IA Explicável em Modelos de Aprendizado de Máquina para Detecção de Intrusão em IoT: uma Análise Comparativa

Matheus Buschermoehle[1], **Fernando Santos**[1]

[1]Departamento de Engenharia de Software
Universidade do Estado de Santa Catarina (UDESC)
Ibirama – SC – Brazil

m.buschermoehle@edu.udesc.br, fernando.santos@udesc.br

Abstract. *The increasing complexity of attacks on IoT networks drives the use of machine learning in intrusion detection systems (IDS), but a lack of interpretability limits its adoption. Explainable Artificial Intelligence (XAI) techniques offer options to overcome this limitation. This paper evaluates the use of the SHapley Additive exPlanations (SHAP) method in three predictive attack detection models: Random Forest, LightGBM, and Multilayer Perceptron (MLP), built from the NSL-KDD dataset. The use of SHAP allowed us to identify the attributes relevant to predictions, demonstrating greater similarity between the Random Forest and LightGBM models, while the MLP presents lower correlation and relies on distinct attributes.*

Resumo. *A crescente complexidade dos ataques em redes IdC impulsiona o uso de aprendizado de máquina em sistemas de detecção de intrusão (SDI), mas a falta de interpretabilidade limita sua adoção. Técnicas de Inteligência Artificial Explicável (IAE) são opções para superar essa limitação. Este artigo avalia o uso do método SHapley Additive exPlanations (SHAP) em três modelos preditivos para detecção de ataques: Random Forest, LightGBM e Multilayer Perceptron (MLP), construídos a partir do dataset NSL-KDD. O uso do SHAP permitiu identificar os atributos relevantes para as predições, evidenciando maior similaridade entre os modelos Random Forest e LightGBM, enquanto o MLP apresenta menor correlação e se apoia em atributos distintos.*

1. Introdução

A Internet das Coisas (IdC) representa um paradigma de integração de tecnologias que conecta objetos físicos à internet, permitindo a comunicação e a automação em setores como saúde, indústria e cidades inteligentes. Seu crescimento acelerado, que deve ultrapassar 30 bilhões de dispositivos em 2025 [Vailshery 2025], traz desafios de segurança cibernética. A heterogeneidade e limitação de processamento dos dispositivos criam vulnerabilidades que podem ser exploradas para ataques de *distributed denial of service* (DDoS) e vazamento de dados [Rehman et al. 2016, Stanko et al. 2024].

Nesse cenário, os Sistemas de Detecção de Intrusão (SDI) surgem como uma camada de defesa essencial, monitorando o tráfego de rede para identificar comportamentos anômalos. Para superar as limitações dos métodos tradicionais, a Inteligência Artificial (IA), especialmente o aprendizado de máquina, tem sido aplicada para aprimorar os SDI, permitindo a identificação de padrões complexos em grandes volumes de dados.

No entanto, muitos desses modelos funcionam como uma caixa-preta, o que dificulta a interpretação das suas decisões e compromete a confiança dos analistas de segurança [Botacin et al. 2021]. Para solucionar essa lacuna de interpretabilidade, a IA Explicável (IAE) oferece técnicas para tornar os modelos transparentes, permitindo compreender como eles classificam ameaças.

Este artigo avalia a aplicação do método *SHapley Additive exPlanations* (SHAP) [Lundberg and Lee 2017] como IAE em três modelos de detecção de intrusão: Random Forest (RF) LightGBM e Multi-Layer Perceptron (MLP). Os modelos foram definidos a partir de trabalhos relacionados, treinadas sobre tráfego de redes IdC e comparados quanto à interpretabilidade das predições. O estudo se justifica pela importância de fornecer explicações claras que validem classificações, reduzam riscos e favoreçam a adoção de sistemas de IA em cenários críticos [Gunning et al. 2019]. A metodologia contemplou a seleção dos modelos e do conjunto de dados, o pré-processamento, o treinamento, a aplicação da camada de IAE e, por fim, a análise individual e comparativa das explicações obtidas. Os resultados obtidos evidenciam uma similaridade entre as explicações dos modelos RF e LightGBM, enquanto o modelo MLP apresenta menor correlação.

2. Materiais e Métodos

Os modelos de detecção foram selecionados a partir da literatura existente. LightGBM e RF foram instanciados com base em [Arreche et al. 2024]. Já o MLP a partir da arquitetura proposta por [de Souza 2023]. Para implementação foi utilizado Scikit-learn, Keras/TensorFlow e LightGBM[1].

