Rede Generativa Adversarial Quântica Semi-Supervisionada (sQGAN) para Detecção de Ataques

Diego Abreu[1], **David Moura** [2], **Christian Rothenberg**[2], **Antônio Abelém**[1]

[1] Universidade Federal do Pará - UFPA

[2]Universidade Estadual de Campinas - Unicamp

Abstract. *The evolution of cybersecurity threats demands efficient and accurate attack detection systems, yet the scarcity of labeled data limits the use of conventional supervised models. This paper proposes a Semi-Supervised Quantum Generative Adversarial Network (sQGAN) for attack detection, combining semi-supervised learning with quantum adversarial architectures to leverage labeled and unlabeled data for improved detection in data-scarce scenarios. Key contributions include (1) a semi-supervised quantum architecture effective with limited labeled data, (2) integration of quantum-based generator and discriminator networks to enhance attack detection, and (3) an experimental study comparing sQGAN's performance with quantum architectures. Results indicate that sQGAN offers a high F1 score and robustness in detecting attacks under challenging labeling conditions.*

Resumo. *A evolução das ameaças cibernéticas exige sistemas de detecção de ataques eficientes e precisos, mas a escassez de dados rotulados limita o uso de modelos supervisionados convencionais. Este artigo propõe a Rede Generativa Adversarial Quântica Semi-Supervisionada (sQGAN) para detecção de ataques, que combina aprendizado semi-supervisionado com arquiteturas adversárias quânticas, aproveitando dados rotulados e não rotulados para melhorar a detecção em cenários de dados escassos. As principais contribuições incluem (1) uma arquitetura quântica semi-supervisionada eficaz com poucos dados rotulados, (2) integração de redes geradoras e discriminativas quânticas para aprimorar a detecção de ataques e (3) um estudo experimental comparando o desempenho da sQGAN com arquiteturas quânticas. Os resultados mostram que a sQGAN oferece F1 score signficativo e robustez para detecção de ataques em condições adversas de rotulagem.*

1. Introdução

A segurança cibernética tem se tornado uma prioridade essencial no atual cenário digital, impulsionada pela crescente sofisticação e frequência de ataques cibernéticos. Um dos principais desafios na detecção de ameaças é a escassez de dados rotulados para treinar modelos supervisionados [Lim et al. 2024]. Grande parte dos dados disponíveis não é rotulada, o que pode limitar a aplicabilidade de técnicas tradicionais de Aprendizado Supervisionado e comprometer a precisão na identificação de atividades maliciosas [Idhammad et al. 2018]. Assim, torna-se fundamental desenvolver soluções capazes de lidar com essa limitação, proporcionando modelos eficientes mesmo em cenários de baixa rotulagem de dados.

As Redes Generativas Adversariais Semi-Supervisionadas (*Semi Supervised Generative Adversarial Networks* - sGANs) [Odena 2016] surgem como uma alternativa promissora para esse problema, ao combinar a geração de dados sintéticos com Aprendizado Semi-Supervisionado. Essa combinação permite que os modelos aprendam a partir de uma quantidade limitada de rótulos, aproveitando informações contidas em dados não rotulados para melhorar a capacidade de generalização [Yang et al. 2022]. No entanto, a arquitetura tradicional das GANs pode não ser suficiente para capturar a complexidade de padrões anômalos em dados de segurança cibernética [Mvula et al. 2023].

Com o avanço da Computação Quântica, novos métodos têm explorado a capacidade dos sistemas quânticos de representar e processar informações de forma diferenciada, possibilitando a aplicação em problemas complexos como a detecção de ataques [Abreu et al. 2024a]. Redes Adversariais Quânticas (*Quantum Adversarial Networks* - QGANs) trazem uma nova perspectiva à análise de dados, oferecendo potencial para uma representação mais diversificada dos padrões nos dados [Nicesio et al. 2023]. Baseando-se nessa premissa, este trabalho propõe o sQGAN, uma Rede Generativa Adversarial Quântica Semi-Supervisionada, voltada para detecção de ataques cibernéticos. A proposta combina Aprendizado Semi-Supervisionado com modelos generativos e discriminativos quânticos, explorando os benefícios de ambas as abordagens.

Dentre as contribuições deste trabalho, destacam-se três pontos principais: (1) a proposição de uma arquitetura quântica semi-supervisionada que possibilita aprendizado eficaz mesmo em cenários com dados rotulados limitados, (2) a utilização de redes geradoras e discriminativas quânticas para aprimorar a identificação de ataques, e (3) um estudo experimental que avalia a eficácia da proposta em diferentes arquiteturas quânticas. O artigo está estruturado da seguinte forma: a Seção 2 apresenta a fundamentação teórica, com um panorama das Redes Generativas Adversariais, incluindo modelos semi-supervisionados e quânticos. A Seção 3 discute os trabalhos relacionados. A Seção 4 detalha a arquitetura da sQGAN e seus componentes. A Seção 5 descreve o estudo de caso, enquanto a Seção 6 apresenta e analisa os resultados experimentais. Por fim, a Seção 7 conclui o trabalho, destacando as contribuições gerais e propondo direções para pesquisas futuras.

2. Fundamentação Teórica

Nesta seção, são apresentados os fundamentos teóricos que embasam a proposta. Discutem-se os conceitos relacionados às Redes Generativas Adversariais e sua variação semi-supervisionada. Além disso, são explorados tópicos fundamentais sobre Aprendizado de Máquina Quântico, bem como as adaptações das GANs no contexto quântico.

