

Explicabilidade de Modelos de Aprendizado de Máquina para Sistemas de Detecção de Intrusão em Redes IoT Usando Fluxo Agregado

Matheus Figueiredo, Rosiane Lima, Frederico Pinagé, Waldir S. S. Júnior e Celso Carvalho

Resumo— Com o crescente uso de dispositivos conectados em redes de Internet das Coisas (IoT), aumentam também as preocupações relacionadas à segurança cibernética. Este trabalho apresenta uma metodologia para detecção de ataques baseada em aprendizado de máquina utilizando dados agregados de fluxo de rede. O conjunto de dados utilizado se chama CICIOT2023. Os resultados demonstraram eficiência na detecção de tráfego malicioso utilizando apenas informações de fluxo. Além disso, técnicas de Inteligência Artificial Explicável (XAI) foram utilizadas para interpretar os modelos e identificar atributos relevantes para a detecção de ataques, mostrando o potencial da abordagem para Sistemas de Detecção de Intrusão (IDS).

Palavras-Chave— Aprendizado de Máquina, Fluxo Agregado, IA Explicável, Internet das Coisas, Sistemas de Detecção de Intrusão.

Abstract— With the increasing use of connected devices in Internet of Things (IoT) networks, concerns regarding cybersecurity have also grown. This work presents a machine learning-based methodology for cyberattack detection using aggregated network flow data. The CICIOT2023 dataset was employed to evaluate the proposed approach. The results demonstrated effective malicious traffic detection using only flow information. Furthermore, Explainable Artificial Intelligence (XAI) techniques were applied to interpret the models and identify relevant features for attack detection, highlighting the potential of the approach for Intrusion Detection Systems (IDS).

Keywords— Machine Learning; Aggregated Flow; Explainable Artificial Intelligence; Internet of Things; Intrusion Detection Systems.

I. INTRODUÇÃO

As necessidades atuais de monitoramento e controle de diferentes dispositivos em organizações exigem uma plataforma de Internet das Coisas (IoT) capaz de integrar elementos, utilizando diferentes protocolos, formatos de dados e tecnologias de comunicação [1]. Nesse sentido, o rápido crescimento de sistemas conectados em redes IoT gera preocupações sobre a segurança e a proteção dos grandes volumes de dados que eles coletam, transmitem, armazenam e processam [2], visto que a grande escala dessas redes abre vulnerabilidades que

podem ser exploradas por potenciais agentes e componentes maliciosos [3].

Nesse contexto, Sistemas de Detecção de Intrusão (IDS) são essenciais para proteger redes IoT contra atividades maliciosas [4]. Embora os IDSs tradicionais permitam a interceptação de todos os ataques conhecidos, Coscia et. al [5] afirmam que eles não são robustos à variação contínua de anomalias ao longo do tempo, o que torna necessário o desenvolvimento de módulos de detecção mais eficientes para a detecção de ataques cibernéticos. Neste cenário, a inteligência artificial (IA), com suas capacidades adaptativas e de autoaprendizagem, emergiu como uma abordagem promissora para aprimorar a detecção de intrusões em ambientes de IoT [4] e modelos de Aprendizado de Máquina (ML) já se mostram eficazes na detecção de ataques cibernéticos [6].

No entanto, a complexidade dos modelos de ML representa um sério desafio à transparência e interpretabilidade dos resultados de detecção produzidos, pois a falta de explicabilidade em IDS orientados por IA mina a confiança do usuário e limita sua implantação prática, especialmente entre as partes interessadas não especializadas [7]. Desta forma, a IA Explicável (XAI) tornou-se cada vez mais importante para interpretar os modelos de ML e aprimorar a gestão da confiança, permitindo que especialistas humanos compreendam as evidências de dados subjacentes e o raciocínio causal [8].

