Comparação do Impacto de Ataques Adversariais Contra Modelo de Classificação baseado em ML

Mateus Pelloso[1,3], **Michele Nogueira** [1,2]

[1]Departamento de Informática
Universidade Federal do Paraná (UFPR)

[2]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)

[3]Instituto Federal Catarinense (IFC)

mateus.pelloso@ifc.edu.br, michele@dcc.ufmg.br

Abstract. *Adversarial attacks pose an imminent risk to solutions based on Artificial Intelligence. The primary characteristic is to induce malfunctioning in machine learning models through the generation of adversarial samples. Cybersecurity solutions benefit from these models to classify Internet-trafficked data, both benign and attack-oriented. In this context, the proposal of this study is to evaluate techniques for generating adversarial samples, understand their dynamics, and the impact caused against a machine learning model. The evaluation was based on indicators such as accuracy, precision, recall, and F1-score. The scenario used consists of training and validating a neural network-based model in association with the CIC-IDS2017 dataset. The evaluated techniques were successful in reducing the robustness of the traffic classifier model, reducing accuracy from approximately 93% to 7%.*

Resumo. *Os ataques adversariais são um risco iminente para soluções baseadas em Inteligência Artificial. A principal característica é provocar o mau funcionamento de modelos de aprendizado de máquina por meio da geração de amostras adversariais. As soluções de cibersegurança se beneficiam desses modelos para classificar dados trafegados na Internet, tanto benignos quanto de ataque. Neste contexto, a proposta deste estudo é avaliar técnicas geradoras de amostras adversárias, compreender sua dinâmica e o impacto causado contra um modelo de aprendizado de máquina. A avaliação tomou por base indicadores como accuracy, precision, recall e f1-score. O cenário utilizado consiste no treinamento e validação de um modelo baseado em rede neural em associação com o conjunto de dados CIC-IDS2017. As técnicas avaliadas se apresentaram efetivas em diminuir a robustez do modelo classificador de tráfego, decrescendo a acurácia de 93% para 7%, aproximadamente.*

1. Introdução

Os ataques adversariais são uma grande ameaça às soluções fundamentadas em Inteligência Artificial (IA), principalmente quando aplicadas em cibersegurança [Comiter 2019]. Esses ataques provocam o mau funcionamento de modelos de Aprendizado de Máquina (do inglês: *Machine Learning* - ML) pela

geração de amostras adversariais submetidas ao modelo ML. Esse ataque é capaz de manipular as soluções baseadas em IA para alterar seu comportamento e servir a objetivos maliciosos [Liu et al. 2021]. À medida que soluções de IA são cada vez mais integradas em componentes críticos da sociedade, estes ataques representam uma vulnerabilidade emergente e sistemática, com potencial para causar danos significativos na cibersegurança [Comiter 2019].

As amostras adversariais têm a capacidade de manipular os classificadores baseados em ML, possibilitando aos atacantes se evadir de sistemas de segurança que implementam tais modelos [Adesina et al. 2022], a exemplo de Sistemas de Detecção de Intrusão (do inglês: *Intrusion Detection Systems* - IDSs) ou Sistemas de Prevenção de Intrusão (do inglês: *Intrusion Prevention Systems* - IPSs). Os modelos de aprendizado de máquina possuem limitações e vulnerabilidades, como a sensibilidade aos exemplos adversários e a transferência de amostras adversariais, que atacantes exploram para comprometer sua robustez, especialmente em tarefas de classificação de tráfego de rede, evadindo-se das ferramentas de defesa que utilizam ML. Assim sendo, o problema objeto deste estudo são os ciberataques baseados em aprendizado de máquina adversarial, em especial as abordagens aplicadas na geração das amostras adversariais e, assim, avaliar os efeitos dos ataques em soluções de segurança cibernética, como ferramentas de detecção de ataques. Compreender como os ataques adversariais podem comprometer a eficácia desses sistemas é fundamental para fortalecer as defesas cibernéticas.

Os estudos [Alshahrani et al. 2022, McCarthy et al. 2023, Shieh et al. 2022a, Shieh et al. 2022b, Shroff et al. 2022] discutem a aplicação de exemplos adversariais contra soluções de cibersegurança. Em geral, demonstram uma das abordagens baseadas em Redes Adversariais Generativas (do inglês: *Generative Adversarial Networks* - GANs) para a geração de amostras adversariais sintéticas. Apenas a pesquisa de [McCarthy et al. 2023] aplica a *Jacobian-based Saliency Map Attack* (JSMA). Nessas, as amostras utilizadas contra modelos de ataques impactam os classificadores e conduzem a erros, reduzindo sua robustez. A partir disso, os autores retreinam os modelos incluindo os exemplos adversariais como entrada e, assim, ocorre alguma recuperação da robustez. Contudo, persiste a lacuna da avaliação do impacto quanto a aplicação de outras abordagens geradoras de amostras adversariais.

Os exemplos adversariais tem como base inúmeras técnicas geradoras, que se dividem em dois grupos: GANs e Non-GANs [Liu et al. 2021]. As técnicas geradoras de amostras adversariais selecionadas, que fazem parte do grupo das Non-GANs, são: *Fast Gradient Sign Method* (FGSM), JSMA e *Carlini & Wagner* (C&W). Essa seleção é motivada com objetivo de compreender um conjunto de comportamentos que poderão evidenciar se (*i*) as amostras adversariais geradas por essas técnicas causam algum impacto contra modelo baseado em rede neural e (*ii*) qual a dimensão do impacto causado pela aplicação dessas técnicas.

Neste contexto, este artigo apresenta os resultados de uma comparação avaliativa entre as abordagens e técnicas evidenciadas para a geração de amostras adversariais, bem como seu impacto junto ao classificador. Entre as contribuições, destaca-se a identificação de vulnerabilidades específicas. O entendimento de como as amostras adversariais podem impactar os modelos abre perspectivas para implementação de contramedidas para mitigar os efeitos e, ainda, fornecer *insights* para a construção de estratégias de defesa mais robus-

tas. Dessa forma, pesquisadores, profissionais de segurança cibernética, desenvolvedores de sistemas baseados em IA e empresas que dependem desses sistemas para proteção contra ameaças cibernéticas estão entre os beneficiados desta pesquisa. A avaliação da aplicação das técnicas geradoras das amostras adversariais ocorre com base em métricas como *Accuracy*, *Precision*, *Recall*, *F1-score*, no uso do classificador implementado por meio de uma rede neural e por meio do conjunto de dados CIC-IDS2017.

