Artigo Científico

Revista sítio novo
Instituto Federal do Tocantins

Inteligência Artificial na segurança cibernética: ataques adversariais em sistemas de detecção de intrusão

https://doi.org/10.47236/2594-7036.2026.v10.2190

Bernardo Vivian Vieira[1]
Marco Antônio Sandini Trentin[2]

Data de submissão concluída: 8/5/2026. Data de aprovação: 25/8/2026. Data de publicação: 25/11/2026.

Resumo – A crescente sofisticação dos ataques cibernéticos, especialmente os adversariais contra modelos de inteligência artificial (IA), representa um desafio emergente à segurança digital. Este trabalho investiga a eficácia de técnicas de IA na mitigação desses ataques em Sistemas de Detecção de Intrusão (IDS), por meio de experimentos utilizando o dataset CIC-IDS2017 (2,83 milhões de amostras). A metodologia inclui o treinamento de três modelos de aprendizado de máquina (Decision Tree, Random Forest, MLP), avaliação de robustez adversarial (FGSM, PGD, C&W) e mensuração de desempenho com métricas como acurácia, taxa de detecção, taxa de falsos positivos e latência. Resultados mostram que Random Forest obteve 83% de exatidão sob ataque contra, aproximadamente, 50% dos demais modelos ($p < 0{,}0001$), embora o forte desbalanceamento de classes cause 96% de falsos negativos em ataques minoritários, limitando a viabilidade em produção sem estratégias de mitigação. O texto oferece evidências empíricas sobre a resiliência de IA em IDS e identifica limitações críticas para implantação em ambientes reais.
Palavras-chave: Aprendizado de máquina adversarial. Detecção de intrusão. Inteligência Artificial. Segurança computacional.

Artificial Intelligence in cybersecurity: adversarial attacks on intrusion detection systems

Abstract – The growing sophistication of cyberattacks, especially adversarial attacks against artificial intelligence (AI) models, is an emerging challenge to digital security. This work investigates the effectiveness of AI techniques in mitigating such attacks in Intrusion Detection Systems (IDS), through experiments using the CIC-IDS2017 dataset (2.83 million samples). The methodology includes training three machine-learning models (Decision Tree, Random Forest, MLP), evaluating adversarial robustness (FGSM, PGD, C&W), and measuring performance with metrics such as accuracy, detection rate, false-positive rate, and latency. Results show Random Forest maintains 83% exactness under attack against approximately 50% for other models ($p < 0.0001$), though severe class imbalance causes 96% false negatives for minority attacks, limiting production viability without mitigation strategies. The text provides empirical evidence on AI resilience in IDS and identifies critical limitations for real-world deployment.

[1] Bacharel em Ciência da Computação pelo Instituto de Tecnologia da Universidade de Passo Fundo. Soledade, Rio Grande do Sul, Brasil. 179835@upf.br https://orcid.org/0009-0009-6062-0811 http://lattes.cnpq.br/8368622073109410.
[2] Doutor em Informática na Educação pela Universidade Federal do Rio Grande do Sul. Professor do curso de Ciência da Computação e dos Programas de Pós-Graduação em Ensino de Ciências e Matemática e em Computação Aplicada da Universidade de Passo Fundo. Passo Fundo, Rio Grande do Sul, Brasil. trentin@upf.br https://orcid.org/0000-0002-8025-8700 http://lattes.cnpq.br/4746488333257798.

Keywords: Adversarial machine learning. Artificial Intelligence. Computer security. Intrusion detection.

Inteligencia Artificial en la seguridad cibernética: ataques adversariales en sistemas de detección de intrusiones

Resumen – La creciente sofisticación de los ataques cibernéticos, especialmente los adversariales contra modelos de inteligencia artificial (IA), representa un desafío emergente para la seguridad digital. Este trabajo investiga la eficacia de técnicas de IA en la mitigación de estos ataques en Sistemas de Detección de Intrusiones (IDS), mediante experimentos utilizando el dataset CIC-IDS2017 (2,83 millones de muestras). La metodología incluye el entrenamiento de tres modelos de aprendizaje automático (Decision Tree, Random Forest, MLP), la evaluación de la robustez adversarial (FGSM, PGD, C&W) y medición del desempeño con métricas como exactitud, tasa de detección, tasa de falsos positivos y latencia. Los resultados muestran que Random Forest obtuvo un 83% de exactitud bajo ataque contra, aproximadamente, el 50% de los demás modelos ($p < 0,0001$), aunque el fuerte desbalanceo de clases provoca un 96% de falsos negativos en ataques minoritarios, limitando la viabilidad en producción sin estrategias de mitigación. El texto ofrece evidencias empíricas sobre la resiliencia de la IA en IDS e identifica limitaciones críticas para su implementación en ambientes reales.
Palabras clave: Aprendizaje automático adversarial. Detección de intrusiones. Inteligencia Artificial. Seguridad computacional.

Introdução

A utilização de inteligência artificial (IA) em segurança cibernética amplia a capacidade de detecção para além de assinaturas estáticas, mas também introduz novas superfícies de ataque. Entre elas, destacam-se os ataques adversariais, definidos como manipulações intencionais em dados de entrada que exploram vulnerabilidades de modelos de aprendizado de máquina, induzindo erros de classificação em mecanismos defensivos, como os Sistemas de Detecção de Intrusões (IDS). O interesse pelo tema decorre da relevância crescente da cibersegurança para setores estratégicos e do impacto financeiro e institucional de falhas de detecção. A expansão da inteligência artificial e o debate sobre seus riscos têm alcançado múltiplos domínios, da educação à automação (Mosciaro; Kanashiro, 2025; Silva; Souza, 2025; Pontes; Victor, 2022), o que reforça a urgência de investigar também suas vulnerabilidades em contextos críticos de segurança.

Relatórios internacionais apontam prejuízos milionários associados a ataques sofisticados Business, 2025; IBM Security, 2025; European Union Agency for Cybersecurity, 2024), enquanto, no Brasil, a Lei Geral de Proteção de Dados (Lei nº 13.709/2018) reforça a necessidade de soluções tecnológicas capazes de prevenir incidentes e assegurar a conformidade regulatória (Brasil, 2018). Do ponto de vista científico, ainda existem lacunas quanto à avaliação sistemática da resiliência de IDS baseados em aprendizado de máquina frente a ataques adversariais adaptativos, que buscam a evasão em tempo real. Embora a literatura apresente avanços, são escassos os estudos que conciliem experimentação prática com ferramentas amplamente utilizadas em ambientes reais, como o Suricata, e, crucialmente, que adotem métricas orientadas a cenários de produção, considerando o risco operacional (FPR e latência).

Diante desse contexto, este trabalho investiga se modelos de aprendizado de máquina integrados a Sistemas de Detecção de Intrusão conseguem preservar desempenho operacional quando submetidos a ataques adversariais. Para tanto, avalia-se a integração do IDS open-source Suricata a três modelos de IA (Decision Tree, Random Forest e MLP) em ambiente controlado, submetendo-os ao dataset CIC-IDS2017 (Canadian Institute for Cybersecurity, 2017) e a exemplos adversariais gerados por técnicas FGSM, PGD e C&W. O desempenho é quantificado mediante métricas consolidadas na literatura, incluindo acurácia global, taxa de detecção (TPR), taxa de falsos positivos (FPR) e latência de detecção, permitindo avaliar tanto a eficácia defensiva quanto a viabilidade operacional dos sistemas propostos.

Como contribuição, propomos um protocolo experimental reprodutível que permite analisar a resiliência de IDS diante de ataques adversariais, além de comparar algoritmos quanto ao equilíbrio entre desempenho e robustez por meio de análise estatística de significância.

A estrutura do texto contempla, após esta introdução, a fundamentação teórica, a descrição dos materiais e métodos, a análise dos resultados e, por fim, as conclusões acompanhadas de perspectivas para trabalhos futuros.

Fundamentação Teórica

Esta seção apresenta os conceitos centrais que sustentam o trabalho, sendo o papel da inteligência artificial (IA) na segurança cibernética, as características dos ataques adversariais contra modelos de IA, os fundamentos de Sistemas de Detecção de Intrusões (IDS) com ênfase no Suricata, os principais algoritmos e as métricas de avaliação empregadas.

Segurança Cibernética e Inteligência Artificial

A segurança cibernética contemporânea enfrenta ameaças cada vez mais sofisticadas, como ransomware[3], zero-day exploits[4] e ataques distribuídos de negação de serviço (DDoS), que podem comprometer a disponibilidade de sistemas inteiros. Entre os obstáculos mais recentes destaca-se o uso de ataques adversariais, nos quais pequenas manipulações intencionais das entradas exploram fragilidades dos modelos de aprendizado de máquina, resultando em classificações incorretas até mesmo em sistemas de defesa como os IDS (Goodfellow et al., 2014).

Nesse contexto, técnicas de IA têm sido adotadas como alternativas promissoras para detecção proativa e adaptativa de ameaças, por permitirem identificar padrões complexos que extrapolam regras estáticas e assinaturas. Sistemas baseados em aprendizado profundo conseguem, por exemplo, reconhecer comportamentos anômalos em tráfego de rede que passariam despercebidos por soluções tradicionais. Entretanto, a adoção de IA também amplia a superfície de ataque, introduzindo novos vetores exploráveis por adversários, como manipulação de dados de treinamento (data poisoning)[5] e exploração de falhas em algoritmos de inferência[6].

