Abordagem Integrada de Deep Learning e P4 para Detecção e Mitigação de Ataques na Borda de Redes IoT

Antonia Mayara de A. da Silva[1] **e Michel Sales Bonfim**[1] **[Orientador]**

[1] Universidade Federal do Ceará - (UFC)
Quixadá – CE – Brasil

mayaraalmeida@alu.ufc.br, michelsb@ufc.br

Abstract. *The Internet of Things (IoT) has expanded its applications across multiple domains, making security issues increasingly critical. This work proposes a deep learning–based solution for attack detection in IoT networks, combined with fast traffic mitigation using the P4 language, which enables data plane programmability. Three deep learning models: CNN, GRU, and LSTM, were trained and evaluated using the Edge-IIoTset dataset, achieving accuracy above 99%. The models were deployed on a Raspberry Pi 4 with P4Pi support and optimized using OpenVINO. In addition, adversarial attacks were evaluated, as well as model retraining with adversarial samples. Based on the overall analysis, the CNN model was selected for the final implementation of the proposed solution.*

Resumo. *A Internet das Coisas (IoT) tem ampliado suas aplicações em diversos cenários, o que torna os aspectos de segurança cada vez mais relevantes. Neste trabalho, é proposta uma solução baseada em aprendizado profundo para a detecção de ataques em redes IoT, em conjunto com à mitigação rápida do tráfego malicioso por meio da linguagem P4, que permite a programabilidade do plano de dados. Três modelos de aprendizado profundo (CNN, GRU e LSTM) foram treinados e avaliados com o conjunto de dados Edge-IIoTset, alcançando acurácia superior a 99%. Os modelos foram implementados em uma Raspberry Pi 4 com suporte à plataforma P4Pi, utilizando a ferramenta OpenVINO para otimização. Além disso, foram avaliados ataques adversários e seu impacto sobre os modelos, bem como o retreinamento com amostras adversárias. A partir da análise das métricas, o modelo CNN foi selecionado para a versão final da solução.*

1. Introdução

A Internet das Coisas (IoT) se consolidou como uma das tecnologias mais utilizadas na atualidade. A diversidade de áreas em que é possível utilizá-la e as funcionalidades disponibilizadas por ela a tornam promissora e, em muitos casos, indispensável. A maioria dos dispositivos presentes em redes IoT são sensores e placas com poder computacional limitado. Além do mais, eles são responsáveis por transmitirem uma grande quantidade de dados e, por isso, tornam-se alvos atrativos para ataques cibernéticos. Portanto, a busca por um sistema eficaz contra diferentes tipos de ataques, desde os mais conhecidos como DoS/DDoS e *malware* até ataques *zero day*, e que funcione em tempo real, sem comprometer o funcionamento da rede, é essencial para garantir a segurança das comunicações IoT [Kaur et al. 2023].

O uso de técnicas de aprendizado de máquina (ML) e, mais recentemente, de aprendizado profundo (DL) tem se mostrado uma alternativa viável para sistemas de detecção de intrusão (IDS) em redes IoT. No entanto, o surgimento de ataques adversários, onde perturbações imperceptíveis são introduzidas nos dados de entrada para enganar os classificadores e induzi-los a inferências erradas, representa uma ameaça à confiabilidade desses sistemas [He et al. 2023]. Conforme destacado por [Qin et al. 2020], o uso isolado de algoritmos de aprendizado pode não ser suficiente diante da heterogeneidade das redes IoT, onde diferentes protocolos e a falta de flexibilidade da rede dificultam a implementação e adaptação de um IDS.

