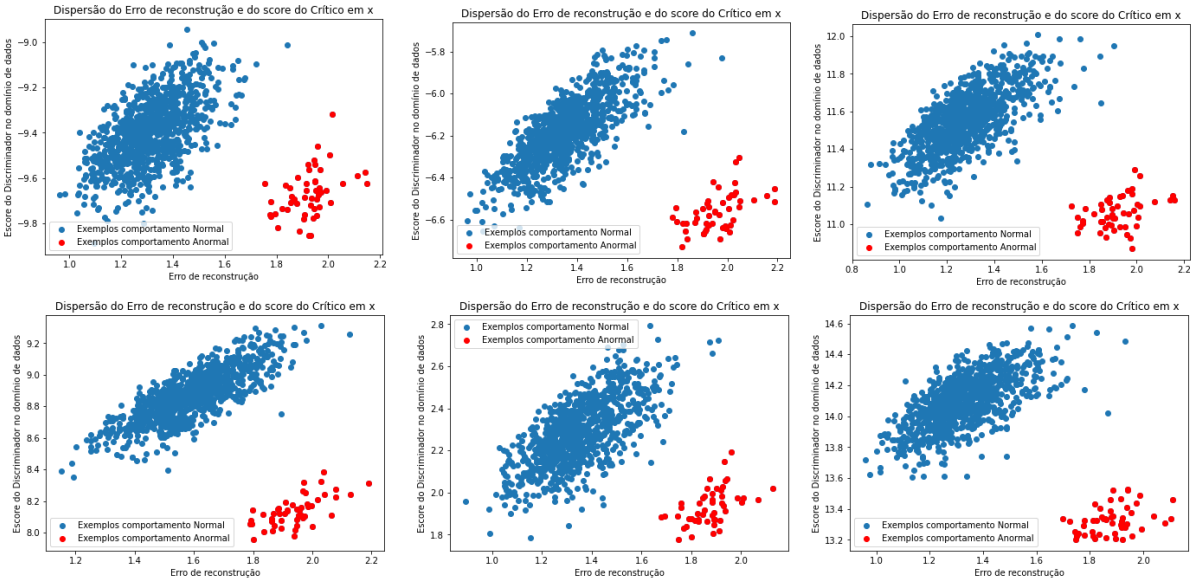

Figura 45 – Dispersão das Janelas no Espaço de *Features* formado.

Fonte: Autor.

5.2.1.2 Processo ARMA

As funções definidas foram implementadas, e as janelas geradas inspecionadas. Um exemplo do sinal gerado pode ser visto na Figura 46, apresentada de forma análoga à Figura 39.

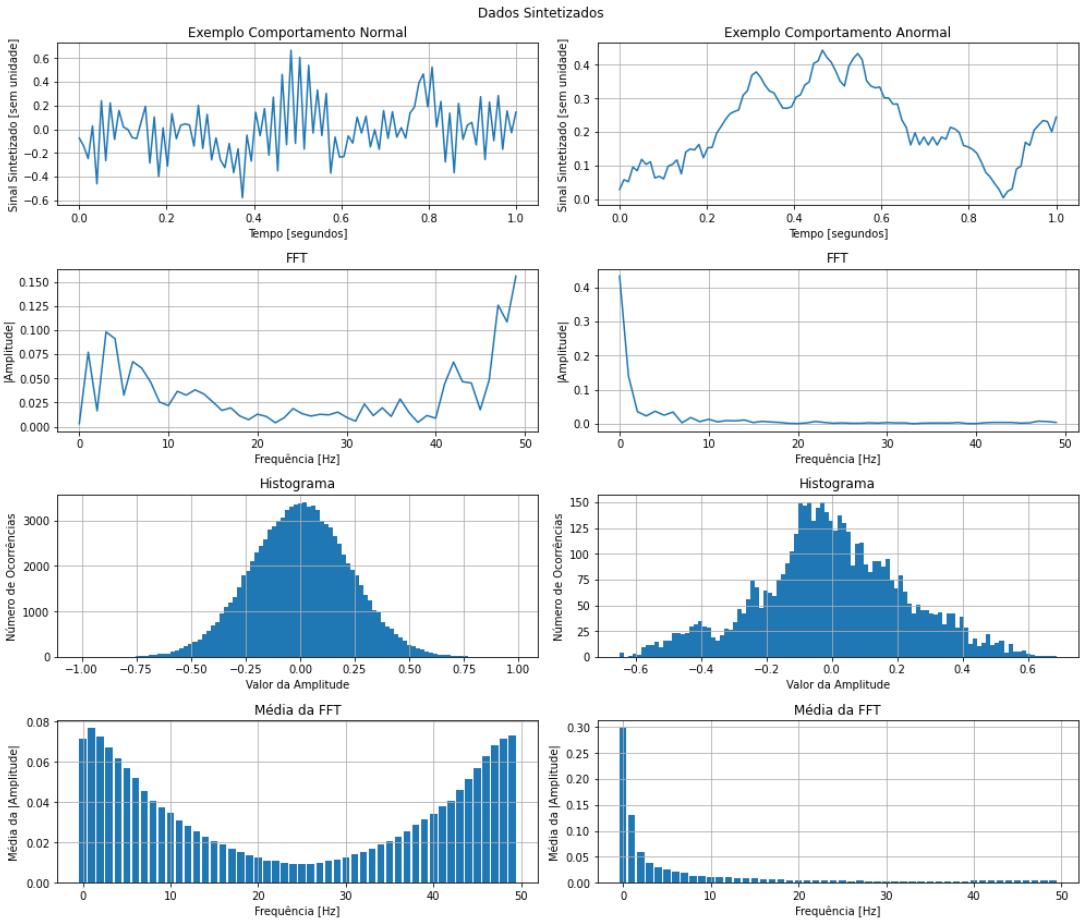

Figura 46 – Exemplo de janelas sintetizadas do processo ARMA(1,1).

Fonte: Autor.

A metodologia adotada para os sinais senoidais gerados foi replicado para o processo ARMA. As curvas de treinamento foram inspecionados, e o conjunto de dados reconstruído gerado pelo método comparado com o original de modo a garantir-se a convergência dos objetivos.

Na Figura 47 vê-se exemplo de janelas reconstruídas para os dois processos bem como as características gerais sobre todas as janelas. Observa-se a modelagem predominante do comportamento majoritário definido como normal.

Figura 47 – Visão geral das janelas reconstruídas no experimento para as senoides.

Fonte: Autor.

Na Tabela 5 pode-se ver os resultados obtidos para o processo ARMA. Vê-se nesse cenário uma menor capacidade de separação obtida com os erros das janelas.

Tabela 5 – Resultados Sumarizados Obtidos.

	Janelas Positivas		**Janelas Anômalas**		**% Janelas**
	Média	**Desvio Padrão**	**Média**	**Desvio Padrão**	**Anormais $>3\sigma$**
Escore do Crítico	0.000	0.099	-2.44	0.968	55.80
Erro Reconstrução (Quadrático)	0.2404	0.0190	0.211	0.0457	8.60%
Erro Reconstrução (DTW)	1.671	0.1276	1.44	0.394	11.12%

Nas Figuras 48, 49 e 50, resultados no formato de gráficos são apresentados para a dispersão das janelas sob o erro de reconstrução quadrático, por DTW e o escore do critico, respectivamente, para cada um dos 10 ensaios.

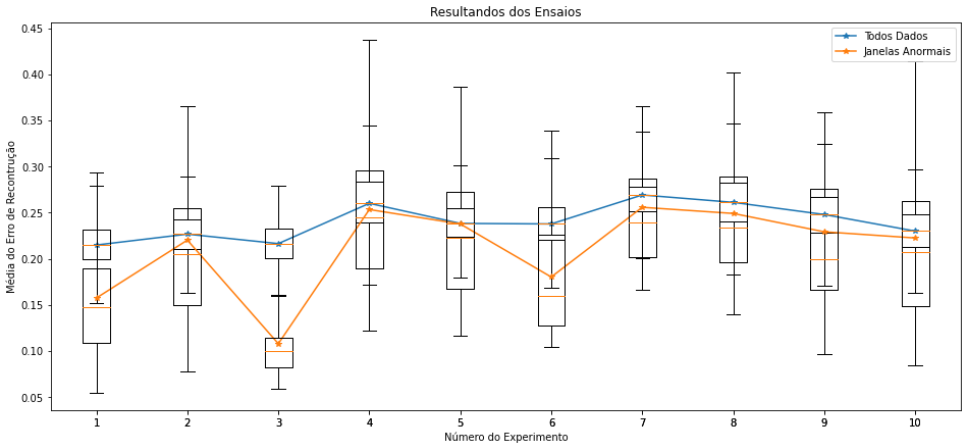

Figura 48 – Erro de Reconstrução quadrático em cada execução.

Fonte: Autor.

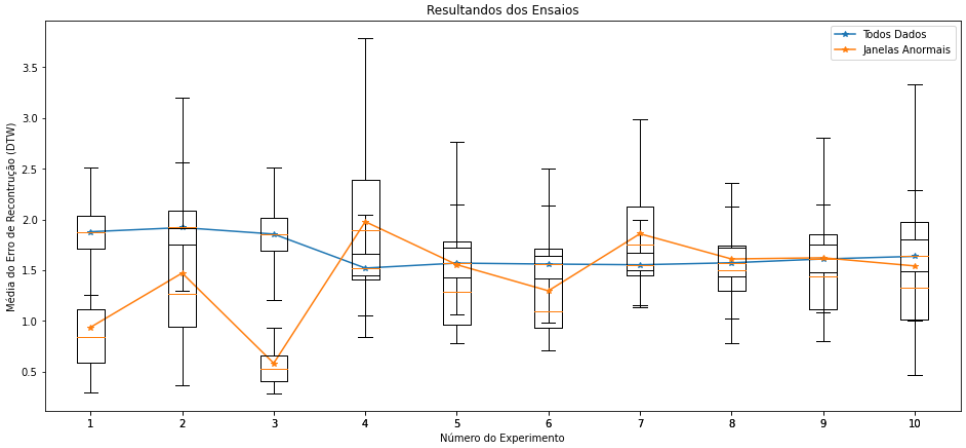

Figura 49 – Erro de Reconstrução DTW em cada execução.

Fonte: Autor.

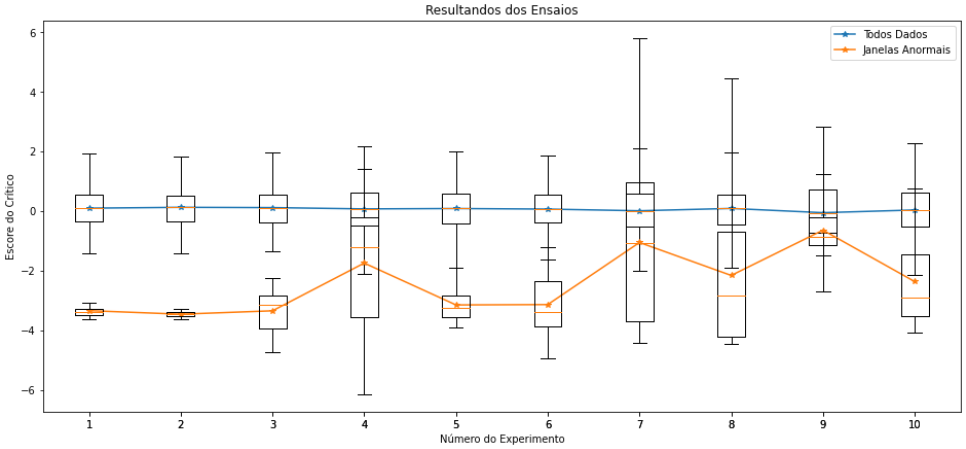

Figura 50 – Escore do Crítico em cada execução.