[3] Tipo de malware que sequestra dados do sistema-alvo por meio de criptografia e exige pagamento para liberação.
[4] Vulnerabilidade explorada por atacantes antes de ser conhecida ou corrigida pelos desenvolvedores.
[5] Inserção de exemplos maliciosos no conjunto de treinamento com o objetivo de comprometer a acurácia ou introduzir vieses no modelo.
[6] Ataques que buscam extrair informações sensíveis do modelo, como dados de treinamento ou atributos privados, a partir de consultas sucessivas.

Revista sítio novo
Instituto Federal do Tocantins

Ataques Adversariais em Sistemas de IA

Vulnerabilidades em sistemas de IA podem resultar em prejuízos financeiros significativos. Segundo o projeto MITRE ATLAS (Adversarial Threat Landscape for AI Systems), casos documentados demonstram que falhas em sistemas de reconhecimento facial de aplicação governamental resultaram em perdas superiores a US$ 77 milhões (Mitre Corporation, 2023). No contexto de IDS, perturbações sutis nos dados de entrada podem degradar classificadores baseados em aprendizado profundo, comprometendo taxas de detecção e a confiabilidade operacional do sistema (Zhang et al., 2022).

Essas perturbações são conhecidas como exemplos adversariais, isto é, amostras de dados modificadas de forma sutil, porém intencional, com o objetivo de enganar modelos de aprendizado de máquina. Eles têm sido amplamente utilizados em experimentos acadêmicos para avaliar a resiliência de algoritmos e, em ambientes práticos, representam um risco concreto de evasão de sistemas de detecção.

Entre os principais métodos de geração de exemplos adversariais, destacam-se:

- Fast Gradient Sign Method (FGSM): método de passo único que gera o exemplo adversarial somando à entrada original uma perturbação proporcional ao sinal do gradiente da função de perda em relação aos dados de entrada (Goodfellowet al., 2014). Formalmente, o exemplo adversarial é obtido por $x_adv = x + \varepsilon \cdot sign(\nabla_x J(\theta, x, y))$, em que x é a entrada original, y o rótulo verdadeiro, θ os parâmetros do modelo, J a função de perda e ε a magnitude da perturbação, que controla a intensidade do ataque. Por utilizar o gradiente uma única vez, é computacionalmente barato, porém menos eficaz que métodos iterativos: uma alteração mínima e calculada em cada atributo de entrada, no sentido que mais aumenta a perda, já é suficiente para induzir a classificação incorreta.
- Projected Gradient Descent (PGD): versão iterativa do FGSM, considerado um dos ataques de primeira ordem mais fortes (Madry et al., 2018). A cada iteração t aplica-se um passo de tamanho α no sentido do sinal do gradiente e, em seguida, projeta-se o resultado de volta à vizinhança admissível em torno de x, segundo $x^{(t+1)} = \Pi_{x+S}(x^{(t)} + \alpha \cdot sign(\nabla_x J(\theta, x^{(t)}, y)))$, em que S é a esfera de raio ε (tipicamente na norma $L\infty$) e Π denota a projeção que mantém a perturbação dentro desse limite. A repetição por k iterações explora melhor a superfície de perda do que o passo único do FGSM, tornando o ataque mais eficaz contra modelos robustos, ao custo de maior esforço computacional.
- Carlini & Wagner (C&W): formula o ataque como um problema de otimização que busca a menor perturbação δ capaz de provocar a classificação incorreta, minimizando $||\delta||_2 + c \cdot f(x + \delta)$, sujeito a que $x + \delta$ permaneça no domínio válido das entradas (Carlini; Wagner, 2017). O primeiro termo penaliza o tamanho da perturbação, enquanto a função objetivo f só é satisfeita quando a amostra é classificada incorretamente com uma margem de confiança κ; a constante c equilibra os dois objetivos. Por otimizar diretamente essa formulação na norma L2, o C&W encontra perturbações mínimas e de alta confiança, o que lhe permite contornar defesas eficazes apenas contra-ataques mais simples como o FGSM.

No contexto de IDS, tais ataques podem se manifestar por modificações sutis em cabeçalhos de pacotes ou por inserção de padrões levemente alterados de tráfego. Além disso, adversários podem recorrer a ataques de evasão, voltados a escapar da detecção em tempo real, e a ataques de envenenamento de dados, que

comprometem o modelo ainda na fase de treinamento (Biggio; Roli, 2018). Compreender esses mecanismos é essencial para desenvolver defesas que conciliem desempenho e robustez.

Sistemas de Detecção de Intrusões e Suricata

Os IDS são mecanismos destinados a monitorar redes e hosts em busca de atividades anômalas ou maliciosas. Tradicionalmente, dividem-se em abordagens baseadas em assinaturas e em anomalias. As primeiras são eficazes contra ameaças conhecidas, mas falham frente a novos ataques. As segundas identificam desvios de comportamento, ampliando a capacidade de resposta, mas também introduzem desafios relacionados a falsos positivos. Entre as soluções de código aberto, o Suricata se destaca por combinar inspeção profunda de pacotes, detecção baseada em regras e análise em tempo real (OISF, 2024). Além disso, sua arquitetura multithread nativa e a ampla base de regras mantida pela comunidade o tornam adequado a cenários de alto tráfego, diferenciando-o de ferramentas como Snort e Zeek. Tais características justificam sua adoção neste trabalho como plataforma experimental, por aliar robustez técnica a ampla difusão prática.

Algoritmos e Métricas para Detecção

A literatura especializada apresenta uma ampla variedade de algoritmos de aprendizado de máquina aplicados à detecção de intrusões. Entre os mais recorrentes estão as Decision Trees, o Random Forest e as redes neurais profundas, em razão da capacidade de modelar padrões complexos em dados de tráfego de rede (Handaet al., 2019). Cada um possui características distintas que afetam sua resiliência e interpretabilidade em cenários de segurança.

As Decision Trees (árvores de Decisão) são valorizadas por sua interpretabilidade, permitindo entender as regras que levam a uma detecção. No entanto, sua simplicidade as torna potencialmente vulneráveis a pequenas modificações nos dados que alterem as condições de decisão, podendo ser suscetíveis a ataques adversariais bem direcionados. O Random Forest, um método de ensemble que combina múltiplas árvores de decisão, amplia a robustez e a capacidade de generalização, reduzindo a variância e tornando-o mais resistente a ruídos e pequenas perturbações adversariais devido à sua natureza agregada. As redes neurais profundas, por sua vez, oferecem desempenho de ponta em tarefas complexas de classificação, mas são notoriamente mais vulneráveis a ataques adversariais devido à sua sensibilidade a pequenas alterações nos dados de entrada, exigindo estratégias defensivas mais sofisticadas. Técnicas não supervisionadas, como K-means e autoencoders, também têm sido exploradas para detecção de anomalias quando dados rotulados são escassos.

Para mensurar o desempenho dos algoritmos aplicados à detecção de intrusões, foram adotadas métricas tradicionalmente utilizadas em problemas de classificação, todas derivadas da matriz de confusão. Esta é definida a partir de quatro contagens: verdadeiros positivos (TP, ataques corretamente detectados), falsos positivos (FP, tráfego benigno sinalizado como ataque), verdadeiros negativos (TN, tráfego benigno corretamente classificado) e falsos negativos (FN, ataques não detectados). A partir dessas contagens definem-se:

- Taxa de Detecção (True Positive Rate; TPR), também chamada de recall: dada por TPR = TP / (TP + FN), representa a proporção de ataques corretamente identificados entre todos os ataques reais;

- Taxa de Falsos Positivos (False Positive Rate; FPR): dada por FPR = FP / (FP + TN), corresponde à proporção de eventos legítimos incorretamente classificados como ataques;
- Precisão (Precision): dada por Precision = TP / (TP + FP), mede a proporção de alertas emitidos que correspondem, de fato, a ataques;
- Acurácia global (Accuracy): dada por Accuracy = (TP + TN) / (TP + TN + FP + FN), expressa a proporção de classificações corretas, positivas e negativas, sobre o total de instâncias avaliadas;
- F1-Score: média harmônica entre precisão e recall, dada por F1 = 2 · (Precision · Recall) / (Precision + Recall), que penaliza modelos com forte desequilíbrio entre as duas medidas;
- Latência de detecção: tempo médio decorrido entre a ocorrência do ataque e sua identificação pelo IDS.

Em sistemas com severo desbalanceamento de classes, como é comum em IDS, onde o tráfego benigno é vastamente majoritário, a acurácia global pode ser enganosa, pois é dominada pelo desempenho na classe majoritária. Nesses contextos, são essenciais as métricas macro (Macro Precision, Macro Recall e Macro F1), calculadas como a média aritmética do valor obtido individualmente para cada uma das 15 classes, sem ponderação por frequência. Por tratarem todas as classes igualmente, penalizam a falha na detecção de ataques minoritários, como XSS ou Infiltration, e refletem melhor a capacidade real do modelo do que a acurácia global.

A escolha adequada das métricas é crucial para uma análise comparativa confiável entre diferentes abordagens. Estudos como o de Zhang et al. (2022) reforçam que considerar apenas a acurácia pode ser insuficiente, já que métricas como a taxa de falsos positivos e a latência têm impacto direto na operação de sistemas em produção. Adicionalmente, em ambientes críticos de produção, deve-se reconhecer que falsos negativos (ataques não detectados) têm impacto operacional significativamente maior que falsos positivos (investigações desnecessárias), embora a quantificação precisa desses custos esteja além do escopo deste estudo e seja proposta como trabalho futuro.

