# Detecção de Ameaças IoT: Abordagem Federada e Explicável

Alessandra Carolina Domiciano & Paulo Otavio Luczensky de Souza

***Abstract*— Technological advancements and increasing digitalization have significantly intensified the occurrence of cyberattacks, especially in distributed and connected systems. In recent years, there has been a significant increase in incidents in Internet of Things (IoT) environments, which have become attractive targets due to the large volume of devices and the continuous exchange of data. In this context, Machine Learning (ML) has emerged as a key approach for threat detection, although challenges related to interpretability and privacy persist. This work aims to present a federated learning application for threat detection, using XAI techniques, in order to reconcile high performance, preservation of data privacy, and greater reliability of the model's decisions. The results demonstrated that the model has optimal performance for binary classification (accuracy and F1 score of 99%), but in multiclass scenarios it presents limitations regarding the discrimination of similar categories of attacks. SHAP analysis complemented the quantitative assessment, revealing the most relevant variables in class discrimination and elucidating the causes of the main limitation observed in multiclass classification—the overlap of features between DDoS and DoS.**

***Index Terms*— Internet of Things, Cybersecurity, Explainable Artificial Intelligence, Federated Learning**

***Resumo*— O avanço tecnológico e a crescente digitalização ampliaram significativamente a ocorrência de ataques cibernéticos, especialmente em sistemas distribuídos e conectados. Nos últimos anos, observou-se um aumento expressivo de incidentes em ambientes de Internet das Coisas (IoT), que se tornaram alvos atrativos devido ao grande volume de dispositivos e à troca contínua de dados. Diante desse cenário, métodos baseados em *Machine Learning* (ML) têm sido amplamente adotados para detecção de ameaças, embora ainda enfrentem desafios relacionados à interpretabilidade e à privacidade. Assim, este trabalho tem como objetivo apresentar uma aplicação de aprendizado federado para detecção de ameaças, com uso de técnicas de XAI, de modo a conciliar alto desempenho, preservação da privacidade dos dados e maior confiabilidade das decisões do modelo. Os resultados demonstraram que o modelo possui desempenho ótimo para classificação binária (acurácia e F1-score de 99%), mas em cenários multiclasses apresenta limitações quanto a discriminação de categorias similares de ataques. A análise SHAP complementou a avaliação quantitativa, revelando as variáveis mais relevantes na discriminação entre as classes e elucidando as causas da principal limitação observada na classificação multiclasses — a sobreposição de características entre DDoS e DoS.**

***Palavras Chave*— Internet das Coisas, Cibersegurança, Inteligência Artificial Explicável, Aprendizado Federado**

## I. Introdução

Desde a Pré-História até os dias atuais, o ser humano tem estado continuamente envolvido com a evolução tecnológica. Esse processo começou com a criação de ferramentas primitivas, como pedras lascadas, e avançou até os modernos sistemas computacionais, incluindo computadores e smartphones, que impulsionam o desenvolvimento social. Paralelamente ao avanço tecnológico, a área de segurança também evoluiu. Atualmente, com grande parte das informações armazenadas em servidores e data centers, a preocupação com a proteção de dados sensíveis tornou-se ainda mais intensa. Nesse contexto, cresce a necessidade de prevenir ataques e evitar o comprometimento dessas informações.

Na tabela abaixo [1], observa-se um crescimento constante no número de ataques cibernéticos entre 2016 e 2021, envolvendo principalmente fraudes de investimentos, golpes com cartões de crédito, extorsão e, de forma destacada, ataques de phishing, um dos métodos mais recorrentes.

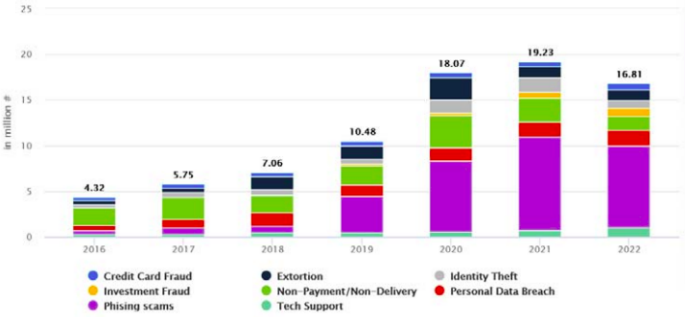


Fig. 1. Número de *cyber* ataques de 2018 a 2022

Apesar da redução significativa registrada em 2022, em 2023 o Brasil voltou a apresentar números elevados de tentativas de ataques cibernéticos, ultrapassando 60 bilhões. Diante desse cenário, a Brasscom estima que os investimentos em cibersegurança possam alcançar R$ 104,6 bilhões entre 2025 e 2028 [2], evidenciando a relevância e a urgência do tema.

