Fernanda Carolina da Silva Pereira

CGANs e Treinamento Adversarial para Classificação de Ameaças em IoT

Belo Horizonte, Minas Gerais
2025

Fernanda Carolina da Silva Pereira

CGANs e Treinamento Adversarial para Classificação de Ameaças em IoT

Relatório apresentado como requisito da disciplina de Monografia em Sistemas de Informação II do Curso de Bacharelado em Sistemas de Informação da UFMG

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Orientador: Prof.ª Dra. Michele Nogueira Lima
Co-orientador: M.Sc. Mateus Pelloso

Belo Horizonte, Minas Gerais
2025

Resumo

Este estudo investiga o campo de Adversarial Machine Learning (AML) aplicado à classificação de ciberataques em redes IoT, partindo dos avanços obtidos na primeira fase desse trabalho, com a Análise Comparativa de Métodos de Geração de Amostras Adversariais na Detecção de Ataques. Esta pesquisa foi conduzida a partir de experimentos em dois conjuntos de dados amplamente utilizados na área (IoT-23 e TON-IoT), empregando tanto abordagens de treinamento convencional quanto adversarial. Para evidenciar o impacto de perturbações intencionais, foram geradas amostras adversariais por técnicas como FGSM e CGAN, demonstrando como pequenas modificações podem comprometer o desempenho de modelos de detecção de ameaças. A partir dessas técnicas, os resultados obtidos antes e após a aplicação do treinamento adversarial foram avaliados a partir de métricas como acurácia e precisão. Dessa forma, as conclusões alcançadas ressaltam a relevância de estratégias defensivas específicas para cibersegurança, visto que as soluções desenvolvidas em outros domínios, como visão computacional, não são imediatamente transferíveis para ambientes de rede.

Palavras-chave: Amostras Adversariais. Aprendizado de Máquina Adversarial. Aprendizado de Máquina. Ataques. Cibersegurança. Inteligência Artificial. IoT.

Abstract

This study investigates the field of Adversarial Machine Learning (AML) applied to the classification of cyberattacks on IoT networks, building on the advances made in the first phase of this work, with the Comparative Analysis of Adversarial Sample Generation Methods in Attack Detection. This research is based on experiments on two datasets widely used in the field (IoT-23 and TON-IoT), using both conventional and adversarial training approaches. To highlight the impact of intentional disturbances, adversarial samples were generated using techniques such as FGSM and CGAN, demonstrating how small modifications can compromise the performance of threat detection models. Using these techniques, the results obtained before and after adversarial training were evaluated using metrics such as accuracy and precision. In this way, In this way, the conclusions reached highlight the relevance of specific defensive strategies for cybersecurity, given that solutions developed in other domains, such as computer vision, are not immediately transferable to network environments.

Key-words: Adversarial Samples. Adversarial Machine Learning. Artificial Intelligence. Attacks. Cybersecurity. IoT. Machine Learning.

Lista de ilustrações

Figura 1 – Arquitetura de uma Rede Adversarial Generativa 13
Figura 2 – Processo de geração de dados sintéticos para o Treinamento Adversarial em GANs 14
Figura 3 – Fluxo de trabalho exemplificado em etapas 1, 2, 3, 4 e 5 19
Figura 4 – Comparação de desempenho de cada técnica no treinamento convencional do conjunto de dados TON-IoT. 22
Figura 5 – Comparação de desempenho de cada técnica no treinamento convencional do conjunto de dados IoT-23. 23
Figura 6 – Comparação de desempenho de cada técnica no treinamento adversarial do conjunto de dados TON-IoT 24
Figura 7 – Comparação de desempenho de cada técnica no treinamento adversarial do conjunto de dados IoT-23. 25

Lista de tabelas

Tabela 1 – Comparação de desempenho para treinamento convencional nos conjuntos de dados IoT-23 e TON-IoT. 22
Tabela 2 – Comparação de desempenho para treinamento adversarial nos conjuntos de dados IoT-23 e TON-IoT. 24

Lista de abreviaturas e siglas

AML	Adversarial Machine Learning
ART	Adversarial Robustness Toolbox
ASCII	American Standard Code for Information Interchange
AT	Adversarial Training
CGAN	Redes Adversariais Generativas Condicionais
C&C	Command and Control
DDoS	Distributed Denial of Service
FGSM	Fast Gradient Signed Method
FN	Falso Negativo
FP	Falso Positivo
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IoT	Internet of Things
IIoT	Industrial Internet of Things
ML	Machine Learning
MNIST	Modified National Institute of Standards and Technology database
MSI	Monografia em Sistemas de Informação
RAM	Random-Access Memory
SSH	Secure Shell Protocol
UNSW	University of New South Wales
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
TON	Telemetry Operating and Network

Sumário

1 INTRODUÇÃO 9
1.1 Objetivos 10
1.1.1 Objetivo Geral 10
1.1.2 Objetivos Específicos 10
1.2 Contribuições 10
1.3 Organização do texto 11

2 FUNDAMENTOS E REFERENCIAL TEÓRICO 12
2.1 Ataques Adversariais 12
2.2 GANs 12
2.2.1 CGANs 13
2.3 FGSM 13
2.4 Treinamento Adversarial 13
2.5 Aplicação 14
2.6 Resumo 15

3 METODOLOGIA 16
3.1 Abordagens 16
3.1.1 CGANs 16
3.1.2 FGSM 16
3.1.3 Treinamento Adversarial 16
3.2 Conjuntos de dados 17
3.2.1 Ataques 17
3.2.2 Pré-processamento de dados 18
3.3 Fluxo de trabalho 18
3.4 Métricas 19
3.5 Resumo 20

4 EXPERIMENTOS 21
4.1 Experimentos 21
4.2 Resultados 21
4.2.1 Avaliação 21
4.3 Resumo 25

5 DESAFIOS E LIMITAÇÕES 27

6 CONCLUSÕES E TRABALHOS FUTUROS 28

REFERÊNCIAS 29

1 Introdução

A Cibersegurança (em inglês, *Cybersecurity*) consiste em métodos de defesa aplicados para detectar e impedir intrusos no espaço cibernético [1]. A partir do uso crescente e cada vez mais amplo de soluções digitais automatizadas, como comunicação pessoal, transações financeiras e serviços de saúde, emergiu a preocupação por soluções robustas e efetivas para o ciberespaço. No entanto, a prática da Cibersegurança enfrenta muitos desafios para o desenvolvimento e aplicação de técnicas que sejam factíveis e apresentem custo-benefício computacional praticável [2]. Entre os desafios enfrentados pela Cibersegurança estão a detecção de intrusão [3] e, com a aplicação de técnicas de Inteligência Artificial neste contexto, os ataques adversariais [4]. A detecção de intrusão visa monitorar recursos de rede para detectar comportamento anômalo e uso indevido na rede, como o reconhecimento de padrões típicos de ataques. Por fim, em ataques adversariais, atacantes se aproveitam de vulnerabilidades nos modelos de aprendizado de máquina, como a falta de robustez desses modelos.

O contexto de ataques adversariais compreende três principais métodos de ataque: envenenamento [5], extração [6] e evasão [7]. Em ataques de envenenamento, o atacante altera os dados de treinamento ou seus rótulos para reduzir o desempenho do modelo. Para realizar a extração ou roubo do modelo, o atacante observa um sistema de aprendizado de máquina (em inglês, *Machine Learning* – ML) para replicar suas funcionalidades ou extrair informações sensíveis. Por fim, em ataques de evasão, o atacante manipula os dados de entrada para levar classificadores a classificações incorretas. Neste caso, o objetivo do atacante é evadir modelos de detecção de *malware*, anomalias e intrusões, por exemplo.