O artigo está estruturado como segue. A Seção 2 contempla os trabalhos relacionados discutindo os principais aspectos relativos à aplicação das técnicas de geração de amostras adversariais. A fundamentação é apresentada na Seção 3 e considera o aprendizado de máquina adversário e a taxonomia dos ataques adversariais juntamente com a descrição dos métodos de geração das amostras adversariais. A Seção 4 dispõe sobre a avaliação por meio da exposição da metodologia, da aplicação das abordagens geradoras das amostras adversariais, das características do conjunto de dados e da rede neural (classificador) e apresenta os resultados.

2. Trabalhos Relacionados

Conforme [Alshahrani et al. 2022], por meio das GANs, ataques adversariais submetidos contra soluções de detecção de intrusão baseadas em aprendizado de máquina causam o declínio da robustez do modelo classificador. O cenário proposto pelos autores consiste em dois tipos de ataques adversariais, envenenamento e evasão, submetidos contra dois modelos de aprendizado de máquina, árvore de decisão e regressão logística. A avaliação de desempenho dos modelos antes e depois dos ataques utilizou o conjunto de dados CIC-IDS2017. Os resultados apontaram que a precisão, tanto dos classificadores de Árvore de Decisão quanto de Regressão Logística, foi afetada, alcançando 94% e 96% no cenário de ataque de evasão, e 97% e 95% no cenário de ataque de envenenamento, respectivamente. Contudo, as redes GANs têm como característica o uso de dois modelos, o generativo e o discriminatório. Como ambos demandam esforço computacional para seus respectivos treinamentos, há a tendência de que esse ataque seja mais custoso. Porém, ainda não é possível afirmar se esse maior custo, considerando o uso do modelo discriminatório, torna-se mais eficiente que outras abordagens.

No estudo de [McCarthy et al. 2023], aplicaram a JSMA como abordagem para gerar exemplos adversariais a serem submetidos contra modelos classificadores de ataques baseados em aprendizado de máquina. De acordo com os autores, o ataque obteve êxito utilizando tanto a abordagem de ataque direcionado (influencia o modelo a errar para uma classe alvo específica) quanto não direcionado. Contudo, o estudo informa apenas que o ataque direcionado obteve 92% de êxito, não mencionando as métricas do uso do JSMA no ataque não direcionado. Dado que a pesquisa propõe o uso da hierarquização dos dados como estratégia para melhorar o desempenho de modelos aplicados na detecção de ataques, é necessário avaliar as demais abordagens para a geração de exemplos adversariais. Dessa forma, haverá possibilidade de avaliar adequadamente a solução proposta neste estudo.

Na pesquisa de [Shieh et al. 2022b], os experimentos demonstraram que os modelos de ML *Random Forest* (RF), *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM) e *Naïve Bayes* (NB) apresentaram resultados positivos ao serem aplicados na detecção de ataques DDoS com base em conjunto de dados livre de amostras adver-

sariais. Porém, ao aplicar a CycleGAN para gerar um conjunto de dados com amostras adversariais e novamente submeter aos modelos citados, os resultados de *True Positive Rate* (TPR) e *F1-score* sofrem significativa degradação. Para fins de comparação, a taxa dos TPRs para RF, KNN, SVM, NB sem os exemplos adversariais foi de 92%, 92%, 91% e 80% respectivamente, enquanto os mesmos modelos apresentaram TPRs de 17%, 15%, 13% e 9% na mesma ordem, ao tentar classificar os dados que continham amostras adversariais. O estudo propôs como solução o uso de uma *Symmetric Defense Generative Adversarial Network* (SDGAN) que se mostrou eficiente, reestabelecendo a classificação próxima dos patamares anteriores aos exemplos adversários. Ainda assim, destaca-se a importância de avaliar outras abordagens geradoras de amostras adversariais com a finalidade de verificar a manutenção de resultados ou ainda identificar potenciais riscos.

De acordo com [Shieh et al. 2022a], através do *framework GAN Dual Discriminators* (GANDD) proposto, os autores demonstraram que é possível restabelecer a robustez dos modelos classificadores de ataques. O cenário de avaliação consiste no treinamento de modelos baseados em aprendizado de máquina com a finalidade de classificar os ataques. Dessa forma, expuseram que ataques com exemplos adversariais são efetivos contra esses modelos e, portanto, reduzem a capacidade de os classificarem corretamente. Nesse estudo foi aplicado o *Wasserstein Generative Adversarial Networks* (WGAN) com *Gradient Penalty* (GP-WGAN) para a geração das amostras adversariais e avaliado contra modelos RF, KNN e SVM, que apresentaram decréscimo de 94% para 9%, 91% para 6% e de 87% para 0,4% aproximadamente em seus respectivos TPRs. Esse é outro exemplo de pesquisa que avaliou o modelo classificador original, sendo depois retreinado com exemplos adversariais com apenas uma abordagem geradora.

Em [Shroff et al. 2022], propõe-se o uso de GANs, em especial a WGAN-GP. A proposta é utilizar a WGAN-GP como ferramenta para gerar instâncias de dados benignas e de dados maliciosos (ex. DDoS) e avaliar o modelo usado para detectar DDoS. Com base nessa avaliação, os modelos detectaram adequadamente os dados benignos e de ataque. Ao manipular características específicas nas instâncias benignas para caracterizá-las como de dados maliciosos, os classificadores se tornaram ineficientes, possibilitando a evasão do modelo. Ao retreinar o modelo, considerando os dados sintéticos tanto benignos quanto maliciosos gerados pela GAN, é reestabelecida a robustez do classificador. Contudo, nesse estudo foi avaliado apenas DDoS volumétrico, e utilizada apenas a variação da rede GAN conhecida como WGAN-GP. Permanecem possibilidades de estudos para além dos ataques DDoS volumétricos, diferentes *features* e, em especial, outras técnicas de Aprendizado de Máquina Adversário (do inglês: *Adversarial Machine Learning* (AML).