O *dataset* NSL-KDD[2] foi utilizado para treinamento e avaliação. Trata-se de um *dataset* amplamente adotado em estudos do gênero. Ele contém 125.973 registros, compostos por 43 atributos descritivos de conexões de rede normais e de diferentes tipos de ataques (DDoS, Probe, R2L e U2R). Para viabilizar a análise, foi realizado o seguinte pré-processamento dos dados: as variáveis categóricas foram transformadas com *LabelEncoder* e as numéricas padronizadas com *StandardScaler*. As múltiplas classes de ataque foram unificadas em uma única categoria "ataque", configurando um problema de classificação binária. Por fim, o *dataset* foi dividido de forma estratificada em 80% para treinamento e 20% para testes. O desempenho dos modelos foi avaliado no conjunto de teste por meio de acurácia, precisão, recall e F1-score, sendo que todos os modelos obtiveram métricas próximas de 1.0, indicando boa capacidade de identificar tráfego anômalo.

Para investigar a interpretabilidade, foi aplicada[3] IAE com a biblioteca SHAP[4]. Conforme recomendado na documentação da biblioteca, aplicou-se *TreeExplainer* para RF e LightGBM e *KernelExplainer* para MLP, permitindo quantificar a contribuição de cada atributo nas decisões dos modelos. Gráficos *beeswarm* foram gerados para analisar as explicações globais e identificar as características mais influentes na classificação.

3. Resultados e Discussão

A aplicação de IAE com SHAP revelou os atributos com maior influência nas decisões de cada modelo, apresentados na Figura 1. Nos modelos RF e LightGBM os atributos

