Rede Adversária Generativa baseada em Transformador e XAI para a Detecção de Intrusões e Anomalias em Redes Definidas por Software.

Vinícius Ferreira Schiavon[1], **Mario Lemes Proença Jr**[1]

[1]Departamento de Computação – Universidade Estadual de Londrina (UEL)
Caixa Postal 10.011 – CEP 86057-970 – Londrina – PR – Brasil

vinicius.schiavon@uel.br, proenca@uel.br

Abstract. *With the increasing integration of devices in the network, it becomes necessary to maintain their confidentiality, integrity, and availability. However, network anomalies can disrupt their operation, making it necessary to develop tools that prevent this inconvenience. A widely studied solution in the literature is the construction of Intrusion Detection Systems in Networks using Deep Learning. This work will implement one of these systems using a combination of two state-of-the-art components: Generative Adversarial Networks and Transformers. This work will also employ Explainable Artificial Intelligence techniques to analyze and justify its behavior to ensure that the proposed model is not treated as a black box.*

Resumo. *Com a crescente integração de dispositivos na rede, torna-se necessário a manutenção da sua confidencialidade, integridade e disponibilidade. Entretanto, anomalias de rede podem atrapalhar seu funcionamento, tornando necessário o desenvolvimento de ferramentas que evitem este inconveniente. Uma solução amplamente estudada na literatura é a construção de Sistemas de Detecção de Intrusão em Redes utilizando Aprendizado Profundo. Este trabalho irá implementar um desses sistemas, utilizando a combinação de dois componentes do estado da arte: as Redes Adversárias Generativas e os Transformadores. Para que o modelo proposto não seja tratado como uma caixa preta, este trabalho também utilizará técnicas de Inteligência Artificial Explicável para analisar e justificar seu comportamento.*

1. Introdução

A utilização de celulares, computadores, sensores e tecnologias vestíveis está cada vez mais presente no dia a dia dos seres humanos [3], [71]. A maioria destes aparelhos possui acesso à internet, implicando com que eles façam parte de uma rede de computadores, possibilitando suas comunicações [42]. O uso destes aparatos passa a ser essencial para atividades como comunicação, estudo, trabalho, locomoção, financeira e entretenimento [65].

Os dispositivos passam a ser dependentes das redes, que devem sempre se manter funcionais para evitar danos e perdas aos usuários ou proprietários [34]. Uma queda de rede está geralmente relacionada a falhas de *hardware*, *software*, humanas ou causadas propositalmente por ataques [32]. Um exemplo real foi a falha de software que ocorreu com a empresa CrowdStrike, resultando num apagão global das máquinas que possuíam seus softwares e gerando gastos para diversas empresas e usuários [16].

As variações na rede podem ser chamadas de anomalias, e há um grande interesse no desenvolvimento de soluções para a sua detecção [21]. Estas soluções auxiliariam o gerente de rede, avisando-o quando uma possível anomalia é detectada. Isto direcionaria qual mitigação deve ser feita, contribuindo para manutenção da confidencialidade, integridade e disponibilidade da rede e dos dados que nela trafegam [13].

Diversos pesquisadores buscam construir um sistema que solucione esse problema das anomalias e ataques, e uma das soluções mais aceitas é a do Sistema de Detecção de Intrusões em Redes [63]. Ele funciona complementarmente ao *firewall*, notificando o gerente de redes quando um possível tráfego anômalo é encontrado. Diversas formas de implementação desse tipo de sistema já foram testadas [15], [24]. A abordagem que obtém os melhores resultados utiliza Aprendizado Profundo [22].

A área de Aprendizado Profundo é uma subárea de Aprendizado de Máquina que possui diversas implementações para detecção de anomalias em redes [27], [49]. Uma destas implementações envolve o uso de Redes Generativas Adversárias [26], uma técnica robusta baseada na Teoria dos Jogos que será utilizada neste trabalho. Além desta técnica, um tipo especial de rede será usado: os Transformadores [64]. Estes, que foram inicialmente desenvolvidos para tradução de texto, se tornaram populares pelo modelo GPT [9].

Um problema recorrente dos modelos de Aprendizado Profundo é que eles são interpretados como caixas pretas. O programador não consegue entender o processo de inferência utilizado pelo modelo para o cálculo dos dados de saída [39]. As técnicas de Inteligência Artificial Explicável são estudadas para resolver este problema, auxiliando na depuração do modelo e na justificativa do seu funcionamento [56].

O restante do documento está organizado da seguinte forma. A seção 2 apresenta os conceitos, métodos, técnicas e revisão do estado da arte fundamentais para o desenvolvimento deste trabalho. A Seção 3 descreve o objetivo que este trabalho visa alcançar, já a Seção 4 explica como estes objetivos serão alcançados. A Seção 5 exibe um cronograma de atividades que serão realizadas para a conclusão do trabalho. Na Seção 6 são descritas quais serão as contribuições que este trabalho trará quando concluído.

2. Fundamentação Teórico-Metodológica e Estado da Arte

2.1. Detecção de Intrusões e Anomalias em Redes de Computadores

Ao observar-se as características do tráfego de uma rede de computadores, é possível definir um comportamento padrão da rede [54]. Este comportamento é chamado de *baseline* [10] e toda observação que estiver fora dela pode ser classificada como uma anomalia de rede [58].

A Figura 1 mostra um exemplo onde a *baseline* é representada pela reta $X = Y$, e os pontos em azul representa dados normais. Já o ponto vermelho é considerado uma anomalia, por estar distante da *baseline*.