Diante desse cenário, surge a oportunidade de integrar DL à programabilidade do plano de dados por meio da linguagem P4, visando a melhoria na segurança em aplicações IoT. A linguagem P4 permite a extração eficiente de características do tráfego diretamente no plano de dados e a aplicação dinâmica de regras de mitigação, sem depender de controladores externos ou comprometer a latência da rede. A execução de modelos de DL em dispositivos de borda com recursos computacionais limitados traz desafios adicionais que precisam ser tratados. O uso de dispositivos SBC (Single Board Computer), como a Raspberry Pi, surge como uma alternativa de baixo custo para a implantação de soluções de segurança na borda da rede IoT, permitindo processamento local sem depender de envio de dados para um ambiente em nuvem. Para viabilizar essa execução em tempo real, técnicas de otimização de modelos, como o OpenVINO, se tornam essenciais para conciliar desempenho e eficiência em ambientes com recursos restritos. A lacuna identificada na literatura é a ausência de soluções que reúnam, em um único sistema embarcado de baixo custo, detecção por DL, mitigação com P4 e defesa contra ataques adversários em ambientes IoT.

1.1. Objetivos

Este trabalho tem por objetivo geral desenvolver uma solução de baixo custo capaz de detectar e mitigar ataques em tempo real na borda de redes IoT, sendo capaz também de identificar ataques adversariais. Para alcançar este objetivo, é feito o uso de algoritmos de DL, de uma ferramenta de otimização e da linguagem P4. Ademais, esta pesquisa possui os seguintes objetivos específicos:

OE1. Identificar, dentre algoritmos de *Deep Learning*, qual possui o melhor desempenho para a detecção de ataques;

OE2. Avaliar o impacto dos ataques adversários nos modelos de inferência selecionados;

OE3. Projetar e avaliar uma arquitetura para a implementação de um IDS em plataformas SBC, utilizando Deep Learning e P4 como base.

1.2. Contribuições do Trabalho

A seguir, são apresentadas as principais contribuições alcançadas com este trabalho[1].

- Desenvolvimento de uma solução para detecção de ataques com DL otimizado para execução em dispositivo SBC. Diferentes algoritmos foram avaliados e submetidos a um processo de otimização com OpenVINO para viabilizar a detecção de ataques internamente na rede IoT;

[1] https://github.com/mayaraalmeid/P4-DL/

- Aplicação da linguagem P4 no processo de mitigação e de extração de features. O P4 foi utilizado para criar regras de bloqueio nas tabelas do switch e para coletar as informações necessárias ao algoritmo de DL para a classificação dos pacotes, sem modificar o tráfego original;
- Avaliação de diferentes técnicas adversariais sobre três algoritmos de DL. As técnicas implementadas (FGSM, PGD, CW e GAN) expuseram os modelos a perturbações e alcançaram redução significativa na acurácia. O trabalho também desenvolveu a mitigação a esse tipo de ataque por meio do retreinamento dos algoritmos com amostras adversárias;
- Uso de hardware de baixo custo e computacionalmente limitado para detecção e mitigação de ataques em tempo real, demonstrando a viabilidade da solução proposta em ambientes IoT com recursos restritos.

2. Trabalhos Relacionados

A Tabela 1 compara os principais trabalhos relacionados quanto aos tópicos centrais desta proposta. Os trabalhos que combinam DL e P4 [Heggi et al. 2022, Musumeci et al. 2022] apresentam apenas à detecção, usando o P4 apenas para extração de recursos. Os que implementaram a mitigação com P4 [Khedr et al. 2023, Zang et al. 2023] não empregam DL e cobrem um conjunto restrito de ataques. No contexto adversarial, apenas [Novaes et al. 2021] e [Reddy et al. 2024] apresentaram estratégias de mitigação, sendo que o segundo é o único a utilizar P4 e defesas adversárias, mas com resultados limitados nessa etapa. Nenhum trabalho anterior implanta a solução em um dispositivo SBC de baixo custo. Este trabalho é o único a reunir todas as cinco dimensões avaliadas.

Tabela 1. Comparação com trabalhos relacionados.