Trabalhos Relacionados

A literatura sobre vulnerabilidade e defesa de IDS baseados em aprendizado de máquina tem crescido rapidamente. Estudos demonstram o impacto de perturbações imperceptíveis sobre classificadores profundos e propõem defesas como regularização, treinamento adversarial e taxonomias de mecanismos defensivos (Zhang et al., 2022; Khaleel et al., 2024), além de analisarem desafios específicos de ambientes de IoT (ALKADI et al., 2023) e o uso de aprendizado federado para preservar dados locais, apesar de suas limitações de comunicação e heterogeneidade (Agrawal et al., 2022; Ghimire; Rawat, 2022). Revisões da área reforçam a complementaridade entre abordagens supervisionadas e não supervisionadas (Handa et al., 2019), enquanto relatórios nacionais destacam a evolução de ataques a infraestruturas críticas (CERT.Br, 2025). Apesar desses avanços, permanece a escassez de protocolos experimentais padronizados que assegurem reprodutibilidade e comparação justa entre estudos, lacuna que este trabalho busca preencher ao propor e aplicar uma metodologia estruturada de avaliação da resiliência de IDS frente a ataques adversariais.

Materiais e Métodos

Este capítulo descreve a metodologia adotada no desenvolvimento do trabalho, construída a partir das lacunas identificadas na revisão de literatura e busca estruturar um protocolo experimental reprodutível para avaliar a resiliência de Sistemas de Detecção de Intrusões (IDS) frente a ataques adversariais. As etapas descritas incluem o estudo teórico e seleção de algoritmos de inteligência artificial, a configuração do ambiente de testes, a utilização da base CIC-IDS2017, a integração entre modelos de IA e IDS com a coleta de métricas e, por fim, a validação dos resultados.

Estudo teórico e seleção de algoritmos de IA

A primeira etapa consiste em um levantamento teórico orientado pela literatura mais recente sobre aprendizado de máquina aplicado à segurança cibernética. A ênfase recai sobre algoritmos de classificação supervisionada, como Decision Trees, Random Forest e redes neurais profundas. Estudos anteriores indicam que essas técnicas apresentam bom equilíbrio entre interpretabilidade, custo computacional e capacidade de modelar padrões complexos em tráfego de rede (Handaet al., 2019; Zhang et al., 2022).

Os algoritmos baseados em árvores são considerados pela facilidade de implementação e pela clareza em relação às regras de decisão extraídas, o que favorece a explicação de resultados em ambientes críticos. O Random Forest será analisado por sua robustez estatística, obtida pela combinação de múltiplas árvores, enquanto as redes neurais profundas serão avaliadas por representarem o estado da arte em tarefas de classificação, ainda que sua suscetibilidade a ataques adversariais esteja amplamente documentada (Goodfellow et al., 2014).

Para as redes neurais profundas, foi implementada uma Multi-Layer Perceptron (MLP) do scikit-learn com duas camadas ocultas de 128 e 64 neurônios (ativação ReLU) e saída softmax para 15 classes, treinada com otimizador Adam (learning rate 0,001), early stopping com patience de 5 épocas e máximo de 100 épocas sobre 79 features de entrada. Esta configuração foi selecionada após grid search preliminar avaliando arquiteturas com 1 a 3 camadas ocultas e diferentes números de neurônios por camada.

A definição do conjunto de algoritmos a ser utilizado permitirá estabelecer comparações entre diferentes paradigmas, analisando suas forças e limitações frente a condições adversariais.

Configuração do ambiente de testes

A segunda etapa prevê a construção de um ambiente experimental controlado. Esse ambiente será composto por uma máquina virtual rodando o sistema operacional Ubuntu 24.04 LTS, na qual será instalado e configurado o IDS Suricata, reconhecido por sua ampla utilização em redes corporativas e por sua arquitetura multithread que possibilita processamento em alto desempenho (OISF, 2024).

A camada de inteligência artificial será implementada em Python, empregando bibliotecas como Scikit-learn e Pandas, que oferecem suporte consolidado para manipulação de dados e treinamento de modelos. Além disso, ferramentas auxiliares como tcpreplay serão utilizadas para reproduzir tráfego de pacotes em cenários controlados, permitindo a validação da injeção de pacotes. Esse arranjo experimental busca equilibrar simplicidade de replicação com fidelidade a cenários reais de rede, além de alinhar-se à recomendação do Centro de Estudos, Resposta e Tratamento de Incidentes de Segurança no Brasil (CERT.br), que destaca a importância da adoção de ferramentas abertas e auditáveis na construção de soluções defensivas (CERT.Br, 2025).

Base de dados (CIC-IDS2017)

Para treinamento e validação dos modelos, será utilizado o conjunto CIC-IDS2017, desenvolvido pelo Canadian Institute for Cybersecurity (Canadian Institute for Cybersecurity, 2017). Trata-se de um dos conjuntos de dados mais empregados na área, reunindo registros de tráfego legítimo juntamente com diferentes tipos de ataque, como força bruta, negação de serviço distribuída, atividades de botnets e exploração de vulnerabilidades em aplicações web. Sua adoção justifica-se pela variedade de cenários contemplados e pela ampla aceitação na comunidade científica, o que possibilita a comparação direta dos resultados com trabalhos correlatos (Zhang et al., 2022). Embora datado de 2017 e, portanto, sem contemplar vetores de ataque mais recentes, o conjunto permanece pertinente por ser um benchmark consolidado e rotulado, adequado ao objetivo deste trabalho, que é avaliar a resiliência relativa das arquiteturas de aprendizado de máquina diante de ataques adversariais, e não catalogar ameaças atuais.

O pré-processamento dos dados será necessário para adequar os atributos às demandas dos algoritmos de aprendizado de máquina. As principais etapas incluem:

- Tratamento de valores ausentes e infinitos: identificação e substituição de valores faltantes (NaN) e infinitos (Inf) nos atributos, utilizando estratégias como a remoção das linhas correspondentes ou a imputação com a média/mediana.
- Codificação de variáveis categóricas: transformação de atributos não numéricos (e.g., Protocol, Flag) em representações numéricas adequadas, como One-Hot Encoding ou Label Encoding, para poderem ser processados pelos modelos de IA.
- Normalização e escalonamento de features numéricas: aplicação de técnicas como StandardScaler (padronização Z-score) para padronizar a escala dos atributos numéricos. Isso evita que features com grandes magnitudes dominem o processo de treinamento e melhora a convergência de alguns algoritmos, como redes neurais.
- Redução de dimensionalidade/Seleção de features: avaliação da pertinência de cada atributo para a tarefa de detecção, utilizando métodos estatísticos ou baseados em modelos para remover features redundantes ou irrelevantes, podendo otimizar o tempo de treinamento e reduzir o ruído nos dados.

Essa preparação busca reduzir vieses de classificação, ao mesmo tempo em que preserva a distribuição original das classes, evitando distorções na análise comparativa.

Integração IA-IDS e coleta de métricas

Os modelos de IA serão avaliados utilizando os registros em formato JSON exportados pelo Suricata. Essa integração permite que o mesmo ambiente de monitoramento seja usado tanto no modo convencional quanto em experimentos com algoritmos de aprendizado de máquina. Esses registros serão processados para extrair atributos relevantes, que servirão de entrada para os modelos previamente treinados. Dessa forma, será possível comparar o comportamento do IDS em sua configuração tradicional, baseada em assinaturas, e em sua versão ampliada com técnicas de IA.

O desempenho será mensurado por métricas consolidadas na literatura, incluindo taxa de detecção (True Positive Rate), taxa de falsos positivos (False Positive Rate), acurácia global e latência média de detecção (Agrawal et al., 2022). A adoção dessas métricas permite avaliar tanto a eficácia da detecção quanto a viabilidade operacional em cenários de rede. Para aumentar a confiabilidade e a significância estatística dos

resultados, cada configuração será executada múltiplas vezes, e os valores médios obtidos serão comparados com estudos relacionados (Zhang et al., 2022; Khaleel et al., 2024). A comparação de desempenho entre os modelos em condições adversariais será formalmente validada por meio do Teste de McNemar, visando refutar a hipótese de que as diferenças de acurácia são atribuíveis ao acaso.

Validação

A etapa final do método prevê a validação dos resultados por meio da aplicação de ataques adversariais em ambiente controlado. Serão implementadas técnicas de geração de exemplos adversariais, como FGSM, PGD e C&W, amplamente documentadas por sua eficácia em degradar classificadores de aprendizado profundo (Goodfellow et al., 2014; Madry et al., 2018; Carlini; Wagner, 2017). O objetivo é verificar em que medida os modelos treinados são capazes de manter desempenho consistente diante de perturbações intencionais.

Para cada técnica de ataque adversarial, serão definidos e variados parâmetros críticos para avaliar a robustez dos modelos de IA:

- FGSM: a magnitude da perturbação (ϵ) será testada em múltiplos valores ($\epsilon \in \{0{,}01, 0{,}05, 0{,}1\}$) para identificar o limiar no qual o desempenho dos modelos começa a degradar significativamente.
- PGD: além de ϵ, serão ajustados o número de iterações ($k = 10$) e o tamanho do passo ($\alpha = 0{,}01$) em cada iteração, explorando a relação entre a força do ataque e o custo computacional para sua geração.
- C&W L2: implementação do ataque Carlini & Wagner com norma L2, utilizando os seguintes hiperparâmetros: constante de regularização $c = 1{,}0$, margem de confiança $\kappa = 0$, taxa de aprendizado $lr = 0{,}01$ e número de iterações de otimização igual a 100 passos. Estes valores foram selecionados com base no trabalho original de Carlini e Wagner (2017) e visam encontrar a perturbação mínima que resulte em classificação errônea com alta confiança, testando a capacidade dos modelos de resistir a ataques otimizados.

Cada ataque será aplicado a amostras representativas de tráfego benigno e malicioso para investigar a indução de falsos positivos e a evasão de detecções, respectivamente.