Diante dessa realidade, uma estratégia defensiva sendo estudada é a geração de amostras adversariais, ao tornar os modelos de aprendizado de máquina mais robustos por meio de técnicas como o Aprendizado de Máquina Adversarial (em inglês, *Adversarial Machine Learning* – AML) ou Treinamento Adversarial (em inglês, *Adversarial Training* - AT). A ideia principal é desenvolver métodos de defesa em que um modelo é treinado recebendo a injeção de amostras adversariais a cada iteração do treinamento. Este, por sua vez, é utilizado para prevenir ataques, como também evitar que dados a respeito do comportamento da rede sejam extraídos.

Almejando melhorar a detecção de intrusão e segurança de sistemas, as Redes Adversariais Generativas (em inglês, *Generative Adversarial Networks* – GANs) têm sido amplamente empregadas. Recentemente, uma pesquisa [2] foi realizada para examinar de forma mais abrangente a aplicação da geração de amostras por GANs na Cibersegurança. Os autores concluíram que esta abordagem tem se mostrado efetiva ao prevenir os problemas dinâmicos em segurança no mundo digital. Entre os principais problemas destacados como prioritários na busca por soluções [3][8][4] estão os ataques adversariais.

Sendo assim, este trabalho propõe experimentar os avanços feitos na primeira etapa da Monografia, como o melhor entendimento da implementação e funcionamento dos ataques adversariais, como recursos para aplicar melhorias nos modelos de ML para evitar evasão por meio de ataques adversariais. Neste sentido, o intuito é aplicar a abordagem AT empregando GANs para melhorar o desempenho de modelos de classificação quando expostos a amostras adversariais. Através desta prática, espera-se contribuir para o desenvolvimento de métodos de defesa contra agentes mal-intencionados no ciberespaço.

1.1 Objetivos

1.1.1 Objetivo Geral

Este trabalho analisa o impacto da utilização das GANs no Treinamento Adversarial para a mitigação da evasão de modelos de detecção de intrusão em um contexto de Internet das coisas (em inglês, *Internet of Things* – IoT). Este trabalho avança os resultados alcançados no MSI I, cujo objetivo foi comparar os métodos de geração de amostras adversariais para analisar seu impacto contra o classificador.

1.1.2 Objetivos Específicos

Os objetivos específicos deste trabalho incluem a avaliação das GANs na geração de amostras adversariais quando essas amostras são utilizadas contra um modelo baseado em ML de detecção de intrusão em um contexto de IoT, bem como avaliar o impacto do Treinamento Adversarial na mitigação da evasão desses modelos. Através desta análise, é esperado contribuir para um aprimoramento do Treinamento Adversarial através das GANs, a fim de mitigar os desafios enfrentados na proteção contra ataques adversariais em modelos de ML e melhorar a robustez do classificador, atuando no desenvolvimento ou aprimoramento de defesas para evitar que modelos sejam levados a classificações errôneas.

1.2 Contribuições

Uma das principais contribuições deste estudo é a avaliação sistemática de cenários de treinamento, na qual foi implementado um pipeline que possibilita a comparação direta entre o treinamento convencional e o treinamento adversarial em dois conjuntos de dados distintos (IoT-23 e TON-IoT). Esse processo permitiu observar como os ataques adversariais afetam o desempenho do modelo em cada conjunto de dados, evidenciando fragilidades e pontos fortes sob diferentes condições.

Outra contribuição relevante foi a proposta de geração de dados sintéticos condicionados. O uso da CGAN para a criação de amostras de ataques representa um avanço no sentido de produzir dados mais balanceados e representativos, principalmente em contextos em que a

coleta de grandes quantidades de dados reais rotulados é inviável ou custosa. Ainda assim, evidenciou-se que a configuração de redes generativas condicionais exige um trabalho rigoroso de calibração de hiperparâmetros, além de uma análise criteriosa da qualidade e diversidade das amostras geradas.

Por fim, o trabalho propõe sugestões de estratégias de defesa embasadas nas descobertas obtidas. Indica-se a adoção de métodos como treinamento adversarial e regularizações específicas, bem como a aplicação de técnicas de ensembling, demonstrando que uma abordagem multifacetada pode fortalecer de maneira significativa a robustez de sistemas de detecção de ciberataques. Dessa forma, o estudo contribui tanto para a discussão teórica quanto para aplicações práticas envolvendo segurança em redes IoT.

1.3 Organização do texto

O primeiro capítulo desde trabalho, Introdução, apresenta o problema a ser investigado e sua contextualização, como também seus objetivos gerais e específicos. O segundo capítulo, Fundamentos e Referencial Teórico, se trata da apresentação dos conceitos pertinentes para o problema e sua importância, embasando-se em referenciais teóricos da literatura. Os passos previstos para a execução do projeto e ferramentas utilizadas para abordar o problema são descritos no capítulo 3, Metodologia. O capítulo 4, Experimentos, aborda os experimentos desenvolvidos e resultados alcançados. Por fim, os desafios e limitações do trabalho desenvolvido são expostos no capítulo 5 e o trabalho é concluído no capítulo 6, demonstrando objetivos alcançados e próximos passos.

2 Fundamentos e Referencial Teórico

Este capítulo apresenta os fundamentos teóricos e as principais referências que embasam a pesquisa sobre GANs Contra Ataques de Evasão nos Modelos de Detecção de Intrusão em IoT. Neste caso, serão explorados conceitos fundamentais a ataques adversariais, sistemas detecção de intrusão, bem como a estrutura de GANs e sua aplicação no Treinamento Adversarial no contexto da detecção de intrusão em IoT.

2.1 Ataques Adversariais

O termo "amostras adversariais"foi estabelecido por [4], ao perceber que aplicando perturbações imperceptíveis aos dados de entrada de um modelo classificador, levavam-no a gerar classificações equivocadas. Logo, outros estudos foram desenvolvidos demonstrando o efeito da aplicação dessas amostras em modelos de detecção de ataques [9][10][11][12].

Inspirado no trabalho [13], ataques adversariais são divididos em subconjuntos de ataques com base em duas propriedades, influência e violação. A influência determina se o ataque é aplicado em tempo de treinamento ou em tempo de teste e a violação representa o tipo de violação de segurança que afeta a disponibilidade ou integridade do sistema [14].

Além disso, de acordo com [15] é necessário considerar algumas características para compreender os ataques, como o objetivo do atacante, o conhecimento do modelo atacado, as ações que os atacantes podem executar e, sua estratégia para alcançar seu objetivo. No que diz respeito ao conhecimento do modelo atacado, os autores distinguem os ataques em ataques caixa-branca (em inglês, *white box attacks*), em que o atacante possui conhecimento completo do modelo atacado, ataques caixa-cinza (em inglês, *grey box attacks*), que abrangem conhecimento parcial do modelo por parte do atacante, e ataques caixa-preta (em inglês, *black box attacks*), em que não há nenhum conhecimento sobre o modelo [14].

2.2 GANs

Uma Rede Adversarial Generativa, (em inglês, *Generative Adversarial Network* – GAN) é um *framework* proposto por [16], em que dois modelos são simultaneamente treinados em uma abordagem adversarial. A geração de dados sintéticos usando GANs se dá por um modelo gerador e um discriminador, que compõem a arquitetura da GAN. O modelo gerador gera amostras durante o processo de treinamento adicionando uma taxa de ruído e, o discriminador distingue a diferença entre as amostras geradas pelo primeiro e amostras originais [17]. A figura 1, adaptada de [18] exemplifica a arquitetura de uma GAN, conforme esquematizado pelos autores do trabalho.