Fonte: Autor.

Observa-se claramente a menor capacidade de diferenciação obtida nesse caso com os erros de reconstrução. Pode-se ver na Figura 51 a dispersão das janelas no espaço de *features* resultante do treinamento.

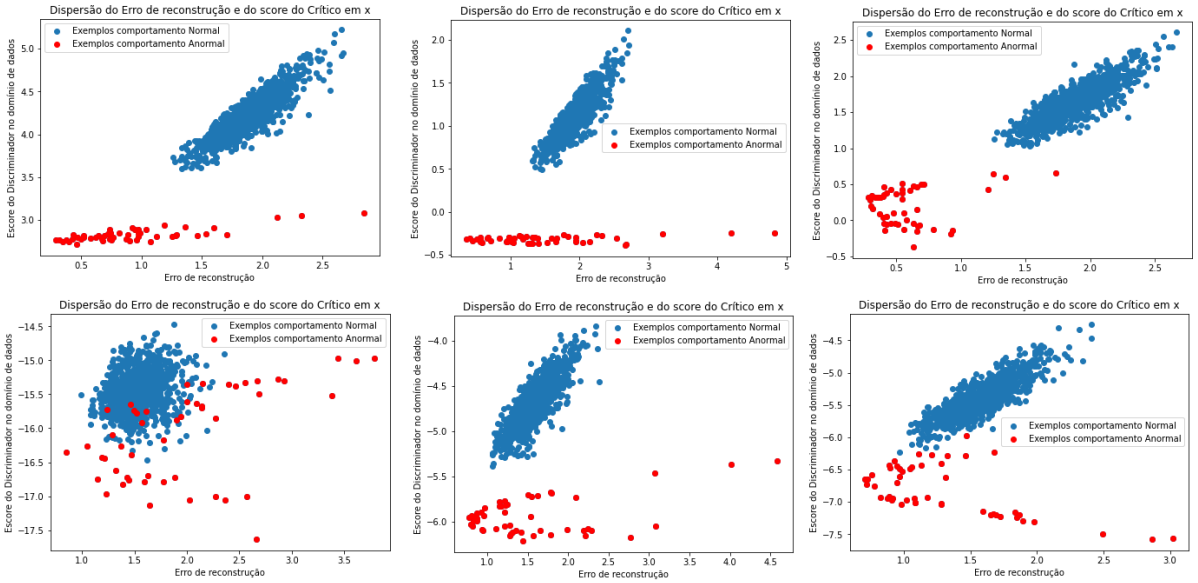

Figura 51 – Espaço de *Features* para processo ARMA.

Fonte: Autor.

Observa-se que apesar da obtenção de um espaço com separabilidade entre os dois processos, o erro de reconstrução é menor para os sinais anômalos, contrariando a expectativa do funcionamento do método. Esse observação apresenta grande interesse, dado que parece se relacionar intimamente com as questões levantadas sobre as limitações do erro de reconstrução para a discriminação das anomalias, discutido na seção 4.3.1.

Constata-se que a entropia do processo não-anômalo é superior, culminando em um erro de reconstrução mínimo teórico maior, dado pela teoria da informação. Além disso, dada a similaridade entre os processos, características extraídas para a normalidade podem ter relevância para os exemplos anormais, como discutido por (BATTIKH *et al.*, 2021). A investigação da influência da entropia dos processos na usabilidade do erro de reconstrução como métrica de detecção de anomalias realizada será apresentada na secções seguintes.

5.2.2 Capacidade do *Encoder*

O conjunto de dados definido foi sintetizado, e inspecionado para o comprimento dos objetivos. Suas propriedades podem ser observada na Figura 52.

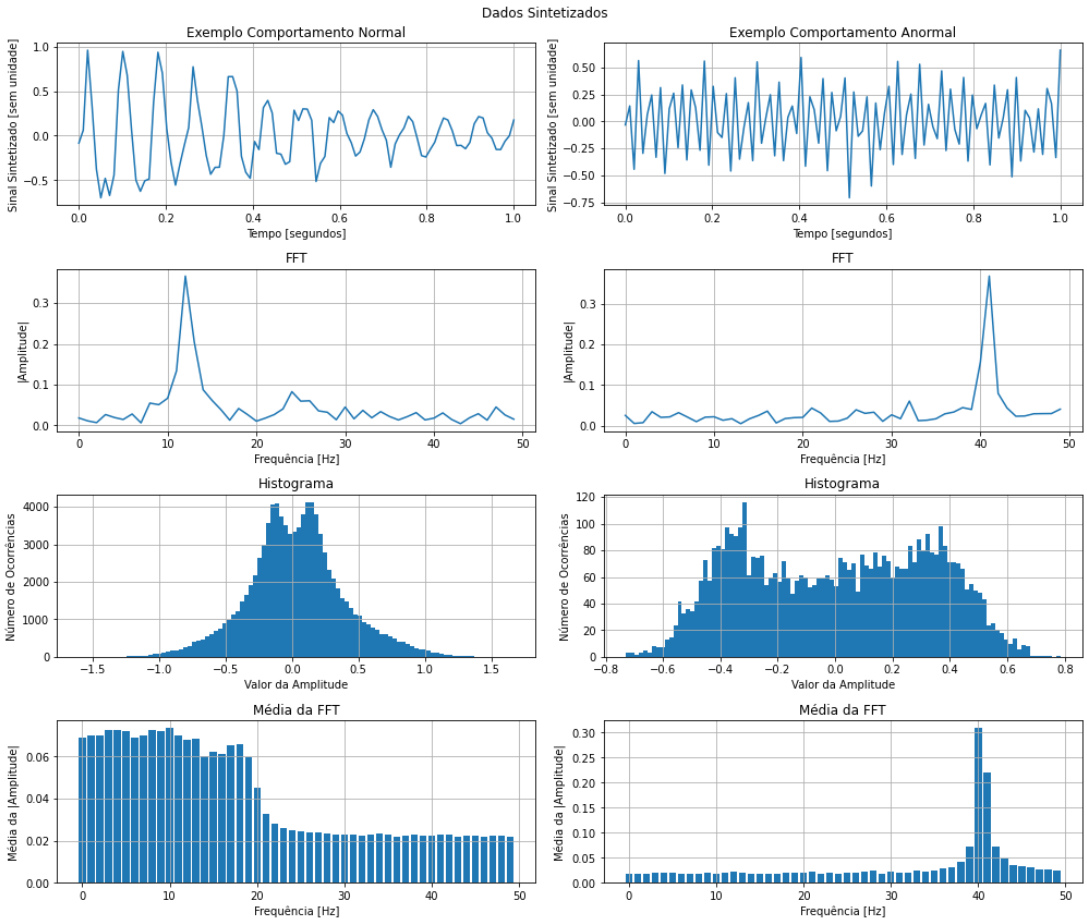

Figura 52 – Dados sintetizados para a verificação da capacidade do encoder

Fonte: Autor.

Observa-se as frequências médias com comportamento não anômalo espalhadas pela faixa definida.

O experimento descrito na seção 4.2.2 foi realizado, com o treinamento da TadGan sobre as janelas sintetizadas por 100 épocas. A dispersão das centroides e frequências fundamentais do sinal reconstruído em relação ao original podem ser vistos na Figuras 53.

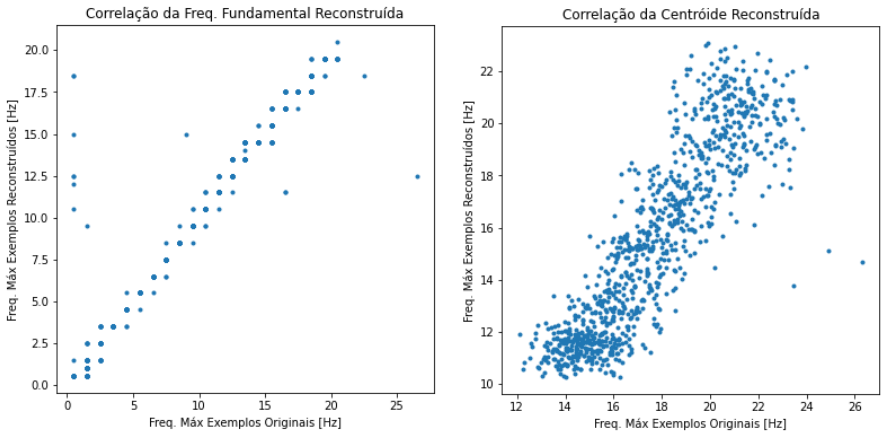

Figura 53 – Correlação entre a centroide dos exemplos originais e reconstruídos.

Fonte: Autor.

Observa-se grande correlação, com 0.971 para a frequência fundamental e 0.876 para a centroide, corroborando com a capacidade de reconstruir as janelas observada. Os resultados observados corroboram com a capacidade do *Encoder* de codificar na representação latente características complexas do processo alvo, como a frequência fundamental.

5.3 LIMITAÇÕES DO MÉTODO

5.3.1 Entropia do Sinal

A partir da implementação obtida para os processos escolhidos, os seis conjuntos de dados foram sintetizados para os valores definidos dos desvios padrões, e o resultado inspecionado. Na Figura 54 pode-se ver o conjunto gerado para o ruído aditivo nos exemplos positivos de $\sigma = 10$. Nas primeiras duas linhas vê-se exemplos de uma janela positiva e anômala, respectivamente, e em sequência o histograma sobre todas janelas e a média dos *bins* do FFTs. Os sinais obtidos corroboram com o objetivo desejado.