Existem diversos tipos de anomalias de redes que possuem características e causas diferentes, entre eles estão as intrusões de rede [15]. Para garantir uma maior eficácia na detecção das anomalias, é necessário escolher o formato mais adequado de coleta de dados. Pois, cada formato de coleta evidenciará diferentes características da rede [72]. Entre os formatos estão: TCP dump, SNMP e IP Flow [22].

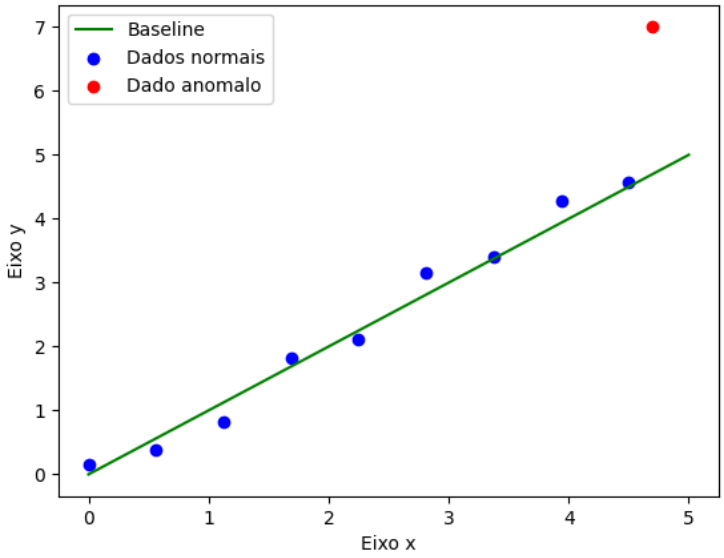

Figura 1. Exemplo de anomalia

Uma solução amplamente aceita no meio científico para detecção de anomalias em redes de computadores é o Sistema de Detecção de Intrusão de Redes (NIDS, do inglês *Network Intrusion Detection System*) [1]. Estes sistemas monitoram o tráfego da rede e notificam o gerente de rede quando uma anomalia é identificada [5]. A função destes sistemas é complementar à do *firewall*, contribuindo para manutenção da confidencialidade, integridade e disponibilidade (Tríade CID) dos dados e da infraestrutura da rede [40].

Um dos tipos de anomalias que visa afetar a Tríade CID são os ataques de rede, que estão em constante mudança e concepção [60], [61]. Esta constante alteração dificulta sua detecção, então o NIDS deve conseguir se adaptar e detectar ataques nunca vistos [18]. O NIDS também deve ter uma resposta rápida, para que os danos causados por um ataque sejam mínimos [20]. Sabendo das características desejáveis de um NIDS, existem dois principais métodos para implementá-lo [22], sendo:

- Baseado em Assinatura [53]: Padrões de anomalias já conhecidos são armazenados em um banco de dados e o NIDS tenta identificar algum desses padrões no tráfego. A vantagem deste método é que ele apresentará uma baixa taxa de falsos positivos, por gerar um alerta apenas quando a anomalia é detectada. Já a desvantagem é que ele não consegue detectar anomalias não conhecidas, exigindo que o banco de dados seja constantemente atualizado.
- Baseado em Anomalia [48]: Usa dados históricos da rede para construir um modelo que representa o tráfego normal. Quando o NIDS identifica um tráfego diferente do normal, uma anomalia é detectada. A vantagem deste método é que ele torna possível a detecção de anomalias desconhecidas. A desvantagem é que a taxa de falsos positivos será elevada, pois nem toda alteração na rede será uma anomalia.

A abordagem mais utilizada na literatura, que contém as características desejáveis descritas é a baseada em anomalia, que será estudada e implementada neste trabalho.

A comunidade científica visa construir um NIDS que atenda todas as necessidades de segurança, publicando diversos trabalhos com diferentes abordagens na área de segurança. Um grupo que se esforça para obter avanços nesta área é o grupo de pesquisa Orion da Universidade Estadual de Londrina, contribuindo desde 2002. Este grupo produziu tanto revisões sistemáticas [22] quanto abordagens práticas para implementação de NIDS [4], [13], [12], [73], [50], [19], [52], [23], [57], [51]. Entre os diversos trabalhos produzidos por este grupo, pode-se destacar Lent *et al.* [6], onde os autores constroem um NIDS baseado em anomalias e não supervisionado, utilizando as Redes Adversárias Generativas. Outro trabalho do grupo é o de Hamamoto *et al.* [28], que propõe um NIDS baseado em anomalias e não supervisionado, utilizando Algoritmos Genéticos e Lógica Difusa. Já o trabalho de Novaes *et al.* [46] apresenta o uso das técnicas de Memória de Curto Longo Prazo e Lógica Difusa para detecção e mitigação de anomalias em redes.

2.2. Aprendizado de Máquina

O Aprendizado de Máquina (ML, do inglês *Machine Learning*) é uma subárea da Inteligência Artificial (AI, do inglês *Artificial Intelligence*). Os estudos da área de AI focam em maneiras de um computador simular a capacidade de resolução de problemas e a inteligência humana [44]. A abordagem da subárea de ML é utilizar dados para treinar um modelo, de regressão ou classificação, e alcançar o objetivo da AI [2]. A Figura 2 mostra a hierarquia das técnicas de AI que serão expostas nas próximas seções.