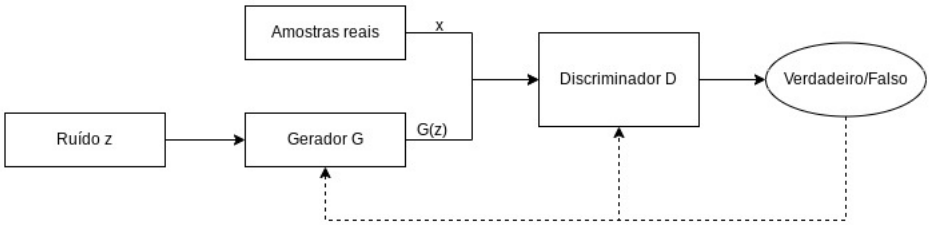

Figura 1 – Arquitetura de uma Rede Adversarial Generativa

2.2.1 CGANs

Rede Adversarial Generativa Condicional, (em inglês, *Conditional Generative Adversarial Network* – CGAN) é um modelo proposto por [19], que tem como base a GAN tradicional condicionada a alguma informação adicional. Nesta estrutura, essa informação condicional é fornecida ao discriminador e ao gerador como uma camada adicional nos modelos. Dessa forma, os dados de treinamento são combinados à informação condicional, que podem ser rótulos, atributos ou descrições textuais, por exemplo, para gerar novos dados sintéticos.

2.3 FGSM

Em [20], os autores propuseram o FGSM, (em inglês, *Fast Gradient Sign Method*) um método baseado em gradiente, simples e rápido, cuja ideia é gerar amostras adversariais. Neste caso, um vetor imperceptivelmente pequeno, cujos elementos são iguais ao sinal dos elementos do gradiente da função de custo, é aplicado aos dados de entrada do modelo, criando perturbações nos dados. Para isso, as perturbações são aplicadas em cada *feature* do conjunto de dados.

2.4 Treinamento Adversarial

O Treinamento Adversarial consiste em uma prática para treinamento de modelos de ML que agrega amostras adversariais e amostras originais, proposta pelos autores do trabalho [4], demonstrando que seria possível incrementar seu desempenho de alguma forma. Baseando-se no *framework* proposto, outros autores [21][22][23] incrementaram a técnica para cenários e necessidades específicas, com foco principalmente no domínio de imagens. Dessa forma, o AT se tornou uma das estratégias mais efetivas para proporcionar mais robustez a modelos submetidos a ataques adversariais [24].

A relevância do treinamento adversarial na mitigação de ciberataques é cada vez maior, considerando o crescente emprego de ML nas soluções de detecção de ataques. Neste contexto, o domínio da geração de amostras adversariais é particularmente interessante para fortalecer modelos utilizados amplamente no espaço cibernético. Além disso, o

ciberespaço está repleto de atacantes (por exemplo, desenvolvedores de *malware* que desejam escapar de antivírus e filtros de spam baseados em aprendizado profundo e aprendizado de máquina) [25].

2.5 Aplicação

Ao explorar a técnica do AT no trabalho [26], os autores demonstram a relação entre Treinamento Adversarial e GANs. Ao incrementar os dados de treinamento do modelo discriminador da GAN adotada, é esperado que este se torne mais robusto ao discernir amostras originais e adversariais. Os resultados encontrados no trabalho [26] demonstram que os objetivos de treinamento da GAN, sejam mínimos quadrados [27] ou lineares [28], são significativamente estabilizados se o discriminador é robusto contra ataques adversariais.

É importante mencionar que uma variedade de aplicações GAN foram usadas na Cibersegurança [2], como detecção de *malware*, invasão de rede e dispositivos móveis, detecção de intrusão e detecção de *botnets*. Os resultados demonstram que GANs possuem potencial de reforçar as defesas de cibersegurança nessas áreas. No entanto, algumas áreas, como a robustez do modelo contra ataques adversariais ainda são limitações enfrentadas para o contexto, demonstrando que o ambiente cibernético continua suscetível a esse tipo de ataque. A Figura 2, apresentada em [29], demonstra a aplicação da estrutura GAN ao processo de geração de dados sintéticos para o Treinamento Adversarial.

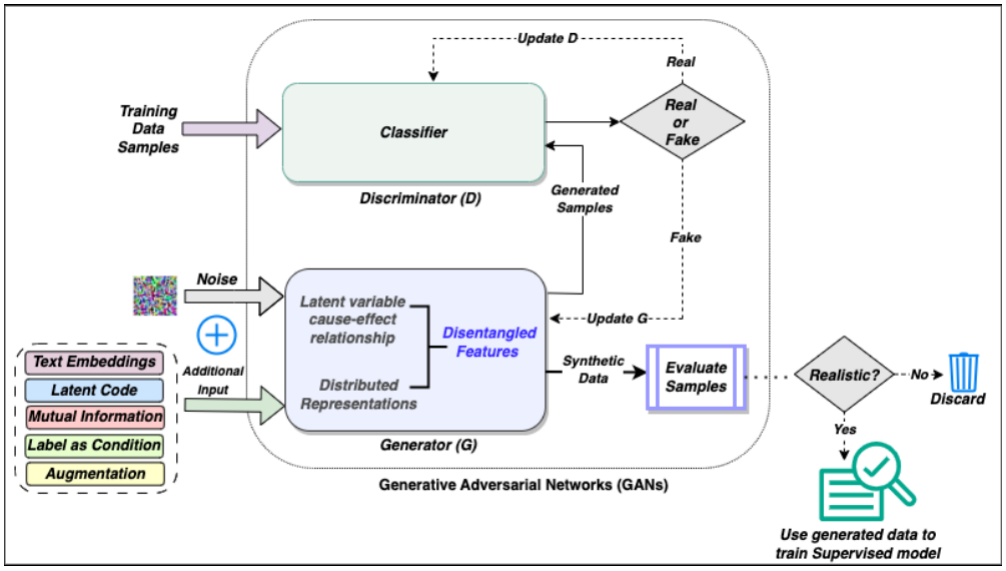

Figura 2 – Processo de geração de dados sintéticos para o Treinamento Adversarial em GANs

2.6 Resumo

Neste capítulo foram explicados e embasados os conceitos de Ataque Adversarial, que objetivam confundir classificadores baseados em ML, assim como seus tipos, que abrangem ataques caixa branca, caixa cinza e caixa preta, determinados pelo nível de conhecimento do modelo por parte do atacante. Na sequência, discutiu-se o treinamento adversarial como uma estratégia defensiva, detalhando como a inclusão de amostras adversariais no treinamento pode aumentar a robustez de modelos de aprendizado de máquina.

Por fim, explorou-se o conceito de redes adversariais generativas (GANs) e redes adversariais generativas condicionais (CGAN), incluindo sua arquitetura básica, composta por gerador e discriminador, e seu potencial no contexto de cibersegurança, seja como ferramenta para geração de amostras adversariais ou como mecanismo de defesa. Por fim, abordou-se a aplicação dos métodos adotados, neste caso, das GANs aplicadas ao Treinamento Adversarial, colaborando para a mitigação de ataques.

3 Metodologia

A proposta para este trabalho é empregar a estrutura de GANs como base para aplicar o Treinamento Adversarial. A finalidade desta prática é visualizar métricas, a serem definidas, resultantes do treinamento para identificar seu impacto quando ataques adversariais são aplicados contra um modelo de detecção de intrusão em um contexto IoT. Essas técnicas adotadas para este trabalho são abordagens apresentadas na literatura dentre um amplo número de técnicas. Logo, essas foram escolhidas conforme lacunas e limitações destacadas por trabalhos na área [3][8][4], o que justifica nossa investigação.

3.1 Abordagens

3.1.1 CGANs

Para avaliar e aumentar a robustez dos modelos, serão utilizadas Redes Adversariais Generativas Condicionais (GANs) para gerar amostras adversariais realistas com base em rótulos das classes fornecidas como treinamento e, o Treinamento Adversarial, para expor o modelo a essas amostras durante o treinamento. As CGANs serão configuradas com uma rede geradora para criar perturbações adversariais e uma rede discriminadora para distinguir entre amostras reais e adversariais. Após o treinamento da CGAN, amostras adversariais serão incorporadas ao processo de treinamento do modelo-alvo.