Além disso, outro setor tecnológico fortemente impactado por esses ataques é o da Internet das Coisas, do inglês *Internet of Things* (IoT), no qual dados coletados e trocados entre diversos dispositivos tornam-se alvos atrativos para invasores. A seguir, apresenta-se uma tabela [1] que evidencia o aumento de ataques a sistemas de IoT em escala mundial entre 2018 e 2022, demonstrando um crescimento significativo: de 32,7 milhões em 2018 para 112,29 milhões em 2022.

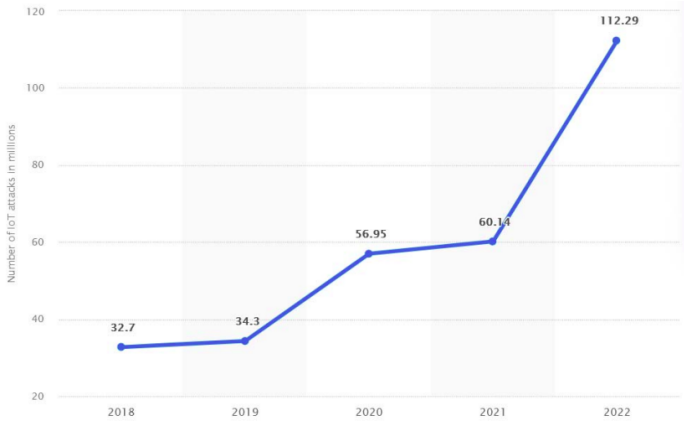


Fig. 2. Número de ataques *IoT* de 2018 a 2022

A segurança em ambientes de IoT tornou-se essencial, especialmente devido ao crescente volume de dispositivos conectados e ao aumento de ataques direcionados a esses sistemas. Nesse contexto, garantir a confidencialidade, a integridade e a disponibilidade dos serviços depende diretamente de mecanismos eficazes de detecção de ameaças.

O uso de Aprendizado de Máquina, do inglês *Machine Learning* (ML) tem se destacado nesse cenário, pois seus algoritmos conseguem aprender padrões em grandes volumes de dados e se adaptar a novos tipos de ataques, tornando-se uma ferramenta fundamental para fortalecer a segurança cibernética.

Ainda assim, compreender as decisões tomadas pelos modelos permanece um desafio. Técnicas de inteligência artificial explicável (*Explainable Artificial Intelligence* – XAI), como SHAP, permitem identificar os fatores que influenciam a classificação de atividades maliciosas, aumentando a transparência e a confiabilidade do processo.

Além disso, sistemas de detecção em IoT podem ser aprimorados com Aprendizado Federado, do inglês *Federated Learning* (FL), que possibilita o treinamento descentralizado dos modelos, preservando a privacidade e favorecendo soluções mais robustas e escaláveis, adequadas à natureza distribuída dos dispositivos IoT.

Diante desse contexto, este trabalho tem como objetivo apresentar uma aplicação de aprendizado de máquina para detecção de ameaças que integre Aprendizado Federado e técnicas de XAI, de modo a conciliar alto desempenho, preservação da privacidade dos dados e maior confiabilidade das decisões do modelo, mesmo em cenários com recursos computacionais limitados, como os ambientes de IoT.

## II. Referencial Teórico

A expansão dos sistemas conectados impulsionada pela Indústria 4.0 intensifica a integração entre dispositivos IoT e serviços digitais, elevando também os riscos de segurança. Com dispositivos cada vez mais limitados em energia e recursos, cresce a necessidade de soluções eficientes e escaláveis. Assim, torna-se essencial explorar abordagens que combinem precisão, explicabilidade e descentralização na segurança para IoT.

1. **The Role of Software Engineering in Industry 4.0 [1]**: O artigo de Samir Lemeš reforça fortemente que a engenharia de software não é apenas um componente de suporte, mas sim o núcleo habilitador da Indústria 4.0. Sem software bem pensado, a promessa de conectividade, automação, análise de dados e segurança não se sustenta. A partir dessa perspectiva, a segurança cibernética em ambientes de IoT ou CPS deve ser tratada desde a concepção do sistema, e não como uma camada adicional. Por outro lado, os desafios práticos de integração de legado, capacitação, escalabilidade de dados apontam para lacunas reais na implementação industrial, o que significa que teoria e prática ainda precisam convergir para viabilizar plenamente o modelo da Indústria 4.0.