2.1. Redes Generativas Adversariais

As Redes Generativas Adversariais são modelos de Aprendizado Profundo que utilizam uma estrutura composta por duas redes neurais, uma geradora e uma discriminativa, em um sistema de competição [de Araujo-Filho et al. 2023]. A rede geradora, G, recebe como entrada um vetor de ruído aleatório e é treinada para produzir dados sintéticos que se assemelham aos dados reais. Em contraste, a rede discriminativa, D, é treinada para distinguir entre os dados reais e os dados gerados, classificando-os como reais ou falsos. O objetivo de G é maximizar as chances de D classificar suas amostras geradas como

reais, enquanto D busca minimizar os erros de classificação. Esse processo configura um jogo *minimax* entre G e D, no qual ambas as redes se aprimoram mutuamente ao longo do treinamento. O treinamento de uma GAN é formulado como uma função de perda de *minimax* na Equação 1.

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{dados}}} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log (1 - D(G(z)))] \tag{1}$$

Na Equação 1, x representa as amostras reais, z é uma variável de ruído amostrada a partir de uma distribuição pré-definida $p(z)$, e $G(z)$ gera dados sintéticos. Durante o treinamento, D é atualizado para maximizar sua capacidade de distinguir entre amostras reais e geradas, enquanto G é atualizado para minimizar sua capacidade de ser detectado por D.

As GANs tradicionais são modelos puramente não supervisionados, o que pode limitar sua aplicabilidade em tarefas onde a presença de dados rotulados pode melhorar a performance do modelo. De outro modo, as GANs Semi-Supervisionadas (sGANs) utilizam uma quantidade limitada de dados rotulados para orientar o treinamento, ampliando significativamente o escopo de aplicação das GANs [Sajun and Zualkernan 2022].

2.2. GANs Semi-Supervisionadas

A arquitetura de GANs semi-supervisionadas estende o modelo padrão de GANs ao adicionar uma camada de classificação no discriminador para lidar com dados rotulados e não rotulados simultaneamente. No modelo semi-supervisionado, o discriminador D é projetado para classificar as amostras em uma das $N + 1$ classes, onde N representa as classes reais dos dados rotulados e a classe adicional representa amostras geradas, ou "falsas".

Formalmente, dado um conjunto de dados x que pertence a uma das classes reais $y \in \{1, \ldots, N\}$ ou à classe gerada, o objetivo do discriminador D é maximizar a probabilidade $p(y|x)$, onde $y = N + 1$ indica que a amostra é gerada pelo gerador G. Assim, o discriminador combina um classificador multiclasses para amostras reais e uma classificação binária (real/falso) para distinguir dados gerados de dados reais. Para otimizar o treinamento, a função de perda do discriminador é dividida em duas partes: a perda binária não supervisionada ($L_{\text{binária}}$) e a perda multiclasse supervisionada ($L_{\text{multiclasse}}$). A perda binária é utilizada para diferenciar entre amostras reais e falsas, como apresenta a Equação 2, onde p_{dados} representa a distribuição dos dados reais e G a do gerador.

$$L_{\text{binária}} = -\mathbb{E}_{x \sim p_{\text{dados}}} \log D(y = \text{real} \mid x) - \mathbb{E}_{x \sim G} \log D(y = \text{falso} \mid x) \tag{2}$$

A perda multiclasse supervisionada é utilizada para classificar corretamente as amostras reais, como apresentada na Equação 3, onde $D(y \mid x)$ é a probabilidade de x pertencer a uma das N classes reais.

$$L_{\text{multiclasse}} = -\mathbb{E}_{(x,y) \sim p_{\text{dados}}} \log D(y \mid x) \tag{3}$$

De outro modo, o gerador G é treinado para maximizar a probabilidade de o discriminador classificar as amostras geradas como reais, o que é expresso na Equação 4, onde z é uma amostra de ruído latente usada pelo gerador para criar uma nova amostra $G(z)$. Isso

permite que o discriminador aprenda tanto a distinguir entre dados reais e falsos quanto a classificar dados reais nas classes corretas, aproveitando tanto os dados rotulados quanto os não rotulados para melhorar a eficácia em tarefas semi-supervisionadas.

$$L_G = -\mathbb{E}_{z\sim p(z)} \log D(y = \text{real} \mid G(z)) \tag{4}$$

2.3. Aprendizado de Máquina Quântico

O Aprendizado de Máquina Quântico busca interligar as áreas de Inteligência Artificial e Computação Quântica [Biamonte et al. 2017]. Esse campo emergente explora como algoritmos quânticos podem ser aplicados a problemas típicos de Aprendizado de Máquina, oferecendo, em potencial, vantagens em eficiência e desempenho.

Nesse contexto, as técnicas híbridas têm se destacado, pois combinam a capacidade de processamento quântico com a computação clássica. Essa abordagem se deve, em parte, às limitações dos computadores quânticos atuais, conhecidos como dispositivos NISQ (*Noisy Intermediate-Scale Quantum*). Os dispositivos NISQ possuem restrições relacionadas ao número de qubits (bits quânticos que utilizam o espaço de Hilbert), ao número de operações realizáveis (como portas lógicas quânticas) e ao tempo em que as propriedades quânticas, como superposição e entrelaçamento, podem ser mantidas. Além disso, há limitações na profundidade dos circuitos quânticos, ou seja, no número de operações consecutivas que podem ser realizadas antes que o ruído comprometa os resultados. Uma discussão mais aprofundada sobre essas limitações pode ser encontrada em [Wang and Liu 2024].

As técnicas híbridas utilizam circuitos parametrizados (circuitos variacionais), que são compostos por portas quânticas cujos parâmetros são ajustáveis. Durante o treinamento, esses parâmetros são otimizados utilizando algoritmos clássicos de Aprendizado de Máquina, como métodos baseados em gradiente. O processo funciona da seguinte maneira: a entrada, que pode consistir em dados clássicos ou quânticos, é processada por um circuito quântico parametrizado que aplica operações quânticas (portas lógicas) nos qubits. Em seguida, o estado final do circuito é medido, e as medições resultam em saídas clássicas. Essas saídas são então avaliadas por meio de uma função de perda, a qual serve como critério para ajuste dos parâmetros do circuito. A otimização é conduzida iterativamente por um algoritmo clássico, buscando que o modelo alcance um desempenho aprimorado ao longo do treinamento [Abreu et al. 2024b].