3.1.2 FGSM

Após o treinamento do modelo, a função *FastGradientMethod* (FGSM) da biblioteca *Adversarial Robustness Toolbox* (ART) foi aplicada. Esta função gera amostras adversariais com base no método FGSM, adicionando pequenas perturbações aos dados de entrada. Essas perturbações são calculadas a partir do gradiente da função de custo em relação aos dados de entrada, visando maximizar o erro de classificação do modelo.

3.1.3 Treinamento Adversarial

O Treinamento Adversarial será empregado para melhorar a robustez do modelo contra ataques adversariais. Para isso, amostras adversariais serão geradas utilizando a estrutura de CGANs e integradas ao conjunto de treinamento, juntamente com amostras limpas. Durante o treinamento, o modelo será exposto a essas amostras adversariais, forçando-o a aprender padrões mais robustos. Ao fim, o treinamento será avaliado a partir das métricas que serão elencadas para melhor avaliar o contexto estudado.

3.2 Conjuntos de dados

Os experimentos foram conduzidos aplicando o conjunto de dados IoT-23 [30] e o conjunto TON-IoT [31]. O IoT-13 é um conjunto de dados que contém tráfego de dispositivos IoT capturado na Universidade CTU, República Tcheca, no período de 2018 a 2019. Este conjunto possui diversas capturas de tráfego de dispositivos IoT reais, além de tráfego normal (benigno) e consiste em 23 capturas (chamadas de cenários) de diferentes amostras. Este conjunto abrange tráfegos de rede agrupados por dispositivos IoT, onde o tráfego malicioso pertence a ataques como C&C-*HeartBeat*, DDoS, *Okiru*, *PartOfAHorizontalPortScan* e *PartOfAHorizontalPortScan-Attack*. Detalhes da utilização do conjunto de dados ainda serão apresentados nas próximas etapas, conforme avanço da implementação do trabalho.

O conjunto TON-IoT, por sua vez, inclui fontes de dados heterogêneas coletadas de conjuntos de dados de telemetria de sensores de IoT e IIoT, conjuntos de dados de sistemas operacionais do Windows 7 e 10, bem como conjuntos de dados de tráfego de rede e TLS do Ubuntu 14 e 18. Os conjuntos de dados foram coletados de uma rede realista e de grande escala projetada no Cyber Range e IoT Labs, na Escola de Engenharia e Tecnologia da Informação, UNSW Canberra the Australian Defence Force Academy. Este conjunto de dados também apresenta várias capturas de tráfego dos dispositivos mencionados para cada cenário, IoT, Linux, Windows e dados de rede, contendo diferentes tipos de dados benignos e de ataque em cada um.

3.2.1 Ataques

Em ambos os conjuntos de dados adotados, alguns ataques foram abordados para classificação pelo modelo e métodos definidos. Portanto, os rótulos aplicados no conjunto[30] são brevemente descritos:

- DDoS: Atacantes sobrecarregam a largura de banda, gerando enorme tráfico de rede;
- C&C-*Heartbeat*: Similar ao *Heartbleed*, o nome descreve tipos de ataques que exploram o mecanismo *Heartbeat* em softwares, enviando pequenos pacotes para rastrear o host infectado pelo servidor C&C;
- *Okiru*: O rótulo indica que as conexões possuem características de um *botnet Okiru*;
- *PartOfAHorizontalPortScan*: O rótulo indica que as conexões são usadas para realizar uma varredura horizontal de portas para coleta de informações para ataques adicionais;
- *Scanning*: técnica de exploração de rede usada para identificar os sistemas conectados à rede de uma organização. Ela fornece informações sobre os sistemas, serviços e recursos acessíveis em um sistema alvo.

3.2.2 Pré-processamento de dados

Para conduzir os experimentos com os dados provenientes do conjunto IoT-23, o arquivo *conn.log.labeled* da captura Kenjiro necessitou de pré-processamento. Este arquivo foi disponibilizado em formato *ASCII text*, portanto, foi necessário aplicar regras específicas para a importação adequada do conteúdo, considerando que a importação padrão de dados *csv* não se adequou ao contexto. Para isso, foram explicitados no código os nomes das *features* esperadas e implementada a leitura dos arquivos linha a linha, ignorando os comentários (apontados com #), separando os dados a serem recebidos por cada *feature* pelo espaço em branco e reunindo em um vetor de dados. A partir dos dados capturados e *features* definidas, conteúdo foi concatenado entre si em um *dataframe*. Além disso, os dados inválidos, rotulados com *'-'* e *'empty'*, foram padronizados como *'NaN'*.

Neste caso, os arquivos foram rotulados aplicando uma *feature* chamada *'label'* e outra *'detailed-label'*, em que *'label'* aponta se o registro é benigno ou maligno e *'detailed-label'*, o tipo de ataque, em caso de dado maligno. Com isso, foi aplicada uma regra a cada registro do novo *dataframe*. Logo, a cada registro maligno, o nome do ataque foi recebido para sobrescrever o rótulo maligno em *'label'*, permitindo descartar a *feature 'detailed-label'*.

Após limpeza dos dados inválidos e reestruturação da *feature 'label'*, os tipos das *features* restantes foram mapeados explicitamente para *'int'* e *'string'*, dependendo do tipo, visto que, após a importação, todos assumiram o mesmo tipo, *'object'*. Por fim, os dados necessitaram de limpeza e remoção de valores inválidos, normalização e padronização de escalas.

Para o conjunto TON-IoT, o processo foi semelhante. Neste contexto foi adotada a captura 1 para os dados de rede, que contém dados benignos e dados de ataque *Scanning*, em formato *csv*, não necessitando de adequações. A partir disso, os tipos de dados foram mapeados explicitamente para *'int'* e *'string'*, dependendo do tipo e o processo de limpeza, remoção de valores inválidos, normalização e padronização de escalas foi aplicado.

3.3 Fluxo de trabalho

A etapa de pré-processamento dos dados no conjunto IoT-23 abordou transformação para adequação do formato de arquivos na exportação, limpeza e remoção de dados inválidos, alteração de tipos, normalização de escalas e reestruturação das features de rotulação para incluir informações sobre ataques em apenas uma *feature*. Este processo foi aplicado na primeira fase deste trabalho, então não se encontra exemplificada na imagem do fluxo.

A partir dos dados pré-processados, o fluxo de desenvolvimento do trabalho abrange 5 fases principais. Na primeira etapa, a CGAN é desenvolvida e treinada. Após construídos ambos os modelos, gerador e discriminador, o modelo gerador recebe um ruído de entrada e gera amostras a partir dele, que o discriminador definirá como reais ou falsas. Esse

processo se repete para todas as amostras geradas, enquanto o gerador tenta gerar amostras falsas que enganam o discriminador. Na segunda etapa, após treinar a CGAN, o ruído e os rótulos falsos são dados como entrada para o gerador gerar amostras, que, na etapa 3, são integradas aos dados de treinamento real, resultando em um novo conjunto de treinamento. Por fim, esse novo conjunto é a entrada para o treinamento do modelo classificador, que é novamente treinado com amostras adulteradas. Dessa forma, ao sofrer ataques adversariais, conforme exemplificado na etapa 5, o classificador sofre menos perda de robustez ao realizar a classificação de ataques. O fluxo desenvolvido é demonstrado na Figura 3.