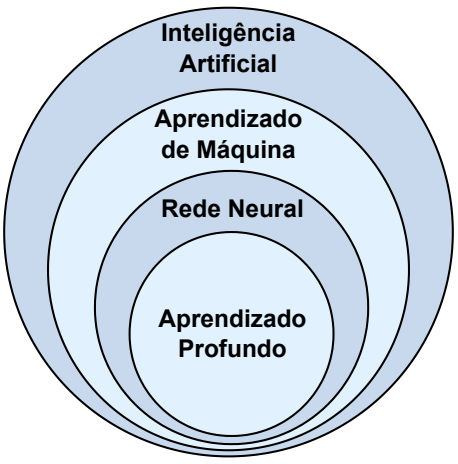

Figura 2. Relação entre as áreas de AI

Um modelo de ML tem como objetivo performar tarefas sobre um conjunto de dados. Por exemplo, é possível desenvolver um modelo que faça previsões sobre dados que o alimentam, diferentes dos que foram usados para treiná-lo [35]. Para cumprir este objetivo, espera-se que o modelo, durante o treinamento, se ajuste automaticamente e aprenda o padrão dos dados utilizados [69]. Modelos de ML suprem a necessidade de análise e interpretação sobre dados, pois são capazes de aprender padrões e relações dos dados que possivelmente não seriam encontrados por métodos tradicionais [68].

2.2.1. Tipos de Aprendizado

O aprendizado de modelos de ML pode ser separado em três principais categorias, divididas em relação aos dados de treinamento [44], sendo:

- Supervisionado [7]: Cada observação *X* do conjunto de dados possui um rótulo *Y* capaz de descrevê-la. O objetivo do modelo que usa este tipo de aprendizado é aprender uma função *f* que mapeie qualquer *X* do conjunto de dados para um *Y*' satisfatoriamente próximo de Y. Este tipo de aprendizado é normalmente utilizado para resolução de problemas de classificação e regressão.
- Semi-supervisionado [37]: O conjunto de dados possui observações com e sem rótulos. Neste caso, o objetivo é utilizar abordagens que se aproveitem dos dados rotulados, mesmo que numa quantidade menor que os não rotulados.
- Não supervisionado [58]: O conjunto de dados não possui rótulos. O modelo que emprega este tipo de aprendizado tem como objetivo aprender padrões e estruturas que estão intrínsecos nos dados. Os usos mais comuns deste aprendizado são para resolução de problemas de clusterização, redução de dimensionalidade e classificação.

A tarefa de rotulação é custosa e os dados normalmente não são extraídos com rótulos [14]. Por isso, aplicações que utilizam o aprendizado não supervisionado terão vantagem ao serem implementadas em meios reais, pois os dados poderão ser utilizados de maneira mais próxima ao que são encontrados no meio, não sendo necessário rotulá-los [6]. A aplicação desenvolvida neste trabalho será não supervisionada.

2.3. Rede Neural

A Rede Neural (NN, do inglês *Neural Network*) é uma especialização de ML inspirada no cérebro humano que utiliza dois principais elementos para efetuar cálculos complexos: os neurônios e suas interconexões [67]. Os neurônios são dispostos em camadas sequenciais, e um neurônio só pode ter ligação com outro se eles estiverem em camadas adjacentes [22]. A primeira camada, chamada de camada de entrada, é responsável por receber as propriedades de uma observação *X* do conjunto de dados. A última camada, chamada de camada de saída, exibe o resultado *Y* dos cálculos da NN. Entre a camada de entrada e a de saída existe ao menos uma camada oculta [43]. A combinação de número de camadas e de neurônios em cada camada é o que define a arquitetura de uma NN, então, é possível construir inúmeras redes neurais ao variar-se esses valores [44].

2.4. Aprendizado Profundo

O Aprendizado Profundo (DL, do inglês *Deep Learning*) é uma especialização de NN, cuja diferença está relacionada ao número de camadas ocultas [25]. A rede neural que possui múltiplas camadas ocultas se enquadra na subárea de DL e recebe o nome de Rede Neural Profunda (DNN, do inglês *Deep Neural Network*) [62]. Por possuir múltiplas camadas, uma DNN consegue modelar funções e resolver problemas mais complexos, porém é mais cara computacionalmente, também necessitando de mais dados de treinamento [38]. O poder apresentado pelas DNN já se provou eficaz na previsão de séries temporais [29], que serão os tipos de dados utilizados neste trabalho.

As DNN possuem diferentes arquiteturas que podem ser classificadas em discriminativas, generativas e híbridas [31]. Os tipos de aprendizado utilizados são: supervisionado, não supervisionado e uma combinação de ambos, respectivamente [1]. Um tipo de

arquitetura híbrida que vem ganhando destaque nos últimos tempos é a Rede Adversária Generativa, que será utilizada neste trabalho e melhor explicada na próxima subseção.

2.5. Rede Adversária Generativa

A Rede Adversária Generativa (GAN, do inglês *Generative Adversarial Network*) é uma arquitetura de DL baseada na Teoria dos Jogos [26]. Esta arquitetura é composta por duas DNN: o Gerador e o Discriminador. O Gerador busca aprender a distribuição das observações de treinamento, conseguindo criar novas observações muito semelhantes às originais. Já o Discriminador busca discernir se a observação é oriunda do Gerador ou do conjunto de dados real [59]. Então, essas duas redes competem, de forma que o Gerador busca enganar o Discriminador e este visa discernir entre dados reais e sintéticos.