Este trabalho propõe uma abordagem para IDS em redes IoT baseada em XAI. Com o objetivo de desenvolver um sistema interpretável para ambientes IoT, foram utilizados atributos agregados de fluxo de rede juntamente com técnicas de seleção e redução de dimensionalidade de dados. Para validar a proposta, foi empregado o dataset CICIOT2023 [9], baseado em fluxo de rede, composto por tráfego IoT benigno e diversos ataques IoT. Além disso, técnicas de XAI baseadas em SHAP foram utilizadas para interpretar o comportamento do modelo e identificar os atributos mais relevantes para o processo de detecção de ataques. As principais contribuições deste trabalho são as seguintes:

- Proposição de uma abordagem para detecção de ataques em redes IoT baseada exclusivamente em atributos agregados de fluxo de rede, reduzindo a dependência de inspeção profunda de pacotes e diminuindo o custo computacional do IDS.
- Aplicação de técnicas de seleção e redução de dimensionalidade de atributos, permitindo a construção de modelos eficientes utilizando um conjunto reduzido de características relevantes do tráfego de rede.

Matheus Figueiredo, Centro de P&D de Tecnologia Eletrônica e da Informação, Universidade Federal do Amazonas, Manaus, Brasil, e-mail: matheusfigueiredo@ufam.edu.br; Rosiane Lima, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal do Amazonas, Manaus, Brasil, e-mail: rosiane.brito@ufam.edu.br; Frederico Pinagé, Waldir S. S. Júnior e Celso Carvalho, Departamento de Eletrônica e Computação, Universidade Federal do Amazonas, Manaus, Brasil, e-mails: fredericopinage@ufam.edu.br, waldirjr@ufam.edu.br, ccarvalho@ufam.edu.br. Este trabalho foi parcialmente financiado pelo Centro de Pesquisa e Desenvolvimento de Tecnologia Eletrônica e da Informação (CETELI-UFAM).

- Emprego de técnicas de XAI baseadas em SHAP para interpretar o comportamento do modelo, identificando os atributos de fluxo mais relevantes para a detecção de ataques em ambientes IoT.

II. TRABALHOS RELACIONADOS

A. IDSs baseados em XAI

Técnicas de XAI para IDS têm sido amplamente utilizadas para detecção de anomalias em tráfego de rede IoT. Esmailyfard et al. [10] propuseram o GELAX, uma arquitetura baseada em Redes Neurais Gráficas, que utiliza dinâmica de grafos e XAI ancorada para detecção de botnets em redes IoT. Esta abordagem depende de representação em grafos e de uma arquitetura GNN-DNN empilhada, o que aumenta a complexidade computacional. Bin Hulayyil et al. [11] desenvolveram um IDS explicável para IoT voltado à identificação de dispositivos vulneráveis ao *Ripple20*, combinando classificadores de ML com SHAP e inferência causal. Apesar da contribuição em interpretabilidade, o estudo é orientado a um caso específico de vulnerabilidade e utiliza uma estrutura explicativa mais complexa, enquanto este trabalho investiga a detecção binária de tráfego malicioso a partir de atributos agregados de fluxo. Em [12], Farhan et al. apresentaram um framework de segurança baseado em XAI para sistemas ciberfísicos na Indústria 4.0, utilizando SHAP e LIME. Apesar da elevada interpretabilidade obtida, o trabalho concentra-se em arquiteturas industriais complexas e sistemas imersivos, não explorando especificamente a detecção de ataques em redes IoT utilizando atributos agregados de fluxo de rede.