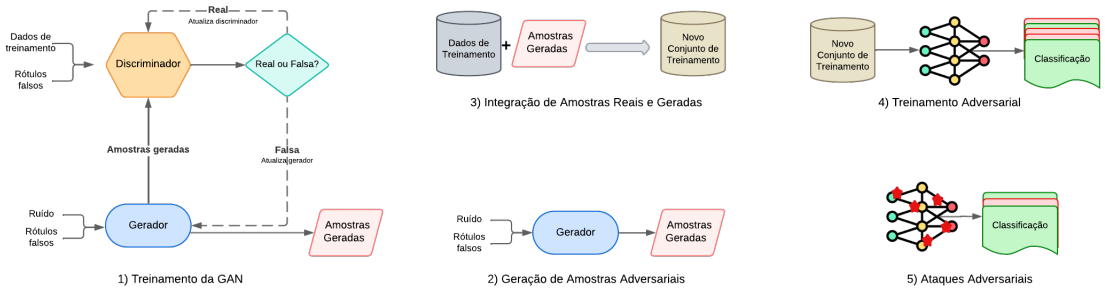

Figura 3 – Fluxo de trabalho exemplificado em etapas 1, 2, 3, 4 e 5

3.4 Métricas

Para a avaliação dos resultados, foram empregadas quatro métricas [32] estatísticas: acurácia, precisão, *recall* e *f1-score*. Acurácia (3.1) mede a proporção de previsões corretas (verdadeiros positivos e verdadeiros negativos) entre o número total de casos examinados. Precisão (3.2) mede a fração de previsões positivas que realmente pertencem ao conjunto de previsões positivas. *Recall* (3.3) quantifica o número de previsões positivas feitas entre todos os exemplos positivos no conjunto de dados. O *f1-score* (3.4) é uma pontuação única que combina precisão e *recall*, definida como a média harmônica da precisão e do *recall* de um modelo. Para o cálculo dessas métricas, consideram-se os seguintes valores: Verdadeiro Positivo (VP) é a amostra de tráfego malicioso classificado como malicioso, Verdadeiro Negativo (VN) é a amostra de tráfego normal classificado como normal, Falso Positivo (FP) é a amostra de tráfego normal classificado como malicioso, e Falso Negativo (FN) é a amostra de tráfego malicioso classificado como normal.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.1}$$

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{3.2}$$

$$\text{Recall} = \frac{VP}{VP + FN} \tag{3.3}$$

$$\text{F1-Score} = \frac{2 \cdot \text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{3.4}$$

3.5 Resumo

Neste capítulo foi apresentada a metodologia adotada para o desenvolvimento dos experimentos. Primeiramente, os conjuntos de dados IoT-23 e TON-IoT foram apresentado e detalhados conforme pré-processamento para aplicação no modelo. Da mesma forma, alguns dos ataques presentes em cada um dos conjuntos foram brevemente descritos. Logo, as abordagens, como treinamento do modelo e geração de amostras adversariais com o método CGAN foram abordadas conforme aplicação.

Além disso, este capítulo também demonstra o fluxo de trabalho envolvido, desde pré-processamento até geração das amostras adversariais e classificações do modelo. Por fim, as métricas empregadas na avaliação dos resultados são apresentadas, sendo acurácia, precisão, *recall* e *f1-score*, considerando valores de verdadeiro positivo e negativo, bem como falso positivo e negativo, que possibilitarão a análise comparativa dos métodos.

4 Experimentos

Neste capítulo são abordadas as configurações necessárias para a condução dos experimentos deste trabalho, bem como da plataforma adotada e recursos empregados. Além disso, são descritos os resultados obtidos com os experimentos e avaliação comparativa usando as métricas quantitativas elencadas.

4.1 Experimentos

No pré-processamento dos conjuntos de dados e treinamento da CGAN, utilizamos uma máquina virtual que dispõe de 24 GB de memória RAM, armazenamento de 300 GB e 24 núcleos virtuais de CPU. Para os treinamentos do modelo, o código utilizado foi inspirado no código disponibilizado pelo estudo [33] e disponível em um repositório público do GitHub[1].

Os modelos foram testados utilizando a plataforma Google Colaboratory, um serviço do Jupyter Notebook hospedado que não requer configuração para uso e oferece acesso gratuito a recursos de computação. A ferramente disponibiliza uma máquina virtual em nuvem, contendo 12,7 GB de memória RAM, 107,72 GB de armazenamento e uma GPU de back-end do Google Compute Engine em Python 3, para conduzir experimentos de Aprendizado de Máquina.

4.2 Resultados

Cada uma das abordagens foi executada duas vezes para avaliação, uma vez para cada um dos conjuntos de dados escolhidos. Para tal, amostras foram coletadas de cada um dos conjuntos, onde 70% dos dados foram utilizados para treino e 30% para testes. Além disso, o modelo classificar foi treinado utilizando 200 épocas. Por fim, as métricas geradas pelos resultados de cada um dos métodos foram extraídas e comparadas entre os diferentes conjuntos de dados.

4.2.1 Avaliação

Os resultados obtidos por cada um dos modelos podem ser visualizados nas Tabelas 1 e 1 e gráficos das Figuras 7, 5, 4 e 6. Ao avaliar estes resultados, é possível perceber que, em ambos os conjuntos de dados, as abordagens FGSM e CGAN deterioraram o modelo de detecção.

[1] https://github.com/mccarthyajb/HL-NTAC

Conjunto de Dados	Abordagem	Acurácia	Precisão	Recall	F1-Score
TON-IoT	Sem amostras adversariais	61,15%	61,63%	75,44%	55,29%
	FGSM	58,08%	58,61%	75,66%	50,05%
	CGAN	50,16%	61,63%	75,44%	55,29%
IoT-23	Sem amostras adversariais	99,97%	71,42%	71,32%	71,37%
	FGSM	5,31%	1,11%	9,16%	1,98%
	CGAN	14,30%	14,30%	9,96%	7,39%

Tabela 1 – Comparação de desempenho para treinamento convencional nos conjuntos de dados IoT-23 e TON-IoT.

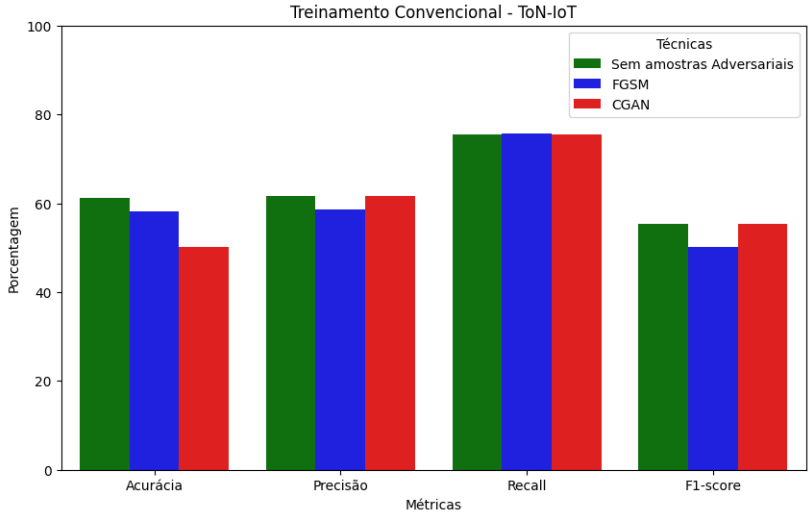

Figura 4 – Comparação de desempenho de cada técnica no treinamento convencional do conjunto de dados TON-IoT.

Os resultados apresentados na Tabela 1 e das Figuras 6 e 5 mostram o desempenho do treinamento convencional do modelo classificador nos conjuntos de dados TON-IoT e IoT-23, em três cenários distintos: sem amostras adversariais, com amostras geradas pelo ataque FGSM e com amostras sintéticas geradas por CGAN. Cada um desses cenários afeta de forma específica as métricas de avaliação, como acurácia, precisão, *recall* e *F1-score*.