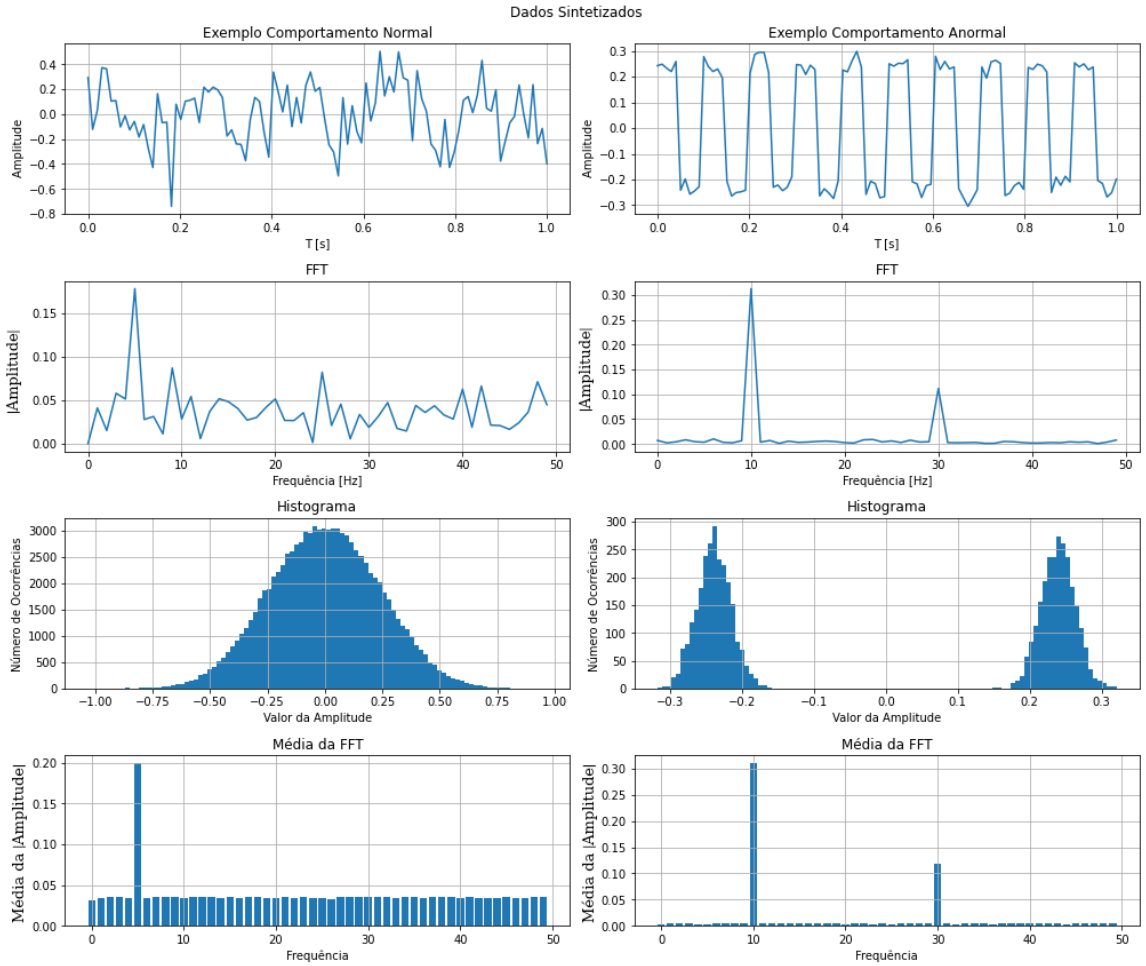

Figura 54 – Exemplos de Sinais sintetizados para $\sigma = 10$.

Fonte: Autor.

Em sequência, para cada um dos conjuntos sintetizados, a TadGan foi treinada com 300 épocas. Os valores obtidos para a média e dispersão do erro de reconstrução dos exemplos totais e anômalos pode ser visto na Figura 55. Observa-se uma clara redução na diferença entre as duas médias com o aumento de ruído, corroborando com a a dependência da utilidade do erro de reconstrução como métrica para com a entropia do sinal.

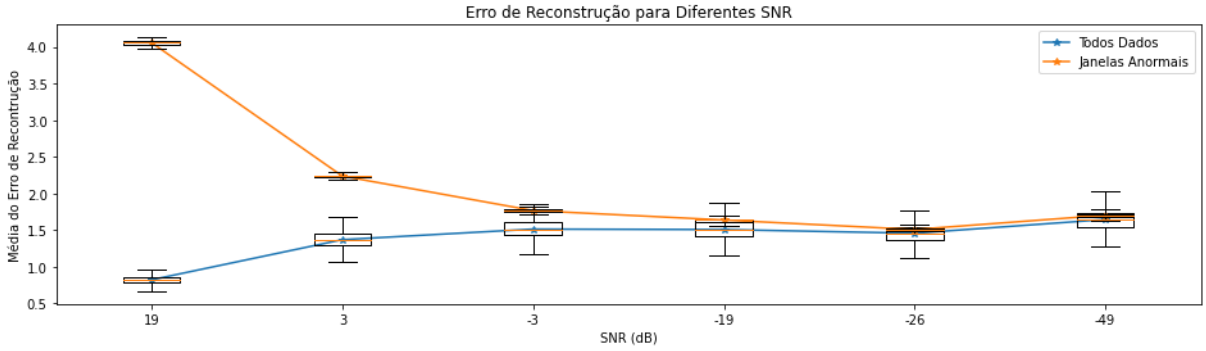

Figura 55 – Comparação do erro de reconstrução dos exemplos anômalos com o conjunto gerado.

Fonte: Autor.

O comportamento do escore do crítico pode ser visto na Figura 56.

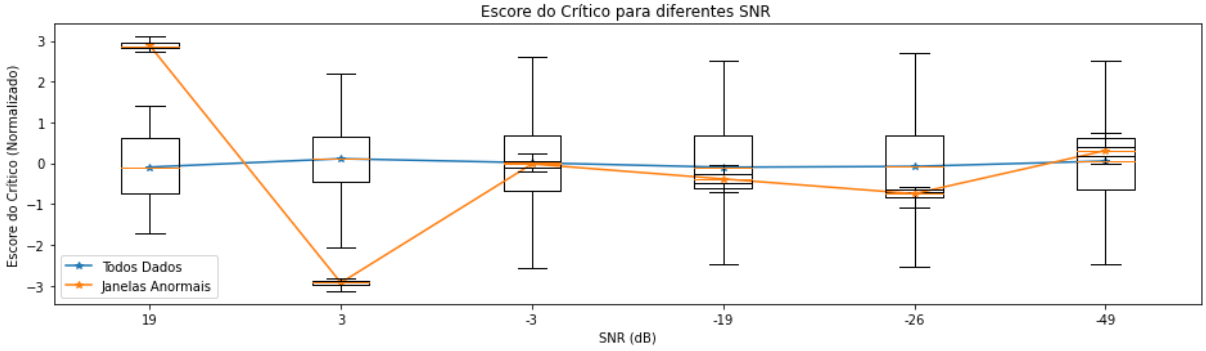

Figura 56 – Comparação do escore do Crítico dos exemplos anômalos com o conjunto gerado.

Fonte: Autor

Constata-se também a diminuição da diferença entre os exemplos não anômalos e anômalos com o escore do crítico. Dado o carácter *adversarial* do treinamento, é possível que impactos nos objetivos de treinamento do gerador tenham impactos no crítico em X.

Por fim, a dispersão das janelas no espaço de *features* resultante pode ser observada na Figura 57 para cada ensaio realizado, corroborando com as observações realizadas com os gráficos anteriores, do colapso das métricas na capacidade de diferenciação entre os dois processos.

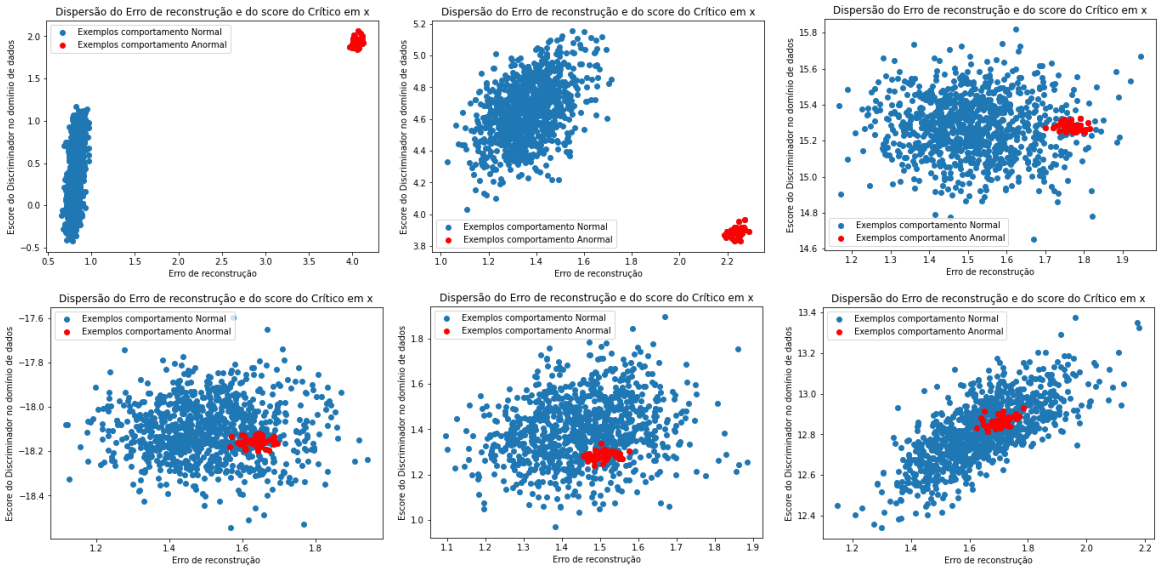

Figura 57 – Dispersão das Janelas no Espaço de *features*formado para diferentes níveis de SNR.

Fonte: Autor.

Em sequência o procedimento de normalização pela *fuzzy entropy* de cada janela foi efetuado. Pode-se ver nas Figuras 58 e 59 o erro de reconstrução corrigido e o espaço de *features* resultante.

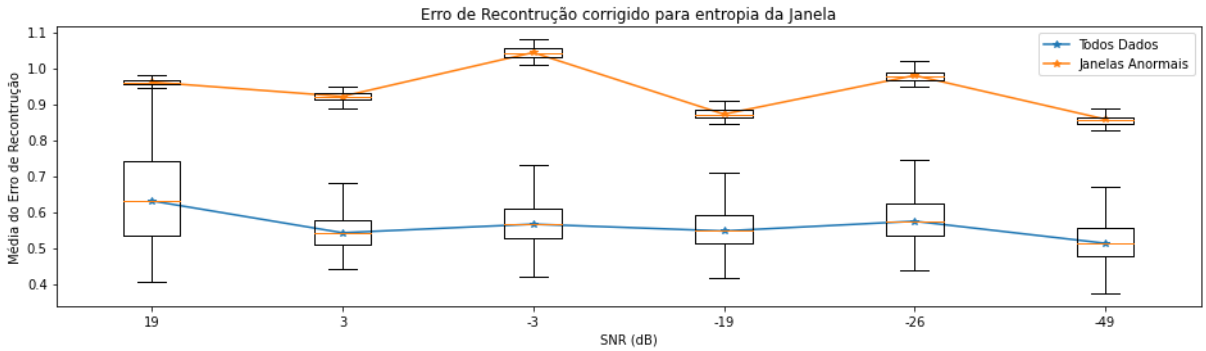

Figura 58 – Comparação do erro de reconstrução compensado para entropia das anomalias em relação à todo conjunto.

Fonte: Autor.

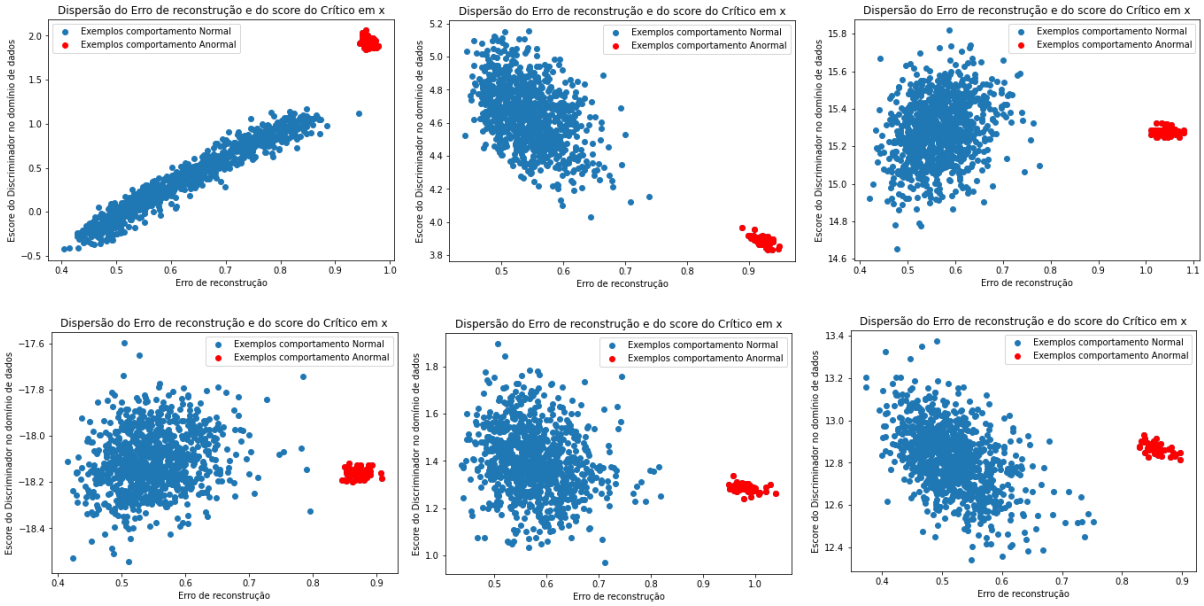

Figura 59 – Dispersão das Janelas no Espaço de *features* formado com erro de reconstrução compensado para entropia.