B. Detecção Baseada em Fluxo

Em [13], é proposto um IDS baseado em especificação para redes IoT, utilizando análise de tráfego em nível de fluxo. A principal ideia consiste em monitorar padrões de comunicação dos dispositivos IoT sem depender de inspeção profunda de pacotes, reduzindo custo computacional e aumentando a escalabilidade do sistema, no entanto, a abordagem é baseada em especificações previamente definidas, exigindo conhecimento especializado para construção manual de regras IDS e padrões esperados de comportamento. Meidan et al. [14] propuseram o N-BaIoT, um sistema de detecção de botnets IoT baseado em autoencoders profundos aplicados ao tráfego de rede, capaz de identificar anomalias associadas a botnets como Mirai e BASHLITE a partir de snapshots estatísticos do fluxo de rede. Já o Kitsune [15] utiliza um ensemble de autoencoders leves para detecção online de anomalias em tráfego de rede, priorizando eficiência computacional e execução em dispositivos com recursos limitados. Por outro lado, Hossain [16] propôs uma abordagem baseada em deep learning utilizando CNN, LSTM, RNN e MLP para classificação multiclasse e detecção binária de ataques em redes IoT a partir de dados flow-based. Apesar dos elevados resultados obtidos, esses trabalhos mantêm características de modelos *black-box*, explorando de forma limitada mecanismos de interpretabilidade e explicabilidade das decisões tomadas pelos IDS.

III. DATASET E DESCRIÇÃO DOS DADOS

Para validação da proposta, foi utilizado o dataset CIIoT2023 [9], desenvolvido pelo Canadian Institute for Cybersecurity (CIC) com o objetivo de representar cenários realistas de ataques cibernéticos em ambientes IoT. O dataset foi construído em um laboratório IoT dedicado composto por 105 dispositivos IoT reais, incluindo câmeras inteligentes, sensores, dispositivos de automação residencial, hubs Zigbee/Z-Wave, smart speakers, plugs inteligentes e dispositivos Raspberry Pi utilizados tanto para comunicação legítima quanto para execução dos ataques.

O dataset contempla 33 ataques distribuídos em sete categorias principais: DDoS, DoS, Reconnaissance, Web-Based, Brute Force, Spoofing e Mirai. Os dados disponibilizados no formato CSV são compostos por atributos extraídos de janelas de pacotes de rede, representando características estatísticas e estruturais dos fluxos de comunicação. Os atributos incluem informações temporais, estatísticas de fluxo, indicadores de protocolos e flags TCP, além de métricas agregadas relacionadas ao comportamento do tráfego de rede. A Tabela I apresenta os principais grupos de atributos utilizados no dataset.

TABELA I: Principais grupos de atributos do CIIoT2023

Categoria	Exemplos de atributos
Fluxo	flow_duration, Rate, Srate, Drate
Protocolos	HTTP, HTTPS, DNS, TCP, UDP, ICMP
Flags TCP	syn_flag_number, ack_flag_number, rst_flag_number, psh_flag_number
Estatísticas de pacotes	Min, Max, AVG, Std, Tot_size
Temporais	IAT, Duration, Header_Length
Estatísticas agregadas	Number, Magnitude, Radius, Covariance, Variance, Weight

IV. METODOLOGIA

A metodologia proposta neste trabalho foi desenvolvida com o objetivo de construir um IDS para ambientes IoT utilizando atributos agregados de fluxo de rede. O pipeline adotado é composto pelas etapas de pré-processamento dos dados, balanceamento das classes, treinamento e seleção de modelos de aprendizado de máquina, redução de dimensionalidade por seleção de atributos e análise de explicabilidade utilizando técnicas de XAI. A Figura 1 apresenta uma visão geral das etapas empregadas.

Pré-processamento dos Dados. Inicialmente, os dados do dataset CIIoT2023 foram submetidos a uma etapa de pré-processamento visando padronização e tratamento de inconsistências presentes nos dados extraídos dos fluxos de rede. Nessa etapa, os nomes dos atributos foram normalizados, removendo espaços e caracteres especiais para evitar incompatibilidades durante o treinamento dos modelos. Além disso, valores ausentes (*NaN*), infinitos e inconsistentes foram substituídos por zero. Como o foco deste trabalho consiste na detecção de tráfego malicioso em ambientes IoT, os rótulos originais do dataset foram convertidos para um problema de classificação binária, agrupando todas as categorias de ataques em uma única classe maliciosa, enquanto o tráfego legítimo foi mantido como classe benigna. Após o pré-processamento, os dados foram divididos com base na estratégia utilizada