2.4. Redes Adversariais Quânticas

As Redes Generativas Adversariais Quânticas são uma extensão das GANs tradicionais que integram circuitos quânticos para aprimorar o desempenho em tarefas de geração e discriminação de dados [Boyle and Nikandish 2024]. A arquitetura híbrida quântico-clássica, como ilustrado na Figura 1, utiliza um gerador e um discriminador quânticos, onde o gerador G recebe um ruído e o transforma em estados quânticos representativos da distribuição de dados, enquanto o discriminador D distingue amostras geradas de amostras reais, executando a tarefa de classificação de real/falso.

O gerador quântico G é composto por uma Rede Neural Quântica que utiliza circuitos quânticos variacionais para gerar amostras sintéticas. Esse processo envolve dois operadores unitários: $U_0(z_i)$, que mapeia o vetor de ruído z para o espaço de Hilbert do

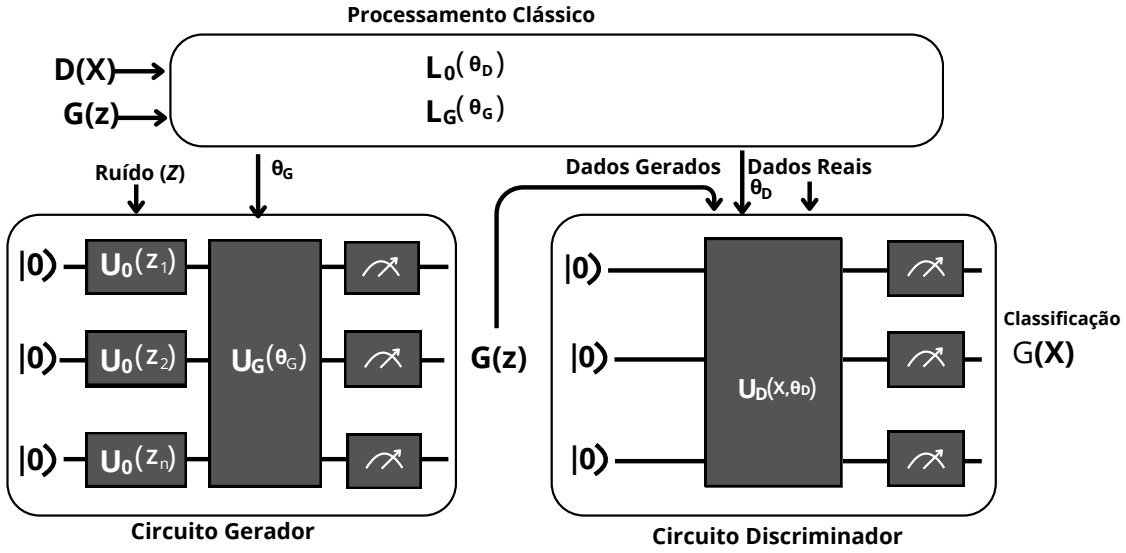


Figura 1. Rede Adversária Híbrida, com gerador e discriminador quântico e otimizador clássico.

sistema quântico, e $U_G(\theta_G)$, que é controlado por parâmetros treináveis θ_G. Esses operadores, combinados com medições, produzem amostras geradas $G(z)$ que são enviadas ao discriminador para avaliação.

O discriminador quântico D também é uma rede neural quântica parametrizada, composta pelo operador $U_D(x, \theta_D)$, onde x é a entrada (amostra real ou gerada) e θ_D são os parâmetros treináveis. O discriminador mede o estado quântico resultante e atribui uma classificação de real ou falso com base na expectativa de um observável, que é escalado para fornecer uma saída entre 0 e 1.

Para treinar essa rede adversarial, o processo de otimização clássico realiza a atualização dos parâmetros θ_G e θ_D através de uma função de perda, como na Equação 2, minimizada por gradiente descendente. Essa perda é reduzida conforme o discriminador melhora sua capacidade de distinguir entre amostras reais e geradas. Já o gerador é treinado para enganar o discriminador, maximizando a probabilidade de $D(G(z))$ classificar suas amostras como reais, como na Equação 4. Essa arquitetura híbrida quântico-clássica permite explorar a representação dos circuitos quânticos para capturar padrões complexos nos dados. A modularidade da abordagem permite ajustar a profundidade dos circuitos e a complexidade do modelo para otimizar o desempenho em computadores quânticos NISQ.

3. Trabalhos Relacionados

Vários trabalhos recentes têm explorado o uso de Aprendizado de Máquina para segurança cibernética, com ênfase na detecção de anomalias e intrusões em redes complexas. Entre essas abordagens, destacam-se as GAN aplicadas para identificar intrusões através da geração e detecção de padrões anômalos em dados de rede [de Araujo-Filho et al. 2023].

No contexto do Aprendizado de Máquina Quântico, Huang et al. [Huang et al. 2021] demonstra experimentalmente o uso de GANs quânticas para

geração de imagens em dispositivos NISQ. A proposta utiliza a estratégia de quantum *patch*, que permite a geração de partes menores de dados (*patches*) através de múltiplos sub-geradores, em contraste com um único gerador (*batch*). Isso viabiliza a modelagem de distribuições de alta dimensionalidade, mesmo com recursos quânticos limitados. Além disso, os autores exploram o uso de circuitos parametrizados e técnicas de medição para transformar os dados latentes em distribuições de probabilidades, mostrando a eficácia dessa abordagem em tarefas de aprendizado generativo. Assim, neste trabalho, as abordagens de *patch* e *batch* são utilizadas na arquitetura geradora e integradas no sQGAN.