Fonte: Autor.

Vê-se corroborado que a relação $\frac{erro\,recontruo}{entropia}$ permanece aproximadamente constante, indicando a relação entre a entropia no sinal e o erro de reconstrução obtido. Esse fato também é reforçado pela observação da separabilidade do espaço, corroborando com a hipótese da relação do colapso da usabilidade do erro de reconstrução pelo aumento da entropia relativa das janelas.

Em geral, pôde-se desenvolver intuições da possibilidade de influência da entropia dos processos na utilização do erro de reconstrução como métrica da anormalidade. Estudos futuros são, entretanto, necessários para melhor delimitar esse problema, suas consequências, e cenários de ocorrência, e devem ser conduzidos sobre metodologias mais abrangentes e robustas.

5.3.2 Estabilidade do Treinamento

A seguir são apresentados os resultados para as duas abordagens de investigação de fonte de variabilidade da performance do método.

5.3.2.1 Quanto ao Número de Épocas

A TadGan foi treinado por 1000 épocas e os dados da distância entre as centroides e dispersão dos exemplos anormais e normais coletados de acordo com metodologia descrita na seção 4.3.2.1. A evolução da distância das centroides com o número de épocas de treinamento pode ser vista na Figura 60.

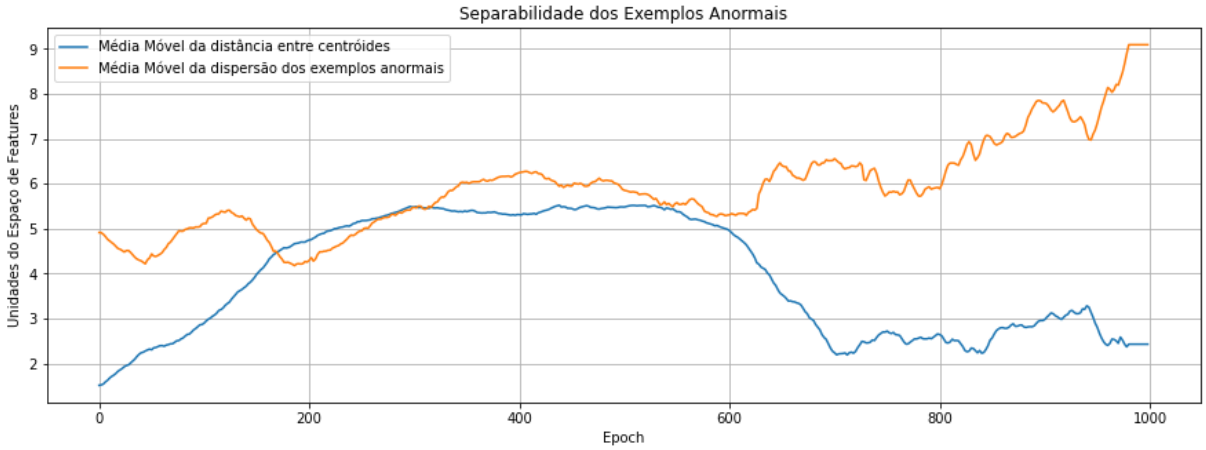

Figura 60 – Separabilidade dos exemplos anormais no espaço de *features* resultante.

Fonte: Autor.

Observa-se uma acentuada diminuição da separabilidade a partir da época 600. Uma possível causa seria a modelagem por parte do modelo gerador do comportamento anormal, o qual passaria a ter um erro de reconstrução próximo ao processo normal, como discutido anteriormente. Vê-se também o aumento da dispersão dos exemplos anormais, prejudicando ainda mais a capacidade de diferenciação entre os dois processos. Essa hipótese pode ser reforçada com a inspeção do comportamento do erro de reconstrução. O erro de reconstrução das janelas anormais durante o experimento pode ser vista na Figura 61.

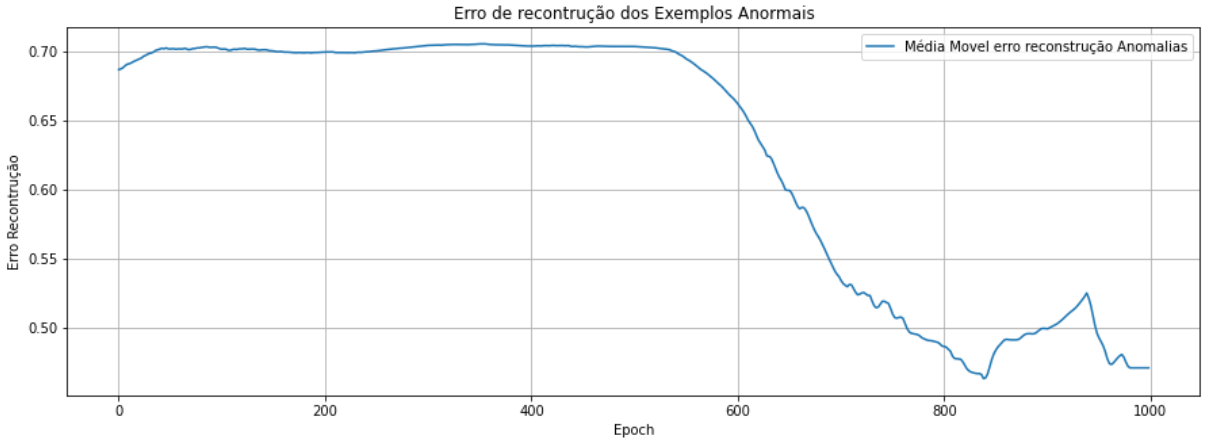

Figura 61 – Erro de reconstrução dos Exemplos Anômalos com o número de épocas.

Fonte: Autor.

Observa-se a diminuição do erro de reconstrução dos exemplos anormais, corroborando com a hipótese da modelagem do processo anormal pelo modelo. A possibilidade de modelagem do processo anormal deve ser levada em consideração na aplicação totalmente não supervisionada do problema, avaliando-se com cautela o número de iterações de treinamento. Sugere-se futuros experimentos para determinação da influência da proximidade

entre os dois processos alvos, bem como em situações em que os processos apresentam grande diferença de complexidade.

5.3.2.2 Quanto a variabilidade

O modelo foi treinado por 200 épocas sobre o mesmo sinal, e a performance avaliada com o F1 amostra-a-amostra e pelo protocolo de GEIGER *et al.,* (2020). Os resultados podem ser vistos nas Figuras 62, 63, 64, para os dois F1 resultantes, para o F1 amostra-a-amostra e seus componentes e pelo F1 (GEIGER *et al.,* 2020) e seus componentes, respectivamente.

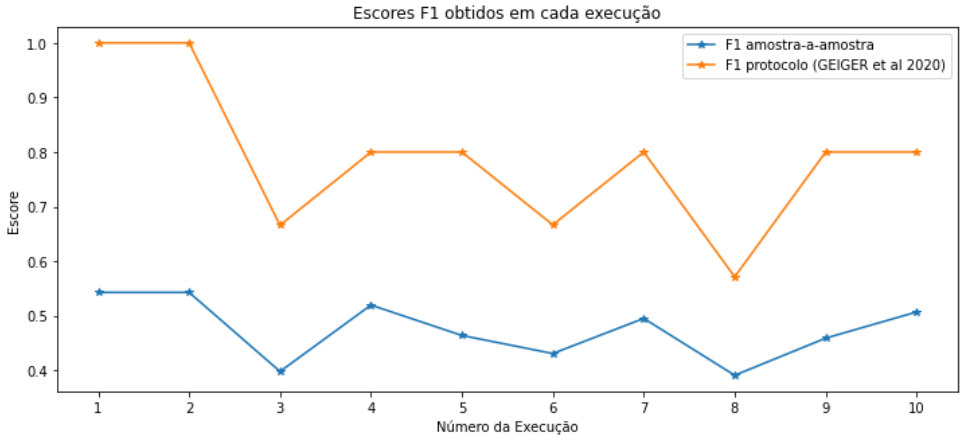

Figura 62 – Escores F1 obtidos em cada execução.

Fonte: Autor.

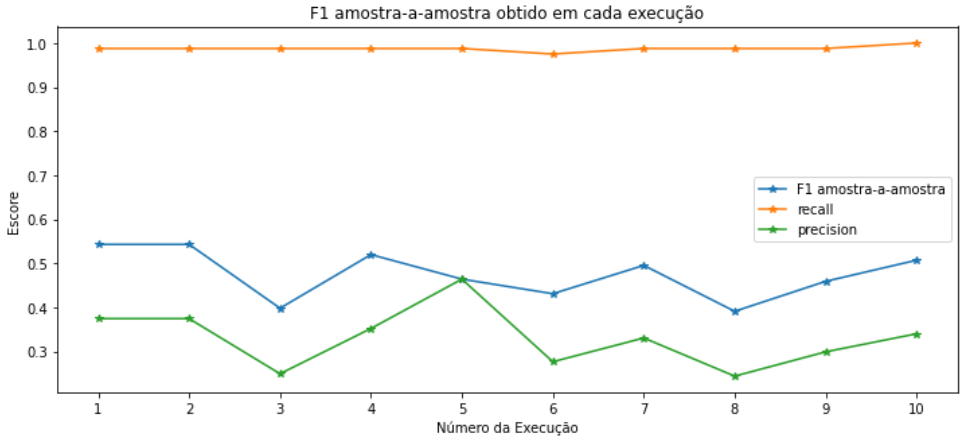

Figura 63 – Escore F1 amostra-a-amostra obtido em cada execução.

Fonte: Autor.

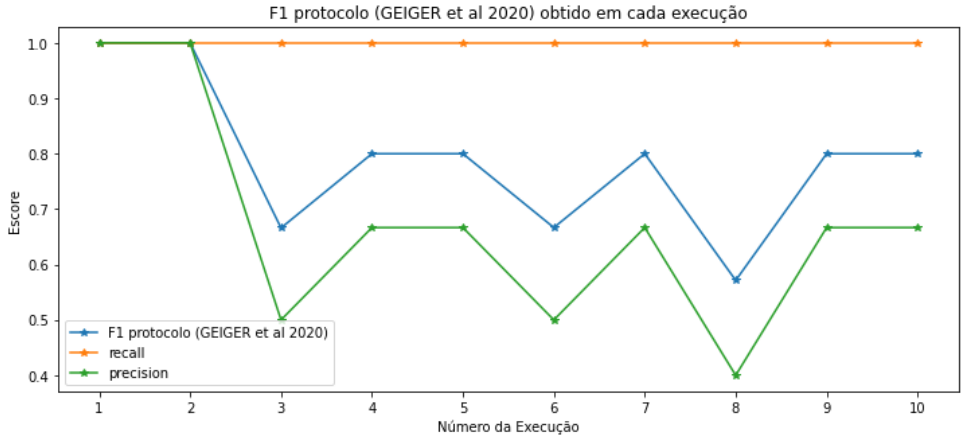

Figura 64 – Escore F1 pelo protocolo de (GEIGER *et al.*, 2020) obtido em cada execução.