Esta arquitetura apresenta uma abordagem generativa, com suas aplicações voltadas para geração de dados. Por exemplo, Wang *et al.* [66] utilizam GAN para construir um modelo focado na restauração de faces em fotos. Já Zhang *et al.* [74] geram imagens de alta qualidade a partir de uma descrição textual. Esta capacidade de geração de dados da GAN pode ser utilizada para a criação de um NIDS [47], onde o Discriminador dirá se o tráfego é anômalo ou não, e o Gerador ajudará a suprir a escassez de dados [41].

2.6. Transformador

O Transformador (do inglês, *Transformer*) é uma arquitetura de rede proposta por pesquisadores da *Google* em Vaswani *et al.* [64]. Ela utiliza o conceito de Atenção (do inglês, *Attention*) que ajuda a manter um contexto enquanto o modelo opera, auxiliando numa melhor produção de saídas [33]. Inicialmente, esta arquitetura foi desenvolvida para trabalhar com a tradução de textos, mas os próprios autores sugeriram que ela fosse testada em outras áreas [36].

O conceito de transformador se popularizou quando aplicado aos grandes modelos de linguagem (LLM, do inglês *Large Language Models*). Um dos exemplos é o Transformador Generativo Pré-treinado (GPT, do inglês *Generative Pre-trained Transformer*), que gera texto a partir de uma entrada em texto [9]. O destaque gerado por esse modelo impulsionou o estudo sobre transformadores, e este trabalho buscará usá-lo em conjunto com uma GAN para construção de um NIDS [70].

2.7. Inteligência Artificial Explicável

As técnicas de Inteligência Artificial Explicável (XAI, do inglês *Explainable Artificial Intelligence*) buscam tornar os sistemas de AI mais compreensíveis para os usuários [11]. Modelos mais complexos de AI, como DNN, são tratados como caixas pretas, pois não se sabe exatamente o que ocorre dentro do modelo, apenas se insere uma entrada e espera-se uma saída [39]. As técnicas de XAI buscam solucionar este problema, atribuindo explicabilidade aos modelos e auxiliando na interpretabilidade dos usuários. Saeed *et al.* [56] definem o uso de técnicas XAI em cinco perspectivas:

- Regulatória: Visa explicar um modelo para fins jurídicos;
- Científica: Para descobrir novos conceitos quando construindo um modelo;
- Industrial: Visando analisar a troca de desempenho por explicabilidade quando um modelo é implementado na indústria;
- Desenvolvimento do modelo: Para auxiliar no desenvolvimento e depuração de um novo modelo;

- Usuário final: Visando convencer o usuário final de que o modelo é confiável e funcional.

As perspectivas abordadas neste trabalho serão a Científica e a de Desenvolvimento do modelo. Elas serão usadas para auxiliar na depuração e para buscar alterações que melhorem o desempenho do modelo proposto [45].

Uma técnica de XAI encontrada na literatura é a das Explicações do Aditivo de Shapley (SHAP, do inglês *Shapley Additive Explanations*) [55]. Esta técnica busca explicar as previsões do modelo por meio de um gráfico, calculando a contribuição de cada característica de uma observação nos resultados obtidos. Outra técnica é a das Explicações Independentes do Modelo Interpretável Local (LIME, do inglês *Local Interpretable Model-agnostic Explanations*) [17]. O objetivo desta técnica é exibir, por meio de um gráfico, como uma previsão local do modelo sobre a observação X é feita. Enquanto o SHAP atua globalmente no modelo, o LIME atua localmente, possibilitando explicações específicas do comportamento do modelo em um determinado contexto de interesse.

A área de XAI está emergindo no momento, principalmente com a popularização de modelos de AI [56]. O objetivo desta área é fazer com que os modelos de AI não sejam mais caixas pretas, tornando eles mais compreensíveis para os diferentes níveis de usuários [30]. O número de publicações acadêmicas nesta área é crescente nos últimos anos, demonstrando haver um futuro promissor [8].

3. Objetivos

Este trabalho tem como objetivo a pesquisa e implementação de um NIDS utilizando GAN com Transformador e técnicas XAI, que fornecerão explicações sobre o modelo, contribuindo para o aumento do desempenho do mesmo.

4. Procedimentos metodológicos/Métodos e técnicas

O trabalho se iniciará com o estudo dos conceitos fundamentais: ML não supervisionado, NIDS, GAN, transformadores e XAI. Com todos esses assuntos bem compreendidos, o escopo será reduzido, agora focando o estudo em trabalhos que combinem o máximo de conceitos fundamentais possíveis. O próximo passo é implementar um NIDS com GAN e transformador, depurando-o com os datasets selecionados. Com o modelo já funcional, serão aplicadas as técnicas de XAI escolhidas para explicar o funcionamento do mesmo.

5. Cronograma de Execução

Esta seção exibe a tabela 1, que contém uma estimativa do período de execução para cada atividade presente no trabalho. As atividades aqui descritas foram estabelecidas com base no que foi dito na seção 4.

Atividades:

1. Levantamento bibliográfico;
2. Estudo dos conceitos fundamentais;
3. Estudo dos trabalhos com o escopo reduzido;
4. Implementação do NIDS baseado em GAN com Transformador;
5. Depuração do NIDS proposto com os datasets selecionados;

6. Aplicação de métodos XAI no modelo proposto;
7. Escrita do TCC (versão preliminar);
8. Escrita do TCC (versão para a banca examinadora).