Além disso, Boyle e Nikandish [Boyle and Nikandish 2024] apresentam um estudo de arquiteturas híbridas quântico-clássica de GANs projetadas especificamente para dispositivos NISQ. A proposta combina geradores e discriminadores baseados em circuitos variacionais parametrizados, destacando uma abordagem modular que equilibra a profundidade dos circuitos com a precisão do modelo. Em particular, os autores exploram diferentes arquiteturas de discriminadores para tarefas de classificação, utilizando múltiplas camadas de circuitos treináveis para distinguir entre dados reais e gerados, o que se mostrou eficiente na detecção de padrões complexos. No trabalho proposto neste artigo, as arquiteturas propostas por Boyle e Nikandish são integradas em uma GAN semi-supervisionada.

Nakaji e Yamamoto [Nakaji and Yamamoto 2021] introduzem uma rede adversária quântica semi-supervisionada com um gerador quântico e um discriminador clássico, destacando a robustez contra ruídos e a alta expressividade dos sistemas quânticos. Os autores demonstram que o gerador quântico, devido à sua alta expressividade, pode atuar como um adversário mais forte do que os geradores clássicos, contribuindo para um discriminador que obtém alta precisão de classificação. Demonstrando o potencial das GANs semi-supervisionada. Em contraste, a proposta apresentada neste artigo integra geradores e discriminadores quânticos, atribuindo a computação clássica apenas a otimização dos circuitos variacionais.

Portanto, a proposta deste trabalho distingue-se ao combinar aspectos complementares dessas abordagens. Foi empregada uma arquitetura semi-supervisionada, inspirada no modelo de GAN proposto por Odena [Odena 2016], aliada à estratégia de *patch*, conforme apresentada por Huang et al. [Huang et al. 2021]. Além disso, foram utilizadas duas das arquiteturas de circuitos discriminadores exploradas por Boyle e Nikandish [Boyle and Nikandish 2024] para realizar a detecção e classificação de ataques. Essa combinação possibilita a geração de padrões de tráfego de rede e a identificação de ataques em cenários com dados rotulados e não rotulados.

4. sQGAN: Rede Geneartiva Adversarial Quântica Semi Supervisionada

Nesta seção, é apresentada a sQGAN, uma Rede Generativa Adversarial Quântica Semi-Supervisionada, que combina os circuitos quânticos com uma arquitetura semi-supervisionada para a detecção de ataques. A sQGAN é uma GAN quântica híbrida, em que o treinamento e a otimização dos parâmetros são realizados em um computador clássico. Essa abordagem possibilita o uso de técnicas quânticas para a geração e classificação de dados, enquanto se aproveita a métodos clássicos de otimização para atualizar os parâmetros da rede.

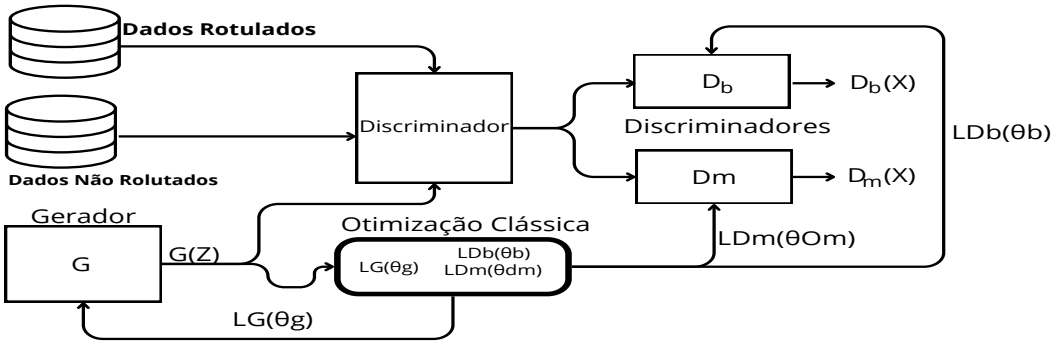


Figura 2. sQGAN- Arquitetura Semi Supervisionada.

4.1. Arquitetura do sQGAN

A sQGAN é estruturada em uma arquitetura semi-supervisionada, conforme ilustrado na Figura 2, o que permite que o modelo aprenda tanto com dados rotulados quanto com dados não rotulados. No gerador quântico, um ruído gaussiano é mapeado para o espaço de Hilbert, criando amostras sintéticas que se assemelham aos dados reais. O discriminador quântico é projetado para realizar duas tarefas de classificação: uma binária e outra multiclasse. O discriminador binário distingue entre amostras reais e geradas (falsas), representando o tráfego normal e o tráfego de ataque, respectivamente.

Conforme ilustrado na Figura 2, o modelo é composto por um gerador (*G*) e dois discriminadores (*Db* e *Dm*). O gerador utiliza o vetor de ruído *Z* para criar amostras sintéticas que são enviadas aos discriminadores. O discriminador binário (*Db*) realiza a tarefa de distinguir entre amostras reais e geradas, enquanto o discriminador multiclasses (*Dm*) classifica as amostras reais em categorias específicas, como tráfego normal e diferentes tipos de ataques.

A função de perda do modelo é composta por três componentes principais. A primeira, *LG*, otimiza o gerador *G* para melhorar a qualidade das amostras geradas. A segunda, *LDb*, treina o discriminador binário *Db* para diferenciar entre amostras reais e geradas. Por fim, a terceira, *LDm*, treina o discriminador multiclasses *Dm* para categorizar corretamente as amostras reais.

O processo de otimização é conduzido em um computador clássico, que é responsável por ajustar os parâmetros θ_G, θ_{Db} e θ_{Dm}. Esses parâmetros são utilizados para atualizar os circuitos variacionais do gerador e dos discriminadores, que são executados em um computador quântico. Durante o treinamento, o computador clássico avalia as funções de perda e otimiza iterativamente os θ utilizando métodos baseados em gradiente. Esses valores são então enviados ao computador quântico, onde os circuitos variacionais são atualizados e utilizados para gerar ou discriminar amostras de dados.