Trabalho	DL	P4	Mitigação	Adversarial	SBC
[Musumeci et al. 2022]	✓	✓	–	–	–
[Heggi et al. 2022]	✓	✓	–	–	–
[Khedr et al. 2023]	–	✓	✓	–	–
[Zang et al. 2023]	–	✓	✓	–	–
[Novaes et al. 2021]	✓	–	✓	✓	–
[Reddy et al. 2024]	✓	✓	✓	✓	–
Este trabalho	✓	✓	✓	✓	✓

3. Arquitetura Proposta

A Figura 1 apresenta o funcionamento da solução, implementada sobre a plataforma P4Pi, que combina o T4P4S como *switch* P4 com o DPDK (Data Plane Development Kit) para processamento acelerado de pacotes em espaço de usuário, proporcionando ganhos significativos de desempenho em hardware de baixo custo. O tráfego, incluindo os ataques, é recebido por uma interface e repassado ao *switch* P4, que extrai campos dos protocolos TCP, UDP, ICMP e MQTT via *digests* baseados em INT-XD, enviando as informações ao módulo de decisão sem modificar os pacotes originais. O módulo de decisão é composto por um modelo de DL e um controlador responsável pela coleta das *features* e adição das regras. Tráfego classificado como benigno segue normalmente na rede; tráfego malicioso é descartado por meio da inserção dinâmica de regras nas tabelas do *switch*.

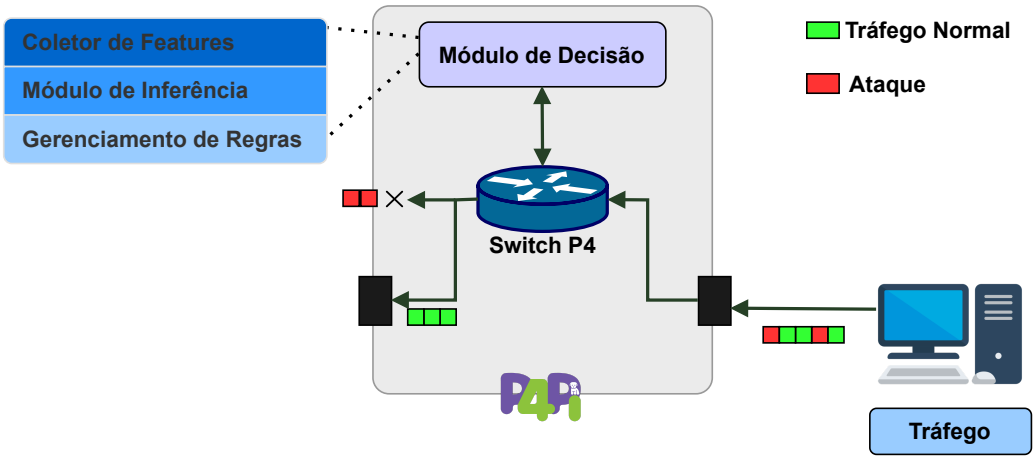


Figura 1. Arquitetura da solução.

Foram selecionados três algoritmos de DL: CNN, LSTM e GRU, eles foram treinados e avaliados com validação cruzada k-fold (k=5) a partir do *dataset* Edge-IIoTset, que contempla 14 classes distintas de tráfego malicioso em ambientes IoT. Os resultados demonstraram acurácia próxima a 99% para os três modelos, conforme apresentado na Tabela 2. Para viabilizar a execução no dispositivo de borda, os modelos foram otimizados com OpenVINO, com redução média de 3-5% na acurácia, considerada aceitável frente aos ganhos de desempenho. A robustez dos modelos foi avaliada com ataques adversários gerados por FGSM, PGD, CW e GAN, com as amostras adversárias posteriormente incorporadas ao conjunto de treinamento para retreinamento.

Tabela 2. Resultado do teste dos algoritmos.

Algoritmo	Acurácia (%)	Precisão (%)	*Recall* (%)	F1 (%)
CNN	99,91	99,62	97,43	98,45
GRU	99,17	98,56	94,57	96,52
LSTM	99,82	99,51	97,63	98,38

4. Resultados

Esta seção apresenta um resumo e uma análise dos principais resultados obtidos ao longo dos experimentos realizados, considerando o desempenho dos modelos de DL no dispositivo SBC, sua robustez frente a ataques adversários e os efeitos do retreinamento como estratégia de mitigação.