Tabela 1. Cronograma de Execução

	jun	jul	ago	set	out	nov	dez	jan	fev
Atividade 1	•	•	•						
Atividade 2	•	•	•	•					
Atividade 3		•	•	•	•				
Atividade 4			•	•	•	•			
Atividade 5				•	•	•	•		
Atividade 6				•	•	•	•	•	
Atividade 7		•	•	•	•				
Atividade 8					•	•	•	•	•

6. Contribuições e/ou Resultados esperados

Espera-se concluir se a combinação de GAN com transformador é uma opção viável para implementar um NIDS. Também pretende-se demonstrar, através da aplicação no modelo proposto, a importância do uso de métodos de XAI na construção de modelos de ML modernos.

7. Espaço para assinaturas

Londrina, *26 de julho de 2024*.

_____ _____

Aluno Orientador

Referências

[1] Arwa Aldweesh, Abdelouahid Derhab, and Ahmed Z Emam. Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. *Knowledge-Based Systems*, 189:105124, 2020. DOI: https://doi.org/10.1016/j.knosys.2019.105124.

[2] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[3] Eirini Anthi, Lowri Williams, Małgorzata Słowińska, George Theodorakopoulos, and Pete Burnap. A supervised intrusion detection system for smart home iot devices. *IEEE Internet of Things Journal*, 6(5):9042–9053, 2019. DOI: https://doi.org/10.1109/JIOT.2019.2926365.

[4] Marcos V.O. Assis, Luiz F. Carvalho, Jaime Lloret, and Mario L. Proença. A gru deep learning system against attacks in software defined networks. *Journal of Network and Computer Applications*, 177:102942, 2021. DOI: https://doi.org/10.1016/j.jnca.2020.102942.

[5] Anurag Bhardwaj, Ritu Tyagi, Neha Sharma, Akhilendra Khare, Manbir Singh Punia, and Vikash Kumar Garg. Network intrusion detection in software defined networking with self-organized constraint-based intelligent learning framework. *Measurement: Sensors*, 24:100580, 2022. DOI: https://doi.org/10.1016/j.measen.2022.100580.

[6] Daniel M. Brandão Lent, Vitor G. da Silva Ruffo, Luiz F. Carvalho, Jaime Lloret, Joel J. P. C. Rodrigues, and Mario Lemes Proença. An unsupervised generative adversarial network system to detect ddos attacks in sdn. *IEEE Access*, 12:70690–70706, 2024. DOI: https://doi.org/10.1109/ACCESS.2024.3402069.

[7] Daniel M. Brandão Lent, Matheus P. Novaes, Luiz F. Carvalho, Jaime Lloret, Joel J. P. C. Rodrigues, and Mario Lemes Proença. A gated recurrent unit deep learning model to detect and mitigate distributed denial of service and portscan attacks. *IEEE Access*, 10:73229–73242, 2022. DOI: https://doi.org/10.1109/ACCESS.2022.3190008.

[8] Saša Brdnik and Boštjan Šumak. Current trends, challenges and techniques in xai field; a tertiary study of xai research. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, pages 2032–2038, 2024. DOI: https://doi.org/10.1109/MIPRO60963.2024.10569528.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. DOI: https://doi.org/10.48550/arXiv.2005.14165.

[10] Jorge Buzzio-García, Jaime Vergara, Santiago Ríos-Guiral, Christian Garzón, Sergio Gutiérrez, Juan F. Botero, Jose Luis Quiroz-Arroyo, and Jesús Arturo Pérez-Díaz. Exploring traffic patterns through network programmability: Introducing sdnflow, a comprehensive openflow-based statistics dataset for attack detection. *IEEE Access*, 12:42163–42180, 2024. DOI: https://doi.org/10.1109/ACCESS.2024.3378271.

[11] Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Claudio Stanzione. Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access*, 10:93575–93600, 2022. DOI: https://doi.org/10.1109/ACCESS.2022.3204171.

[12] Luiz F. Carvalho, Gilberto Fernandes, Joel J. P. C. Rodrigues, Leonardo S. Mendes, and Mario Lemes Proença. A novel anomaly detection system to assist network management in sdn environment. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6, 2017. DOI: https://doi.org/10.1109/ICC.2017.7997214.

[13] Luiz Fernando Carvalho, Taufik Abrão, Leonardo de Souza Mendes, and Mario Lemes Proença. An ecosystem for anomaly detection and mitigation in software-defined

networking. *Expert Systems with Applications*, 104:121–133, 2018. DOI: https://doi.org/10.1016/j.eswa.2018.03.027.

[14] Luiz Fernando Carvalho, Sylvio Barbon Jr, Leonardo de Souza Mendes, and Mario Lemes Proenca Jr. Unsupervised learning clustering and self-organized agents applied to help network management. *Expert Systems with Applications*, 54:29–47, 2016. DOI: https://doi.org/10.1016/j.eswa.2016.01.032.

[15] Dylan Chou and Meng Jiang. A survey on data-driven network intrusion detection. *ACM Comput. Surv.*, 54(9), oct 2021. DOI: https://doi.org/10.1145/3472753.

[16] CNN. Microsoft power outage and crowdstrike incident, 2024. URL: https://edition.cnn.com/2024/07/22/us/microsoft-power-outage-crowdstrike-it/index.html. Acessado em 22 de julho de 2024. URL.