Essa estrutura modular permite que o modelo aprenda uma representação robusta das características dos dados, utilizando tanto os dados rotulados quanto os não rotulados para ajustar os seus parâmetros. A abordagem híbrida quântico-clássica, como destacado na Figura 2, possibilita que o treinamento e a otimização sejam conduzidos em uma infraestrutura clássica, enquanto os circuitos quânticos realizam a geração e classificação dos dados com alto grau de precisão.

4.2. Arquitetura Generativa

Na abordagem proposta, o gerador quântico segue a estratégia de *Patch* [Huang et al. 2021], em que a geração de amostras ocorre através de uma estrutura de sub-geradores G_i. Cada sub-gerador G_i é uma rede parametrizada que transforma o ruído de entrada z em uma parte específica das características do dado final. Essa abordagem, é particularmente vantajosa no contexto quântico, pois permite o uso eficiente de recursos limitados em dispositivos NISQ. Ao dividir o gerador em múltiplos sub-geradores, a estratégia de *Patch* explora a modularidade dos circuitos quânticos, possibilitando que mesmo dispositivos com poucos qubits e circuitos de baixa profundidade possam ser utilizados de maneira eficaz. As saídas geradas por cada sub-gerador são combinadas para reconstruir a amostra completa, possibilitando a representação de distribuições complexas com maior flexibilidade e adaptabilidade. A proposta também pode ser implementada utilizando a estratégia *Batch*, em que o gerador é projetado como um único circuito parametrizado responsável por gerar a amostra sintética completa em uma única iteração.

A arquitetura do circuito gerador é apresentada na Figura 3, e segue a abordagem proposta por Huang et al. 2021 [Huang et al. 2021]. O Circuito pode ser pode ser dividido em quatro etapas principais: incorporação dos estados (*state embedding*), camadas parametrizadas, transformação não-linear e pós-processamento.

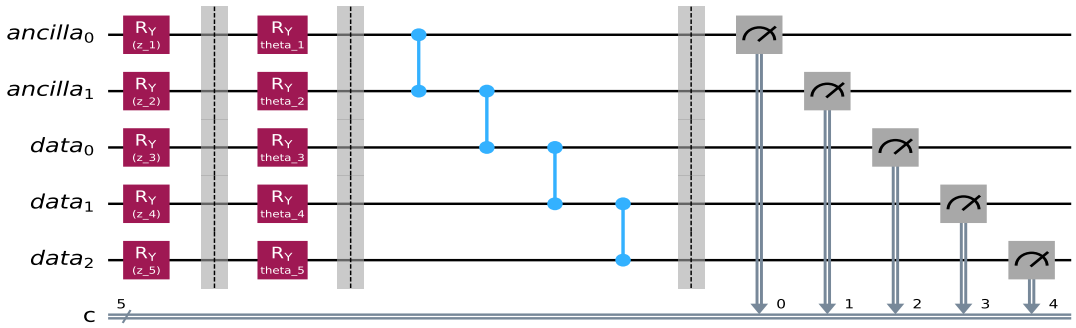


Figura 3. Arquitetura Geradora adaptada de Huang et al. 2021.

Na etapa de *state embedding*, um vetor latente $z \in \mathbb{R}^N$ é amostrado no intervalo $[0, \pi/2)$. O mesmo vetor latente é fornecido a todos os sub-geradores, sendo incorporado ao sistema através de portas R_Y, como mostrado na primeira seção do circuito da Figura 3. Em seguida, são aplicadas as camadas parametrizadas compostas por portas R_Y com parâmetros treináveis θ, seguidas por portas CZ de controle, responsáveis por introduzir o entrelaçamento entre os qubits.

Após as camadas parametrizadas, ocorre a transformação não-linear, viabilizada pela utilização de qubits ancilares (qubits adicionais, que não representam dados de entrada clássica). O estado quântico ($|\Psi(z)\rangle$) gerado pelo circuito, antes da medição, é descrito na Equação 5 onde $U_G(\theta)$ representa a operação unitária aplicada às camadas parametrizadas.

$$|\Psi(z)\rangle = U_G(\theta)|z\rangle \tag{5}$$

Para introduzir não-linearidade, é feita uma medição parcial no subsistema ancilar, representada por Π. O estado resultante nos qubits restantes é apresentado na Equação 6, onde Tr_A indica o traço parcial sobre os qubits ancilares.

$$\rho(z) = \mathrm{Tr}_A \left(\frac{\Pi \otimes I |\Psi(z)\rangle\langle\Psi(z)|}{\langle\Psi(z)|\Pi \otimes I|\Psi(z)\rangle} \right) \tag{6}$$

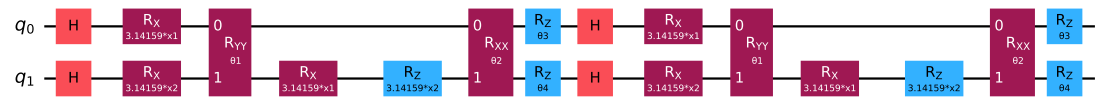
Esse procedimento resulta em um estado misto $\rho(z)$, cuja dependência do vetor latente z está presente tanto no numerador quanto no denominador. Essa característica garante que a transformação realizada no sistema seja não-linear. A partir do estado resultante $\rho(z)$, são feitas medições nos qubits de dados restantes em todas as bases computacionais j. As probabilidades $P(j)$ associadas a cada base representam as amplitudes do estado $\rho(z)$ no espaço de Hilbert. A saída do sub-gerador $G^{(i)}$ é definida como o vetor de probabilidades de medição, apresentado na Equação 7 onde $P(j)$ é a probabilidade de medir o estado j nos qubits de dados.

$$G^{(i)} = [P(0), P(1), \ldots, P(2^{N-N_A} - 1)] \tag{7}$$

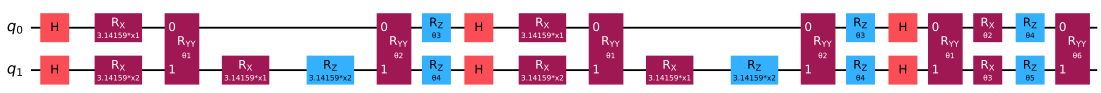
A saída $G^{(i)}$ corresponde, portanto, à projeção das amplitudes de $\rho(z)$ no espaço computacional, convertendo a informação do vetor latente z em uma distribuição de probabilidades. Esse procedimento estabelece uma ponte entre o vetor latente clássico e a saída quântica gerada, permitindo que o sistema generativo capture características complexas e não-lineares dos dados.