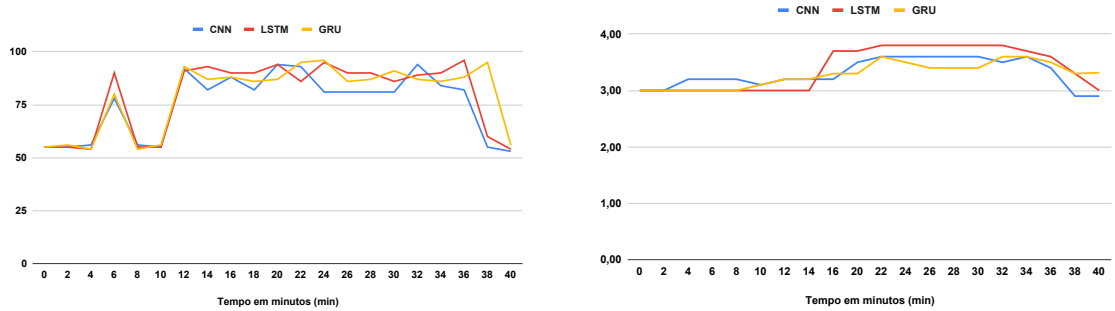
4.1. Desempenho dos Modelos na Raspberry 4 Pi

Os experimentos conduzidos na Raspberry Pi 4 demonstraram que os três modelos avaliados (CNN, LSTM e GRU) são viáveis para execução em dispositivos do tipo SBC, especialmente quando combinados com técnicas de otimização, como o uso do OpenVINO. Conforme apresentado na Tabela 3, todos os modelos alcançaram taxas de acurácia superiores a 93%, com destaque para o LSTM, que obteve a maior acurácia (96,93%).

Tabela 3. Avaliação na Raspberry Pi.

Métrica	CNN	LSTM	GRU
Acurácia (%)	95,60	**96,93**	93,95

Apesar da boa precisão, a análise do consumo de recursos evidenciou diferenças relevantes entre os modelos. As Figuras 2a e 2b mostram que os modelos recorrentes (LSTM e GRU) demandaram maior uso de CPU e memória RAM durante a execução do tráfego, especialmente em cenários de maior carga gerada pelo Tcpreplay. O CNN, por sua vez, apresentou um consumo mais estável e inferior, refletindo sua menor complexidade computacional.



(a) Uso da CPU. **(b) Uso de RAM.**

Figura 2. Consumo de recursos na Raspberry Pi 4.