Fonte: Autor

Apesar da metodologia pouco robusta empregada, constata-se a ocorrência de grande variância entre diferentes execuções. Sugere-se a investigação em trabalhos futuros do comportamento levantado, através de uma metodologia mais robusta, para entender e delimitar a variância de performance. Pode-se também verificar a dependência dessa variância para o número de épocas de treinamento, de modo a sugerir a existência de uma convergência final.

5.3.3 Considerações Gerais da Limitações

A partir do experimentos conduzidos, bem como da revisão bibliográfica, pôde-se elencar algumas limitações do método:

- estabilidade do treinamento: observadas variância em diferentes execuções, bem como dependência para com numero de interações;
- presunção da estacionaridade da normalidade: as premissas de detecção por janela e modelagem indiscriminadas das janelas não permitem a mudança dinâmica do processo definidor do comportamento normal e anômalo;
- colapso do erro de reconstrução como métrica de anormalidade: como investigado, o erro de reconstrução como métrica da anormalidade parece possuir dependência de características do sinal, como a entropia, e da capacidade do modelo;
- instabilidade do escore do Crítico como métrica de anormalidade: verificou-se grande variância nos escores fornecidos pelo crítico, como também reportado por (GEIGER *et al.* 2020);
- análise por janela: dada a modelagem do processo por janelas, as características e propriedades capturadas estão restritas nas que podem ser observadas dentro do número de amostras definido;

- elevado custo computacional.

5.4 AVALIAÇÃO DISCRIMINADA POR PROBLEMA

5.4.1 Anomalias Sintetizadas

As anomalias definidas foram implementadas e os sinais sintetizados. Os sinais gerador para cada tipo de anomalia podem ser vistos na Figura 65. Dado o carácter sintético dos sinais, nem os valores de amplitude nem de tempo possuem unidades. Nenhum das amostras anômalos gerados está à mais de 3 desvios padrões da média, $x_n > 3\sigma$, não possibilitando a detecção trivial por amplitude.

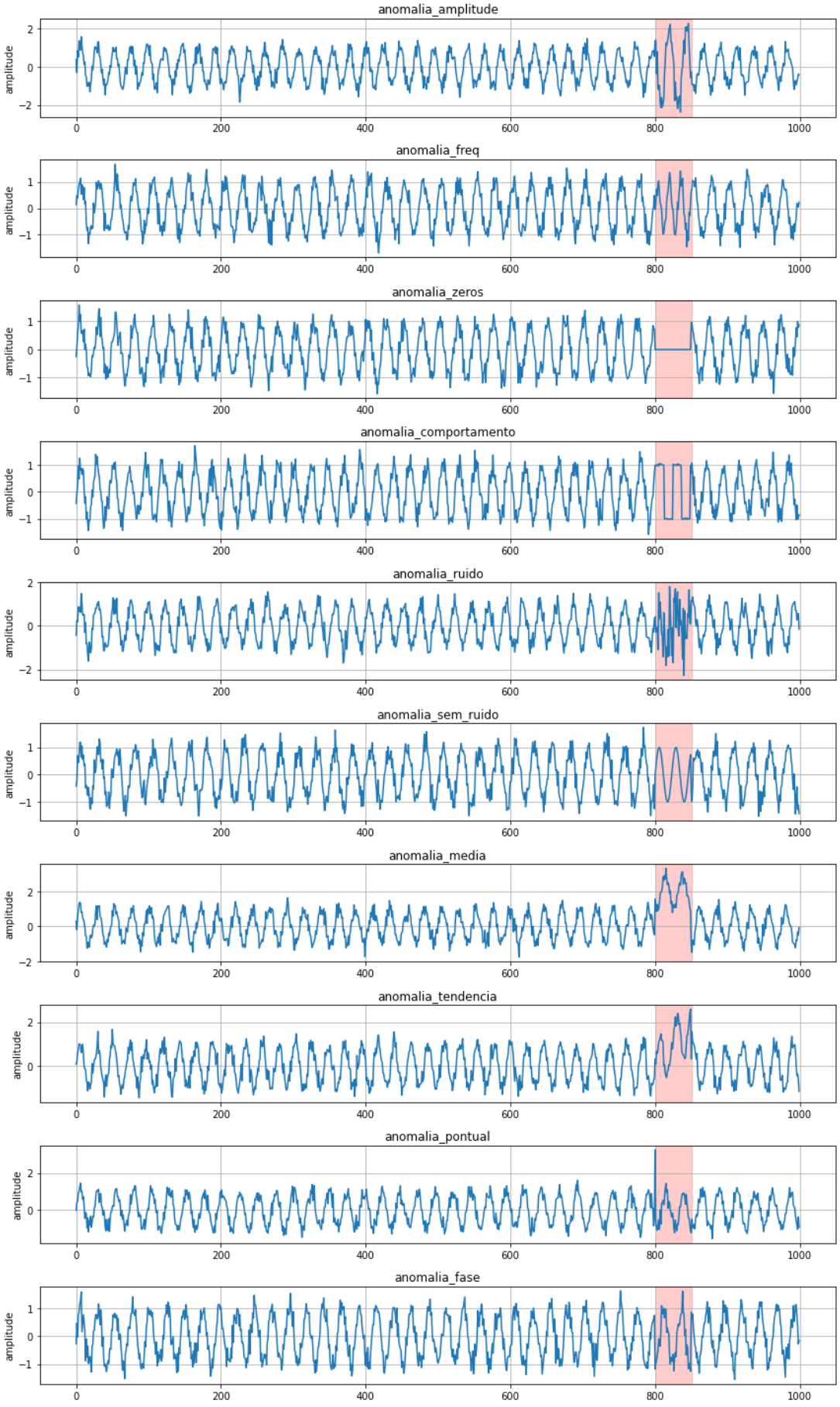

Figura 65 – Séries com Anomalias sintetizadas.

Fonte: Autor

O método foi aplicado ao dados gerados, e o erro de reconstrução e escore do crítico calculados para cada sinal. Na Figura 66 pode-se ver a representação gráfica das amostras anormais e de todas amostras para o erro de reconstrução. Observa-se a notável diferença entre as médias dos exemplos anômalos para com o resto do conjunto. Constata-se menor separação nas *anomalia_sem_entropia* e na *anomalia_fase*.

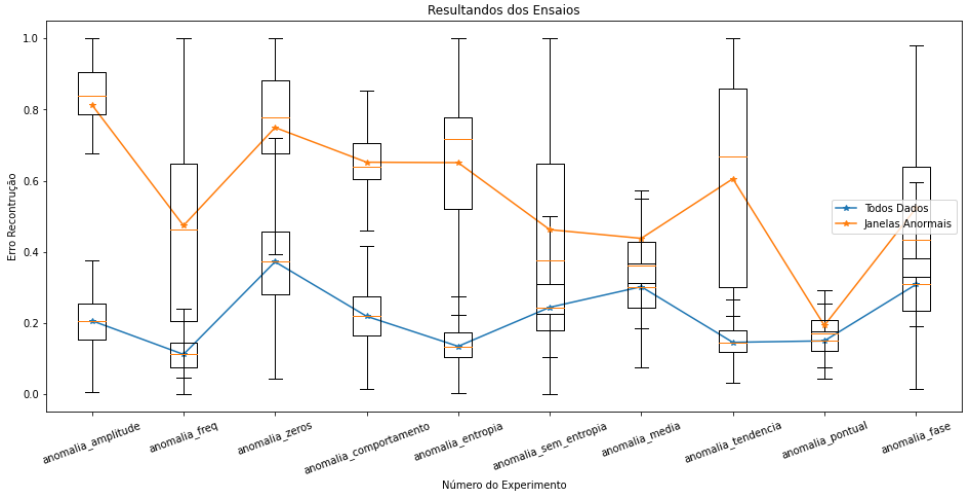

Figura 66 – Erro de reconstrução nas séries sintetizadas.

Fonte: Autor.

A mesma representação foi obtida para o escore do crítico, que pode ser visto na Figura 67. Vê-se que apesar da flutuação da posição relativa entre o sinal das amostras anômalas e não anômalas, a diferença entre as médias é mantida.

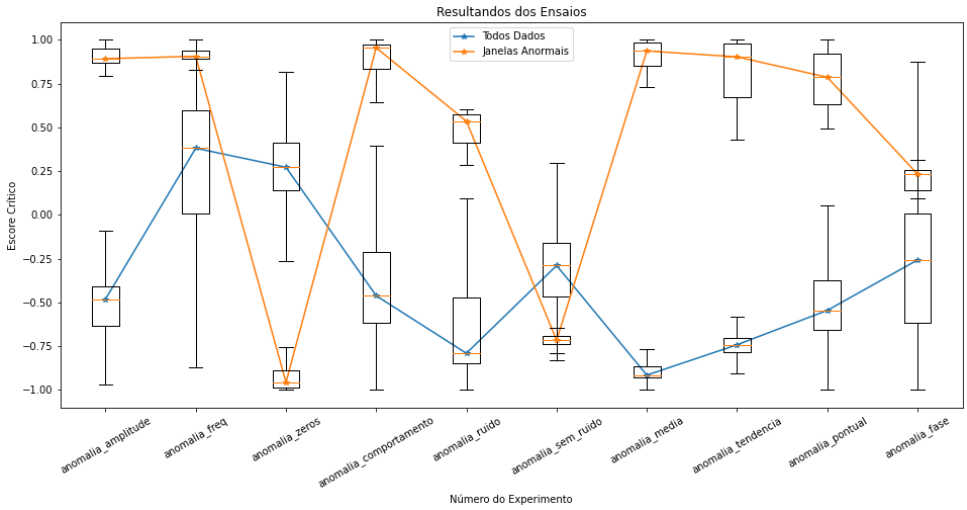

Figura 67 – Escore do Crítico sobre as séries sintetizadas.

Fonte: Autor.

Um exemplo do comportamento das duas métricas no tempo, bem como do escore combinado e o sinal reconstruído pode ser vista na Figura 68, onde observa-se a não

reconstrução efetiva dos pontos anormais.