Os trabalhos discutidos demonstram que o AML, por meio das GANs, impacta nos modelos baseados em ML prejudicando a robustez. Esses modelos estão presentes em diversas tecnologias baseadas em IA para aplicação na área de Cibersegurança. Em geral, os estudos focam apenas uma abordagem geradora de exemplos adversários, e por isso é necessário ampliar os esforços para desenvolver estudos por meio de experimentos que demonstrem outras abordagens e seus respectivos impactos. E, a partir disso, avaliar adequadamente soluções e contramedidas aos ataques adversariais.

3. Fundamentação

Esta seção apresenta a fundamentação dos principais conceitos referentes às métricas avaliadas, ao Aprendizado de Máquina Adversarial, à taxonomia e às características elementares dos ataques adversariais, junto aos métodos de geração de amostras adversariais.

3.1. Indicadores de desempenho

Os indicadores Matriz de Confusão, *Accuracy*, *Precision*, *Recall* e *F1-score* do modelo alvo baseados em aprendizado de máquina foram usados para medir o desempenho do ataque adversário. A matriz de confusão consiste em uma abordagem de avaliação de desempenho de modelos de aprendizado utilizada para verificar a classificação de um conjunto de dados realizado por um modelo. A ideia é contabilizar a quantidade de vezes que um dado que pertence a uma classe, por exemplo, de ataque DoS, é classificado corretamente e/ou como pertencente a outra classe de forma equivocada. Portanto, a matriz de confusão é uma tabela que permite analisar o desempenho do modelo, analisando de que forma cada dado foi classificado, se certo ou errado [Géron 2019].

Existem algumas formas de representar uma matriz de confusão. As duas mais usuais são como mapa de calor, por meio da plotagem dos valores da matriz usando diferentes intensidades de cores (especialmente ao se tratar de múltiplas classes), ou por meio de uma matriz em que os elementos são preenchidos com o número de vezes que as amostras foram classificadas em uma classe específica. Ainda assim, se faz necessário utilizar métricas mais concisas, ou seja, que possibilitem comparações mais diretas por um único valor numérico, como exemplo, *Accuracy* e *F1-score*. A *Accuracy* é definida como a proporção de predições totais corretas (ver a Equação 1). A *Precision* refere-se a uma taxa de amostras positivas de todas as amostras previstas como positivas pelos classificadores (ver a Equação 2). A *Recall* refere-se à taxa de amostras positivas que são previstas corretamente a partir de todas as amostras previstas corretamente (ver a Equação 3) e *F1-score* é a média ponderada de *Precision* e *Recall*, que é construída usando a média harmônica de *Precision* e *Recall* (ver Equação 4) [Géron 2019].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

3.2. Aprendizado de Máquina Adversário

De acordo com [Liu et al. 2021], o Aprendizado de Máquina Adversário é uma subárea do Aprendizado de Máquina. Ou ainda, pode-se entender que AML é uma técnica do ML. Em geral, o AML é uma abordagem que busca provocar o mau funcionamento de um determinado modelo de aprendizado de máquina, por meio da aplicação/submissão de informações incorretas ou enganosas. Ou seja, compromete o funcionamento de determinado modelo por meio da inserção de perturbações, conhecidas também como amostras ou exemplos adversariais. Conforme [Shroff et al. 2022], a possibilidade de confundir determinados modelos de aprendizado de máquina baseados em classificação se deve a

tais algoritmos serem implementados para atuar em conjuntos de problemas específicos, em que os dados de treinamento e teste são produzidos a partir da mesma distribuição estatística.

No momento em que são aplicados em ambientes reais, os adversários disponibilizam dados que quebram o modelo em uso. Com isso, os dados utilizados pelos adversários são organizados para explorar vulnerabilidades específicas dos modelos, comprometendo os resultados da classificação. Neste contexto, o AML possibilita o estudo dos efeitos da inserção de amostras no fluxo de dados com a finalidade de avaliar a robustez dos modelos de ML e, assim, garantir o adequado funcionamento. Existem diversos métodos aplicados pelo AML, entre os quais pode-se citar a GAN, WGANs, *Zeroth order optimization* (ZOO), *Limited-Memory Broyden-Fletcher-Goldfarb-Shanno* (L-BFGS), FGSM, JSMA, Deepfool, C&W, entre outros.

3.2.1. Taxonomia dos Ataques Adversariais

Segundo [Liu et al. 2021, Rosenberg et al. 2021, Zhou et al. 2022], os ataques baseados em AML são classificados de acordo com objetivo, direcionamento e informações que o atacante possui sobre o alvo. A nomenclatura dos ataques, conforme o objetivo, pode ser de Inferência, *Membership Inference*, Evasão, Envenenamento e Trojan. Inferência refere-se ao adversário predizer o comportamento do modelo alvo; *Membership Inference* o adversário prevê as amostras utilizadas no treinamento do modelo alvo; Evasão consiste no atacante adicionar perturbações capazes de fazer o alvo classificar como dados benignos; Evenenamento implica na injeção de amostras adversariais ou alterar amostras legítimas durante o processo de treinamento do modelo; Trojan indica um ataque iniciado como envenenamento, e que também inclui a inserção de uma *backdoor* no alvo.