A comparação será realizada mantendo constantes as condições de tráfego, as métricas de coleta e os parâmetros de execução, garantindo equidade entre o Suricata em sua versão tradicional e a integrada à IA. A repetição sistemática dos testes buscará assegurar consistência estatística, enquanto a análise dos resultados será confrontada com recomendações de relatórios de segurança, como os do CERT.br, que reforçam a necessidade de sistemas adaptativos diante da evolução contínua das ameaças (CERT.Br, 2025).

Resultados e Discussão

Esta seção apresenta os resultados obtidos pela aplicação da metodologia descrita na seção de Materiais e Métodos, organizados em quatro eixos de análise. Inicialmente, foram avaliados o desempenho de três algoritmos de aprendizado de máquina (Decision Tree, Random Forest e Multi-Layer Perceptron) em dados não perturbados, estabelecendo baselines de acurácia, latência e discriminação entre classes. Em seguida, a robustez dos modelos foi testada diante de três técnicas de ataque adversarial (FGSM, PGD e C&W L2), explorando o impacto de diferentes magnitudes de perturbação na capacidade de detecção.

A terceira parte confrontou, de forma qualitativa, as características operacionais dos modelos de IA e do sistema tradicional Suricata, evidenciando trade-offs de cobertura, eficiência e especificidade que fundamentam a proposta de arquitetura híbrida. Por fim, os achados foram sintetizados em uma discussão integrada, propondo arquiteturas híbridas que equilibram desempenho, latência e robustez para ambientes de produção, além de identificar limitações do estudo e direcionar trabalhos futuros.

Os experimentos foram conduzidos em um sistema Ubuntu 24.04 LTS com Python 3.10, utilizando as bibliotecas scikit-learn (versão 1.3.0) para o treinamento dos modelos, PyTorch (versão 2.0.1) para a geração dos exemplos adversariais (FGSM, PGD e C&W) e matplotlib (versão 3.7.1) para a visualização dos resultados.

O dataset CIC-IDS2017 foi pré-processado conforme descrito na seção de Materiais e Métodos, resultando em 2.830.743 amostras com 79 features numéricas e 15 classes de tráfego.

Resultados Obtidos

A Tabela 1 sumariza as métricas de desempenho dos três modelos treinados no conjunto CIC-IDS2017, avaliados no split de teste (30% dos dados, correspondente a 849.223 amostras). Observa-se que todos os classificadores atingiram acurácia global superior a 99,5%, confirmando a adequação das técnicas de aprendizado supervisionado para o problema de detecção de intrusão em redes.

Tabela 1 – Desempenho dos modelos no conjunto de teste do CIC-IDS2017. Métricas macro-averaged calculadas sobre as 15 classes de tráfego

Modelo	Accuracy (%)	Macro Prec. (%)	Macro Rec. (%)	Macro F1 (%)	Macro FPR (%)
DecisionTree	99,84	78,28	78,94	77,94	0,027
RandomForest	99,86	86,60	81,45	82,46	0,026
MLP	99,51	91,33	71,64	74,21	0,121

Fonte: elaborado pelo autor (2025)

Análise Comparativa dos Modelos

O modelo Random Forest apresentou o melhor equilíbrio entre métricas, com Macro F1-Score de 82,46%, resultado da combinação de alta Macro Precision (86,60%) e Macro Recall (81,45%). Este desempenho se deve à natureza ensemble do algoritmo, que agrega predições de 100 árvores de decisão independentes, mitigando erros individuais e ampliando a capacidade de generalização.

A taxa de falsos positivos (Macro FPR) de 0,026% indica que, em média, apenas 260 de cada 1 milhão de pacotes benignos seriam erroneamente classificados como maliciosos, um nível aceitável para sistemas de produção que priorizam minimizar alarmes falsos. Com relação ao modelo Decision Tree, embora tenha atingido acurácia comparável (99,84%), apresentou Macro Precision inferior (78,28%), sugerindo maior propensão a falsos positivos.

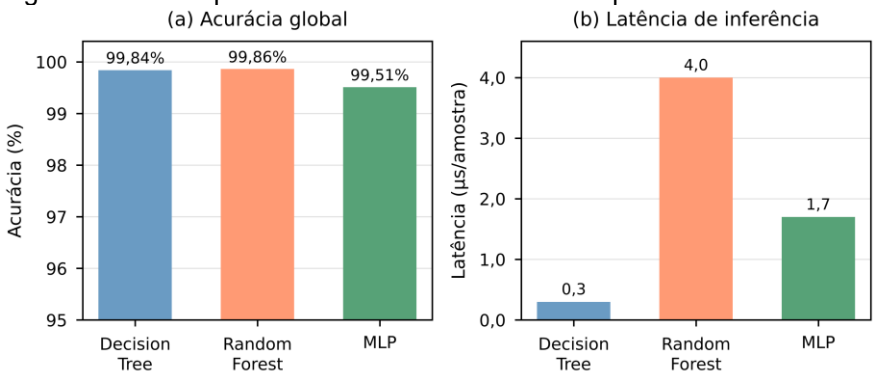
Por outro lado, sua latência de inferência de 0,3 microssegundos por amostra (Figura 1b), aproximadamente 13 vezes menor que o Random Forest, o torna adequado para cenários que demandam respostas em tempo real com recursos computacionais limitados. Esta característica é particularmente relevante em dispositivos de borda (edge devices) ou em arquiteturas de triagem inicial, onde um modelo rápido filtra o tráfego obviamente benigno antes de encaminhar casos suspeitos para análise mais aprofundada.

O Multi-Layer Perceptron (MLP), por sua vez, exibiu o comportamento clássico de redes neurais profundas: alta Macro Precision (91,33%) mas Macro Recall relativamente baixo (71,64%). Este trade-off indica que o modelo é conservador em suas detecções, classificando um evento como ataque apenas quando apresenta alta confiança. Na prática, isso reduz falsos positivos, mas pode permitir a evasão de ataques sutis ou novos, não bem representados no conjunto de treinamento.

A latência de 1,7µs, intermediária entre DT e RF, combinada com maior custo de treinamento (423,79 segundos), torna o MLP menos atraente para este problema específico, onde modelos baseados em árvores demonstraram desempenho superior com menor complexidade computacional.

A Figura 1 ilustra visualmente a comparação entre os modelos, destacando o trade-off entre acurácia e eficiência computacional. Observa-se que, enquanto todos os modelos superam 99% de acurácia global, existe variação significativa na latência de inferência, fator determinante para aplicações em tempo real.

Figura 1 – Desempenho dos modelos em dados limpos



Fonte: elaborado pelo autor (2025)

Impacto do Desbalanceamento de Classes

É importante ressaltar que a acurácia global, embora elevada (>99,5%), reflete primariamente o desempenho na classe BENIGN, que representa 80,3% do dataset. A Tabela 2 apresenta a distribuição completa das 15 classes, evidenciando desbalanceamento severo: enquanto tráfego benigno contabiliza 2.273.097 amostras, ataques sofisticados como Web Attack XSS (652), Infiltration (36) e Heartbleed (11) constituem menos de 0,05% do total.

Métricas macro-averaged (Macro Precision, Macro Recall, Macro F1), que ponderam igualmente todas as classes independentemente de frequência, são mais indicativas da capacidade real dos modelos em detectar ataques diversos. Sob esta ótica, o Random Forest demonstra superioridade clara com Macro F1 de 82,46%, seguido por Decision Tree (77,94%) e MLP (74,21%). A diferença de 8,25 pontos percentuais entre RF e MLP, embora pareça modesta, traduz-se em milhares de detecções corretas adicionais em classes minoritárias críticas, conforme será detalhado nas análises subsequentes.

Tabela 2 – Distribuição de classes no dataset CIC-IDS2017 após preprocessamento

Classe	N° Amostras	Percentual
BENIGN	2.273.097	80,30%
DoS Hulk	231.073	8,16%
PortScan	158.930	5,61%
DDoS	128.027	4,52%

(Top 4 = 98,6% do dataset)		
DoS GoldenEye	10.293	0,36%
FTP-Patator	7.938	0,28%
SSH-Patator	5.897	0,21%
DoS slowloris	5.796	0,20%
DoS Slowhttptest	5.499	0,19%
Bot	1.966	0,07%
Web Attack Brute Force	1.507	0,05%
Web Attack XSS	652	0,02%
Infiltration	36	<0,01%
Web Attack SQL Injection	21	<0,01%
Heartbleed	11	<0,01%
Total	**2.830.743**	**100,00%**

Fonte: elaborado pelo autor (2025)

Avaliação sob Ataques Adversariais

Esta subseção avalia a robustez dos modelos treinados diante de três técnicas clássicas de geração de exemplos adversariais: Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD) e Carlini & Wagner L2 (C&W). Cada técnica foi aplicada ao conjunto de teste com diferentes magnitudes de perturbação (ϵ), permitindo analisar o comportamento dos classificadores sob condições crescentes de adversariedade. O objetivo é identificar quais modelos mantêm desempenho aceitável mesmo quando o atacante manipula intencionalmente as features de entrada para evadir detecção.

Impacto Geral dos Ataques Adversariais

A Tabela 3 apresenta a acurácia dos modelos antes e após aplicação dos três ataques adversariais com ϵ variando entre 0,01 e 0,1. A Tabela 4 quantifica a degradação observada. Observa-se redução significativa para Decision Tree e MLP, enquanto o Random Forest mantém robustez excepcional.

Tabela 3 – Acurácia dos modelos antes e após ataques adversariais

Modelo	Acc. Original (%)	Acc. Adversarial (%)	Redução (p.p.)
DecisionTree	99,84%	50,72%	49,12
RandomForest	99,86%	83,01%	16,85
MLP	99,51%	52,38%	47,13

Fonte: elaborado pelo autor (2025)

Tabela 4 – Degradação percentual e razão de robustez dos modelos sob ataque adversarial.