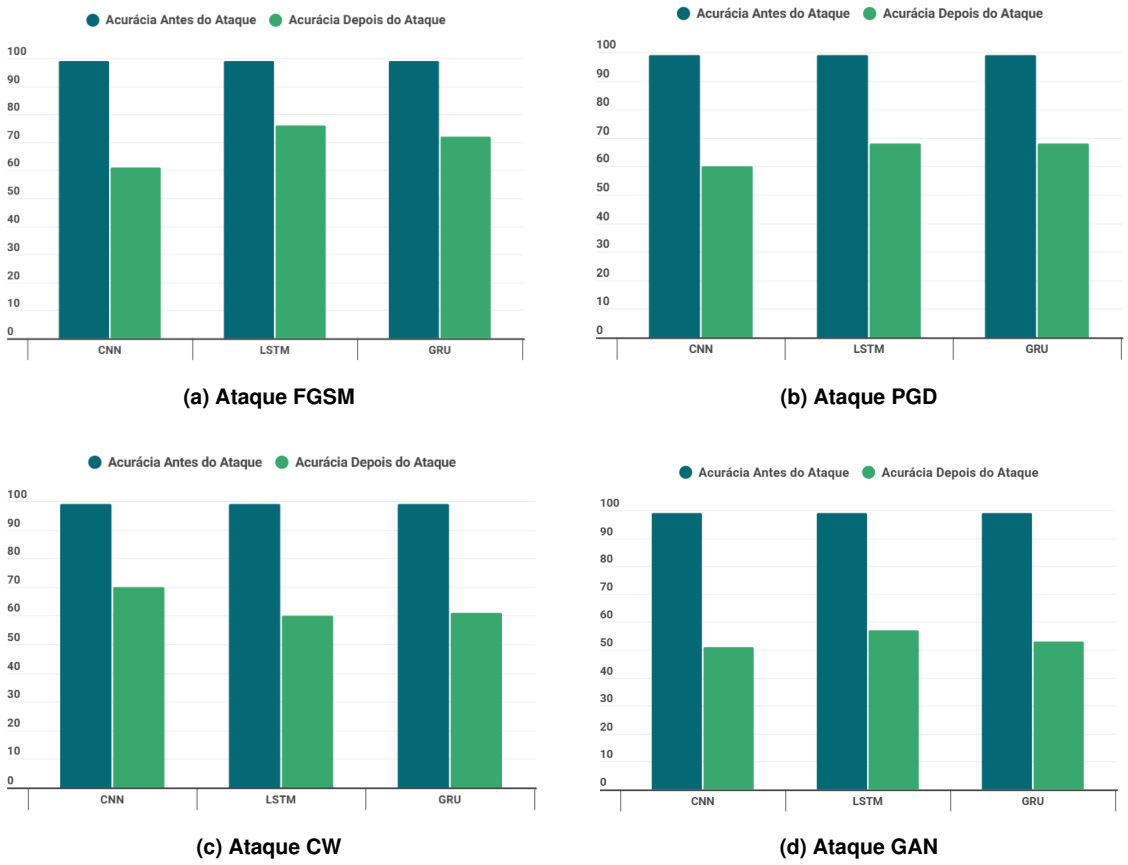
Em termos de tempo médio necessário para a inferência, o CNN destacou-se novamente, apresentando o menor tempo médio de inferência por pacote (3,66 ms), enquanto GRU teve o tempo médio foi de 149 milissegundos e o LSTM levou, em média, 184 milissegundos para classificar. Esses resultados indicam que, embora os modelos recorrentes apresentem ligeira vantagem em acurácia, o CNN é mais adequado para cenários que exigem respostas rápidas e baixo consumo de recursos.

4.2. Impacto dos Ataques Adversários

A introdução das técnicas adversárias (FGSM, PGD, CW e GAN) revelou vulnerabilidades importantes nos modelos avaliados. Conforme ilustrado nas Figuras 3a, 3b, 3c e 3d, todos os algoritmos sofreram reduções expressivas de acurácia quando expostos a amostras adversárias, independentemente da arquitetura utilizada.

Os ataques baseados em gradiente (FGSM e PGD) afetaram de forma mais intensa o modelo CNN, reduzindo sua acurácia para valores próximos de 60%. Já os modelos recorrentes demonstraram maior resiliência relativa a esses ataques, devido às suas estruturas de memória, que capturam dependências temporais mais longas.

Por outro lado, a técnica CW mostrou-se mais eficaz contra os modelos LSTM e GRU, enquanto o CNN apresentou menor degradação nesse cenário. O ataque baseado em GAN foi o mais severo entre os avaliados, resultando nas menores taxas de acurácia para todos os modelos, conforme sintetizado na Tabela 4. Esses resultados evidenciam que ataques mais sofisticados e adaptativos conseguem explorar vulnerabilidades específicas de cada arquitetura.



(a) Ataque FGSM **(b) Ataque PGD**

(c) Ataque CW **(d) Ataque GAN**

Figura 3. Impacto dos ataques adversários na acurácia dos modelos avaliados.