4.3. Arquitetura Discriminativa

O papel do discriminador na sQGAN é duplo: distinguir entre amostras reais e geradas (classificação binária) e categorizar as amostras reais em normais ou em diferentes tipos de ataques (classificação multiclasse). Conforme ilustrado na Figura 2, a abordagem utiliza dois discriminadores distintos, mas com a mesma arquitetura subjacente: um responsável pela **classificação binária** e outro pela **classificação multiclasse**. Essa estratégia permite que a mesma arquitetura seja aplicada de maneira flexível a ambas as tarefas. A Figura 4 apresenta as arquiteturas dos circuitos discriminadores, baseadas nas arquiteturas exploradas por Boyle e Nikandish 2024 [Boyle and Nikandish 2024].



(a) Circuito Discriminador de Arquitetura 1 (Arq.1).



(b) Circuito Discriminador de Arquitetura 2 (Arq.2).

Figura 4. Arquiteturas dos Circuitos Discriminadores.

As arquiteturas dos circuitos discriminadores, apresentadas na Figura 4, combinam sub-circuitos variacionais com parâmetros ajustáveis por um otimizador clássico. Ambas as arquiteturas incluem portas de rotação RX e RZ, responsáveis por manipular as amplitudes dos qubits, além de portas de entrelaçamento, como RYY e RXX. Enquanto a **Arquitetura 1** utiliza portas RYY e RXX para promover o entrelaçamento entre os qubits, a **Arquitetura 2** emprega exclusivamente portas RYY. Uma discussão extensiva dessas arquiteturas é encontrada em Boyle e Nikandish 2024 [Boyle and Nikandish 2024].

4.4. Detecção de Ataques com o sQGAN

A detecção de ataques utilizando o sQGAN se baseia em uma arquitetura semi-supervisionada que explora a combinação de dados rotulados e não rotulados. O gerador, estruturado segundo a estratégia de *patch*, cria amostras sintéticas que imitam o tráfego normal da rede. Essa estratégia modular, implementada com múltiplos sub-geradores, permite uma divisão das características do tráfego em partes menores, permitindo o uso de dispositivos quânticos NISQ. O discriminador binário, por sua vez, é responsável por distinguir entre amostras reais (tráfego legítimo) e amostras falsas (potenciais ataques), enquanto o discriminador multiclasse classifica as amostras reais entre tráfego normal e diferentes categorias de ataques. Ambas as tarefas são executadas usando circuitos discriminadores com arquiteturas parametrizadas, que integram sub-circuitos variacionais ajustados por otimizadores clássicos.

A abordagem híbrida quântico-clássica da sQGAN oferece uma representação robusta e não-linear dos dados, permitindo a captura de padrões complexos presentes no tráfego de rede. O modelo aproveita os circuitos geradores parametrizados para transformar o vetor latente clássico em estados quânticos, cujas medições produzem distribuições de probabilidade que representam as características dos dados gerados. Com isso, o sQGAN apresenta potencial para lidar com cenários desafiadores, onde os dados rotulados são escassos, aprimorando a detecção de anomalias e ataques em sistemas de segurança cibernética.

5. Estudo de Caso

Para avaliar o desempenho do modelo sQGAN, são utilizadas as bases de dados NRC-IoMT24 e NRC-ACI-IOT23 [Sasi et al. 2024]. As bases IoMT24 [Dadkhah et al. 2024] e IOT23 [Bastian et al. 2023], amplamente reconhecidas para detecção de intrusões em redes IoT, incluem uma combinação de tráfego de rede normal e ataques diversos, como DDoS, DoS, *Spoofing*, Reconhecimento, MITM e Força Bruta. Neste trabalho, utilizam-se as versões modificadas das bases, conforme proposto por Sasi et al. (2024). Essas versões apresentam aprimoramentos significativos, incluindo a extração de mais características de fluxo de rede, adição de valores ausentes em colunas e correção de erros existentes nas bases originais. Para o estudo de caso, foram selecionadas as 10 melhores características para cada base de dados, de acordo com o apresentado em Sasi et al. (2024).

As bases modificadas são divididas em conjuntos de dados rotulados e não rotulados, com percentuais de dados rotulados variando entre 10%, 20% e 30%. Essa organização permite avaliar o impacto da quantidade de dados rotulados no desempenho

do modelo em um cenário semi-supervisionado, no qual o sQGAN aprende a partir de informações limitadas.

O experimento foi configurado para treinar e testar, considerando a natureza híbrida quântica-clássica do modelo. Na etapa de pré-processamento, os dados de entrada foram normalizados e divididos em amostras rotuladas e não rotuladas. No treinamento do gerador e discriminador, o gerador foi configurado para operar em duas estratégias: *Patch*, com quatro sub-geradores independentes, e *Batch*, com apenas um gerador responsável pela geração completa. Os discriminadores foram implementados utilizando as duas arquiteturas (arq.1 e arq.2) apresentadas na Seção 4. Após o treinamento, o modelo foi avaliado em um conjunto de dados de teste, e o F1 Score, foi coletado para cada configuração. A arquitetura do modelo sQGAN foi implementada utilizando o *framework* IBM Qiskit[1]. O backend `ibm_sherbrooke` foi utilizado como dispositivo quântico devido ao seu baixo nível de ruído em comparação com os outros dispositivos disponíveis na plataforma IBM Quantum. O otimizador clássico usado foi o SGD (*Stochastic Gradient Descent*) assim como em [Boyle and Nikandish 2024] e [Huang et al. 2021].