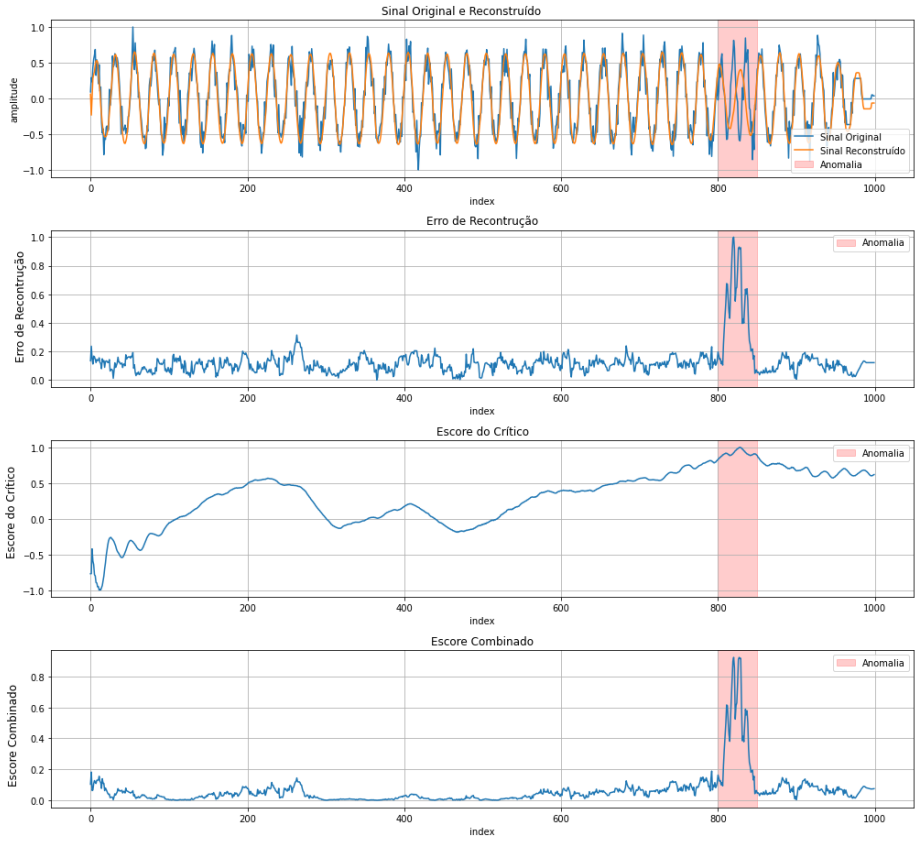

Figura 68 – Exemplo da detecção em um sinal sintetizado.

Fonte: Autor.

Com o objetivo de quantificar a capacidade de detecção de anomalias pela métrica resultante, aplicou-se à um detector ingênuo, considerando como anômalo simplesmente toda amostra tal que $x_n > 3\sigma$. Os resultados podem ser visto na Tabela 6.

Tabela 6 – Resultados obtidos com o detector ingênuo nos dados sintetizados.

	F1 amostra-a-amostra	**F1 GEIGER *et al.*, 2020**
anomalia_amplitude	0.95	1.00
anomalia_freq	0.667	1.00
anomalia_zeros	0.51	1.00
anomalia_comportamento	0.73	1.00
anomalia_ruido	0.76	1.00
anomalia_sem_ruido	0.45	1.00
anomalia_media	0.26	1.00
anomalia_tendencia	0.72	1.00
anomalia_pontual	0.1	1.00
anomalia_fase	0.40	1.00

Observa-se a capacidade do método de identificar os diferentes tipos de anomalias, corroborando com sua aplicabilidade em domínios diversos.

5.4.2 Diferentes Problemas

A TadGan foi treinada sobre os sinais selecionados por 200 épocas. Em sequência, o escore resultante foi utilizado pelo detector trivial descrito na secção anterior para a obtenção dos escores de F1. Os resultados são exibidos na Tabela 7.

Tabela 7 – Resultados obtidos com o detector ingênuo nos dados sintetizados.

	F1 GEIGER *et al.*, 2020	F1 amostra-a-amostra
DISTORTEDInternalBleeding	0.5	0.10
WalkingAceleration	1.0	0.14
Lab2Cmac011215EPG1	0.5	0.21
PowerDemand2	Não Detectado	Não Detectado
WalkingAceleration1	Não Detectado	Não Detectado
GP711MarkerLFM5z	1.0	0.5
CIMIS44AirTemperature	1.0	0.5

Observa-se a capacidade de detecção de anomalias nos diferentes domínios compilados corroborando com a aplicabilidade do método. Foi notável a influência do tamanho da janela na capacidade de efetivamente detectar as anomalias, refletindo-se na necessidade de realizar *downsampling* de algumas séries de modo que coubesse em 100 amostras informações relevantes do processo.

Na Figura 69 replica-se a representação com gráficos para cada sinal.

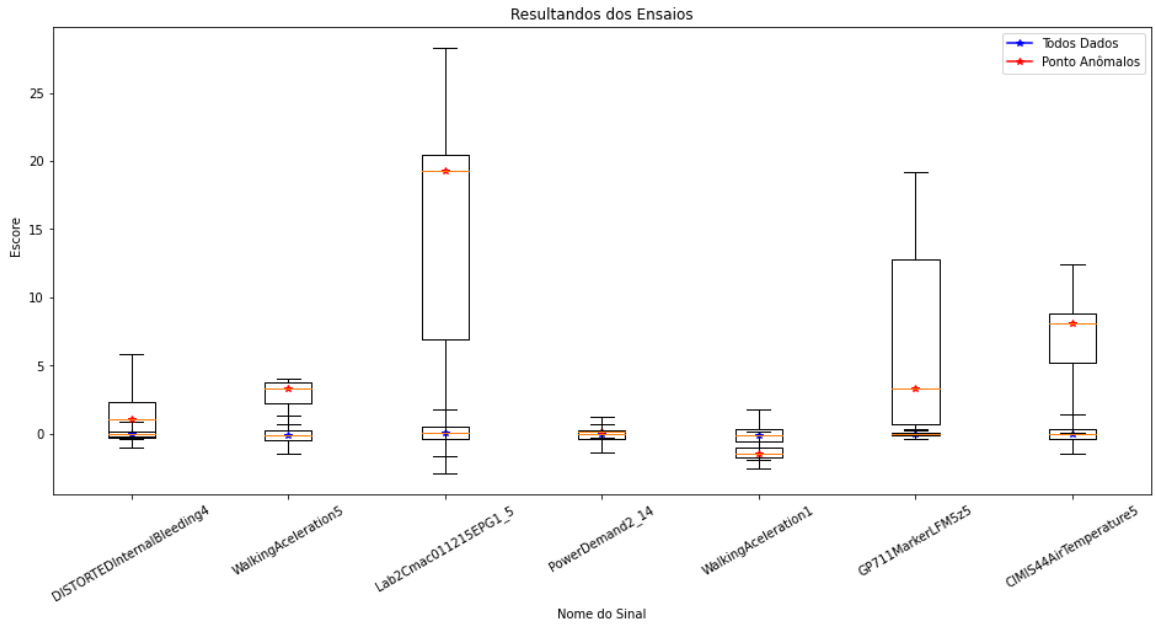

Figura 69 – Dispersão das amostras com o escore de anormalidade para cada sinal.

Fonte: Autor.

5.5 PROPOSTA DE MODIFICAÇÕES

As métricas alternativas propostas foram implementadas e inicialmente validadas sobre os sinais com anomalias sintéticos gerados na seção 4.2.1.1. Um exemplo do comportamento das métricas sugeridas para um dos sinais sintéticos gerados pode se vista na Figura 70. Observa-se a sua capacidade de quantificar o comportamento anômalo.

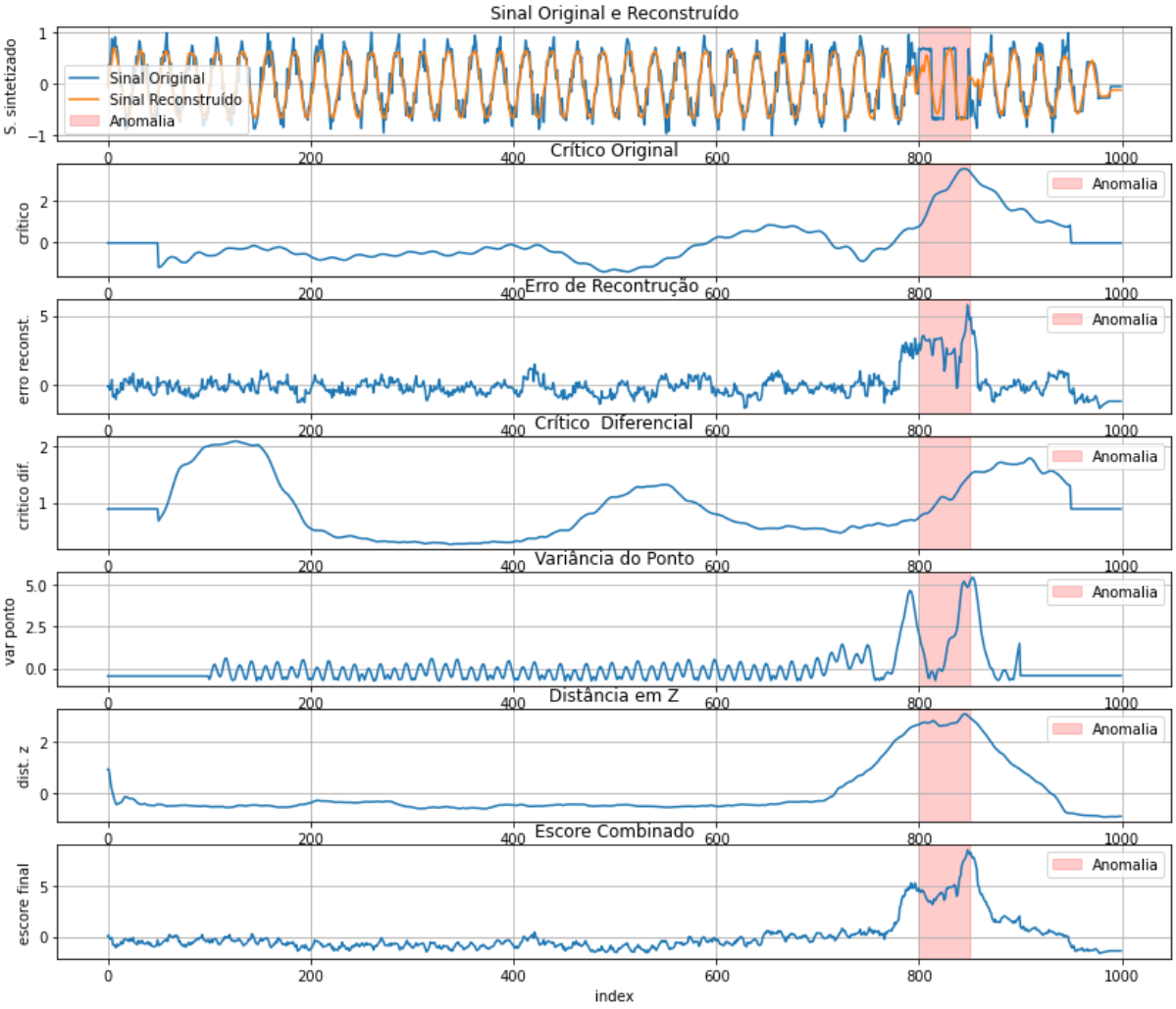

Figura 70 – Exemplo do comportamento das novas métricas sugeridas.

Fonte: Autor.

Em sequência, as novas métricas foram aplicadas à coleção de sinais descritas em 4.4.2 a fim de corroborar com sua aplicabilidade em aplicações reais. Os resultados podem ser vistos compilados na Tabela 8, bem como a representação gráfica na Figura 71.

Tabela 8 – Resultados obtidos com o detector ingênuo nos dados sintetizados.