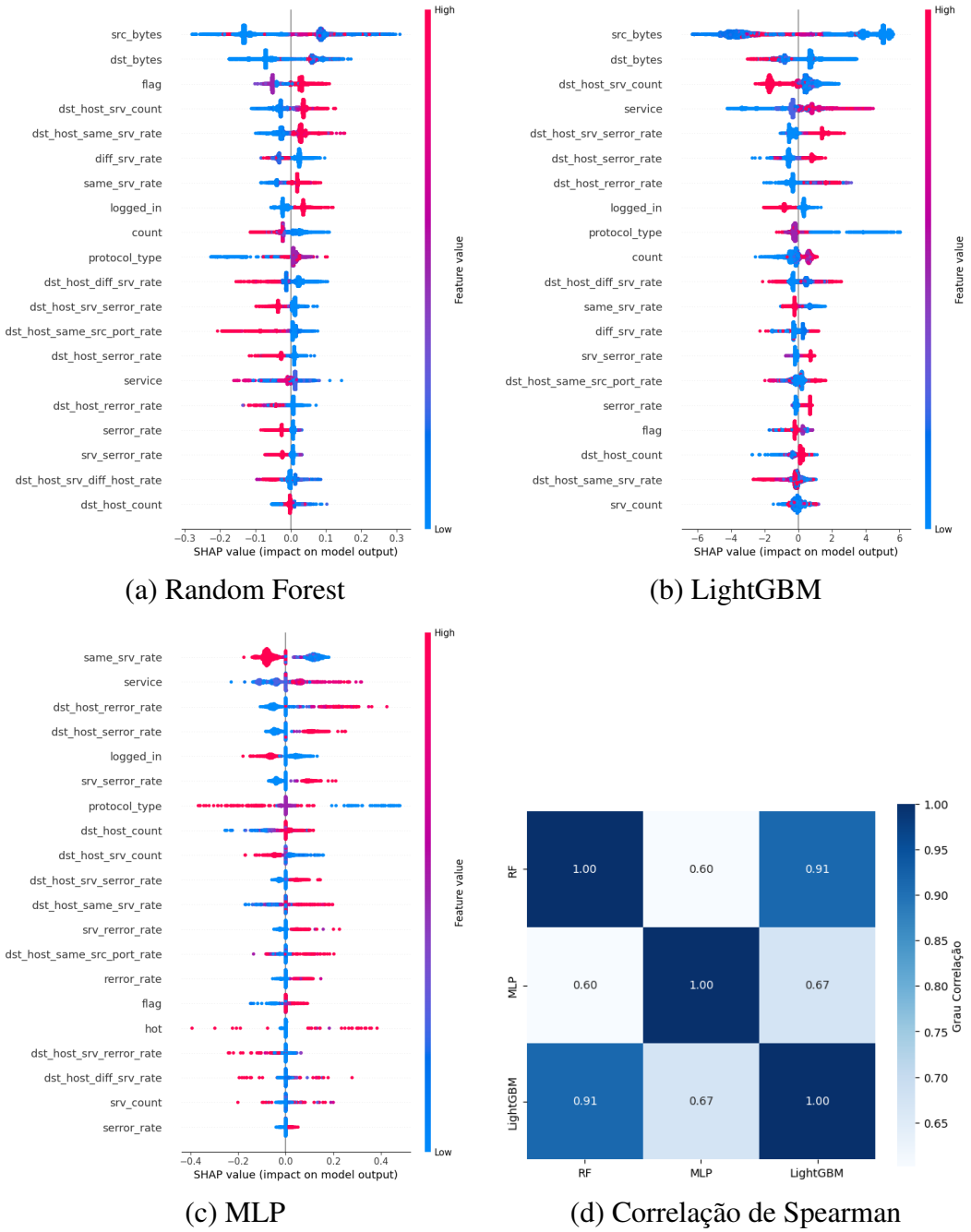
[1]https://lightgbm.readthedocs.io/

[2]https://www.unb.ca/cic/datasets/nsl.html

[3]https://github.com/matheusbus/xai-in-ids

[4]https://shap.readthedocs.io

relacionados ao volume de dados (i.e., *src_bytes* e *dst_bytes*) foram apontados como os mais relevantes para classificação. Já no MLP os atributos mais relevantes foram aqueles que descrevem o padrão dos serviços (i.e., *same_srv_rate* e *service*). A divergência mais notável foi observada no atributo *logged_in* que indica se o usuário da conexão teve sucesso ou não na autenticação: enquanto o RF o associou a um indicativo de ataque, o LightGBM e o MLP o interpretaram como um sinal de tráfego normal, evidenciando como modelos podem gerar decisões semelhantes considerando atributos diferentes.



(a) Random Forest (b) LightGBM

(c) MLP (d) Correlação de Spearman

Figura 1. Atributos relevantes conforme o SHAP e correlação entre modelos

A correlação de Spearman foi calculada para avaliar a similaridade nas decisões dos modelos. A correlação entre os rankings de importância dos atributos do RF e do

LightGBM foi relativamente alta (0,91), confirmando que suas estratégias de classificação são bastante similares. No entanto, o MLP apresentou uma correlação menor com os modelos RF e LightGBM (0,60 e 0,67), sugerindo que sua arquitetura de rede neural captura padrões de forma distinta. Este resultado evidencia que, embora todos os modelos sejam eficazes na detecção de intrusões, a natureza das suas explicações é variada.

4. Considerações finais

Este trabalho buscou analisar diferentes arquiteturas de modelos de detecção de intrusão que atuam no contexto de detecção de intrusões baseadas no *dataset* NSL-KDD, evidenciando diferenças de interpretabilidade entre RF, LightGBM e MLP. Enquanto atributos relacionados ao volume de dados tem grande influência nas predições dos modelos RF e LightGBM, o MLP deu maior relevância para atributos que descrevem o padrão dos serviços. Também foi possível evidenciar maior correlação das explicações entre os modelos baseados RF e LightGBM enquanto o MLP apresentou menor correlação.

Para estudos futuros, sugere-se avaliar a camada IAE em ambientes reais de produção, desenvolvendo interfaces interpretativas que traduzam as explicações do SHAP em textos acessíveis para operadores de segurança. Também sugere-se e realizar estudos para verificar a aplicabilidade das explicações.

Fernando Santos agradece à FAPESC pelo apoio financeiro recebido (TO2023TR246).

Referências

Arreche, O., Guntur, T., and Abdallah, M. (2024). XAI-IDS: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems. *Applied Sciences*, 14(10):4170.

Botacin, M., Ceschin, F., Sun, R., Oliveira, D., and Grécio, A. (2021). Challenges and pitfalls in malware research. In *Computers & Security vol. 106*. Elsevier B.V.

de Souza, C. A. (2023). *Detecção e prevenção de intrusão em computação de nevoeiro e Internet das Coisas*. Doutorado em ciência da computação, Universidade Federal de Santa Catarina (UFSC).

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G. (2019). XAI—explainable artificial intelligence. *American Association for the Advancement of Science*, 4(37).

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Rehman, S. U., Khan, I. U., Moiz, M., Hasan, S., et al. (2016). Security and privacy issues in IoT. *International Journal of Communication Networks and Information Security*, 8(3):147.

Stanko, A., Duda, O., Mykytyshyn, A., Totosko, O., and Koroliuk, R. (2024). Artificial intelligence of things (AIoT): Integration challenges, and security issues. In *Proceedings of the Bioinformatics and applied information technologies*.

Vailshery, L. S. (2025). Internet of things (IoT) - statistics & facts. Disponível em: `https://www.statista.com/topics/2637/internet-of-things`. Acesso em: 21 ago. 2025.