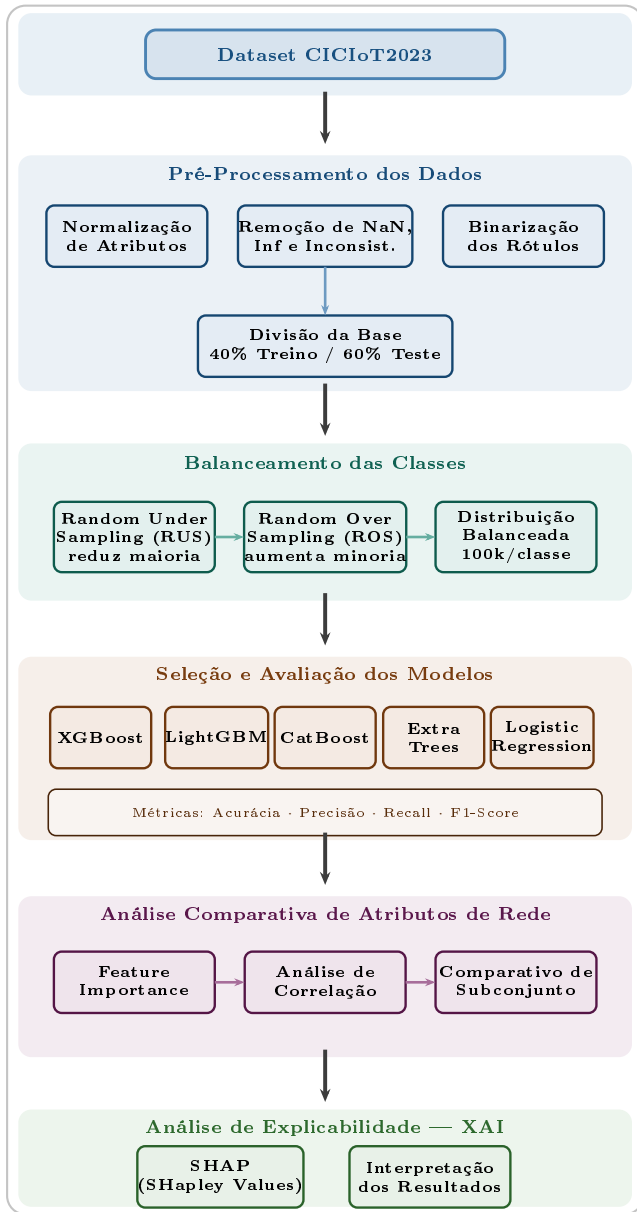


Fig. 1: Etapas da metodologia proposta.

em [5], que consiste em conjuntos de treinamento e teste utilizando proporções de 40% e 60%, respectivamente, de forma estratificada.

Balanceamento das Classes. O dataset CICIOT2023 apresenta forte desbalanceamento entre tráfego benigno e diferentes categorias de ataques. Neste trabalho, foi utilizada uma abordagem híbrida de balanceamento baseada em técnicas de subamostragem e sobreamostragem. Inicialmente, aplicou-se Random Under Sampling (RUS) para reduzir a quantidade de amostras das classes majoritárias, evitando dominância excessiva durante o treinamento. Em seguida, foi utilizado Random Over Sampling (ROS) para aumentar artificialmente o número de amostras das classes minoritárias, permitindo uma distribuição mais equilibrada entre os diferentes tipos de tráfego. A Tabela II mostra o resultado do balanceamento das

classes.

TABELA II: Distribuição das classes antes e após o balanceamento

Classe	Original	Balanceado
Benign	439276	300000
DDoS	793283	300000
Recon	276214	276214
DoS	256439	256439
Spoofing	170642	170642
Mirai	42220	100000
Web	8644	100000
BruteForce	5226	100000
Malware	1287	100000

Seleção e Avaliação dos Modelos. Após o balanceamento dos dados, diferentes algoritmos de aprendizado de máquina foram avaliados para a tarefa de detecção binária de ataques. Os modelos analisados foram XGBoost, LightGBM, CatBoost, ExtraTrees e Logistic Regression. A escolha dos modelos considerou algoritmos amplamente utilizados em aplicações de detecção de intrusão e classificação de tráfego de rede, incluindo métodos baseados em árvores de decisão, *boosting* e modelos lineares. A avaliação dos modelos foi realizada utilizando métricas clássicas de classificação, incluindo Accuracy, Precision, Recall, F1-Score e Balanced Accuracy.