No conjunto de dados TON-IoT, o modelo obteve uma acurácia de 61,15%, com *recall* de 75,44% e *F1-score* de 55,29% quando não foram utilizadas amostras adversariais. Esses valores indicam que o modelo consegue recuperar boa parte dos ataques (*recall*), porém tem dificuldade em manter um equilíbrio entre precisão e *recall*, o que se reflete em um *F1-score* moderado. Quando se introduziram amostras adversariais geradas via FGSM, a acurácia caiu para 58,08%, e o *F1-score* para 50,05%. Essa redução mostra que o modelo é vulnerável a pequenas perturbações, embora ainda conserve algum nível de robustez. Já as amostras geradas por CGAN resultaram em uma acurácia de 50,16%, mantendo, curiosamente, os mesmos valores de precisão e *recall* obtidos sem as amostras adversariais. Isso sugere que os dados gerados sinteticamente podem ter confundido o modelo, mas não a ponto de alterar todos os indicadores de maneira uniforme.

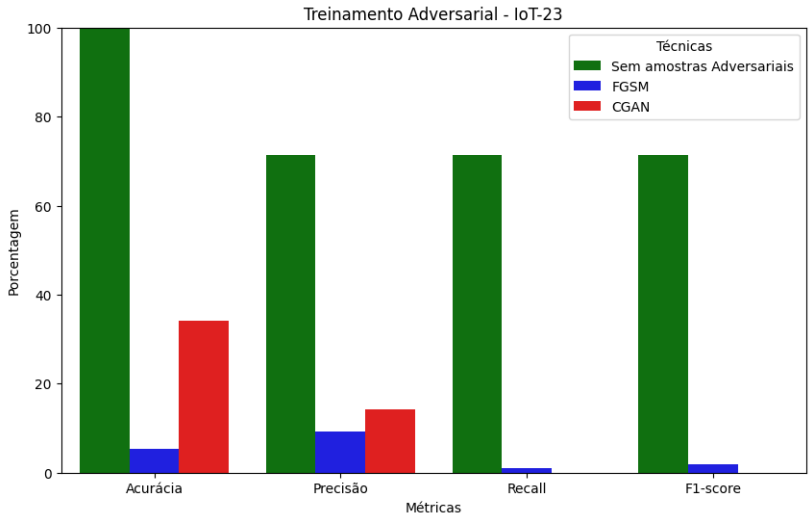

Figura 5 – Comparação de desempenho de cada técnica no treinamento convencional do conjunto de dados IoT-23.

Em contraposição, o conjunto de dados IoT-23 apresenta resultados extremos em cada cenário. Sem amostras adversariais, o modelo atinge 99,97% de acurácia, porém o *F1-score* de 71,37% indica um possível desbalanceamento de classes ou sobreajuste a determinados padrões do *dataset*. Com a introdução de amostras adversariais geradas por FGSM, houve uma queda drástica na acurácia para 5,31% e *F1-score* de 1,98%. Esses números evidenciam que o modelo é extremamente sensível a perturbações adversariais e que o alto desempenho anterior pode mascarar uma falta de robustez. Já a abordagem via CGAN afetou a classificação de forma menos severa em relação ao FGSM, mas ainda assim reduziu a acurácia para 14,30%, com *F1-score* de 7,39%.

Esses resultados mostram que os ataques adversariais podem minar de forma considerável a capacidade de generalização do modelo. No caso do IoT-23, a queda de desempenho diante do FGSM revela uma fragilidade significativa, possivelmente devido a um sobreajuste ou forte dependência de padrões estatísticos específicos. Enquanto isso, no TON-IoT, o modelo apresenta maior resiliência, apesar de ainda ser impactado pelas amostras adversariais.

Por fim, a comparação entre as diferentes abordagens de geração de amostras adversariais sugere que o FGSM é particularmente efetivo em explorar vulnerabilidades do modelo, resultando em quedas mais expressivas de acurácia e *F1-score*. A CGAN, por sua vez, introduz confusão de maneira mais distribuída, o que, em alguns casos, gera um impacto menos agressivo, mas ainda relevante. Tais observações evidenciam a importância de se considerar estratégias defensivas, como o treinamento adversarial e métodos de regularização, para tornar os modelos de detecção de ciberataques mais robustos a variações e manipulações artificiais dos dados.

Os resultados exibidos na Tabela 2 e nas Figuras 4 e 7 referem-se ao treinamento adversarial do modelo classificador de ciberataques nos conjuntos de dados TON-IoT e

Conjunto de Dados	Abordagem	Acurácia	Precisão	Recall	F1-Score
TON-IoT	Sem amostras adversariais	50,58%	51,21%	73,30%	35,76%
	FGSM	66,39%	66,80%	77,67%	62,90%
	CGAN	49,32%	49,96%	45,72%	33,18%
IoT-23	Sem amostras adversariais	99,97%	71,42%	71,32%	71,37%
	FGSM	5,31%	1,11%	9,16%	1,98%
	CGAN	34,22%	14,28%	0%	0%

Tabela 2 – Comparação de desempenho para treinamento adversarial nos conjuntos de dados IoT-23 e TON-IoT.

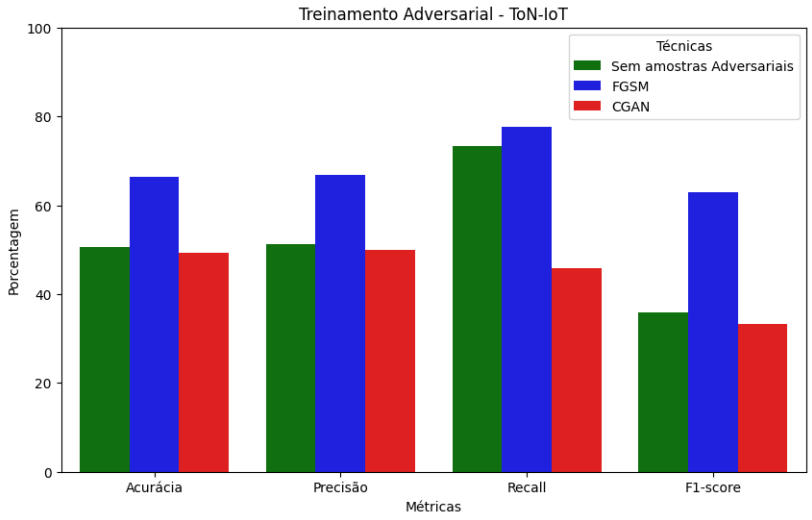

Figura 6 – Comparação de desempenho de cada técnica no treinamento adversarial do conjunto de dados TON-IoT

IoT-23, avaliados em três cenários: sem amostras adversariais, com amostras geradas pelo ataque FGSM e com amostras sintéticas geradas por CGAN. A partir da análise das métricas geradas, é possível entender em que medida o modelo, após ser exposto a técnicas de treinamento adversarial, se torna mais ou menos robusto a perturbações adversariais.