	F1 GEIGER *et al.*, 2020	F1 amostra-a-amostra
DISTORTEDInternalBleeding	1.0	0.489
WalkingAceleration	1.0	0.37
Lab2Cmac011215EPG1	0.4	0.10
PowerDemand2	Não Detectado	Não Detectado
WalkingAceleration1	Não Detectado	Não Detectado
GP711MarkerLFM5z	1.0	0.6363
CIMIS44AirTemperature	1.0	0.25

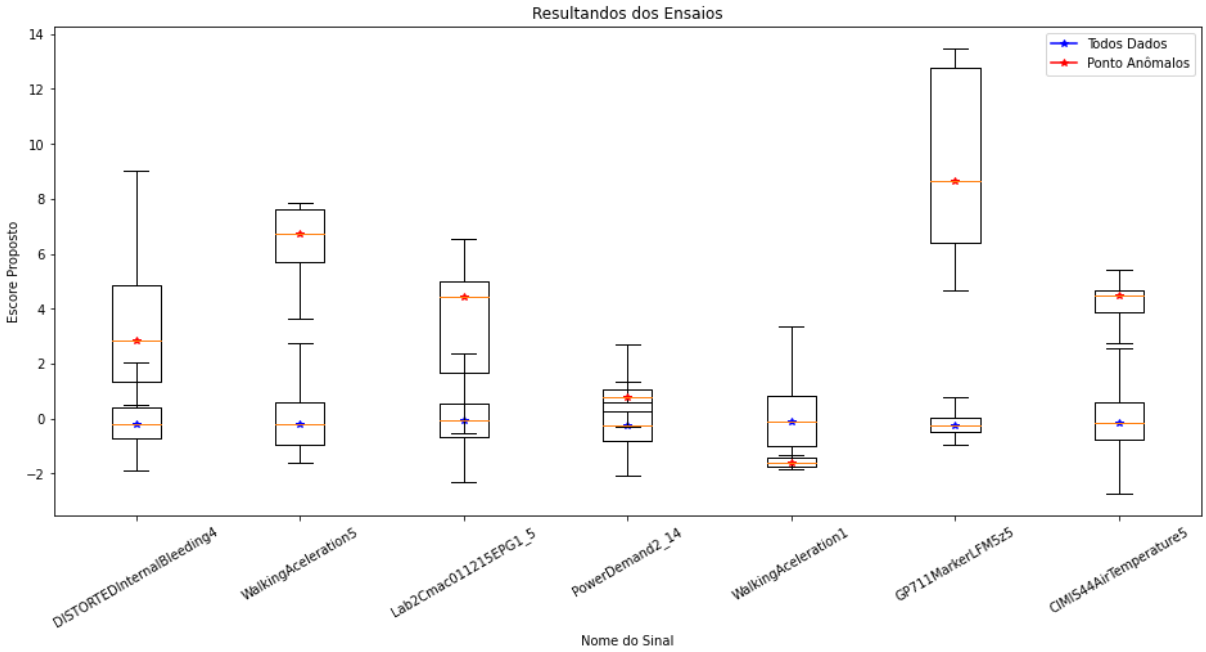

Figura 71 – Dispersão das amostras com o novo escore de anormalidade obtido para cada sinal.

Fonte: Autor.

Constata-se a capacidade de detecção de anomalias do procedimento alternativo apresentado. Apesar da performance marginalmente superior à aplicação original do método, uma metodologia de avaliação mais robusta é necessária para conclusões definitivas. Sugere-se a execução de trabalhos futuros nessa direção, explorando outras métricas alternativas que possam mitigar as limitações do método, bem como formas aprimoradas de combiná-las para formação de um escore único. A proposta dessas métricas alternativas teve como objetivo apenas apontar direções possíveis para futuros desenvolvimentos.

6 CONCLUSÕES

O trabalho desenvolvido teve como objetivo realizar um estudo da aplicação de GANs para detecção de anomalias de forma não supervisionada em séries temporais. A literatura foi revisada para obtenção de fundamentos quanto ao funcionamento das redes GANs, bem como para os princípios de detecção de anomalias e suas particularidades na aplicação das séries temporais. Em sequência, os principais trabalhos envolvendo soluções baseadas em GANs foram compilados e apresentados. Uma topologia, proposta por (GEIGER *et al.,* 2020), foi selecionada para estudo mais detalhado, pela sua ilustração didática dos princípios fundamentais da aplicação de GANs à séries temporais e bom resultados reportados em conjuntos de avaliação.

Observou-se as sutilezas da delimitação do escopo do problema de detecção de anomalias, e suas consequências negativas no desenvolvimento do campo e em especial de soluções agnósticas à aplicação. Conjuntos de dados compilados são frequentemente alvos de críticas da literatura especializada, em parte pela ausência de taxonomias objetivas que permitam categorizar os tipos de anomalias e sub áreas de aplicação. Questões levantadas por (WU *et al.*, 2021), da trivialidade de muitos banco de dados compilados, são originadas em parte das definições imprecisas do conceito de anomalia e da falta de critérios objetivos para análise dos tipos de anomalias no contexto de séries temporais. Conclui-se que o desenvolvimento do campo teria grande benefício de uma metodologia unificada que permitisse discriminar a performance dos métodos propostos em diferentes tipos de anomalias e subproblemas da área.

A partir da revisão realizada, uma implementação do método foi obtida e verificada sobre conjuntos de dados compilados, reproduzindo-se resultados parciais do trabalho original de (GEIGER *et al.*, 2020). Em sequência, com o objetivo de verificar os princípios fundamentais do método, conjuntos de dados sintéticos foram gerados a partir de funções analíticas. Avaliou-se a capacidade do Gerador de modelar processos no tempo e de gerar sinais pertencentes à esse processo, do *Encoder* de encontrar representações para os processos com a extração de características relevantes do sinal original, do sistema formado pelo *Encoder*-Gerador de efetivamente reconstruir sinais do processo modelado, e da relevância dos espaço de *features* formado pelas métricas derivadas do método em diferenciar diferentes processos, e consequentemente, detectar anomalias. Os resultados obtidos para os dois conjuntos de dados sintetizados, de senoides atenuadas e processo ARMA, corroboraram com os princípios do método, e possibilitaram a verificação das suas capacidades para diferenciar diferentes processos no tempo. Observou-se a ocorrência de um erro de reconstrução maior para o processo contaminante no conjunto de dados, assim como diferentes escores do crítico para o processo positivo e anormal. A observação da dispersão das janelas nesse espaço de *features* criado sugere a possibilidade de separabi-

lidade, e portanto sua relevância para detecção de anomalias. Foi também observada a capacidade do *Encoder* de extrair e encontrar representações relevantes para o processo modelado, em especial com a frequência fundamental das janelas geradas, corroborando com a aplicabilidade do método para a modelagem de processos com variância no comportamento. Além disso, verificou-se a utilidade da experimentação sobre o método com sinais analiticamente compreendidos, uma vez que conclusões sobre as capacidades e limitações puderam ser feitas de forma mais confiável, sem a influência de fatores imprevisíveis muitas vezes presentes em sinais experimentais.

Baseado nas observações realizadas nos experimentos, bem como de reportadas na literatura, um conjunto de possíveis limitações para o método foram compiladas, e experimentos propostos para sua investigação. Investigou-se a estabilidade do treinamento do método, da perspectiva da escolha do número de interações e da variabilidade entre diferentes execuções. Foi verificada a possibilidade de variação de performance para diferentes iterações de treinamento, situação na qual levantou-se e corroborou-se a hipótese da ocorrência de modelagem pelo método do comportamento anômalo em conjuntos de dados contaminados para número de interações muito altos. Esse comportamento apresenta acentuada relevância em aplicações onde o comportamento positivo apresenta alta complexidade, que demandaria um alto número de iterações de treino, e ao mesmo tempo o conjunto de dados possui contaminação significativa com anomalias. Observou-se também a ocorrência de variância de performance entre diferentes execuções, atribuídas em parte pela instabilidade do treinamento de GANs e pelas características estocásticas gerais do métodos de *Deep Learning*. Os experimentos conduzidos sugeriram ainda que a maior fonte de variância aparenta ser o surgimento de falsos positivos, enquanto a detecção de determinados comportamentos anômalos permanece mais constantes entre todas as execuções.Verificou-se a necessidade de investigações futuras mais criteriosas para a elucidação dessas questões levantadas.

Investigou-se também a possibilidade da influência de propriedades dos processos na utilização da métrica do erro de reconstrução para medição da anormalidade. Observou-se que para dois processos arbitrários definidos, a capacidade de diferenciar esses processos através a métrica de erro de reconstrução parece diminuir com o aumento de entropia do processo positivo. Esse fato foi atribuído à influência da entropia no erro de reconstrução mínimo teórico, visto que entropia e capacidade de reconstrução estão intimamente relacionados. Conclui-se a possibilidade da relação, e mais uma vez da necessidade de estudos mais aprofundados de um ponto de vista formal e experimental para delimitação real do problema, suas consequências e contornos, visto a posição fundamental ocupada por essa métrica na detecção de anomalias com GANs.

Em sequência, séries temporais com anomalias de tipos específicos foram sintetizadas, e o método aplicado para a detecção, de modo a demonstrar-se o potencial e aplicabilidade na detecção de diferentes tipos de anomalias, bem como identificar situação limitantes.

Observou-se, para os sinais simples sintetizados, grande capacidade das métricas derivadas em detectar as amostras anômalas geradas, demostrando a generalidade da aplicação, e a sua independência para com o tipo e a natureza da anomalia presente. Séries temporais de aplicações diversas não presentes na avaliação realizada por (GEIGER *et al* 2020) também foram compiladas, do *UCR time series archive*, e o método aplicado, de modo a analisar sua performance em aplicações distintas e em dados coletados com metodologias diferentes. Verificou-se novamente a capacidade do método de detectar anomalias na grande maioria dos sinais, corroborando ainda mais com a sua aplicabilidade em domínio diversos

Por fim, com base nas observações e experimentos realizados, métricas alternativas para quantificação da anormalidade foram sugeridas, implementadas e superficialmente validadas nas séries sintetizadas e compilado de séries experimentais. Verificou-se sua potencialidade na detecção de anomalias, e seu caráter complementar em determinadas situações às métricas originais do método.

Como conclusões gerais, através da realização do trabalho, pôde-se explorar as aplicações existentes baseadas em GANs para detecção de anomalias, e através da experimentação com um método específico selecionado, investigar suas propriedades e possíveis limitações. Verificou-se a utilidade de experimentação sobre sinais sintetizados para formulação e teste de hipóteses, bem como para o desenvolvimento de demonstrações didáticas do funcionamento do método. Em geral, constatou-se o grande potencial de aplicabilidade de GANs à detecção de anomalias em séries temporais, dado pela sua capacidade de geração de *features* e representações dos processos alvos relevantes à detecção de anomalias de forma agnóstica ao processo de interesse e sem a necessidade de engenharia de *features* e especialista no processo. Constatou-se sua capacidade de detectar diversos tipos de anomalias distintas em aplicações variadas. Observou-se também a grande complexidade do método, que abre um grande espaço para investigação e explorações futuras decorrentes dos inúmeros processos, parâmetros e procedimentos adotados.

6.1 PROPOSTA PARA TRABALHOS FUTUROS

Dada a grande aplicabilidade do campo, assim como da complexidade do método, a gama de trabalhos futuros na aplicação de GANs para detecção de anomalias em séries temporais é vasta. Trabalhos na direção de formalizar o problema geral de detecção de anomalias em séries temporais, e de disponibilizar metodologias de avaliação discriminada para tipos de anomalias e subproblemas trariam grandes benefícios ao desenvolvimento do campo como um todo, aumentando a confiabilidade geral dos métodos desenvolvidos e das metodologias de avaliação empregadas. Além disso, as limitações levantadas devem ser investigadas mais rigorosamente, de modo a serem delimitadas e, por consequência, propostas de melhorias elaboradas. As consequências da influência de características

do processo na utilização do erro de reconstrução, como a entropia, apresenta grande interesse, visto ser a principal abordagem baseada em GANs para detecção de anomalias. A exploração da utilização de métricas alternativas, bem como de novas maneiras de combinar as métricas existentes também apresentou possibilidades de progresso, e devem ser explorados mais rigorosamente.