[17] Sara Cuéllar, Matilde Santos, Fernando Alonso, Ernesto Fabregas, and Gonzalo Farias. Explainable anomaly detection in spacecraft telemetry. *Engineering Applications of Artificial Intelligence*, 133:108083, 2024. DOI: https://doi.org/10.1016/j.engappai.2024.108083.

[18] Marcos V.O. de Assis, Luiz F. Carvalho, Joel J.P.C. Rodrigues, Jaime Lloret, and Mario L. Proença Jr. Near real-time security system applied to sdn environments in iot networks using convolutional neural network. *Computers & Electrical Engineering*, 86:106738, 2020. DOI: https://doi.org/10.1016/j.compeleceng.2020.106738.

[19] Marcos VO De Assis, Anderson H Hamamoto, Taufik Abrão, and Mario Lemes Proença. A game theoretical based system using holt-winters and genetic algorithm with fuzzy logic for dos/ddos mitigation on sdn networks. *IEEE Access*, 5:9485–9496, 2017. DOI: https://doi.org/10.1109/ACCESS.2017.2702341.

[20] Marcos VO De Assis, Matheus P Novaes, Cinara B Zerbini, Luiz F Carvalho, Taufik Abrãao, and Mario L Proença. Fast defense system against attacks in software defined networks. *IEEE Access*, 6:69620–69639, 2018. DOI: https://doi.org/10.1109/ACCESS.2018.2878576.

[21] Ayesha S. Dina and D. Manivannan. Intrusion detection based on machine learning techniques in computer networks. *Internet of Things*, 16:100462, 2021. DOI: https://doi.org/10.1016/j.iot.2021.100462.

[22] Gilberto Fernandes, Joel JPC Rodrigues, Luiz Fernando Carvalho, Jalal F Al-Muhtadi, and Mario Lemes Proença. A comprehensive survey on network anomaly detection. *Telecommunication Systems*, 70:447–489, 2019. DOI: https://doi.org/10.1007/s11235-018-0475-8.

[23] Gilberto Fernandes Jr, Joel JPC Rodrigues, and Mario Lemes Proenca Jr. Autonomous profile-based anomaly detection system using principal component analysis and flow analysis. *Applied Soft Computing*, 34:513–525, 2015. DOI: https://doi.org/10.1016/j.asoc.2015.05.019.

[24] Xianwei Gao, Chun Shan, Changzhen Hu, Zequn Niu, and Zhen Liu. An adaptive ensemble machine learning model for intrusion detection. *IEEE Access*, 7:82512–82521, 2019. DOI: https://doi.org/10.1109/ACCESS.2019.2923640.

[25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: http://www.deeplearningbook.org.

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. DOI: https://doi.org/10.48550/arXiv.1406.2661.

[27] Neha Gupta, Vinita Jindal, and Punam Bedi. Cse-ids: Using cost-sensitive deep learning and ensemble algorithms to handle class imbalance in network-based intrusion detection systems. *Computers & Security*, 112:102499, 2022. DOI: https://doi.org/10.1016/j.cose.2021.102499.

[28] Anderson Hiroshi Hamamoto, Luiz Fernando Carvalho, Lucas Dias Hiera Sampaio, Taufik Abrão, and Mario Lemes Proença Jr. Network anomaly detection system using genetic algorithm and fuzzy logic. *Expert Systems with Applications*, 92:390–402, 2018. DOI: https://doi.org/10.1016/j.eswa.2017.09.013.

[29] Zhongyang Han, Jun Zhao, Henry Leung, King Fai Ma, and Wei Wang. A review of deep learning models for time series prediction. *IEEE Sensors Journal*, 21(6):7833–7848, 2021. DOI: https://doi.org/10.1109/JSEN.2019.2923982.

[30] AKM Bahalul Haque, A.K.M. Najmul Islam, and Patrick Mikalef. Explainable artificial intelligence (xai) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change*, 186:122120, 2023. DOI: https://doi.org/10.1016/j.techfore.2022.122120.

[31] William Grant Hatcher and Wei Yu. A survey of deep learning: Platforms, applications and emerging research trends. *IEEE Access*, 6:24411–24432, 2018. DOI: https://doi.org/10.1109/ACCESS.2018.2830661.

[32] Sergio Iglesias Pérez, Santiago Moral-Rubio, and Regino Criado. A new approach to combine multiplex networks and time series attributes: Building intrusion detection systems (ids) in cybersecurity. *Chaos, Solitons & Fractals*, 150:111143, 2021. DOI: https://doi.org/10.1016/j.chaos.2021.111143.

[33] Saidul Islam, Hanae Elmekki, Ahmed Elsebai, Jamal Bentahar, Nagat Drawel, Gaith Rjoub, and Witold Pedrycz. A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, 241:122666, 2024. DOI: https://doi.org/10.1016/j.eswa.2023.122666.

[34] Sana Ullah Jan, Saeed Ahmed, Vladimir Shakhov, and Insoo Koo. Toward a lightweight intrusion detection system for the internet of things. *IEEE Access*, 7:42450–42471, 2019. DOI: https://doi.org/10.1109/ACCESS.2019.2907965.

[35] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015. DOI: https://doi.org/10.1126/science.aaa8415.

[36] Hyeongwon Kang and Pilsung Kang. Transformer-based multivariate time series anomaly detection using inter-variable attention mechanism. *Knowledge-Based Systems*, 290:111507, 2024. DOI: https://doi.org/10.1016/j.knosys.2024.111507.