O ataque adversarial é direcionado (*targeted attack*) quando tem o objetivo de fazer o classificador apresentar uma resposta especificada pelo atacante e comprometer sua robustez. O ataque não-direcionado (*untargeted attack*) objetiva que o classificador não rotule corretamente, independente do rótulo atribuído [Liu et al. 2021, Rosenberg et al. 2021]. Outra abordagem é relacionada ao nível de conhecimento que o atacante possui a respeito do alvo. Esses níveis são denominados *black-box*, *gray-box* ou *white-box*, conforme caracterização a seguir:

- *Black-box*: é desenvolvido após o *deploy* do modelo, ou seja, quando está em produção, o atacante realiza consultas ao alvo visando determinar as vulnerabilidades existentes a explorar. O atacante não possui informações prévias sobre o modelo alvo, como parâmetros ou pesos, abordagem do treinamento, dados ou arquitetura. Como exemplo, o modelo realiza um ataque contra um IDS alvo usando o retorno (falha ou sucesso) para treinar o gerador sem utilizar outras informações do alvo [Liu et al. 2021, Rosenberg et al. 2021].
- *Gray-box*: assume que o atacante possui um conhecimento parcial sobre o modelo alvo. Ou seja, quando o adversário conhece alguns elementos como as características (*features*) utilizadas durante o processo de treinamento, mas não todos os elementos relativos ao alvo [McCarthy et al. 2023, Rosenberg et al. 2021].
- *White-box*: o ataque somente é considerado *white-box* se o atacante estiver consciente dos parâmetros e pesos do modelo. Além disso, em algumas ocorrências,

o atacante conhece também o processo de treinamento, bem como os dados utilizados. Ou seja, os detalhes das etapas de pré-processamento, do modelo ML e hiperparâmetros. Assim sendo, para o *white-box* é necessário acesso ao fluxo de dados normal da rede [Liu et al. 2021, Rosenberg et al. 2021].

3.3. Métodos de Geração de Amostras Adversariais

Além das maneiras de classificar ataques adversariais, é importante observar que as técnicas utilizadas para gerar ruídos nas amostras estão divididas em dois conjuntos. De acordo com [Liu et al. 2021] esses métodos do AML são denominados GANs ou Non-GANs. A principal distinção entre eles é que as GANs são abordagens baseadas em redes neurais generativas, enquanto as Non-GANs não estão relacionadas ao ML ou DL.

A fim de ilustrar, de forma não exaustiva, alguns dos métodos geradores de amostras adversariais classificados como GANs na literatura estão GANs, WGANs, *Auto Encoder* (AE), *Variational Auto Encoder* (VAE). São identificadas também as Non-GANs, C&W, FGSM, JSMA, L-BFGS e ZOO. Segundo [Goodfellow et al. 2014, Géron 2019], uma GAN consiste em duas redes neurais, uma generativa e uma discriminatória. A rede generativa tem como objetivo gerar dados semelhantes aos de treinamento, enquanto a discriminatória busca distinguir os dados reais dos dados sintéticos.

As técnicas C&W, FGSM e JSMA foram selecionadas para compor este estudo. As motivações desta seleção incluem (*i*) a facilidade de aplicação, dado que estão implementados em bibliotecas de linguagens bastante difundidas, como Python - o que favorece o uso por atacantes em potencial, (*ii*) baixo custo computacional quando comparado com as GANs ao considerar que estas precisam de duas redes neurais para sua implementação, (*iii*) ser aplicada em pesquisas da literatura, oportunizando verificações, e (*iv*) disponibilidade de forma pública, viabilizando reproduções ou expansão dos resultados.

4. Avaliação

Este artigo apresenta uma avaliação dos métodos geradores de amostras adversariais relevantes da literatura, FGSM, JSMA e C&W. Esta avaliação consiste na implementação do modelo classificador de ataques baseado em uma rede neural artificial. A verificação do desempenho do modelo utiliza os indicadores *Accuracy*, *Precision*, *Recall*, *F1-score* associados a matrizes de confusão criadas com base nos dados originais e dados contendo amostras adversariais pelas técnicas mencionadas. O treinamento e a validação do modelo baseia-se no conjunto de dados CIC-IDS2017, utilizado também para gerar as amostras adversariais. O conjunto de dados possui dados benignos e de diversos ataques, a exemplo de *Bot*, DoS, DDoS, PortScan, SSH, Web (*brute force*, XSS, SQL *Injection*), entre outros. Neste contexto, o estudo evidencia as seguintes questões relativas à aplicação das técnicas FGSM, JSMA e C&W: (*i*) as amostras adversariais geradas por essas técnicas causam algum impacto contra modelo baseado em rede neural e (*ii*) qual a dimensão do impacto das técnicas aplicadas.

A metodologia adotada para o desenvolvimento deste estudo é descrita na subseção seguinte. Nela, são tratados detalhes em relação aos principais aspectos do conjunto de dados utilizado, identificadas as ferramentas aplicadas na manipulação e tratamento dos dados associadas à dinâmica do processo realizado, assim como os resultados obtidos juntamente às análises sobre os impactos do desempenho do classificador, com base nas amostras adversárias geradas pelas abordagens e técnicas selecionadas.

4.1. Metodologia

A avaliação do impacto causado por amostras adversariais contra modelos classificadores de ataques passou pela seleção das técnicas geradoras das amostras e a definição de uma rede neural simples, mas efetiva na classificação de ataques. Com a seleção dessas técnicas e a implementação da rede neural, foi selecionado o conjunto de dados utilizado no processo de avaliação. Em uma primeira etapa, foi elaborado um *pipeline* (ver Figura 1) para a implementação do cenário, análise e avaliação dos resultados. Essa *pipeline* consiste da construção do fluxo da rede, extração das características, rotulação, pré-processamento (manipulação e transformações necessárias) bem como implementar, treinar e avaliar o modelo de ML.

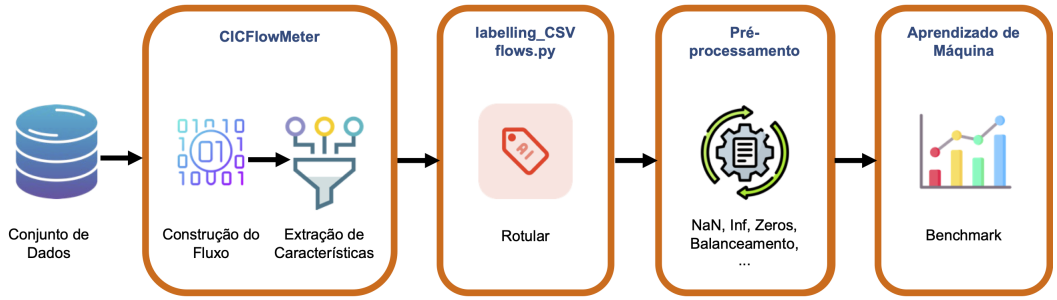


Figura 1. Pipeline dos experimentos

A construção do fluxo da rede corresponde a identificação e organização dos fluxos do conjunto de dados, dado que estes, quando registrados no binário em formato pcap, não necessariamente estão ordenados conforme os fluxos. A extração de características compreende a indicação de quais elementos dos fluxos serão extraídos, a exemplo de tamanho do pacote, duração do fluxo, protocolo, flags, entre outros. A etapa de rotulação corresponde a associar a classificação do dado, ou seja, se é benigno ou a respectiva classe de ataque quando maligno. Além dos procedimentos de pré-processamento para remover dados inconsistentes, NaN (*Not a Number*), zeros ou infinitos normalmente realizados para algoritmos de ML, também foi realizado o balanceamento do conjunto de dados original. E por fim, a etapa de Aprendizado de Máquina consiste no treinamento e avaliação dos modelos, sendo assim denominada de *Benchmark*, considerando que subsidia as comparações entre as técnicas adversariais.