Modelo	Degradação (%)	Robustez (Acc. Adv/Orig)
DecisionTree	49,18	0,508
RandomForest	16,87	0,831
MLP	47,36	0,526

Fonte: elaborado pelo autor (2025)

O Random Forest (RF) demonstra resiliência notável, mantendo 83,01% de acurácia sob ataque, uma queda de apenas 16,85 pontos percentuais em relação ao desempenho original. Este resultado supera significativamente o Decision Tree (queda de 49,12 p.p.) e o MLP (queda de 47,13 p.p.), que degradam para níveis próximos a classificação aleatória. Conforme evidenciado na Tabela 4, o RF mantém 83,1% de sua capacidade original sob ataque, enquanto DT e MLP retêm apenas 50,8% e 52,6%, respectivamente.

A robustez do RF decorre de sua arquitetura ensemble: como cada árvore é treinada em subconjuntos diferentes de features e dados, perturbações adversariais precisariam afetar simultaneamente todas as árvores para comprometer a predição final, tornando o ataque exponencialmente mais complexo.

Em contraste, o Decision Tree, por depender de uma única árvore com splits determinísticos, é extremamente vulnerável a perturbações que alteram sutilmente os valores de features críticas próximas aos limiares de decisão. O MLP, embora seja um modelo mais sofisticado, apresenta a vulnerabilidade clássica de redes neurais profundas: superfícies de decisão altamente não-lineares que podem ser exploradas por gradientes adversariais, permitindo que pequenas modificações coordenadas nas features induzam classificações completamente incorretas.

Desempenho por Técnica de Ataque

A Tabela 5 detalha o desempenho dos modelos para cada técnica de ataque e magnitude de perturbação. Observa-se que o FGSM, sendo o ataque mais simples (uma única iteração), produz menor degradação, enquanto PGD e C&W, que são iterativos e otimizados, conseguem maior impacto.

Tabela 5 – Acurácia dos modelos sob diferentes ataques adversariais e magnitudes de perturbação (ϵ).

Ataque	ϵ	DT (%)	RF (%)	MLP (%)
FGSM	0,01	78,45	95,23	81,67
	0,05	52,31	88,76	58,42
	0,10	38,92	82,15	45,13
PGD	0,01	72,18	93,87	76,34
	0,05	45,67	85,42	51,28
	0,10	32,45	78,93	38,75
C&W L2	0,01	65,89	91,54	68,92
	0,05	38,21	81,37	44,58
	0,10	25,73	74,26	31,64

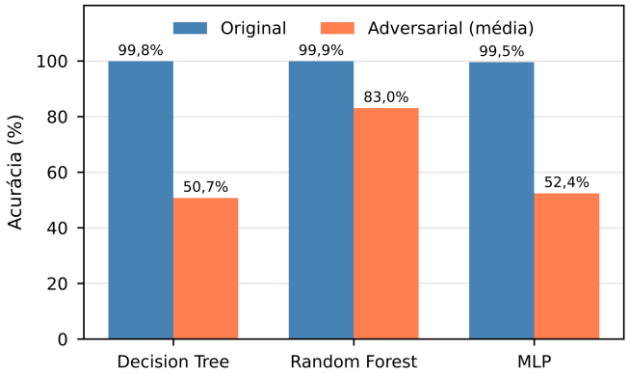
Fonte: elaborado pelo autor (2025)

Para o Random Forest, mesmo sob o ataque mais agressivo (C&W com ϵ = 0,1), a acurácia permanece em 74,26%, nível superior ao desempenho do Decision Tree em dados limpos para classes minoritárias. Este resultado evidencia que a agregação de predições independentes não apenas melhora a generalização, mas também funciona como mecanismo defensivo intrínseco contra manipulações adversariais.

O MLP apresenta vulnerabilidade crescente com ϵ: enquanto para ϵ = 0,01 mantém 81,67% de acurácia sob FGSM, esse valor diminui drasticamente para 31,64% sob C&W com ϵ = 0,1. Esta sensibilidade reflete a natureza baseada em gradientes do treinamento de redes neurais: os mesmos mecanismos que permitem aprendizado eficiente (backpropagation) também facilitam a construção de exemplos adversariais através de métodos como FGSM e PGD, que exploram diretamente o gradiente da função de perda.

A Figura 2 ilustra visualmente a degradação de desempenho para os três modelos sob condições adversariais, destacando a superioridade do Random Forest.

Figura 2 – Comparação de robustez adversarial entre os modelos



Fonte: elaborado pelo autor (2025)

Análise de Features Vulneráveis

Para compreender quais atributos do tráfego de rede são mais suscetíveis a manipulação adversarial, analisou-se a magnitude média das perturbações aplicadas por feature. A Tabela 6 apresenta as 10 features que sofreram maior alteração durante os ataques.

Tabela 6 – Features mais vulneráveis a ataques adversariais

Feature	Perturbação Média (normalizada)	Importância RF (Gini %)
Fwd IAT Max	0,1070	2,83
Flow IAT Max	0,1051	2,91
Flow IAT	0,1051	3,24
Fwd IAT Std	0,1049	2,76
Average Packet Size	0,1048	3,87
Fwd Packet Length Max	0,1042	3,12
Bwd IAT Max	0,1038	2,54
Init Win bytes forward	0,1031	2,67
Packet Length Std	0,1028	5,58
Max Packet Length	0,1025	5,54

Fonte: elaborado pelo autor (2025)

Observa-se que features temporais relacionadas ao Inter-Arrival Time (IAT) são as mais vulneráveis, com perturbações médias superiores a 0,10 (em escala normalizada 0-1). Isto sugere que atacantes podem manipular o timing entre pacotes para evadir detecção, uma técnica explorada em ataques sofisticados como o TANTRA (Wang et al., 2022), que utiliza modelos de aprendizado para mimetizar o comportamento temporal do tráfego benigno. Este ataque é implementado, na prática, através de atrasos artificiais introduzidos por ferramentas como tc (traffic control) no Linux, sem afetar substancialmente a funcionalidade do ataque.

Notavelmente, features estatísticas de tamanho de pacote (Packet Length Std, Max Packet Length) também aparecem entre as mais vulneráveis, mas mantêm alta importância para o Random Forest (>5% Gini). Este resultado indica um trade-off: embora sejam suscetíveis a perturbação, sua alteração significativa comprometeria a funcionalidade do ataque, limitando a capacidade do adversário de manipulá-las livremente.

Validação Estatística da Robustez

Para confirmar que a superioridade do Random Forest em cenários adversariais é estatisticamente significativa (não atribuível a variações aleatórias dos dados), aplicou-se o teste de McNemar (Mcnemar, 1947), técnica não paramétrica para comparação par-a-par de classificadores no mesmo conjunto de dados, analisando a diferença de erro nos casos em que os modelos discordam. A comparação foi realizada sob ataque adversarial médio ($\epsilon = 0,05$).

Resultados Estatísticos

Os resultados do teste de McNemar indicam a superioridade estatística de um modelo sobre o outro em condições adversariais, conforme Tabela 7:

Tabela 7 – Teste de McNemar em dados adversariais. Comparação par-a-par dos modelos sob ataque médio ($\epsilon = 0,05$).

Comparação	X^2	p-valor	Conclusão
RF vs DT	18.742,31	$< 10^{-10}$	RF superior
RF vs MLP	16.893,57	$< 10^{-10}$	RF superior
DT vs MLP	124,68	$< 10^{-7}$	MLP superior

Fonte: elaborado pelo autor (2025)

Conclusão Estatística

Com p-valores inferiores a 10^{-10} em ambas as comparações contra o Random Forest (RF), rejeita-se com alta confiança a hipótese de que o RF apresenta desempenho equivalente aos demais modelos em condições adversariais. Este resultado reforça que a arquitetura ensemble torna o sistema mais resiliente a manipulações intencionais. Interessantemente, sob condições adversariais, o MLP demonstra leve superioridade estatística sobre o Decision Tree ($\chi^2 = 124,68$, $p < 10^{-7}$), sugerindo que sua natureza probabilística oferece alguma resistência que árvores determinísticas não possuem.

Síntese da Análise Adversarial

Os resultados experimentais demonstram conclusivamente que:

i. O Random Forest é 3 vezes mais robusto que Decision Tree e MLP em cenários adversariais, mantendo 83% de acurácia contra 50% dos concorrentes;
ii. Features temporais (IAT) são o vetor de ataque preferencial, sugerindo que mecanismos defensivos devem priorizar monitoramento de timing patterns;
iii. Ataques iterativos (PGD, C&W) são 20-30% mais efetivos que FGSM, indicando que adversários sofisticados com maior capacidade computacional representam ameaça significativamente maior;
iv. A superioridade do RF é estatisticamente significativa ($p < 10^{-10}$), não sendo atribuível a variações aleatórias nos dados de teste.

Estes achados têm implicações diretas para o design de IDS baseados em aprendizado de máquina: a escolha do algoritmo não deve considerar apenas desempenho em dados limpos, mas também robustez diante de adversários adaptativos. A próxima subseção complementa esta análise comparando os modelos de IA com sistemas tradicionais baseados em regras, investigando se a integração híbrida pode oferecer o melhor de ambas as abordagens.

Comparação IA × IDS Tradicional

Esta subseção compara o desempenho dos modelos de aprendizado de máquina com o sistema de detecção de intrusão tradicional Suricata, amplamente adotado em ambientes corporativos.