[37] Pushpajit Khaire and Praveen Kumar. A semi-supervised deep learning based video anomaly detection framework using rgb-d for surveillance of real-world critical environments. *Forensic Science International: Digital Investigation*, 40:301346, 2022. DOI: https://doi.org/10.1016/j.fsidi.2022.301346.

[38] Farrukh Aslam Khan, Abdu Gumaei, Abdelouahid Derhab, and Amir Hussain. A novel two-stage deep learning model for efficient network intrusion detection. *IEEE Access*, 7:30373–30385, 2019. DOI: https://doi.org/10.1109/ACCESS.2019.2899721.

[39] Rafał Kozik, Massimo Ficco, Aleksandra Pawlicka, Marek Pawlicki, Francesco Palmieri, and Michał Choraś. When explainability turns into a threat - using xai to fool a fake news detection method. *Computers & Security*, 137:103599, 2024. DOI: https://doi.org/10.1016/j.cose.2023.103599.

[40] Sang-Woong Lee, Mokhtar Mohammadi, Shima Rashidi, Amir Masoud Rahmani, Mohammad Masdari, Mehdi Hosseinzadeh, et al. Towards secure intrusion detection systems using deep learning techniques: Comprehensive analysis and review. *Journal of Network and Computer Applications*, 187:103111, 2021. DOI: https://doi.org/10.1016/j.jnca.2021.103111.

[41] Willone Lim, Kelvin Sheng Chek Yong, Bee Theng Lau, and Colin Choon Lin Tan. Future of generative adversarial networks (gan) for anomaly detection in network security: A review. *Computers & Security*, 139:103733, 2024. DOI: https://doi.org/10.1016/j.cose.2024.103733.

[42] Wei Lo, Hamed Alqahtani, Kutub Thakur, Ahmad Almadhor, Subhash Chander, and Gulshan Kumar. A hybrid deep learning based intrusion detection system using spatial-temporal representation of in-vehicle network traffic. *Vehicular Communications*, 35:100471, 2022. DOI: https://doi.org/10.1016/j.vehcom.2022.100471.

[43] Panos Louridas and Christof Ebert. Machine learning. *IEEE Software*, 33(5):110–115, 2016. DOI: https://doi.org/10.1109/MS.2016.114.

[44] Mohssen Mohammed, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashier. *Machine learning: algorithms and applications*. Crc Press, 2016. DOI: https://doi.org/10.1201/9781315371658.

[45] Subash Neupane, Jesse Ables, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, and Maria Seale. Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities. *IEEE Access*, 10:112392–112415, 2022. DOI: https://doi.org/10.1109/ACCESS.2022.3216617.

[46] Matheus P. Novaes, Luiz F. Carvalho, Jaime Lloret, and Mario Lemes Proença. Long short-term memory and fuzzy logic for anomaly detection and mitigation in

software-defined network environment. *IEEE Access*, 8:83765–83781, 2020. DOI: https://doi.org/10.1109/ACCESS.2020.2992044.

[47] Matheus P. Novaes, Luiz F. Carvalho, Jaime Lloret, and Mario Lemes Proença. Adversarial deep learning approach detection and defense against ddos attacks in sdn environments. *Future Generation Computer Systems*, 125:156–167, 2021. DOI: https://doi.org/10.1016/j.future.2021.06.047.

[48] Yazan Otoum and Amiya Nayak. As-ids: Anomaly and signature based ids for the internet of things. *Journal of Network and Systems Management*, 29:1–26, 2021. DOI: https://doi.org/10.1007/s10922-021-09589-6.

[49] Hamed Haddad Pajouh, Reza Javidan, Raouf Khayami, Ali Dehghantanha, and Kim-Kwang Raymond Choo. A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in iot backbone networks. *IEEE Transactions on Emerging Topics in Computing*, 7(2):314–323, 2019. DOI: https://doi.org/10.1109/TETC.2016.2633228.

[50] Eduardo HM Pena, Luiz F Carvalho, Sylvio Barbon Jr, Joel JPC Rodrigues, and Mario Lemes Proença Jr. Anomaly detection using the correlational paraconsistent machine with digital signatures of network segment. *Information Sciences*, 420:313–328, 2017. DOI: https://doi.org/10.1016/j.ins.2017.08.074.

[51] M. Lemes Proença, C. Coppelmans, M. Bottoli, A. Alberti, and L. S. Mendes. The hurst parameter for digital signature of network segment. In José Neuman de Souza, Petre Dini, and Pascal Lorenz, editors, *Telecommunications and Networking - ICT 2004*, pages 772–781, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. DOI: https://doi.org/10.1007/978-3-540-27824-5_103.

[52] Mario Lemes Proença, Bruno Bogaz Zarpelão, and Leonardo S Mendes. Anomaly detection for network servers using digital signature of network segment. In *Advanced Industrial Conference on Telecommunications/Service Assurance with Partial and Intermittent Resources Conference/E-Learning on Telecommunications Workshop (AICT/SAPIR/ELETE'05)*, pages 290–295. IEEE, 2005. DOI: https://doi.org/10.1109/AICT.2005.26.

[53] Mario Lemes Proença Jr., Gilberto Fernandes Jr., Luiz F. Carvalho, Marcos V. O. de Assis, and Joel J. P. C. Rodrigues. Digital signature to help network management using flow analysis. *International Journal of Network Management*, 26(2):76–94, 2016. DOI: https://doi.org/10.1002/nem.1892.