2. **Improving IoT Security With Explainable AI: Quantitative Evaluation of Explainability for IoT Botnet Detection [3]**: O estudo de Kalakoti, Bahsi e Nõmm (2024) destaca avanços significativos no uso de técnicas de *Explainable AI* (XAI) para fortalecer a detecção de botnets em ambientes de Internet das Coisas (IoT). O trabalho avalia quantitativamente a qualidade das explicações geradas por métodos pós-hoc, como LIME e SHAP, aplicados a modelos de classificação treinados em diferentes conjuntos de dados de ataques IoT. Além disso, utiliza técnicas de redução de atributos para tornar os modelos mais leves, aspecto crucial para dispositivos com capacidade computacional limitada. Os resultados mostram que o SHAP fornece explicações mais fiéis, consistentes e menos sensíveis, contribuindo para maior transparência das decisões do modelo. A pesquisa reforça que a combinação entre desempenho, interpretabilidade e otimização de recursos é fundamental para o desenvolvimento de soluções de segurança viáveis em ambientes IoT, evidenciando a importância de integrar XAI como parte estruturante na detecção de ameaças.

3. **Federated XAI IDS: An Explainable and Safeguarding Privacy Approach to Detect Intrusion Combining Federated Learning and SHAP [4]**: O trabalho de Fatema et al. (2025) propõe um sistema de detecção de intrusões para IoT que combina FL e técnicas de XAI, formando o framework Federated XAI IDS. O estudo destaca que, embora modelos tradicionais de detecção apresentem alta acurácia, eles frequentemente negligenciam questões de privacidade e interpretabilidade — aspectos críticos em ambientes IoT descentralizados. A abordagem federada permite treinar modelos de forma colaborativa sem compartilhamento direto dos dados, preservando informações sensíveis, enquanto o uso de SHAP fornece explicações claras sobre as características que influenciam cada decisão do modelo. Os resultados demonstram que a integração entre FL e XAI melhora a transparência, mantém o desempenho competitivo e possibilita diagnósticos mais confiáveis em redes IoT heterogêneas. Esse trabalho reforça a importância de soluções que combinem privacidade, explicabilidade e eficiência operacional em sistemas modernos de detecção de ameaças.

## III. Metodologia

Nesta seção são descritos os processos de implementação do sistema proposto, em que cada tópico corresponde a uma etapa

do desenvolvimento. São apresentados a arquitetura do projeto, o dataset utilizado, o modelo de aprendizado de máquina adotado, a aplicação de aprendizado federado e da ferramenta de otimização de hiperparâmetros, além do uso de técnicas de XAI para interpretação das decisões do modelo.

### A. *Arquitetura do Projeto*

Inicialmente, os dados são segmentados para aplicação de Aprendizado Federado (FL), permitindo distribuir o processo de treinamento entre vários clientes. Em seguida, o modelo de aprendizado de máquina é treinado localmente em cada cliente. Para aprimorar ainda mais o desempenho, foi incorporado o *framework* Optuna, responsável pela automatização da busca pelos valores ideais dos hiperparâmetros do modelo.

Após o término do treinamento federado, o desempenho do modelo é avaliado sob as métricas de acurácia, precisão e recall. A avaliação foi conduzida em dois cenários distintos: classificação multiclasse e binária. Por fim, aplicou-se uma técnica de Inteligência Artificial Explicável, para obter uma compreensão mais profunda sobre o impacto de cada variável no processo de decisão do modelo, visando um sistema mais transparente e confiável.Todo esse processo é executado tanto para classificação multiclasses quando para classificação binária.

A seguir, apresenta-se a figura que sintetiza o fluxo completo da arquitetura proposta.