No caso do TON-IoT, observamos que o cenário "sem amostras adversariais" apresentou valores menores de acurácia (50,58%) e *F1-score* (35,76%) quando comparado ao mesmo cenário de treinamento convencional. Isso pode ocorrer porque o treinamento adversarial tende a penalizar o modelo por memorizar padrões muito específicos dos dados originais, fazendo com que ele seja, em geral, menos "sobreajustado" (*overfitted*). Por outro lado, quando o teste é realizado com amostras FGSM, a acurácia aumenta consideravelmente para 66,39%, enquanto a *F1-Score* chega a 62,90%. Esse incremento sugere que o modelo efetivamente aprendeu a lidar melhor com as manipulações introduzidas pelo FGSM durante o processo de treinamento adversarial, tornando-o mais robusto. Já o cenário com CGAN resultou em uma queda de desempenho (49,32% de acurácia e 33,18% de *F1-Score*), possivelmente porque as amostras sintéticas geradas não seguiram um padrão de perturbação tão sistemático quanto o FGSM, dificultando a adaptação do modelo ou

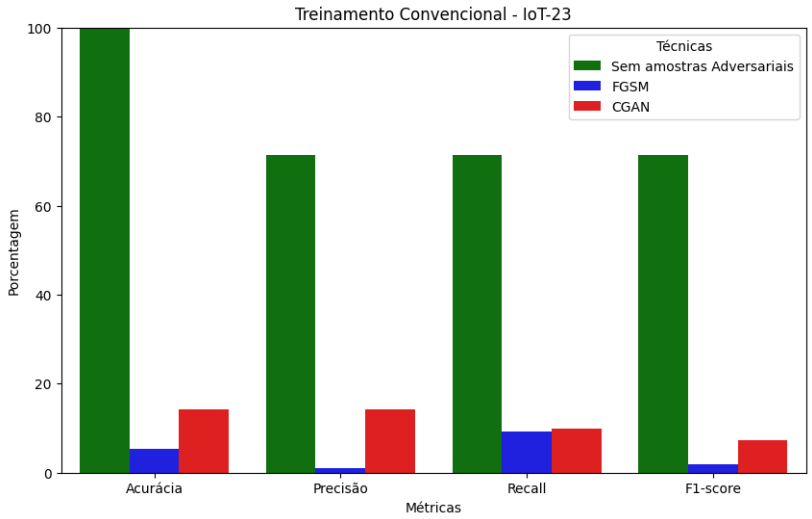

Figura 7 – Comparação de desempenho de cada técnica no treinamento adversarial do conjunto de dados IoT-23.

confundindo a rede de forma mais aleatória.

Para o IoT-23, o modelo mantém a acurácia de 99,97% e valores de *F1-Score* em torno de 71% no cenário sem amostras adversariais, repetindo o comportamento do treinamento convencional. No entanto, a introdução de amostras FGSM faz o desempenho cair para 5,31% de acurácia e 1,98% de *F1-Score*, demonstrando novamente que esse conjunto de dados continua extremamente sensível a esse tipo de ataque, mesmo após o treinamento adversarial. Já no cenário com CGAN, a acurácia cai para 34,22%, e o *recall* é 0%, resultando em *F1-Score* também de 0%. Isso indica que o modelo praticamente não reconheceu os exemplos de uma das classes (ou de mais de uma classe), possivelmente devido a um desequilíbrio severo ou por não aprender adequadamente a distribuição das amostras sintéticas geradas.

Os resultados demonstram que o treinamento adversarial aplicado consegue, em alguns casos, melhorar a robustez do modelo (como ocorreu no TON-IoT com o FGSM), mas não garante proteção ampla contra todos os tipos de ataques. No IoT-23, apesar de o desempenho ser ótimo sem ataques, o FGSM e o CGAN ainda expõem fragilidades do modelo. Esse contraste reforça a importância de avaliar a estratégia de geração de amostras adversariais (tipo de ataque, intensidade das perturbações e características do *dataset*) e de adotar mecanismos de defesa adicionais, como regularizações específicas ou métodos de *ensembling*, para lidar com diferentes ameaças no contexto de detecção de ciberataques.

4.3 Resumo

Este capítulo abordou as configurações necessárias para a condução dos experimentos, como do servidor utilizado, além do código tomado como inspiração. Também foram

descritos os experimentos para cada um dos treinamentos, convencional e adversarial, bem como os métodos aplicados, CGANs e FGSM. Assim, o modelo foi treinado para detecção de ataques, geração de amostras adversariais e novo treinamento, a partir de cada abordagem.

As métricas elencadas, acurácia, precisão, *recall* e *f1-score*, apoiaram a análise comparativa da CGAN na geração de dados sintéticos, bem como do Treinamento Adversarial. As tabelas e os gráficos possibilitaram a visualização dos valores alcançados e sua comparação ao modelo sem amostras adversariais.

5 Desafios e Limitações

A geração de dados de rede com uma distribuição próxima àquela observada em ambientes reais representa um dos principais desafios nos experimentos de cibersegurança, principalmente pela complexidade e diversidade dos cenários de tráfego. Assim, recriar artificialmente todas essas variáveis de forma realista requer o desenvolvimento de métodos de geração robustos, capazes de capturar nuances estatísticas e comportamentais dos dados originais. Além disso, a falta de rótulos acarreta complicações tanto no treinamento de classificadores supervisionados quanto na avaliação dos métodos de detecção de ataques.

Outra etapa que gerou dificuldades durante o desenvolvimento do trabalho foi a aplicação de bibliotecas para lidar com ciência de dados e aprendizado de máquina, como *numpy*, *pandas*, *torch*, *keras* e *tensorflow*. Por se tratar de novas tecnologias, além da integração com o contexto de cibersegurança, não foi trivial encontrar compatibilidade entre todas as ferramentas, o que gerou muitas inconsistências durante o processo.

No que diz respeito ao uso de modelos generativos, a aplicação de redes GAN tradicionais (sem condicionalidade) se mostrou desafiadora neste contexto, pois a geração de dados não rotulados ou com rótulos implícitos demanda abordagens de classificação semi-supervisionadas ou não supervisionadas.

Diante dessas limitações, a adoção do modelo condicional com a Conditional GAN (CGAN), surgiu como uma alternativa mais apropriada para a geração de amostras rotuladas de ataques. Nessa abordagem, é possível direcionar o processo de geração com base nos rótulos de ataques, facilitando a criação de dados mais representativos e balanceados. Porém, configurar adequadamente essas CGANs envolve diversos parâmetros, como a escolha de arquiteturas de rede, funções de perda e hiperparâmetros de treinamento (taxa de aprendizado, tamanho de lote, etc.). Ajustes inadequados podem levar a problemas como baixa diversidade de amostras ou discrepâncias significativas em relação à distribuição original dos dados.

Em suma, os obstáculos encontrados, como a dificuldade em reproduzir o comportamento de tráfego real e complexidade de configuração de redes GAN convencionais ou condicionais, mostram a necessidade de pesquisas mais aprofundadas e metodologias mais elaboradas, que combinem técnicas de geração, rotulagem e validação cuidadosa dos dados para garantir a eficácia e a confiabilidade de sistemas de detecção de ciberataques.

6 Conclusões e Trabalhos Futuros

Este trabalho possibilitou a investigação sobre a classificação de ciberataques em ambientes de IoT, explorando diferentes conjuntos de dados (IoT-23 e TON-IoT) e abordagens de treinamento, incluindo métodos convencionais e adversariais, assim como solidificar os conhecimentos sobre diferentes ataques cibernéticos, tais como ataques convencionais e ataques específicos de ML. A aplicação do Treinamento Adversarial, permite conhecer o desempenho do modelo em cenários de pior caso, conhecendo pontos de vulnerabilidade dos atacantes.

Neste contexto, buscamos avaliar o desempenho de modelos classificadores de ciberataques sem a inserção de amostras adversariais e posteriormente a influência das amostras geradas por técnicas adversariais, como FGSM e CGANs, revelando que perturbações e manipulações específicas podem reduzir drasticamente o desempenho do classificador. Os resultados evidenciam a importância de aplicar treinamento adversarial para tornar o modelo mais robusto a ataques, embora a eficácia varie conforme a técnica de geração utilizada e as características do *dataset*.