Otimização de Hiperparâmetros. Após a etapa de seleção do modelo, realizou-se uma etapa de otimização de hiperparâmetros utilizando a biblioteca Optuna [17]. O processo de otimização teve como objetivo identificar configurações mais adequadas para o classificador. Além disso, foi realizada uma análise de importância dos atributos utilizando os valores de importância fornecidos pelo modelo selecionado para avaliar a performance do modelo com menos atributos de rede.

Explicabilidade do Modelo. Para interpretação do comportamento do modelo, foram utilizadas técnicas de XAI baseadas em SHAP¹ (*SHapley Additive exPlanations*). O SHAP permite quantificar a contribuição individual de cada atributo para a decisão do modelo, fornecendo explicações locais e globais sobre o processo de classificação.

V. RESULTADOS

Nesta seção, são apresentados os resultados obtidos para detecção de ataques em redes IoT utilizando atributos agregados de fluxo de rede.

A. Avaliação dos Modelos

Nesta etapa, diferentes algoritmos de ML foram avaliados para a tarefa de detecção binária de ataques em redes IoT utilizando atributos agregados de fluxo de rede. Os modelos analisados foram XGBoost, LightGBM, CatBoost, ExtraTrees e Logistic Regression. Para garantir uma comparação justa entre os classificadores, todos os modelos foram inicialmente avaliados utilizando suas configurações padrão (*default*), sem ajuste de hiperparâmetros. A avaliação foi conduzida utilizando as métricas Accuracy, Precision, Recall, F1-Score

¹Disponível em: <https://shap.readthedocs.io/en/stable/index.html>

TABELA III: Comparação entre os modelos avaliados (média $\pm$ desvio padrão em 10 execuções)

Modelo	Acc.	Bal. Acc.	Prec. Atk	Rec. Atk	F1 Atk
XGB	0.9262 $\pm$ 0.0003	0.8601 $\pm$ 0.0008	0.9306 $\pm$ 0.0004	0.9783 $\pm$ 0.0002	0.9539 $\pm$ 0.0002
LGBM	0.9215 $\pm$ 0.0003	0.8478 $\pm$ 0.0008	0.9242 $\pm$ 0.0004	0.9797 $\pm$ 0.0002	0.9511 $\pm$ 0.0002
CatBoost	0.9335 $\pm$ 0.0002	0.8779 $\pm$ 0.0006	0.9398 $\pm$ 0.0004	0.9774 $\pm$ 0.0002	0.9582 $\pm$ 0.0001
ET	0.9296 $\pm$ 0.0001	0.8823 $\pm$ 0.0003	0.9442 $\pm$ 0.0002	0.9669 $\pm$ 0.0002	0.9554 $\pm$ 0.0001
LR	0.8195 $\pm$ 0.0021	0.6101 $\pm$ 0.0042	0.8200 $\pm$ 0.0016	0.9847 $\pm$ 0.0010	0.8948 $\pm$ 0.0011

e Balanced Accuracy. A Tabela III apresenta os resultados obtidos pelos modelos avaliados.

Observa-se que os modelos baseados em árvores apresentaram desempenho superior em comparação ao modelo linear LR. Entre os modelos avaliados, o ET apresentou o melhor equilíbrio entre desempenho geral e capacidade discriminativa considerando a métrica de Balanced Accuracy, sendo este selecionado para as etapas subsequentes da metodologia.