Trabalhos voltados para a diminuição do custo computacional geral do método também trariam grande benefício para futuras melhorias, permitindo a experimentação mais extensiva das suas propriedades em aplicações distintas.

REFERÊNCIAS

AGGARWAL, C. C. Outlier Analysis, Springer, Segunda Edição 2017;

AHMED, F.; COURVILLE, A. Detecting semantic anomalies, 2019, Université de Montréal.

AKCAY, S.; ABARGHOUEI, A. A.; BRECKON, T. P. GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training. abs/1805.06725, 2018.

ARJOVSKY, M.; CHINTALA, S.; BOTTOU, L. , "Wasserstein generative adversarial networks," International Conference on Machine Learning, pp. 214–223, 2017.

ARORA, S.; RISTESKI, A.; ZHANG, Y. Theoretical Limitations of Encoder-Decoder GAN architectures, Princeton University, 2017

AYED, F.; STELLA, L.; JANUSCHOWSKI, T.; GASTHAUS, J. Anomaly Detection at Scale: The Case for Deep Distributional Time Series Models, Amazon Research 2020

BASHAR, M. A.; NAYAK, R. "TAnoGAN: Time series anomaly detection with generative adversarial networks," in Proc. IEEE Symp. Comput. Intell. (SSCI), Canberra, ACT, Australia, Dec. 2020, pp. 1778–1785, doi: 10.1109/SSCI47803.2020.9308512.

BATTIKH, M.; LENSKIY, A. Latent-Insensitive Autoencoders for Anomaly Detection, arXiv 2021.

BLÁZQUEZ-GARCIA, A.; CONDE, A.; MORI, U.; LOZANO, J. A.; A Review on Outlier/Anomaly Detection in Time Series Data, ACM Computing Surveys, Vol. 54, No. 3, 2021;

BOUGUESSA, M.; CHOUCHANE, A. A Statistical Framework for Handling Network Anomalies. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly Detection: A Survey, University of Minnesota, 2017

CHAUAN, S.; VIG, L. "Anomaly detection in ECG time signals via deep long short-term memory networks,", IEEE International Conference on Data Science and Advanced Analytics, 2015, pp. 1–7.

CHEN, X.; DUAN, Y.; HOUTHOOFT, R.; SCHULMAN, J.; SUTSKEVER, I.; ABBEEL, P. “Infogan: Interpretable representation learning by information maximizing generative adversarial nets,” in Neural Information Processing Systems, pp. 2172–2180, 2016.

CHEN, Y. C.; GENOVESE, C. R.; WASSERMAN, L. Density Level Sets: Asymptotics, Inference, and Visualization, 2016;

DI MATTIA, F.; GALEONE, P.; DE SIMONI, M.; GHELFI, E. A Survey on GANs for Anomaly Detection, Zuru Tech, Modena, Italy, 2021

DONHAHUE, J.; KRAHENBUHL, P.; DARRELL, T. “Adversarial feature learning,” arXiv preprint arXiv:1605.09782, 2016.

DU, B.; SUN, X.; YE, J.; CHENG, K.; WANG, J.; SUN, L. "GAN-Based Anomaly Detection for Multivariate Time Series Using Polluted Training Set,"in IEEE Transactions on Knowledge and Data Engineering, doi: 10.1109/TKDE.2021.3128667.

ESTEBAN, C.; HYLAND, S. L.; RATSCH, G. “Real-valued (medical) time series generation with recurrent conditional gans,” arXiv preprint arXiv:1706.02633, 2017.

FROGNER, C.; ZHANG, C.; MOBAHI, H.; ARAYA-POLO, M.; POGGIO, T. Learning with a Wasserstein Loss, arXiv, 2015

GASPERIN, M.; BOSKOSKI, P.; JURICIC, D. Prognosis of gear health using stochastic dynamical models with online parameter estimation. Annual Conference of the PHM Society, 1(1). Retrieved from http://www.papers.phmsociety.org/index.php/phmconf/article/view/1634 (2021)

GEIGER, A.; LIU, D.; ALNEHEIMISH, S.; CUESTA-INFANTE, A.; VEERAMACHANENI, K. TadGAN: Time Series Anomaly Detection Using Generative Adversarial Networks, IEEE, 2020

GHERBI, E.; HANCZAR, B.; JANODET, J. C.; KLAUDEL, W. Construction d’espace latent pour la détection d’anomalies par apprentissage adversarial. Conférence sur l’Apprentissage automatique (CAP 2019), Jul 2019, Toulouse, France. ffhal-02473865f

GOLDSTEIN, M.; UCHIDA, S. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. PloS one. 11. e0152173. 10.1371/journal.pone.0152173.2016

GOODFELLOW, I.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative Adversarial Nets. pp. 2672–2680, 2014.

GOODFELLOW, I. NIPS 2016 Tutorial: Generative Adversarial Networks, 2017;

GULRAJANI *et al.*; Improved Training of Wasserstein GANs, arXiv, 2017;

GUI, J.; SUN, Z.; WEN, Y.; TAO, D.; YE, J. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications, 2020;

HAWKINS, D. M. Identification of outliers. London: Chapman and Hall, 1980.

HEUSEL, M.; RAMSAUER, H.; UNTERTHINER, T.; NESSLER, B.; HOCHREITER, S. "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in Neural Information Processing Systems, pp. 6626–6637, 2017.

HUNDMAN, K.; CONSTANTINOU, V.; LAPORTE, C.; COLWELL, I.; SODERSTROM, T. "Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding," in Proc. of the 24th ACM SIGKDD, 2018.

ISOLA, P.; ZHU, J. Y.; ZHOU, T.; EFROS, A. A. Image-to-Image Translation with Conditional Adversarial Networks, Berkeley AI Research (BAIR) Laboratory, UC Berkeley, 2018

JIANG, W.; HONG, Y.; ZHOU, B.; HE, X.; CHENG, C. A GAN-Based Anomaly Detection Approach for Imbalanced Industrial Time Series

KHOSHNEVISAN, F.; FAN, Z.; CARVALHO, V. R., Improving Robustness on Seasonality-Heavy Multivariate Time Series Anomaly Detection. MileTS San Diego, California, USA, 2020,

KIM, S.; CHOI, K.; CHOI, H.; LEE, B.; YOON, S. Towards a Rigorous Evaluation of Time-series Anomaly Detection, arXiv 2021.

LAI, K. H.; ZHA, D.; XU, J.; ZHAO, Y.; WANG, G.; e HU, X. Revisiting Time Series Outlier Detection: DefinitionsCand Benchmarks.2021

LEE, C. K.; CHEON, Y.; HWANG, W. Y. Studies on the GAN-Based Anomaly Detection Methods for the Time Series Data 2021, IEEE Access

LI, D.; CHEN, D.; JIN, B.; SHI, L.; GOH, J.; Ng, S. K. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks, in Proc. Int. Conf. Artif. Neural Netw. Cham, Switzerland:Springer, Sep. 2019, pp. 703–716.

LI, Y.; PENG, X.; ZHANG, J.; LI, Z.; WEN, M. "DCT-GAN: Dilated Convolutional Transformer-based GAN for Time Series Anomaly Detection,"IEEE Transactions on Knowledge and Data Engineering, doi: 10.1109/TKDE.2021.3130234, 2021.

LIAO, H. J.; LIN,C. H. R.; Lin,Y. C.; TUNG, K.Y. “Intrusion detection system: A comprehensive review,” Journal of Network and Computer Applications, vol. 36, no. 1, pp. 16–24, 2013

MEHOTRA, K. G.; MOHAN, C. K.; HUANG, H. Anomaly Detection Principles and Algorithms, Springer, 2017

NIU, Z.; YU, K.; WU, X. LSTM-Based VAE-GAN for Time-Series Anomaly Detection. Sensors. 2020; 20(13):3738. https://doi.org/10.3390/s20133738

OUYANG, X.; ZHANG, X.; MA, D.; AGAM, G. Generating Image Sequence from Description with LSTM Conditional GAN, Illinois Institute of Technology, 2018

RABATEL, J.; BRINGAY, S.; PONCELET, P. Anomaly Detection in Monitoring Sensor Data for Preventive Maintenance. Expert Systems with Applications, Elsevier, 2011, 38 (6), pp.7003-7015. 10.1016/j.eswa.2010.12.014 lirmm-00670917

RADFORD, A.; METZ, L. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, 2016

RUFF, L.; KAUFFMANN, J. R.; VANDERMEULEN, R. A.; MONTAVON, G.; SAMEK, W.; KLOFT, M.; DIETTERICH, T.; MULLER, K. Unifying Review of Deep and Shallow Anomaly Detection, IEEE, 2021;

SALIMANS, T.; GOODFELLOW, I.; ZAREMBA, W.; CHEUNG, V.; RADFORD, A.; CHEN, X. “Improved techniques for training gans,” in Neural Information Processing Systems, pp. 2234–2242, 2016.

SCHLEGL, T.; SEEBOCK, P.; WALDSTEIN, S. M.; SCHMIDT-ERFURTH, U.; LANGS, G. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. abs/1703.05921, 2017.

SONG, D.; XIA, N.; CHENG, W.; CHEN, H.; TAO, D. Deep r-th root of rank supervised joint binary embedding for multivariate time series retrieval. In Proceedings of the 24th ACM SIGKDD International Conference on Knowl-edge Discovery Data Mining, 2229–2238. ACM. 2018

SUN, Y.; YU, W.; CHEN, Y.; KADAM, A. Time Series Anomaly Detection Based on GAN Power-train Vehicle Research and Development Isuzu Technical Center of America Plymouth, USA, 2019

WANG, Z.; BOVIK, A. C.; SHEIKH, H. R.; SIMONCELLI, E. P. *et al.* “Image quality assessment: from error visibility to structural similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

WEITING, CHEN *et al.* Characterization of surface EMG signal based on fuzzy entropy, IEEE Transactions on neural systems and rehabilitation engineering, 15.2 (2007): 266-272.

WENIG, P.; SCHMIDL, S.; PAPENBROCK, T. TimeEval: A Benchmarking Toolkit for Time Series Anomaly Detection Algorithms. PVLDB, 15(12): 3678 - 3681, 2022. doi:10.14778/3554821.3554873

WU, R.; KEOGH, E. Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress, Institute of Electrical and Electronics Engineers (IEEE), 2021

XU, L.; ZHENG, L.; Li, W. *et al* NVAE-GAN Based Approach for Unsupervised Time Series Anomaly Detection,arXiv, 2021

YANG, J.; ZHOU, K.; LI, Y.; LIU, Z. Generalized Out-of-Distribution Detection: A Survey, 2021

YIN, C.; ZHANG, S.; WANG, J.; XIONG, N. N. "Anomaly Detection Based on Convolutional Recurrent Autoencoder for IoT Time Series,"in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 52, no. 1, pp. 112-122, Jan. 2022, doi: 10.1109/TSMC.2020.2968516.

ZENATI, H.; FOO, C. S.; LECOUAT, B.; MANEK, G.; CHANDRASEKHAR, V. R. Efficient GAN-Based Anomaly Detection. abs/1802.06222, 2018.

ZHANG, C. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In The Thirty-Third AAAI Conference on Artificial Intelligence, 1409–1416. 2019

ZHU, J. Y.; PARK, T.; ISOLA, P.; Efros, A. A. , “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks,” in IEEE Int. Conf. on Computer Vision (ICCV), oct 2017, pp. 2242–2251.