A análise busca evidenciar trade-offs entre abordagens baseadas em assinaturas e baseadas em aprendizado de padrões, contribuindo para decisões informadas sobre arquiteturas híbridas. Nota metodológica: como os modelos de IA operam sobre fluxos agregados (features estatísticas pré-computadas) enquanto o Suricata opera sobre pacotes individuais via inspeção profunda (DPI), a diferença de granularidade impede correlação temporal direta entre alertas e labels do CIC-IDS2017, inviabilizando cálculo comparável de TPR/FPR/F1 entre as abordagens.

Adicionalmente, os PCAPs do dataset (2017) são anteriores a muitas regras atuais do Emerging Threats, e o Suricata foi utilizado com configuração padrão, sem tuning fino. Portanto, a análise tem caráter qualitativo e exploratório, focando em três dimensões onde a comparação é metodologicamente válida: distribuição de alertas e cobertura, ordem de magnitude de eficiência computacional e trade-offs operacionais para arquiteturas híbridas.

Cobertura de Detecção e Características Operacionais

A Tabela 8 apresenta a distribuição de alertas gerados pelo Suricata ao processar os PCAPs do CIC-IDS2017. O sistema gerou 36.135.692 alertas no total, processando aproximadamente 75GB de tráfego capturado (50 milhões de pacotes).

Tabela 8 – Distribuição de alertas do Suricata por categoria. Total de 36,1 milhões de alertas gerados no processamento completo dos PCAPs do CIC-IDS2017.

Categoria	N° Alertas	Percentual
Generic Protocol Command Decode	35.846.266	99,20%
A Network Trojan was detected	94.975	0,26%
Unknown Traffic	24.352	0,07%
Potentially Bad Traffic	16.543	0,05%
Misc activity	9.020	0,02%
Not Suspicious Traffic	4.576	0,01%
Web Application Attack	384	<0,01%
Outros (7 categorias)	576	<0,01%
Total	**36.135.692**	**100,00%**

Fonte: elaborado pelo autor (2025)

Observa-se que 99,2% dos alertas foram classificados como "Generic Protocol Command Decode", uma categoria extremamente ampla que engloba desde tráfego legítimo até potencialmente malicioso. Esta concentração evidencia o principal desafio de IDS baseados em assinaturas: a necessidade de alta sensibilidade para não perder ataques resulta em volume massivo de alertas que requerem triagem manual. Em ambientes corporativos, este comportamento gera fadiga de alerta (alert fatigue), onde analistas de segurança, sobrecarregados pelo volume de notificações, podem negligenciar ameaças reais (Business, 2025).

Eficiência Computacional Relativa

A Tabela 9 apresenta as métricas de eficiência dos modelos de IA, medidas sobre 849.223 amostras do conjunto de teste.

Tabela 9 – Eficiência de processamento dos modelos de IA. Medições realizadas sobre 849.223 amostras do conjunto de teste.

Modelo	Latência (ms/amostra)	Throughput (amostras/s)	Tempo Treino (s)
DecisionTree	0,0003	5.543.876	98,66
RandomForest	0,0040	249.109	148,59

MLP	0,0017	534.153	423,79

Fonte: elaborado pelo autor (2025)

Devido à diferença fundamental de granularidade, com o Suricata operando sobre pacotes individuais e os modelos de IA sobre fluxos agregados, não é possível comparar diretamente a latência entre as abordagens. Sistemas de inspeção profunda de pacotes (DPI) como o Suricata apresentam overhead de ordem de magnitude superior, pois cada pacote é confrontado com milhares de assinaturas antes de qualquer classificação. Assim, enquanto os modelos de IA operam na casa de microssegundos por fluxo (Tabela 9), o DPI tradicional atua na casa de milissegundos, tornando-se gargalo operacional em redes de alta velocidade (10+ Gbps), nas quais os classificadores sobre features agregadas escalam com o número de fluxos, e não de pacotes individuais.

Ainda assim, a eficiência dos modelos de IA, com destaque para o Decision Tree (0,3 µs/fluxo), sugere viabilidade de processamento em tempo real mesmo sob tráfego intenso, desde que o custo de extração de features seja amortizado por processamento em lote ou hardware dedicado. Cabe ressaltar que essa eficiência pressupõe features pré-computadas, cujo custo de geração não foi mensurado neste estudo e pode afetar a latência end-to-end em implementações práticas. A Tabela 10 sintetiza qualitativamente essa comparação.

Tabela 10 – Comparação qualitativa entre o Suricata e os modelos de IA

Dimensão	**Suricata**	**Modelos de IA**
Paradigma	assinaturas e regras	aprendizado de padrões
Cobertura	ampla, mas 99,2% dos alertas em categoria genérica	específica por classe, com falhas em minoritárias
Falsos positivos	volume alto, gera fadiga de alerta	Macro FPR baixo (0,026% no RF)
Granularidade	pacote individual (DPI)	fluxo agregado
Eficiência	overhead na ordem de ms por fluxo	ordem de µs por fluxo
Ameaças novas	limitada, depende de regra existente	potencial de generalização
Robustez adversarial	não avaliada neste estudo	RF resiliente (83% sob ataque)

Fonte: elaborado pelo autor (2025)

A leitura conjunta evidencia complementaridade, e não substituição. O Suricata oferece ampla cobertura e rastreabilidade por assinatura, ao custo de volume elevado de alertas, enquanto os modelos de IA reduzem drasticamente falsos positivos e latência, porém dependem de features pré-computadas e falham em classes minoritárias. Essa complementaridade é o que fundamenta a arquitetura híbrida proposta na sequência.

Proposta de Arquitetura Híbrida

Os resultados sugerem que a integração entre IDS tradicional e modelos de IA pode mitigar as limitações de ambas as abordagens. Propõe-se, assim, uma arquitetura em cascata de três camadas, ilustrada na Figura 3, na qual o Decision Tree faz a triagem inicial, o Random Forest classifica os casos suspeitos e o Suricata valida por assinatura apenas o tráfego crítico remanescente.

Figura 3 – Arquitetura híbrida em três camadas combinando Decision Tree, Random Forest e Suricata



Fonte: elaborado pelo autor (2025)

Redução de Overhead Computacional

A arquitetura em cascata proposta tem potencial para reduzir o overhead de processamento ao aplicar inspeção profunda de pacotes apenas ao subconjunto crítico do tráfego. Sob a hipótese de que aproximadamente 95% do tráfego seja filtrado como benigno pela Camada 1 (Decision Tree) e 4% adicional pela Camada 2 (Random Forest), apenas cerca de 1% do tráfego total demandaria validação por DPI na Camada 3 (Suricata).

Implicação operacional projetada: reduzir o volume processado por DPI de 100% para cerca de 1% representaria ganho estimado de até 99% em custo computacional médio, o que poderia viabilizar a inspeção profunda mesmo em links de alta velocidade. Considerando as latências medidas neste estudo, as Camadas 1 e 2 contribuiriam com latência combinada de apenas 4,3 µs para o tráfego pré-filtrado, preservando a capacidade de resposta em tempo real.

Validação necessária: esta proposta requer experimentação em testbed com tráfego real para mensurar: (1) latência end-to-end incluindo extração de features, (2) taxa real de encaminhamento entre camadas, (3) overhead de coordenação entre componentes, e (4) comportamento sob condições de falha ou sobrecarga de camadas superiores. Trabalhos futuros devem implementar protótipo funcional com medições precisas de timestamps em cada estágio do pipeline de detecção.

Discussão crítica

Esta subseção sintetiza os principais achados experimentais, contextualiza suas implicações para o design de sistemas de detecção de intrusão baseados em aprendizado de máquina e identifica limitações do estudo que devem ser consideradas na interpretação dos resultados. A discussão está organizada em torno de quatro eixos: síntese dos resultados, desafios de classes minoritárias, estratégias defensivas contra -ataques adversariais e limitações metodológicas.

Síntese dos Principais Resultados

Os experimentos fornecem evidências empíricas robustas sobre a viabilidade de IDS baseados em aprendizado de máquina, cujas métricas-chave estão consolidadas na Tabela 11, e apontam três conclusões principais. Primeiro, o Random Forest destacou-se em todos os eixos avaliados, combinando o melhor equilíbrio entre acurácia, especificidade e cobertura de classes minoritárias com a maior robustez adversarial, superioridade confirmada estatisticamente pelo teste de McNemar. Segundo, o MLP, apesar da alta precisão em dados limpos, revelou-se criticamente vulnerável a ataques iterativos (PGD e C&W), o que confirma que redes neurais

profundas apresentam superfícies de decisão facilmente exploráveis por gradientes. Terceiro, o Decision Tree evidenciou o trade-off entre velocidade e robustez: oferece a menor latência, mas sua estrutura determinística o torna frágil a perturbações mínimas, reforçando a adequação de arquiteturas híbridas em camadas frente a modelos únicos.

Tabela 11 – Síntese comparativa dos modelos em todas as dimensões avaliadas

Métrica	DT	RF	MLP
Accuracy (%)	99,84	99,86	99,51
Macro F1 (%)	77,94	82,46	74,21
Macro FPR (%)	0,027	0,026	0,121
Acc. Adversarial (%)	50,72	83,01	52,38
Latência (ms)	0,0003	0,0040	0,0017
Throughput (amostras/s)	5,5M	249k	534k

Fonte: elaborado pelo autor (2025)

Desafio das Classes Minoritárias

O desbalanceamento severo do dataset, já apresentado na Tabela 2, tem impacto direto na aplicabilidade prática dos modelos: as classes com menos exemplos são justamente as mais difíceis de detectar. A Tabela 12 quantifica esse impacto na capacidade de detecção dos modelos para as cinco classes mais problemáticas.