B. Otimização do Melhor Modelo e Análise Comparativa dos Atributos de Rede

Após a etapa de seleção do modelo, realizou-se a otimização dos hiperparâmetros do classificador ExtraTrees utilizando a biblioteca Optuna. Durante a etapa de busca, foram avaliadas combinações envolvendo parâmetros estruturais do classificador, isto é: número de árvores, profundidade máxima, número mínimo de amostras para divisão dos nós, número mínimo de amostras por folha, estratégia de seleção de atributos e utilização de balanceamento de classes. Os melhores hiperparâmetros obtidos pelo processo de otimização foram: `n_estimators=416`, `max_depth=32`, `min_samples_split=10`, `min_samples_leaf=7`, `max_features=None`, `bootstrap=True` e `class_weight=balanced`.

Após a otimização dos hiperparâmetros, o modelo apresentou melhoria significativa na capacidade de classificação, alcançando Accuracy de 0.9301 e Balanced Accuracy de 0.9268. Além disso, foi realizada uma etapa de seleção de atributos utilizando os valores de importância fornecidos pelo próprio modelo ExtraTrees. Com base nesses valores, foram avaliados subconjuntos contendo os 15, 10 e 5 atributos mais relevantes, respectivamente, com o objetivo de investigar o impacto da redução de dimensionalidade no desempenho do IDS. A Tabela IV apresenta a comparação entre o modelo completo e os modelos reduzidos utilizando os atributos mais relevantes selecionados pelo ExtraTrees.

TABELA IV: Impacto da redução de dimensionalidade

Modelo	Acc.	Bal. Acc.	Prec. Atk	Rec. Atk	F1 Atk	FP	FN
Completo (39)	0.9301	0.9268	0.9766	0.9327	0.9541	52190	156763
Top-15	0.9258	0.9237	0.9762	0.9274	0.9512	52707	169144
Top-10	0.9213	0.9208	0.9761	0.9216	0.9481	52660	182762
Top-5	0.9134	0.9129	0.9735	0.9137	0.9427	57926	201083

C. Explicabilidade do Modelo

Após a etapa de treinamento e seleção dos atributos mais relevantes, foi realizada uma análise de explicabilidade utilizando SHAP (*SHapley Additive exPlanations*). A Figura 2

apresenta três visualizações complementares geradas para o modelo ExtraTrees treinado com os 15 atributos mais relevantes: (a) o *summary plot*, que oferece uma visão global da importância dos atributos; (b) o *dependence plot*, que detalha a relação entre o valor do atributo mais relevante e seu impacto na predição; e (c) o *waterfall plot*, que ilustra a explicação local de uma instância individual classificada como ataque.

Na Figura 2(a), a posição horizontal representa o impacto do atributo na saída do classificador, enquanto as cores indicam a magnitude do valor da feature, variando entre valores baixos (azul) e altos (vermelho). Observa-se que o atributo *Number* apresentou contribuição significativamente superior aos demais atributos, o que evidencia que a quantidade de pacotes presentes no fluxo exerce forte influência na decisão do modelo. Esse comportamento indica que padrões volumétricos do tráfego têm elevada capacidade discriminativa no cenário avaliado. O atributo *HTTPS* também apresentou elevada relevância global. Entretanto, a distribuição dos valores SHAP indica comportamento mais heterogêneo, sugerindo que sua influência depende do contexto do fluxo analisado. Dessa forma, embora o protocolo HTTPS contribua para a classificação, ele não representa isoladamente um forte indicador de tráfego malicioso. Já atributos relacionados às estatísticas do fluxo, como *Header_Length*, *Max*, *Min*, *AVG* e *Std*, apresentaram impacto moderado no modelo, indicando que características associadas ao tamanho e variabilidade dos pacotes contribuem parcialmente para a separação entre tráfego benigno e malicioso.

A Figura 2(b) detalha o comportamento do atributo *Number*, evidenciando uma relação não linear entre o valor do atributo e seu impacto na predição: valores elevados (próximos de 100, em vermelho) deslocam a saída do modelo consistentemente na direção da classe maliciosa, enquanto valores baixos (próximos de 10, em azul) exercem influência oposta, favorecendo a classificação como tráfego benigno.