Durante o processo, foi possível identificar alguns desafios e limitações que dificultam a reprodução e avaliação de cenários de ciberataques em redes IoT. Dentre eles, destacamos a dificuldade na coleta e rotulação de dados reais em larga escala, a necessidade de metodologias para garantir que os dados sintéticos reflitam situações de ataque factíveis e, principalmente, a complexidade de configurar modelos generativos condicionais que sejam efetivamente compatíveis com a distribuição dos dados originais. Esses aspectos apontam para a necessidade de pesquisas mais aprofundadas na intersecção entre cibersegurança, ciência de dados e aprendizado de máquina.

Por fim, com este trabalho reforçamos que a aplicação de técnicas adversariais e métodos de geração de dados sintéticos desempenha um papel relevante na busca por modelos mais resilientes a ataques. Embora haja ganhos práticos e teóricos, ainda há muito espaço para aprimoramentos, indicando que a investigação acerca da robustez e eficácia de classificadores em diferentes cenários de IoT permanece em aberto para novas descobertas e inovações.

Referências

[1] KEMMERER, R. A. Cybersecurity. In: IEEE. *25th International Conference on Software Engineering, 2003. Proceedings.* [S.l.], 2003. p. 705–715.

[2] ARIFIN, M. M. et al. A survey on the application of generative adversarial networks in cybersecurity: Prospective, direction and open research scopes. *arXiv preprint arXiv:2407.08839*, 2024.

[3] KHRAISAT, A. et al. Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity*, Springer, v. 2, n. 1, p. 1–22, 2019.

[4] SZEGEDY, C. et al. *Intriguing properties of neural networks*. 2014.

[5] TIAN, Z. et al. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, ACM New York, NY, v. 55, n. 8, p. 1–35, 2022.

[6] GONG, X. et al. Model extraction attacks and defenses on cloud-based machine learning models. *IEEE Communications Magazine*, IEEE, v. 58, n. 12, p. 83–89, 2020.

[7] AYUB, M. A. et al. Model evasion attack on intrusion detection systems using adversarial machine learning. In: IEEE. *2020 54th annual conference on information sciences and systems (CISS)*. [S.l.], 2020. p. 1–6.

[8] ASLAN, A.; SAMET, R. A comprehensive review on malware detection approaches. *IEEE Access*, v. 8, p. 6249–6271, 2020.

[9] CHEN, L.; YE, Y. Secmd: Make machine learning more secure against adversarial malware attacks. In: PENG, W.; ALAHAKOON, D.; LI, X. (Ed.). *AI 2017: Advances in Artificial Intelligence*. Cham: Springer International Publishing, 2017. p. 76–89. ISBN 978-3-319-63004-5.

[10] BARRENO, M. et al. The security of machine learning. *Mach. Learn.*, Springer Science and Business Media LLC, v. 81, n. 2, p. 121–148, nov. 2010.

[11] CHEN, L.; YE, Y.; BOURLAI, T. Adversarial machine learning in malware detection: Arms race between evasion attack and defense. In: *2017 European Intelligence and Security Informatics Conference (EISIC)*. [S.l.: s.n.], 2017. p. 99–106.

[12] LIU, J. et al. Adversarial machine learning: A multilayer review of the state-of-the-art and challenges for wireless and mobile systems. *IEEE Communications Surveys Tutorials*, v. 24, n. 1, p. 123–159, 2022.

[13] HUANG, L. et al. Adversarial machine learning. In: *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*. New York, NY, USA: Association for Computing Machinery, 2011. (AISec '11), p. 43–58. ISBN 9781450310031. Disponível em: <https://doi.org/10.1145/2046684.2046692>.

[14] APRUZZESE, G. et al. Addressing adversarial attacks against security systems based on machine learning. In: *2019 11th International Conference on Cyber Conflict (CyCon)*. [S.l.: s.n.], 2019. v. 900, p. 1–18.

[15] BIGGIO, B. et al. Security evaluation of support vector machines in adversarial environments. *CoRR*, abs/1401.7727, 2014. Disponível em: <http://arxiv.org/abs/1401.7727>.

[16] GOODFELLOW, I. J. et al. *Generative Adversarial Networks*. 2014. Disponível em: <https://arxiv.org/abs/1406.2661>.

[17] PAN, Z. et al. Recent progress on generative adversarial networks (gans): A survey. *IEEE Access*, v. 7, p. 36322–36333, 2019.

[18] WANG, K. et al. Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, v. 4, n. 4, p. 588–598, 2017.

[19] MIRZA, M.; OSINDERO, S. *Conditional Generative Adversarial Nets*. 2014. Disponível em: <https://arxiv.org/abs/1411.1784>.

[20] GOODFELLOW, I. J.; SHLENS, J.; SZEGEDY, C. *Explaining and Harnessing Adversarial Examples*. 2015.

[21] GOODFELLOW, I. J.; SHLENS, J.; SZEGEDY, C. *Explaining and Harnessing Adversarial Examples*. 2015. Disponível em: <https://arxiv.org/abs/1412.6572>.

[22] MADRY, A. et al. *Towards Deep Learning Models Resistant to Adversarial Attacks*. 2019. Disponível em: <https://arxiv.org/abs/1706.06083>.

[23] ZHANG, H. et al. Theoretically principled trade-off between robustness and accuracy. In: CHAUDHURI, K.; SALAKHUTDINOV, R. (Ed.). *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019. (Proceedings of Machine Learning Research, v. 97), p. 7472–7482. Disponível em: <https://proceedings.mlr.press/v97/zhang19p.html>.

[24] PANG, T. et al. Bag of tricks for adversarial training. *arXiv preprint arXiv:2010.00467*, 2020.

[25] ROSENBERG, I. et al. *Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain*. 2021.

[26] ZHOU, B.; KRÄHENBÜHL, P. Don't let your discriminator be fooled. In: *International conference on learning representations*. [S.l.: s.n.], 2018.

[27] MAO, X. et al. Least squares generative adversarial networks. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2017. p. 2794–2802.

[28] ARJOVSKY, M.; CHINTALA, S.; BOTTOU, L. *Wasserstein GAN*. 2017. Disponível em: <https://arxiv.org/abs/1701.07875>.

[29] AGRAWAL, G.; KAUR, A.; MYNENI, S. A review of generative models in generating synthetic attack data for cybersecurity. *Electronics*, v. 13, n. 2, 2024. ISSN 2079-9292. Disponível em: <https://www.mdpi.com/2079-9292/13/2/322>.

[30] GARCIA, S.; PARMISANO, A.; ERQUIAGA, M. J. *IoT-23: A labeled dataset with malicious and benign IoT network traffic*. [S.l.]: Zenodo, 2020.

[31] ALSAEDI, A. et al. $\mathrm{Ton}_i ottelemetrydataset : Anewgenerationdatasetofiotandiiotfordata-drivenintrusiondetectionsystems.IEEE\ Access, v.$ $-165150, 2020.$

[32] GERON, A. *Hands-on machine learning with scikit-learn, keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. 2. ed. Sebastopol, CA: O'Reilly Media, 2019.

[33] MCCARTHY, A. et al. Defending against adversarial machine learning attacks using hierarchical learning: A case study on network traffic attack classification. *Journal of Information Security and Applications*, Elsevier, v. 72, p. 103398, 2023.

[34] S, K.; DURGADEVI, M. Generative adversarial network (gan): a general review on different variants of gan and applications. In: *2021 6th International Conference on Communication and Electronics Systems (ICCES)*. [S.l.: s.n.], 2021. p. 1–8.

[35] AKHTAR, N.; MIAN, A. Threat of adversarial attacks on deep learning in computer vision: A survey. *CoRR*, abs/1801.00553, 2018. Disponível em: <http://arxiv.org/abs/1801.00553>.

[36] ASHOOR, A. S.; GORE, S. Importance of intrusion detection system (ids). *International Journal of Scientific and Engineering Research*, v. 2, n. 1, p. 1–4, 2011.

[37] BACE, R. G. *Intrusion detection*. [S.l.]: Sams Publishing, 2000.