6. Resultados

Nesta seção, são apresentados e discutidos os resultados obtidos pela sQGAN na detecção e classificação de ataques, utilizando diferentes configurações de geradores e discriminadores. O desempenho é avaliado tanto para análise binária (real/falso) quanto para classificação multiclasse dos diferentes tipos de ataques presentes nas bases de dados NRC-IoMT24 e NRC-ACI-IOT23.

6.1. Resultados da Detecção do Ataque (Análise Binária)

Na Tabela 1, são apresentados os resultados da detecção de ataques na base de dados NRC-IoMT24 para análise binária, utilizando o F1 Score como métrica de desempenho. Observa-se que o gerador com a abordagem Patch, combinado com o discriminador de arquitetura 1 (Arq.1), apresenta o melhor desempenho em todas as porcentagens de dados rotulados (10%, 20% e 30%), atingindo um F1 Score de até 94.55% com 30% de dados rotulados. Em comparação, a abordagem *Batch* do gerador mostrou-se menos eficaz, especialmente quando combinada com o discriminador de arquitetura 2 (Arq.2), alcançando um F1 Score máximo de 88.11%.

Na Tabela 2, os resultados para a base de dados NRC-ACI-IOT23 seguem uma tendência similar. A configuração *Patch* com o discriminador Arq.1 também obteve os melhores resultados, atingindo um F1 Score de 89.59% com 30% de dados rotulados. Em contraste, a configuração *Batch* do gerador e o discriminador Arq.2 apresentaram desempenhos inferiores, confirmando a eficácia superior da estratégia Patch combinada com o discriminador Arq.1 na detecção binária de ataques.

6.2. Resultados da Detecção do Ataque (Análise Multiclasse)

Os resultados apresentados a seguir utilizam exclusivamente a rede Geradora com *Patch* e o Discriminador Arquitetura 1 (Arq. 1), pois essa configuração apresentou os melhores resultados na análise binária. Na Tabela 3, são exibidos os resultados de classificação de diferentes tipos de ataques na base de dados NRC-IoMT24 para a análise multiclasse.

[1]https://www.ibm.com/quantum/qiskit

Tabela 1. Detecção de Ataques (Binário) na base de dados NRC-IoMT24. (Métrica: F1 Score).

Gerador	Discriminador	10%	20%	30%
Batch	Arq.1	85.47 ± 0.51	87.23 ± 0.48	91.82 ± 0.59
Batch	Arq.2	81.69 ± 0.36	86.45 ± 0.54	88.11 ± 0.40
Patch	Arq.1	$\mathbf{90.22} \pm 0.46$	$\mathbf{93.04} \pm 0.37$	$\mathbf{94.55} \pm 0.41$
Patch	Arq.2	87.36 ± 0.60	89.78 ± 0.53	92.14 ± 0.34

Tabela 2. Detecção de Ataques (Binário) na base de dados NRC-ACI-IOT23. (Métrica: F1 Score).

Gerador	Discriminador	10%	20%	30%
Batch	Arq.1	80.13 ± 0.45	81.76 ± 0.52	82.34 ± 0.48
Batch	Arq.2	80.29 ± 0.50	81.58 ± 0.47	82.63 ± 0.42
Patch	Arq.1	$\mathbf{86.72} \pm 0.44$	$\mathbf{88.34} \pm 0.39$	$\mathbf{89.59} \pm 0.37$
Patch	Arq.2	84.21 ± 0.46	86.77 ± 0.53	88.12 ± 0.40

Observa-se que o modelo obteve valores elevados de F1 Score para ataques como DDoS e DoS, atingindo até 97.18% e 95.12% com 30% de dados rotulados, respectivamente. Por outro lado, ataques como *Spoofing* e MQTT apresentaram valores mais baixos, sugerindo que esses tipos de ataque são mais desafiadores para detecção.

Na Tabela 4, os resultados de classificação na base de dados NRC-ACI-IOT23 mostram que o modelo obteve F1 Score mais elevado para ataques como *Reconnaissance* e DoS, chegando a 88.12% e 86.14% com 30% de dados rotulados. Em comparação, os ataques MITM e *Brute Force* apresentaram desempenho ligeiramente inferior, mas ainda em níveis satisfatórios para a detecção de ataques em redes IoT. Esses resultados confirmam a eficácia do modelo sQGAN, com a configuração Geradora *Patch* e Arquitetura 1, na tarefa de classificação multiclasse. A abordagem se mostra particularmente eficaz na detecção de ataques como DDoS e *Reconnaissance*.

Tabela 3. Classificação de Ataques (Multiclasse) na base de dados NRC-IoMT24. (Métrica: F1 Score).

Ataque	10%	20%	30%
DDoS	92.15 ± 0.21	95.43 ± 0.33	$\mathbf{97.18 \pm 0.27}$
DoS	89.24 ± 0.30	92.57 ± 0.19	$\mathbf{95.12 \pm 0.24}$
Recon	78.42 ± 0.28	79.34 ± 0.25	85.29 ± 0.18
Spoofing	$\mathit{65.73} \pm 0.22$	73.16 ± 0.27	78.41 ± 0.31
MQTT	$\mathit{76.89} \pm 0.29$	79.08 ± 0.15	84.54 ± 0.23

7. Conclusão e Trabalhos Futuros

A sQGAN demonstrou-se uma abordagem eficaz para a detecção e classificação de ataques cibernéticos em cenários com dados rotulados limitados. Os resultados obtidos evidenciam que a arquitetura híbrida quântico-clássica, combinada com a estratégia de patch no gerador e o discriminador de Arquitetura 1, alcança alto desempenho em tarefas de detecção binária e classificação multiclasses. Em particular, foram registrados F1 scores

Tabela 4. Classificação de Ataques (Multiclasse) na base de dados NRC-ACI-IOT23. (Métrica: F1 Score).