Por fim, a Figura 2(c) apresenta a explicação local de uma instância corretamente classificada como ataque. Nesse exemplo, o atributo *HTTPS* foi o principal responsável por deslocar a predição do modelo em direção à classe maliciosa, seguido por contribuições menores de *Header_Length* e outros atributos estatísticos do fluxo. O atributo *Number*, apesar de apresentar globalmente a maior importância média, exerceu nesse caso específico uma contribuição em sentido oposto, indicando que sua influência, embora relevante em termos agregados, pode variar conforme as características de cada fluxo individual. Essa análise local complementa a interpretação global do modelo, demonstrando a capacidade do SHAP de fornecer explicações consistentes tanto em nível agregado

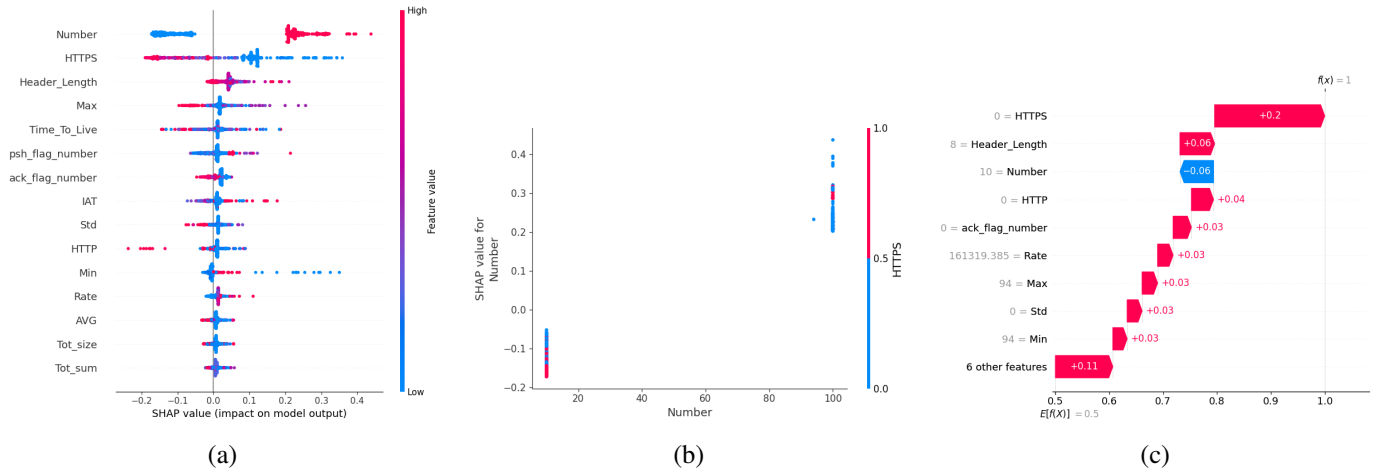


Fig. 2: Análise de explicabilidade via SHAP para o modelo ExtraTrees (Top 15 features): (a) importância global das features (*summary plot*); (b) relação entre o valor da feature Number e seu impacto na predição (*dependence plot*); (c) explicação local de uma instância classificada como ataque (*waterfall plot*).

quanto em nível de instâncias específicas.

VI. CONCLUSÕES

Este trabalho apresentou uma abordagem para detecção de ataques em redes IoT utilizando atributos agregados de fluxo de rede e técnicas de XAI. Os resultados demonstraram que modelos baseados em árvores possuem elevada capacidade discriminativa para identificação de tráfego malicioso, mesmo após redução de dimensionalidade do conjunto de atributos. Além disso, a utilização do SHAP permitiu interpretar o comportamento do modelo e identificar os atributos mais relevantes para o processo de classificação. Como trabalhos futuros, pretende-se investigar métodos de geração automática de assinaturas de ataques para IDS a partir da explicabilidade dos modelos sob atributos agregados de rede.