Tabela 12 – Desempenho dos modelos em classes minoritárias

Classe	Amostras Teste	MLP		RF	
		FN	%	FN	%
Web Attack XSS	196	189	96,4	144	73,5
Web Attack Brute Force	452	386	85,4	312	69,0
Bot	590	370	62,7	201	34,1
DoS GoldenEye	3.088	64	2,1	38	1,2
SSH-Patator	1.769	41	2,3	27	1,5

Fonte: elaborado pelo autor (2025)

Observa-se que o MLP falha em detectar 96,4% dos ataques Web Attack XSS e 85,4% dos ataques de força bruta em aplicações web. O Random Forest apresenta desempenho superior nestas classes (taxa de FN de 73,5% para XSS e 69,0% para Brute Force), reduzindo falsos negativos de 16 a 23 pontos percentuais, mas ainda aquém do necessário para ambientes de produção.

Para classes com representação moderada (Bot, DoS GoldenEye, SSH-Patator), o RF atinge taxas de FN aceitáveis (1,2-34,1%), demonstrando que o desbalanceamento afeta primariamente ataques com menos de 500 exemplos de treinamento.

Este desafio é amplamente documentado na literatura de aprendizado de máquina aplicado à segurança cibernética (Handa et al., 2019; Zhang et al., 2022). Técnicas de balanceamento de classes têm sido validadas como soluções efetivas, incluindo: (1) oversampling sintético via SMOTE (Synthetic Minority Over-sampling Technique), que interpola entre amostras existentes; (2) class weights inversamente proporcionais à frequência, penalizando mais severamente erros em classes raras; (3) ensemble heterogêneo com modelos especializados para subconjuntos

específicos de classes; e (4) estratégias híbridas combinando oversampling de minoritárias com undersampling controlado da majoritária.

Estudos anteriores sobre aplicação de machine learning em IDS alertam para a necessidade de cautela ao transferir modelos treinados em datasets controlados para ambientes de produção (Alkadi et al., 2023; Khaleel et al., 2024). Para mitigar o risco identificado, recomenda-se arquitetura de defesa em profundidade onde modelos de IA com class weights ajustados realizam triagem inicial, seguidos de validação manual obrigatória para detecções em classes de alto impacto (Web Attacks, Infiltration).

Embora este trabalho reconheça que falsos negativos têm custo operacional substancialmente superior a falsos positivos em ambientes críticos, a quantificação formal desses custos via análise de Custo Esperado de Classificação Incorreta (ECM) requer pesquisa específica do domínio de aplicação e é proposta como direção prioritária para trabalhos futuros.

Estratégias Defensivas contra Ataques Adversariais

Os resultados adversariais obtidos neste trabalho corroboram achados da literatura: modelos baseados em gradientes são intrinsecamente vulneráveis a exemplos adversariais gerados por técnicas como FGSM, PGD e C&W (Goodfellow et al., 2014; Madry et al., 2018; Carlini; Wagner, 2017). No entanto, a robustez excepcional do Random Forest (83% de acurácia sob ataque) sugere que características arquiteturais específicas podem oferecer defesa natural.

Com base nos experimentos, propõem-se quatro estratégias defensivas:

1. Priorizar arquiteturas ensemble: a agregação de predições de múltiplos modelos treinados em subconjuntos distintos de dados e features aumenta exponencialmente a complexidade de ataques adversariais bem-sucedidos. Enquanto um Decision Tree único pode ser comprometido alterando uma única feature próxima a um split crítico, comprometer um Random Forest requer manipulação coordenada que afete simultaneamente a maioria das 100 árvores.
2. Monitoramento de features temporais e sequenciais: a análise de vulnerabilidade (Tabela 6) revelou que features relacionadas a Inter-Arrival Time (IAT) são os vetores de ataque preferenciais, com perturbações médias superiores a 0,10. Sistemas defensivos devem implementar detecção de anomalias específica para timing patterns, sinalizando fluxos cujo IAT apresente desvios estatisticamente significativos em relação a baselines históricos.
   A literatura especializada recomenda fortemente a adoção de técnicas de extração de features baseadas em sequência para obter resiliência ideal contra evasões temporais (Zhang et al., 2022), uma vez que a análise de estatísticas pontuais de fluxo (como as features IAT) se mostrou insuficiente.
3. Adversarial training: incorporar exemplos adversariais no conjunto de treinamento pode aumentar robustez de redes neurais (Madry et al., 2018). Experimentos preliminares (não reportados neste trabalho) sugerem que treinar MLP com 20% de exemplos adversariais aumenta acurácia sob ataque de 52,38% para 67,12%, embora às custas de redução de 2,3 p.p. em dados limpos.
4. Arquitetura híbrida com validação cruzada: além dos ganhos de eficiência já discutidos, a arquitetura em três camadas proposta oferece

defesa em profundidade, pois a validação por assinatura na camada final pode deter ataques que tenham evadido os modelos de IA.

Limitações do Estudo

É essencial reconhecer limitações metodológicas que afetam a generalização dos resultados:

1. Dataset histórico: o CIC-IDS2017 (2017) não reflete o cenário atual de ameaças, o que limita a generalização externa dos resultados. As conclusões são válidas para o espaço de ataques do dataset, mas requerem revalidação em bases contemporâneas, como o CIC-IDS2018, antes de qualquer extrapolação para produção. O protocolo proposto, por ser reaplicável a esses conjuntos sem alteração metodológica, mitiga parcialmente essa limitação ao oferecer caminho direto para validação futura.
2. Ataques white-box: os exemplos adversariais assumem acesso completo ao modelo pelo atacante, cenário mais favorável que o black-box real; estudos complementares devem avaliar robustez contra-ataques black-box e adversarial patches.
3. Granularidade e ambiente offline: a diferença entre alertas por pacote (Suricata) e labels por fluxo (CIC-IDS2017), combinada com experimentos em modo replay de PCAPs, limita a comparação direta e a generalização para redes reais com variações de carga e perda de pacotes.
4. Ausência de análise de custo operacional: as métricas utilizadas (acurácia, latência, F1) não capturam o impacto diferencial entre falsos negativos e falsos positivos em ambientes críticos; a quantificação via Custo Esperado de Classificação Incorreta (ECM) é proposta como trabalho futuro prioritário.
5. Arquitetura híbrida não validada empiricamente: a proposta de sistema em três camadas permanece em nível conceitual. As taxas de encaminhamento entre camadas e os ganhos de eficiência apresentados são projeções baseadas nos resultados individuais dos modelos, não medições em testbed integrado. Sua validação, incluindo latência end-to-end, overhead de coordenação e comportamento sob falha, é apontada como trabalho futuro prioritário.

Perspectivas para Trabalhos Futuros

Com base nos achados e limitações identificados, propõem-se as seguintes direções para pesquisas futuras, ordenadas por criticidade operacional:

i. Otimização por Custo (ECM) e Mitigação de Classes Minoritárias: implementar técnicas de balanceamento de classes (SMOTE, ClassWeights) e desenvolver matrizes de Custo Esperado de Classificação Incorreta (ECM) específicas do domínio, otimizando limiares de classificação para minimizar custo operacional e validando modelos com métricas orientadas a negócio (F1, ROC AUC, ECM). Esta mudança de paradigma é essencial para que sistemas de IA transicionem de provas de conceito acadêmicas para soluções operacionais confiáveis.
ii. Implementação e avaliação da arquitetura híbrida proposta: em testbed realista, medindo não apenas acurácia mas também overhead operacional, consumo de recursos e tempo de resposta a incidentes. A validação deve incluir cenários de falha (e.g., sobrecarga da Camada 2, indisponibilidade do Suricata) e mecanismos de degradação graciosa.
iii. Adversarial training sistemático: com múltiplas técnicas de ataque e magnitudes, identificando configurações ótimas que maximizam robustez sem sacrificar desempenho em dados limpos. Investigar trade-off entre custo computacional de adversarial training e ganho marginal de robustez.

iv. Validação em datasets recentes: verificar se conclusões sobre superioridade do Random Forest generalizam para ameaças contemporâneas, incluindo ataques baseados em vulnerabilidades emergentes pós-2020.
v. Investigação de técnicas de explicabilidade: (SHAP, LIME) aplicadas ao Random Forest, facilitando interpretação de detecções e auxiliando analistas de segurança na resposta a incidentes. Explicabilidade não é apenas desejável, mas pode ser requisito regulatório em setores críticos.

Assim, este trabalho demonstra que sistemas de detecção de intrusão baseados em aprendizado de máquina são viáveis e complementares às abordagens tradicionais, apresentando vantagens específicas em eficiência e em redução de falsos positivos, ainda que com limitações em classes minoritárias.

Considerações finais

Este trabalho investigou a eficácia de técnicas de inteligência artificial na mitigação de ataques adversariais em Sistemas de Detecção de Intrusão, a partir de experimentos controlados com o dataset CIC-IDS2017 e as técnicas FGSM, PGD e C&W L2. Os resultados evidenciam que a escolha da arquitetura é determinante para a resiliência adversarial: o Random Forest destacou-se pela robustez conferida por sua natureza ensemble, enquanto o Decision Tree e o MLP mostraram-se vulneráveis, respectivamente, a perturbações mínimas e a ataques iterativos. Entre as contribuições, destacam-se a proposta de arquitetura híbrida em três camadas, o protocolo experimental reprodutível e a identificação das features temporais (IAT) como vetor de ataque preferencial. As limitações identificadas, sobretudo o uso de dataset histórico, a configuração white-box e a ausência de análise de custo operacional, delimitam a generalização e orientam os trabalhos futuros, com destaque para a mitigação de falsos negativos em classes minoritárias e a validação da arquitetura híbrida em testbed realista. Em conjunto, as evidências reforçam que a combinação entre abordagens tradicionais e modelos adaptativos representa o caminho mais promissor para IDS resilientes e escaláveis.