[54] Markus Ring, Sarah Wunderlich, Deniz Scheuring, Dieter Landes, and Andreas Hotho. A survey of network-based intrusion detection data sets. *Computers & Security*, 86:147–167, 2019. DOI: https://doi.org/10.1016/j.cose.2019.06.005.

[55] Khushnaseeb Roshan and Aasim Zafar. Using kernel shap xai method to optimize the network anomaly detection model. In *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 74–80, 2022. DOI: https://doi.org/10.23919/INDIACom54597.2022.9763241.

[56] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*,

263:110273, 2023. DOI: https://doi.org/10.1016/j.knosys.2023.110273.

[57] Gustavo F Scaranti, Luiz F Carvalho, Sylvio Barbon, and Mario Lemes Proença. Artificial immune systems and fuzzy logic to detect flooding attacks in software-defined networks. *IEEE Access*, 8:100172–100184, 2020. DOI: https://doi.org/10.1109/ACCESS.2020.2997939.

[58] Gustavo Frigo Scaranti, Luiz Fernando Carvalho, Sylvio Barbon, Jaime Lloret, and Mario Lemes Proença. Unsupervised online anomaly detection in software defined network environments. *Expert Systems with Applications*, 191:116225, 2022. DOI: https://doi.org/10.1016/j.eswa.2021.116225.

[59] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, 2019. DOI: https://doi.org/10.1016/j.media.2019.01.010.

[60] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *International Conference on Information Systems Security and Privacy*, 2018. URL: https://api.semanticscholar.org/CorpusID:4707749.

[61] Iman Sharafaldin, Arash Habibi Lashkari, Saqib Hakak, and Ali A. Ghorbani. Developing realistic distributed denial of service (ddos) attack dataset and taxonomy. In *2019 International Carnahan Conference on Security Technology (ICCST)*, pages 1–8, 2019. DOI: https://doi.org/10.1109/CCST.2019.8888419.

[62] Ajay Shrestha and Ausif Mahmood. Review of deep learning algorithms and architectures. *IEEE Access*, 7:53040–53065, 2019. DOI: https://doi.org/10.1109/ACCESS.2019.2912200.

[63] Bayu Adhi Tama, Marco Comuzzi, and Kyung-Hyune Rhee. Tse-ids: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system. *IEEE Access*, 7:94497–94507, 2019. DOI: https://doi.org/10.1109/ACCESS.2019.2928048.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. DOI: https://doi.org/10.48550/arXiv.1706.03762.

[65] R. Vinayakumar, Mamoun Alazab, K. P. Soman, Prabaharan Poornachandran, Ameer Al-Nemrat, and Sitalakshmi Venkatraman. Deep learning approach for intelligent intrusion detection system. *IEEE Access*, 7:41525–41550, 2019. DOI: https://doi.org/10.1109/ACCESS.2019.2895334.

[66] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. DOI: https://doi.org/10.48550/arXiv.2101.04061.

[67] Junfeng Xie, F Richard Yu, Tao Huang, Renchao Xie, Jiang Liu, Chenmeng Wang, and Yunjie Liu. A survey of machine learning techniques applied to software defined networking (sdn): Research issues and challenges. *IEEE Communications Surveys & Tutorials*, 21(1):393–430, 2018. DOI: https://doi.org/10.1109/COMST.2018.2866942.

[68] Yang Xin, Lingshuang Kong, Zhi Liu, Yuling Chen, Yanmiao Li, Hongliang Zhu, Mingcheng Gao, Haixia Hou, and Chunhua Wang. Machine learning and deep learning methods for cybersecurity. *Ieee access*, 6:35365–35381, 2018. DOI: https://doi.org/10.1109/ACCESS.2018.2836950.

[69] Linli Xu, Martha White, and Dale Schuurmans. Optimal reverse prediction: A unified perspective on supervised, unsupervised and semi-supervised learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 1137–1144, New York, NY, USA, 2009. Association for Computing Machinery. DOI: https://doi.org/10.1145/1553374.1553519.

[70] Liyan Xu, Kang Xu, Yinchuan Qin, Yixuan Li, Xingting Huang, Zhicheng Lin, Ning Ye, and Xuechun Ji. Tgan-ad: Transformer-based gan for anomaly detection of time series data. *Applied Sciences*, 12(16), 2022. DOI: https://doi.org/10.3390/app12168085.

[71] Li Yang, Abdallah Moubayed, and Abdallah Shami. Mth-ids: A multitiered hybrid intrusion detection system for internet of vehicles. *IEEE Internet of Things Journal*, 9(1):616–632, 2022. DOI: https://doi.org/10.1109/JIOT.2021.3084796.

[72] Zhen Yang, Xiaodong Liu, Tong Li, Di Wu, Jinjiang Wang, Yunwei Zhao, and Han Han. A systematic literature review of methods and datasets for anomaly-based network intrusion detection. *Computers & Security*, 116:102675, 2022. DOI: https://doi.org/10.1016/j.cose.2022.102675.

[73] Cinara Brenda Zerbini, Luiz Fernando Carvalho, Taufik Abrão, and Mario Lemes Proenca Jr. Wavelet against random forest for anomaly mitigation in software-defined networking. *Applied Soft Computing*, 80:138–153, 2019. DOI: https://doi.org/10.1016/j.asoc.2019.02.046.

[74] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. DOI: https://doi.org/10.48550/arXiv.1612.03242.