Os experimentos foram conduzidos utilizando a linguagem Python 3, as bibliotecas de aprendizado de máquina, e outras que são requeridas para processar os experimentos por meio dela, como: cicflowmeter [Hieu 2023], Labelling_CSV_flows.py [Engelen et al. 2021], Pandas, Numpy, Matplotlib and Sklearn e, em especial, Keras, aplicadas em atividades de pré-processamento, representação de dados e implementação da rede neural.

4.2. Conjunto de dados

O conjunto de dados selecionado para o experimento foi o CIC-IDS2017. Esse conjunto consiste de aproximadamente 2,28 milhões de registros de fluxo de rede capturados ao longo de cinco dias. Nos dados capturados são identificados os ataques *brute force*, *denial of service* (DoS), *distributed denial of service* (DDoS), *heartbleed*, *web*, *infiltration*

e *Secure Socket Shell* (SSH). Nesse cenário, os dados benignos e de ataques representam aproximadamente 83% e 17% respectivamente [Alshahrani et al. 2022]. A descrição dos arquivos pcap que compõem o conjunto de dados estão listados na Tabela 1. Diversas etapas e tarefas de pré-processamento são necessárias para adequar os conjuntos de dados aos algoritmos de aprendizado de máquina e a seus respectivos treinamentos e testes. Essas operações envolvem a exclusão de registros incompletos e valores infinitos, reorganização dos dados por tipo de ataques, *labelling encoder* e normalização das características contidas no conjunto.

Tabela 1. Descrição conjunto de dados

PCAP	Tamanho
Monday-WorkingHours.pcap	11GB
Tuesday-WorkingHours.pcap	11GB
Wednesday-WorkingHours.pcap	13GB
Thursday-WorkingHours.pcap	7.8GB
Friday-WorkingHours.pcap	8.3GB

A motivação para seleção de tal conjunto de dados se deve ao fato de (*i*) conter ataques rotulados, possibilitando melhor compreensão dos dados para análises e conclusões; (*ii*) utilizar o padrão de formato de arquivos pcap, empregado por diversas ferramentas de redes, oportunizando o manuseio dos registros; (*iii*) ser amplamente usado em outras pesquisas da literatura, viabilizando verificações; e (*iv*) estar disponibilizado de forma pública na Internet, permitindo a reprodução dos resultados.

4.3. Classificador

O classificador implementado para a realização destes experimentos é caracterizado como rede neural - modelo de aprendizado de máquina baseado em aprendizado supervisionado, isto é, os dados utilizados para o treinamento, avaliação e validação do modelo possuem rótulos. Isso permite avaliar de forma objetiva se o modelo classifica corretamente ou não o fluxo de dados da rede. Assim, é possível apontar os verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos que o classificador possa gerar. A partir disso, se definem as taxas de acertos e erros do classificador. A arquitetura do modelo que compõe a rede neural pode ser observada na Tabela 2.

Tabela 2. Estrutura da rede neural

Modelo: sequencial Camada (tipo)	Formato de Saída	Parâmetro #
dense (Densa)	(None, 128)	8320
dense_1 (Densa)	(None, 64)	8256
dense_2 (Densa)	(None, 15)	975
activation (Ativação)	(None, 15)	0
Total de parâmetros: 17,551 Parâmetros treináveis: 17,551 Parâmetros não-treináveis: 0		

Essa rede neural foi utilizada como classificador inicial. O modelo consiste em três camadas densas com 200 *epochs* para o treinamento. As amostras foram pré-

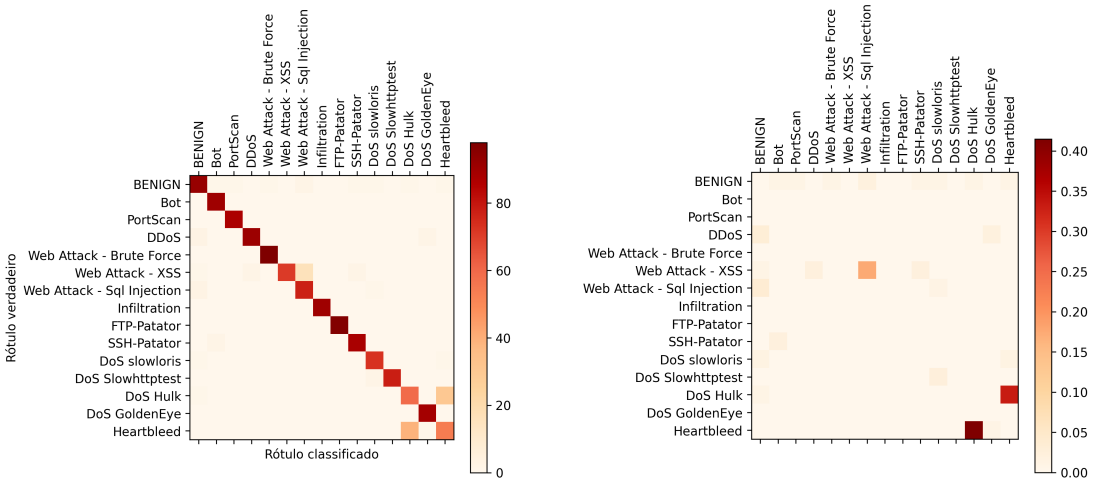
processadas, dimensionadas e divididas no conjunto de dados reamostrado em treinamento e teste, 70% e 30% respectivamente. O otimizador Adam foi utilizado para treinar o modelo. A Tabela 3 mostra o desempenho do classificador. Usamos as métricas *Accuracy*, *Precision*, *Recall* e *F1-score*. A coluna de suporte indica o número de ocorrências da classe na amostra especificada. A definição pelo uso da rede neural está relacionada a características como a capacidade de (*i*) lidar com tarefas de classificação de múltiplas classes, (*ii*) aprender relações complexas nos dados, ou seja, de aprendizado não linear, (*iii*) manipular conjuntos de dados de diversos tamanhos e complexidades e (*iv*) da capacidade de generalização [Géron 2019]. Neste contexto, está fora de escopo avaliar comparativamente outros modelos de ML, dado que a análise comparativa tem a finalidade de determinar o impacto das técnicas geradoras de amostras adversariais neste modelo.