4.3. Retreinamento com Amostras Adversárias

O retreinamento com amostras adversárias demonstrou ser eficaz como estratégia de mitigação, conforme ilustrado na Figura 4: a inclusão dessas amostras no conjunto de treinamento permitiu que todos os modelos recuperassem grande parte da acurácia perdida durante os ataques. A CNN, apesar de ter sido o modelo mais impactado, recuperou a maior acurácia após o retreinamento (98,57%), mostrando que sua arquitetura, embora mais vulnerável a perturbações, possui boa capacidade de adaptação quando exposta a amostras adversárias durante o treinamento. Os modelos recorrentes também se recuperaram de forma consistente. Esses resultados indicam que a inclusão de amostras adversárias no conjunto de treinamento é uma abordagem viável para restaurar o desempenho dos modelos em ambientes IoT sujeitos a esse tipo de ameaça. É importante destacar, contudo, que a métrica reportada reflete a recuperação da acurácia em dados não perturbados; o retreinamento aumenta a resistência sobretudo aos ataques empregados nessa etapa, podendo haver diferenças na detecção em modelos adversariais ainda não avaliados.

5. Conclusão

Este trabalho propôs e avaliou uma arquitetura integrada de DL e P4 para detecção e mitigação de ataques em redes IoT, implantada em uma Raspberry Pi 4. Os resultados demonstraram a viabilidade da solução em ambiente embarcado, com todos os modelos alcançando acurácia superior a 93% no dispositivo e a CNN se destacando pelo tempo de

Tabela 4. Acurácia (%) antes e após ataques adversários.

Cenário	CNN	LSTM	GRU
Acurácia inicial	**99,91**	99,82	99,17
Após FGSM	61,57	**76,18**	72,79
Após PGD	60,15	**68,80**	68,66
Após CW	**71,32**	60,63	62,01
Após GAN	51,10	**57,12**	53,42
Após retreinamento	**98,57**	98,44	97,24

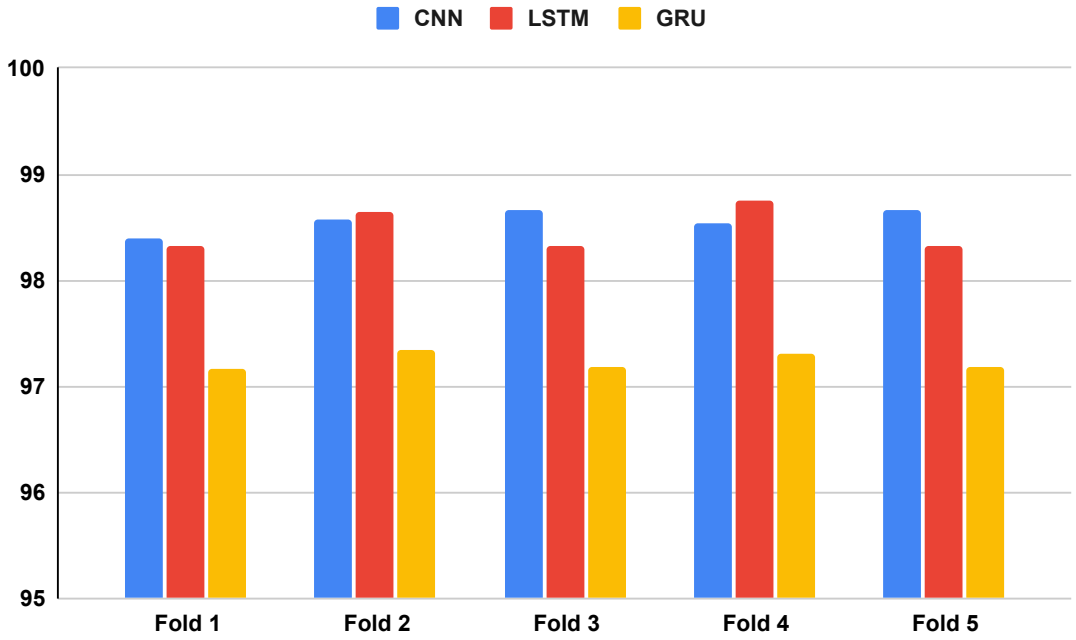


Figura 4. Acurácia após o retreinamento.

inferência de apenas 3,66 ms/pacote. A avaliação adversária revelou reduções de acurácia de até 48%, sendo o ataque GAN o mais crítico; contudo, o retreinamento com amostras adversárias se mostrou eficiente, restaurando a acurácia da CNN a 98,57%. Considerando o conjunto de métricas avaliados, a CNN foi considerada a alternativa mais adequada para a versão final da arquitetura.