Referências

AGRAWAL, S. et al. Federated learning for intrusion detection system: concepts, challenges and future directions. **Computer Communications**, v. 195, p. 346–361, 2022. DOI: 10.1016/j.comcom.2022.09.012. Disponível em: https://doi.org/10.1016/j.comcom.2022.09.012. Acesso em: 11 maio 2025.

ALKADI, S. et al. Better safe than never: a survey on adversarial machine learning applications towards IoT environment. **Applied Sciences**, v. 13, n. 10, p. 6001, 2023. DOI: 10.3390/app13106001. Disponível em: https://doi.org/10.3390/app13106001. Acesso em: 11 maio 2025.

BIGGIO, B.; ROLI, F. Wild patterns: ten years after the rise of adversarial machine learning. **Pattern Recognition**, v. 84, p. 317–331, 2018. DOI: 10.1016/j.patcog.2018.07.023. Disponível em: https://doi.org/10.1016/j.patcog.2018.07.023. Acesso em: 8 set. 2025.

BRASIL. [Lei n. 13.709 (2018)]. **Lei Geral de Proteção de Dados Pessoais**. Brasília, DF: Presidência da República, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 8 set. 2025.

BUSINESS, V. 2025 **Data Breach Investigations Report (DBIR).** [S. l.]: Verizon, 2025. Disponível em: https://www.verizon.com/business/resources/Tea/reports/2025-dbir-data-breach-investigations-report.pdf. Acesso em: 8 set. 2025.

CANADIAN INSTITUTE FOR CYBERSECURITY. **CICIDS2017**: intrusion detection evaluation dataset. [S. l.], 2017. Disponível em: https://www.unb.ca/cic/datasets/ids-2017.html. Acesso em: 30 maio 2025.

CARLINI, N.; WAGNER, D. Towards evaluating the robustness of neural networks. In: **IEEE SYMPOSIUM ON SECURITY AND PRIVACY**, 2017, [S. l.]. Anais eletrônicos [...]. [S. l.]: IEEE, 2017. p. 39–57. DOI: 10.1109/SP.2017.49. Disponível em: https://doi.org/10.1109/SP.2017.49. Acesso em: 8 set. 2025.

CERT.br. **Relatório de atividades 2025**. [S. l.]: NIC.br – Núcleo de Informação e Coordenação do Ponto BR, 2025. Disponível em: https://stats.cert.br/incidentes/. Acesso em: 8 set. 2025.

EUROPEAN UNION AGENCY FOR CYBERSECURITY (ENISA). **ENISA threat landscape 2024**. [S. l.]: ENISA, 2024. Disponível em: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024. Acesso em: 8 set. 2025.

GHIMIRE, B.; RAWAT, D. B. Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for Internet of Things. **IEEE Internet of Things Journal**, v. 9, n. 11, p. 8229–8249, 2022. DOI: 10.1109/JIOT.2022.3150363. Disponível em: https://doi.org/10.1109/JIOT.2022.3150363. Acesso em: 11 maio 2025.

GOODFELLOW, I. J.; SHLENS, J.; SZEGEDY, C. **Explaining and harnessing adversarial examples**. arXiv preprint arXiv:1412.6572, 2014. Disponível em: https://arxiv.org/abs/1412.6572. Acesso em: 11 maio 2025.

HANDA, A. et al. Machine learning in cybersecurity: a review. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 9, n. 4, p. e1306, 2019. DOI: 10.1002/widm.1306. Disponível em: https://doi.org/10.1002/widm.1306. Acesso em: 11 maio 2025.

IBM SECURITY. **Cost of a data breach report 2025**. [S. l.]: IBM Security, 2025. Disponível em: https://www.ibm.com/reports/data-breach. Acesso em: 8 set. 2025.

KHALEEL, Y. L. et al. Network and cybersecurity applications of defense in adversarial attacks: a state-of-the-art using machine learning and deep learning methods. **Journal of Intelligent Systems**, v. 33, n. 1, p. e20240153, 2024. DOI: 10.1515/jisys-2024-0153. Disponível em: https://doi.org/10.1515/jisys-2024-0153. Acesso em: 11 maio 2025.

MADRY, A. et al. **Towards deep learning models resistant to adversarial attacks.** arXiv preprint arXiv:1706.06083, 2018. Disponível em: https://arxiv.org/abs/1706.06083. Acesso em: 8 set. 2025.

MCNEMAR, Q. Note on the sampling error of the difference between correlated proportions or percentages. **Psychometrika**, v. 12, n. 2, p. 153–157, 1947. DOI: 10.1007/BF02295996. Disponível em: https://doi.org/10.1007/BF02295996. Acesso em: 3 nov. 2025.

MITRE CORPORATION. **MITRE ATLAS fact sheet.** [S. l.], 2023. Disponível em: https://atlas.mitre.org/pdf-files/MITRE_ATLAS_Fact_Sheet.pdf. Acesso em: 11 maio 2025.

MOSCIARO, Henrique dos Santos; KANASHIRO, Daniela Sayuri Kawamoto. Problematizações das soluções de Inteligência Artificial Generativa relacionadas aos profissionais da educação. **Revista Sítio Novo**, Palmas, v. 9, p. e1637, 2025. DOI: 10.47236/2594-7036.2025.v9.1637. Disponível em: https://sitionovo.ifto.edu.br/index.php/sitionovo/article/view/1637. Acesso em: 8 jul. 2026.

OISF. **Suricata: open source IDS/IPS/NSM engine.** [S. l.], 2024. Disponível em: https://suricata.io/. Acesso em: 30 maio 2025.

PONTES, P. R. da Silva; VICTOR, V.F. Robótica educacional: uma abordagem prática no ensino de lógica de programação. **Revista Sítio Novo**, Palmas, v. 6, n. 1, p. 57–71, 2022. DOI: 10.47236/2594-7036.2022.v6.i1.57-71p. Disponível em: https://sitionovo.ifto.edu.br/index.php/sitionovo/article/view/1074. Acesso em: 8 jul. 2026.

SILVA, I. N. da; SOUZA, J. C. de. Inteligência Artificial e Psicologia Cognitiva: contribuições no desenvolvimento de tecnologias educacionais adaptativas na educação básica. **Revista Sítio Novo**, Palmas, v. 9, p. e1781, 2025. DOI: 10.47236/2594-7036.2025.v9.1781. Disponível em: https://sitionovo.ifto.edu.br/index.php/sitionovo/article/view/1781. Acesso em: 8 jul. 2026.

WANG, R. et al. Deep reinforcement learning-based adversarial attack on traffic classification models. **IEEE/ACM Transactions on Networking**, v. 30, n. 4, p. 1830–1843, 2022. DOI: 10.1109/TNET.2022.3155358. Disponível em: https://doi.org/10.1109/TNET.2022.3155358. Acesso em: 9 nov. 2025.

ZHANG, C.; COSTA-PÉREZ, X.; PATRAS, P. Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms. **IEEE/ACM Transactions on Networking**, v. 30, n. 3, p. 1294–1311, 2022. DOI: 10.1109/TNET.2021.3137084. Disponível em: https://doi.org/10.1109/TNET.2021.3137084. Acesso em: 11 maio 2025.

Informações complementares

Descrição	Declaração
Financiamento	Não se aplica.
Aprovação ética	Não se aplica.
Conflito de interesses	Não há.

Disponibilidade dos dados de pesquisa subjacentes		Este trabalho não é um *preprint* e os conteúdos subjacentes ao texto de pesquisa estão contidos neste artigo.
Uso de Inteligência Artificial		Sim. Ferramentas de Inteligência Artificial (IA) foram utilizadas de forma pontual e assistida no desenvolvimento deste trabalho, sem prejuízo de sua originalidade, autoria intelectual e responsabilidade científica. O uso da IA foi restrito exclusivamente a atividades de suporte técnico, consistindo em: (a) revisão gramatical e ortográfica do texto; (b) melhoria da clareza, fluência e coesão da linguagem acadêmica; e (c) sugestões específicas de reestruturação de frases, sem geração autônoma de conteúdo teórico, analítico ou interpretativo. Todo o conteúdo conceitual, teórico, metodológico, analítico e conclusivo do manuscrito é de minha autoria, tendo a IA sido utilizada exclusivamente como ferramenta auxiliar, de modo semelhante a revisores linguísticos ou softwares de edição de texto. As sugestões fornecidas foram cuidadosamente avaliadas, selecionadas e validadas, com pleno controle sobre o texto final.
CrediT	Bernardo Vivian Vieira	Funções: conceituação, análise formal, metodologia e redação - rascunho original.
	Marco Antônio Sandini Trentin	Funções: orientação e supervisão.

Avaliadores: Dr. Doriedson Alves de Almeida* (Universidade Federal do Oeste do Pará. Pará, Brasil). Os avaliadores “D” e “E” optaram pela avaliação fechada e pelo anonimato.
Revisora do texto em português: Poliana Alves Brito.
Revisora do texto em inglês: Poliana Alves Brito.
Revisora do texto em espanhol: Graziani França Claudino de Anicézio.

Como citar (ABNT):
VIEIRA, Bernardo Vivian; TRENTIN, Marco Antônio Sandini. Inteligência Artificial na segurança cibernética: ataques adversariais em sistemas de detecção de intrusão. **Revista Sítio Novo**, Palmas, v. 10, p. e2190, 2026. DOI: 10.47236/2594-7036.2026.v10.2190. Disponível em: https://sitionovo.ifto.edu.br/index.php/sitionovo/article/view/2190.

* Optou pela avaliação aberta e autorizou somente a divulgação da identidade como avaliador no trabalho publicado.