4.4. Resultados

Esta subseção demonstra os resultados alcançados com os experimentos e apresenta os indicadores de desempenho do modelo. A finalidade desses experimentos é avaliar o desempenho das abordagens e técnicas geradoras de amostras adversariais não generativas conhecidas como Non-GANs, em especial FGSM, JSMA e C&W. A Tabela 3 associada a matriz de confusão (ver Figura 2(a)) ilustra o resultado da classificação do conjunto de dados original, ou seja, o experimento base (*baseline*). Os indicadores demonstram que o modelo apresentou índices elevados de acerto ao atribuir as amostras benignas e de ataques nas suas respectivas classes, com uma acurácia global de 93%. Além disso, *Precision*, *Recall* e *F1-score* também obtiveram o índice de 93%. As Figuras 2(a) e 2(b) expõem o comportamento do classificador por meio da matriz de confusão com os dados originais. A matriz de confusão 2(a) demonstra, através da diagonal principal, a elevada taxa de assertividade do classificador considerando os dados originais, ou seja, sem amostras adversariais. E, a Figura 2(b) apresenta o erro percebido nestes mesmos dados classificados.

Tabela 3. Classificação

	precision	recall	f1-score	support
BENIGN	0.92	0.95	0.93	98.00
Bot	1.00	0.97	0.98	93.00
PortScan	1.00	1.00	1.00	87.00
DDoS	0.97	0.98	0.97	95.00
Web Attack - Brute Force	1.00	1.00	1.00	98.00
Web Attack - XSS	0.86	0.98	0.91	81.00
Web Attack - Sql Injection	0.95	0.86	0.90	90.00
Infiltration	1.00	1.00	1.00	90.00
FTP-Patator	1.00	1.00	1.00	97.00
SSH-Patator	0.98	0.98	0.98	90.00
DoS slowloris	1.00	0.96	0.98	78.00
DoS Slowhttptest	1.00	1.00	1.00	80.00
DoS Hulk	0.63	0.60	0.62	95.00
DoS GoldenEye	1.00	1.00	1.00	89.00
Heartbleed	0.60	0.63	0.61	89.00
accuracy	0.93	0.93	0.93	0.93
macro avg	0.93	0.93	0.93	1350.00
weighted avg	0.93	0.93	0.93	1350.00



(a) Matriz de confusão da classificação

(b) Erro de classificação

Figura 2. Matriz de confusão original

Em relação ao desempenho das técnicas FGSM, JSMA e C&W, avaliadas para a geração das amostras adversariais, todas se mostraram efetivas ao causar redução significativa na eficiência da rede neural classificadora dos ataques. Com os resultados dos exemplos adversários pelas respectivas técnicas e a visualização das Tabelas 4, 5 e 6, é possível verificar o impacto no desempenho do classificador. A acurácia da FGSM, JSMA e C&W foi de 7%, 41% e 29%, respectivamente.

As técnicas geradoras de exemplos adversários utilizam parâmetros para controlar a intensidade e a força do ataque. Esse controle é necessário para evitar que o ataque seja detectado, e, portanto, ineficiente. Assim, foi utilizado um conjunto de parâmetros que manteve a magnitude das perturbações limitadas. O parâmetro eps da FGSM foi definido em 0.5, enquanto os parâmetros theta e gamma foram definidos em 0.05 e 0.02 respectivamente para JSMA.

Tabela 4. FGSM

	precision	recall	f1-score
BENIGN	0.90	0.14	0.25
Bot	0.00	0.00	0.00
PortScan	0.00	0.00	0.00
DDoS	0.00	0.00	0.00
Web Attack - Brute Force	0.00	0.00	0.00
Web Attack - XSS	0.00	0.00	0.00
Web Attack - Sql Injection	0.00	0.00	0.00
Infiltration	0.00	0.00	0.00
FTP-Patator	0.00	0.00	0.00
SSH-Patator	0.02	0.00	0.01
DoS slowloris	0.00	0.00	0.00
DoS Slowhttptest	0.00	0.00	0.00
DoS Hulk	0.00	0.00	0.00
DoS GoldenEye	0.00	0.00	0.00
Heartbleed	0.00	0.00	0.00
accuracy	0.07	0.07	0.07
micro avg	0.07	0.07	0.07
macro avg	0.06	0.01	0.02
weighted avg	0.43	0.07	0.12
samples avg	0.07	0.07	0.07