Ataque	10%	20%	30%
Reconnaissance	84.21 ± 0.24	86.67 ± 0.18	$\mathbf{88.12 \pm 0.29}$
DoS	83.35 ± 0.30	84.58 ± 0.21	$\mathbf{86.14 \pm 0.33}$
MITM	81.42 ± 0.27	82.73 ± 0.14	86.29 ± 0.19
Brute Force	82.58 ± 0.25	84.12 ± 0.32	86.88 ± 0.23

superiores a 97% para ataques como DDoS na base de dados NRC-IoMT24 e aproximadamente 89% para *Reconnaissance* na base NRC-ACI-IOT23.

Assim, a proposta revelou-se robusta na análise binária e multiclasse, conseguindo diferenciar tráfego legítimo e tipos variados de ataques, oferecendo uma capacidade avançada de detecção em condições adversas de rotulagem. No entanto, limitações como a pouca disponibilidade de infraestrutura quântica atual, os desafios na integração com sistemas legados e a alta complexidade de implementação podem reduzir a escalabilidade da solução em ambientes com alta dimensionalidade no curto prazo. Além disso, a dependência de dados rotulados, ainda que mínima, restringe sua aplicabilidade em contextos com ausência completa de rótulos. Outra limitação refere-se ao consumo computacional do treinamento híbrido, que pode ser um entrave em ambientes com recursos computacionais limitados e com requisitos de baixa latência.

Essas considerações destacam a relevância da sQGAN para detecção de ataques em segurança cibernética e apontam para a necessidade de futuras pesquisas para ampliar sua aplicabilidade e robustez. Assim, como trabalhos futuros, sugere-se explorar a integração de técnicas de detecção de ataques clássicas, e investigar abordagens para reduzir o custo computacional associado ao treinamento híbrido. Ademais, estudos futuros poderiam focar na adaptação da sQGAN a arquiteturas quânticas emergentes, otimizando seu desempenho em ambientes de alta dimensionalidade e explorando a aplicação em novos cenários de ameaças.

Agradecimentos

Este trabalho foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), e pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) projeto 2023/00811-0, projeto 2023/00673-7, projeto 2021/00199-8 (CPE SMARTNESS), projeto 2020/04031-1, e projeto 2018/23097-3.

Referências

Abreu, D., Rothenberg, C., and Abelém, A. (2024a). qids: Sistema de detecção de ataques baseado em aprendizado de máquina quântico híbrido. In *Anais do XLII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 295–308, Porto Alegre, RS, Brasil. SBC.

Abreu, D., Rothenberg, C. E., and Abelém, A. (2024b). Qml-ids: Quantum machine learning intrusion detection system. In *2024 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE.

Bastian, N., Bierbrauer, D., McKenzie, M., and Nack, E. (2023). Aci iot network traffic dataset 2023.

Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., and Lloyd, S. (2017). Quantum machine learning. *Nature*, 549(7671):195–202.

Boyle, A. O. and Nikandish, R. (2024). A hybrid quantum-classical generative adversarial network for near-term quantum processors. *IEEE Access*.

Dadkhah, S., Neto, E. C. P., Ferreira, R., Molokwu, R. C., Sadeghi, S., and Ghorbani, A. A. (2024). Ciciomt2024: A benchmark dataset for multi-protocol security assessment in iomt. *Internet of Things*, 28:101351.

de Araujo-Filho, P. F., Naili, M., Kaddoum, G., Fapi, E. T., and Zhu, Z. (2023). Unsupervised gan-based intrusion detection system using temporal convolutional networks and self-attention. *IEEE Transactions on Network and Service Management*, 20(4):4951–4963.

Huang, H.-L., Du, Y., Gong, M., Zhao, Y., Wu, Y., Wang, C., Li, S., Liang, F., Lin, J., Xu, Y., et al. (2021). Experimental quantum generative adversarial networks for image generation. *Physical Review Applied*, 16(2):024051.

Idhammad, M., Afdel, K., and Belouch, M. (2018). Semi-supervised machine learning approach for ddos detection. *Applied Intelligence*, 48:3193–3208.

Lim, W., Chek, K. Y. S., Theng, L. B., and Lin, C. T. C. (2024). Future of generative adversarial networks (gan) for anomaly detection in network security: A review. *Computers & Security*, page 103733.

Mvula, P. K., Branco, P., Jourdan, G.-V., and Viktor, H. L. (2023). A systematic literature review of cyber-security data repositories and performance assessment metrics for semi-supervised learning. *Discover Data*, 1(1):4.

Nakaji, K. and Yamamoto, N. (2021). Quantum semi-supervised generative adversarial network for enhanced data classification. *Scientific reports*, 11(1):19649.

Nicesio, O. K., Leal, A. G., and Gava, V. L. (2023). Quantum machine learning for network intrusion detection systems, a systematic literature review. In *2023 IEEE 2nd International Conference on AI in Cybersecurity (ICAIC)*, pages 1–6. IEEE.

Odena, A. (2016). Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*.

Sajun, A. R. and Zualkernan, I. (2022). Survey on implementations of generative adversarial networks for semi-supervised learning. *Applied Sciences*, 12(3):1718.

Sasi, T., Lashkari, A. H., Lu, R., Xiong, P., and Iqbal, S. (2024). An efficient self attention-based 1d-cnn-lstm network for iot attack detection and identification using network traffic. *Journal of Information and Intelligence*.

Wang, Y. and Liu, J. (2024). A comprehensive review of quantum machine learning: from nisq to fault tolerance. *Reports on Progress in Physics*.

Yang, X., Song, Z., King, I., and Xu, Z. (2022). A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954.