Como trabalhos futuros, o foco será aprimorar o gerenciamento das regras de mitigação no switch P4, adotando estratégias como monitoramento de frequência de uso, priorização por tipo de ataque e agregação de regras redundantes. Do ponto de vista do hardware, o objetivo é alcançar uma maior otimização dos modelos para reduzir o consumo de CPU e RAM na Raspberry Pi, bem como a avaliação da arquitetura em diferentes pontos da rede, considerando latência, consumo de recursos e assertividade na detecção.

6. Produção Científica Associada

Os principais resultados desta dissertação foram publicados nos seguintes trabalhos:

- Silva, A., Bonfim, M., Callado, A., e Gonçalves, E. (2025). Detection and Mitigation of Attacks at the Edge of IoT Networks Using Deep Learning and P4. *Anais do XXI Simpósio Brasileiro de Sistemas de Informação (SBSI 2025)*, pp. 595–604. SBC.
- Silva, A., Rego, P., e Bonfim, M. (2025). Avaliação e Mitigação de Ataques Adversários em Sistema de Detecção de Intrusão IoT. *Anais do XXV Simpósio Brasileiro de Cibersegurança (SBSeg 2025)*, pp. 987–994. SBC.

A dissertação completa está disponível em: `https://repositorio.ufc.br/handle/riufc/80317`

7. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES - Código de Financiamento 001

Referências

He, K., Kim, D. D., and Asghar, M. R. (2023). Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 25(1):538–566.

Heggi, S. R., Sukarno, P., and Mugitama, S. A. (2022). Lstm-nb: Dos attack detection on sdn with p4 programmable dataplane. In *2022 International Conference on Advanced Creative Networks and Intelligent Systems (ICACNIS)*, pages 1–6. IEEE.

Kaur, B., Dadkhah, S., Shoeleh, F., Neto, E. C. P., Xiong, P., Iqbal, S., Lamontagne, P., Ray, S., and Ghorbani, A. A. (2023). Internet of things (iot) security dataset evolution: Challenges and future directions. *Internet of Things*, page 100780.

Khedr, W. I., Gouda, A. E., and Mohamed, E. R. (2023). P4-hldmc: A novel framework for ddos and arp attack detection and mitigation in sd-iot networks using machine learning, stateful p4, and distributed multi-controller architecture. *Mathematics*, 11(16):3552.

Musumeci, F., Fidanci, A. C., Paolucci, F., Cugini, F., and Tornatore, M. (2022). Machine-learning-enabled ddos attacks detection in p4 programmable networks. *Journal of Network and Systems Management*, 30:1–27.

Novaes, M. P., Carvalho, L. F., Lloret, J., and Proença Jr, M. L. (2021). Adversarial deep learning approach detection and defense against ddos attacks in sdn environments. *Future Generation Computer Systems*, 125:156–167.

Qin, Q., Poularakis, K., and Tassiulas, L. (2020). A learning approach with programmable data plane towards iot security. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*, pages 410–420. IEEE.

Reddy, S. S., Nishoak, K., Shreya, J., Reddy, Y. V., and Venkanna, U. (2024). A p4-based adversarial attack mitigation on machine learning models in data plane devices. *Journal of Network and Systems Management*, 32(1):5.

Zang, M., Zheng, C., Dittmann, L., and Zilberman, N. (2023). Towards continuous threat defense: In-network traffic analysis for iot gateways. *IEEE Internet of Things Journal*.