Tabela 5. JSMA

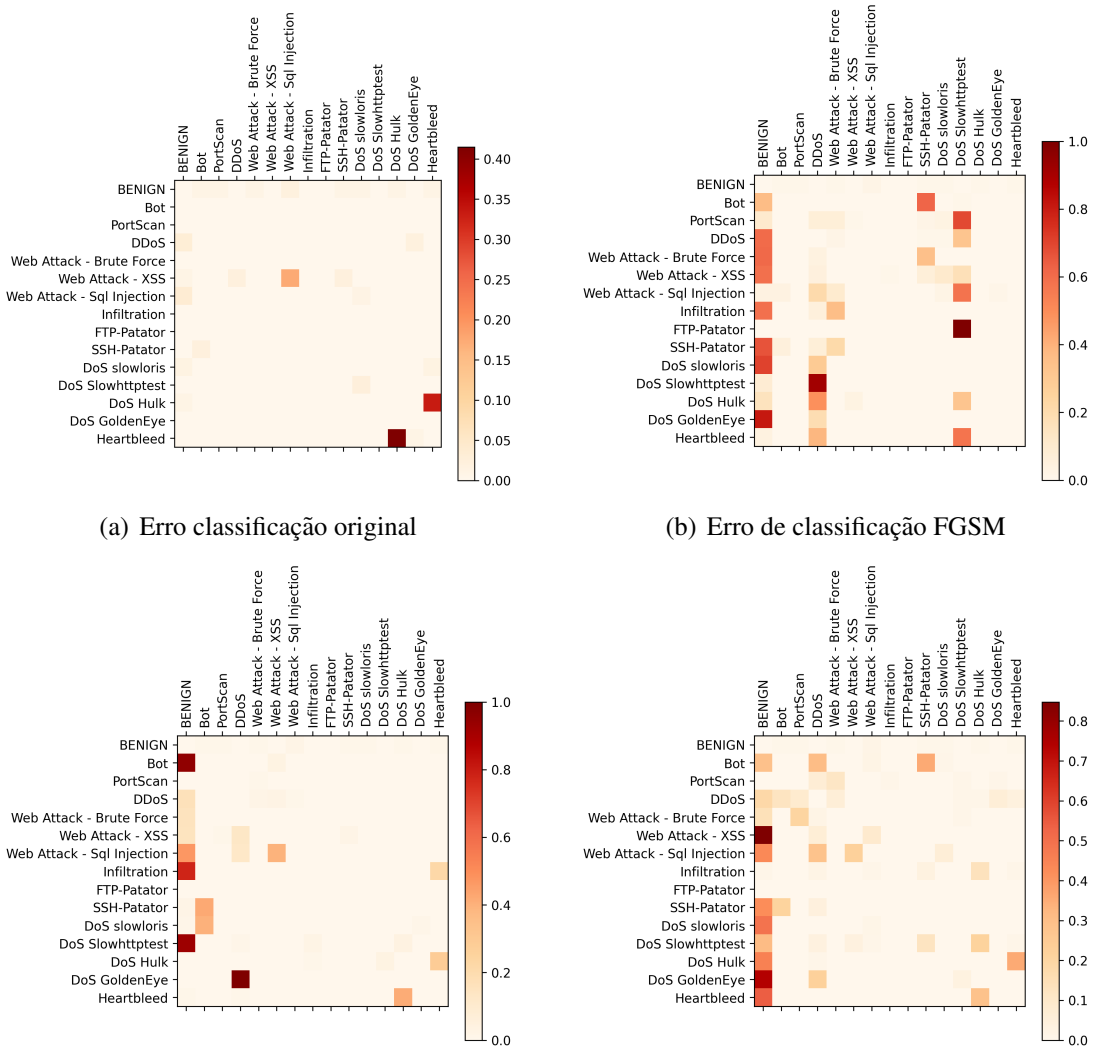
	precision	recall	f1-score
BENIGN	0.90	0.22	0.35
Bot	0.14	0.81	0.25
PortScan	0.98	1.00	0.99
DDoS	0.77	0.35	0.48
Web Attack - Brute Force	0.89	1.00	0.94
Web Attack - XSS	0.62	0.34	0.44
Web Attack - Sql Injection	0.09	0.23	0.13
Infiltration	0.17	1.00	0.29
FTP-Patator	1.00	1.00	1.00
SSH-Patator	0.56	0.65	0.60
DoS slowloris	0.00	0.00	0.00
DoS Slowhttptest	0.00	0.00	0.00
DoS Hulk	0.00	0.00	0.00
DoS GoldenEye	0.00	0.00	0.00
Heartbleed	0.00	0.00	0.00
accuracy	0.41	0.41	0.41
micro avg	0.43	0.45	0.44
macro avg	0.41	0.44	0.36
weighted avg	0.73	0.45	0.50
samples avg	0.43	0.43	0.43

Tabela 6. C&W

	precision	recall	f1-score
BENIGN	0.90	0.28	0.43
Bot	0.01	0.05	0.02
PortScan	0.86	0.72	0.79
DDoS	0.31	0.30	0.30
Web Attack - Brute Force	0.50	0.74	0.60
Web Attack - XSS	0.01	0.03	0.02
Web Attack - Sql Injection	0.01	0.05	0.02
Infiltration	0.92	1.00	0.96
FTP-Patator	1.00	1.00	1.00
SSH-Patator	0.39	0.37	0.38
DoS slowloris	0.08	0.21	0.12
DoS Slowhttptest	0.14	0.48	0.21
DoS Hulk	0.26	0.31	0.28
DoS GoldenEye	0.00	0.00	0.00
Heartbleed	0.09	0.16	0.11
accuracy	0.29	0.29	0.29
micro avg	0.38	0.45	0.41
macro avg	0.37	0.38	0.35
weighted avg	0.61	0.45	0.48
samples avg	0.38	0.38	0.38

4.5. Discussão

Os resultados apresentados demonstraram que FGSM, JSMA e C&W são técnicas capazes de reduzir significativamente o desempenho de um modelo de ML baseado em



(a) Erro classificação original

(b) Erro de classificação FGSM

(c) Erro classificação JSMA

(d) Erro de classificação C&W

Figura 3. Matriz de confusão erro Original, FGSM, JSMA e C&W

rede neural. De acordo com os índices resultantes, obtivemos um decréscimo de 93% da acurácia para 7%, 41% e 29%, respectivamente. Considerando apenas esse indicador, tem-se um impacto negativo na capacidade de classificação do modelo baseado em rede neural na ordem de 13.28, 2.26 e 3.2 vezes em relação à classificação com base nos dados originais - sem amostras adversariais.

Os diversos indicadores, incluindo *Accuracy*, *Precision*, *Recall* e *F1-score* e a matriz de confusão, associados uns aos outros, são ferramentas que auxiliam na avaliação deles próprios. Foram plotadas as matrizes apresentadas na sequência (ver Figuras 3(a), 3(b), 3(c) e 3(d)), nas quais apresenta-se o conjunto de figuras que contêm as matrizes de erros do conjunto de dados original e aqueles com as amostras adversariais geradas com base nas técnicas FGSM, JSMA e C&W.