VII. INFORMAÇÕES SUPLEMENTARES

O desenvolvimento deste trabalho foi realizado em linguagem Python (versão 3.10) e o código-fonte utilizado para o pré-processamento dos dados, balanceamento das classes, treinamento e avaliação dos modelos, otimização de hiperparâmetros e análise de explicabilidade está disponível publicamente em: <https://github.com/matheus-fig/explainable-ids-ciciot2023-sbrt2026>.

REFERÊNCIAS

- [1] T. Domínguez-Bolaño, O. Campos, V. Barral, C. J. Escudero, and J. A. García-Naya, "An overview of iot architectures, technologies, and existing open-source projects," *Internet of Things*, vol. 20, p. 100626, Nov. 2022.
- [2] T. Slaibi, N. Ivaki, and M. Vieira, "Iot security assessment: A systematic literature review," *Journal of Systems and Software*, vol. 237, p. 112833, Jul. 2026.
- [3] A. Kanellopoulos and K. G. Vamvoudakis, "A moving target defense control framework for cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 65, no. 3, p. 1029–1043, Mar. 2020.
- [4] M. M. Rahman, S. A. Shakil, and M. R. Mustakim, "A survey on intrusion detection system in iot networks," *Cyber Security and Applications*, vol. 3, p. 100082, Dec. 2025.
- [5] A. Coscia, V. Dentamaro, S. Galantucci, A. Maci, and G. Pirlo, "Automatic decision tree-based nids ruleset generation for dos/ddos attacks," *Journal of Information Security and Applications*, vol. 82, p. 103736, May 2024.
- [6] M. Figueiredo, D. P. Sodre, R. Medeiros, V. F. Lucena, and I. Bessa, "Detection of cyberattacks in iot networks using artificial intelligence: A comparative study," in *2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, Sep. 2024, p. 1–4.
- [7] Y. Wang, M. A. Azad, M. Zafar, and A. Gul, "Enhancing ai transparency in iot intrusion detection using explainable ai techniques," *Internet of Things*, vol. 33, p. 101714, Sep. 2025.
- [8] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model," *Complexity*, vol. 2021, no. 1, Jan. 2021.
- [9] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "Ciciot2023: A real-time dataset and benchmark for large-scale attacks in iot environment," *Sensors*, vol. 23, no. 13, p. 5941, Jun. 2023.
- [10] R. Esmailyfard, Z. Shoaie, and R. Javidan, "Gelax: Iot botnet detection using dynamic graph pruning and anchored explainable ai," *Cybersecurity*, vol. 9, no. 1, Apr. 2026.
- [11] S. B. hulayyil, S. Li, and N. Saxena, "Explainable ai-based intrusion detection in iot systems," *Internet of Things*, vol. 31, p. 101589, May 2025.
- [12] M. Farhan, M. R. Naeem, A. Almadhor, A. K. Bashir, Z. Zhu, and T. R. Gadekallu, "Explainable ai-driven security framework for cyber-physical production systems in industry 4.0: Leveraging immersive embedded ciot," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 4, p. 10446–10453, Nov. 2025. [Online]. Available: <http://dx.doi.org/10.1109/TCE.2025.3602792>
- [13] L. Santos, R. Gonçalves, C. Rabadão, and J. Martins, "A flow-based intrusion detection framework for internet of things networks," *Cluster Computing*, vol. 26, no. 1, p. 37–57, Feb. 2021.
- [14] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, D. Breitenbacher, A. Shabtai, and Y. Elovici, "N-baiot: Network-based detection of iot bot-net attacks using deep autoencoders," *arXiv preprint arXiv:1805.03409*, 2018.
- [15] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An ensemble of autoencoders for online network intrusion detection," *arXiv preprint arXiv:1802.09089*, 2018.
- [16] M. A. Hossain, "Deep learning-based intrusion detection for iot networks: a scalable and efficient approach," *EURASIP Journal on Information Security*, vol. 2025, no. 1, Sep. 2025.
- [17] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19. ACM, Jul. 2019.