Nos experimentos, utilizou-se as respectivas matrizes de confusão geradas com base na submissão do conjunto de dados original (ver Figuras 2(a) e 2(b)) e, a seguir, foram geradas as amostras adversariais por meio das técnicas FGSM, JSMA e C&W

(ver Figuras 3(b), 3(c), 3(d)). Assim, foi possível verificar por meio da visualização e análise as diferenças resultantes do classificador quando submetido ao conjunto de dados manipulado e contendo amostras adversariais. Além disso, foram incluídos os valores resultantes de indicadores como *Precision*, *Recall* e *F1-score*, com o objetivo de demonstrar a qualidade do classificador original e subsidiar as análises e a discussão.

A análise da acurácia, individualmente, deixa lacunas sobre a veracidade dos valores demonstrados, já que possui características que podem impactar seus resultados com potencial de distorção, tanto positiva como negativamente. Entre estas, destaca-se a assimetria do conjunto de dados. Por isso, os indicadores foram avaliados em conjunto com as respectivas matrizes de confusão. Ao analisar detalhadamente os valores apresentados nas Tabelas 3, 4, 5 e 6, é possível perceber que os indicadores demonstram que não há itens que possam indicar distorções à acurácia. Outra motivação para entender a acurácia das amostras como coerente se deve ao pré-processamento dos dados, em que foi realizado o *resample* para minimizar o risco de *overfitting* ou *underfitting*, uma vez que no conjunto original há desbalanceamento inerente. Assim sendo, os indicadores demonstraram a efetividade do uso das amostras adversariais ao causar impacto contra o modelo classificador, sendo que o menor deles foi uma redução de 93% da acurácia para 41% no caso do JSMA e o mais efetivo foi o FGSM, em que a acurácia do classificador atingiu apenas 7%. Isso demonstra que os modelos classificadores são muito suscetíveis às amostras adversariais e que medidas adicionais podem ser necessárias para fortalecer sua robustez contra ataques desse tipo.

5. Conclusão

A pesquisa demonstrou que exemplos adversariais são capazes de reduzir a robustez de modelos implementados em soluções de cibersegurança baseadas em IA. Neste contexto, este estudo avaliou comparativamente os métodos FGSM, JSMA e C&W (geradores de exemplos adversariais) com a finalidade de investigar (*i*) se amostras geradas por essas técnicas causam algum impacto contra modelo baseado em redes neurais e (*ii*) qual a dimensão do impacto das técnicas. Assim, a avaliação realizada respondeu às questões ao demonstrar que a rede neural sofreu um decréscimo com as técnicas avaliadas e que o impacto na robustez foi significativo. Cabe destacar que, de maneira empírica, verificou-se que FGSM foi a geradora de amostras adversariais executada com menor custo computacional, dado o cenário em estudo. Mesmo sendo uma pesquisa ainda em estágio inicial, é possível apontar trabalhos futuros. Neste mesmo cenário, é possível avaliar outras abordagens e técnicas geradoras de amostras adversariais, bem como incluir as GANs de forma direta, ampliando a comparação e o entendimento dos impactos causados. Avaliar exemplos adversariais de múltiplos modelos de forma cruzada, tanto em termos de técnicas ainda não exploradas como também diversificando os conjuntos de dados (relativos a cibersegurança) e também a redução e/ou ampliação do número de características. Diferentes classificadores devem ser aplicados, a exemplo do RF, SVM, entre outros. Adicionalmente, há oportunidades de estudos quanto à transferência de amostras adversariais, hiperparâmetros e modelos pré-treinados. Além disso, é necessário avançar a pesquisa na indicação de contramedidas para evitar e/ou mitigar os ataques cibernéticos baseados em amostras adversariais.

Referências

Adesina, D., Hsieh, C.-C., Sagduyu, Y. E., and Qian, L. (2022). Adversarial machine learning in wireless communications using rf data: A review. *IEEE Communications Surveys & Tutorials*.

Alshahrani, E., Alghazzawi, D., Alotaibi, R., and Rabie, O. (2022). Adversarial attacks against supervised machine learning based network intrusion detection systems. *Plos one*, 17(10):e0275971.

Comiter, M. (2019). Attacking artificial intelligence. *Belfer Center Paper*, 8:2019–08.

Engelen, G., Rimmer, V., and Joosen, W. (2021). Troubleshooting an intrusion detection dataset: the cicids2017 case study. In *2021 IEEE Security and Privacy Workshops (SPW)*, pages 7–12. IEEE.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. "O'Reilly Media, Inc.".

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Hieu, L. (Último acesso em Ago/2023). cicflowmeter. `https://gitlab.com/hieulw/cicflowmeter`.

Liu, J., Nogueira, M., Fernandes, J., and Kantarci, B. (2021). Adversarial machine learning: A multi-layer review of the state-of-the-art and challenges for wireless and mobile systems. *IEEE Communications Surveys & Tutorials*.

McCarthy, A., Ghadafi, E., Andriotis, P., and Legg, P. (2023). Defending against adversarial machine learning attacks using hierarchical learning: A case study on network traffic attack classification. *Journal of Information Security and Applications*, 72:103398.

Rosenberg, I., Shabtai, A., Elovici, Y., and Rokach, L. (2021). Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, 54(5):1–36.

Shieh, C.-S., Nguyen, T.-T., Lin, W.-W., Huang, Y.-L., Horng, M.-F., Lee, T.-F., and Miu, D. (2022a). Detection of adversarial ddos attacks using generative adversarial networks with dual discriminators. *Symmetry*, 14(1):66.

Shieh, C.-S., Nguyen, T.-T., Lin, W.-W., Lai, W. K., Horng, M.-F., and Miu, D. (2022b). Detection of adversarial ddos attacks using symmetric defense generative adversarial networks. *Electronics*, 11(13):1977.

Shroff, J., Walambe, R., Singh, S. K., and Kotecha, K. (2022). Enhanced security against volumetric ddos attacks using adversarial machine learning. *Wireless Communications and Mobile Computing*, 2022.

Zhou, S., Liu, C., Ye, D., Zhu, T., Zhou, W., and Yu, P. S. (2022). Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. *ACM Computing Surveys*, 55(8):1–39.