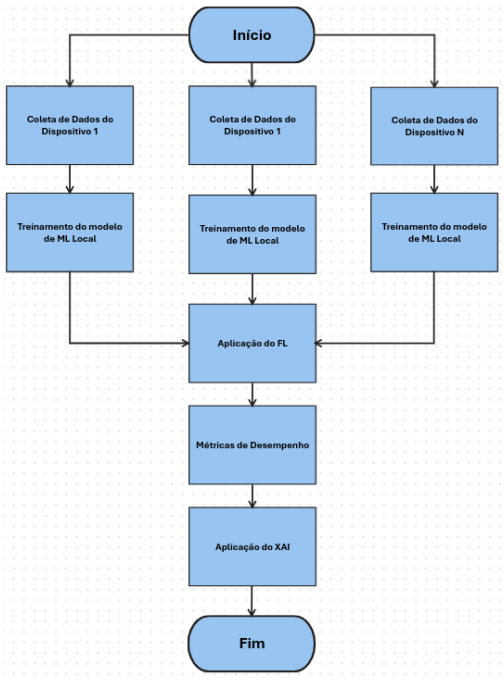


Fig. 3. Arquitetura proposta para a implementação do sistema de detecção de ameaças.

### B. *Dataset*

O dataset utilizado foi o CICIoT2023 [5], composto por dados gerados em um cenário IoT com 115 dispositivos reais. Ele abrange 33 tipos de ataques, organizados em sete categorias, além de tráfego benigno. Para este trabalho, foram considerados quatro tipos de tráfego: benigno, DDoS, DoS e reconhecimento. A relação entre as classes, as categorias de tráfego e seus respectivos ataques é apresentada na tabela I. No total, foram utilizadas 664.743 amostras, cuja distribuição é apresentada na tabela II. Além disso, os dados foram particionados em três conjuntos — treinamento, validação e teste — seguindo a proporção de 70%, 15% e 15%, respectivamente.

TABELA I
MAPEAMENTO DE CLASSES, CATEGORIAS E TIPOS DE ATAQUE

| Classe | Categoria | Ataque |
|---|---|---|
| 0 | Benign | - |
| 1 | DDoS | PSHACK FLOOD<br>ICMP FLOOD<br>TCP FLOOD<br>SYN FLOOD<br>UDP FLOOD<br>SYNONYMOUSIP FLOOD<br>RSTFIN FLOOD<br>SLOWLORIS<br>ICMP FRAGMENTATION<br>ACK FRAGMENTATION<br>UDP FRAGMENTATION<br>HTTP FLOOD |
| 2 | DoS | UDP FLOOD<br>TCP FLOOD<br>SYN FLOOD<br>HTTP FLOOD |
| 3 | Recon | HOST DISCOVERY<br>VULNERABILITY SCAN<br>PORT SCAN<br>OS SCAN<br>PING SWEEP |

TABELA II
DISTRIBUIÇÃO DAS CLASSES NO CONJUNTO DE DADOS

| Classe | Percentual |
|---|---|
| 0 | 2,5% |
| 1 | 77,5% |
| 2 | 18,4% |
| 3 | 1,6% |

Cada fluxo de rede é descrito por 39 features, derivadas de metadados, estatísticas temporais, comportamento de pacotes e informações de protocolo, como mostra a tabela III.

TABELA III
DESCRIÇÃO DAS PRINCIPAIS FEATURES DO DATASET

| Feature | Descrição |
|---|---|
| Number | Número total de pacotes no fluxo |
| psh flag number | Valor da flag PSH |
| IAT | Diferença de tempo em relação ao pacote anterior |
| fin flag number | Valor da flag FIN |
| Header Length | Tamanho do cabeçalho |
| syn flag number | Valor da flag SYN |
| Rate | Taxa de transmissão de pacotes em um fluxo |
| ICMP | Indica se o protocolo de rede é ICMP |
| UDP | Indica se o protocolo de transporte é UDP |
| HTTPS | Indica se o protocolo de aplicação é HTTPS |
| rst flag number | Valor da flag RST |
| AVG | Tamanho médio dos pacotes no fluxo |
| TCP | Indica se o protocolo de transporte é TCP |
| Max | Tamanho máximo do pacote no fluxo |
| ack flag number | Valor da flag ACK |
| Duration | Tempo de vida (TTL — Time to Live) |
| syn count | Número de pacotes com flag SYN ativa no mesmo fluxo |

### C. Treinamento com HPO

Nesta seção, apresenta-se o modelo de aprendizado de máquina selecionado para a classificação dos ataques cibernéticos, bem como a técnica empregada para a otimização dos hiperparâmetros, visando identificar a melhor combinação de valores e, consequentemente, maximizar o desempenho do modelo.

#### *c.1 XGBoost*

O modelo de Aprendizado de Máquina utilizado neste trabalho foi o *XGBoost*. O qual baseia-se em árvores de decisão e é conhecido pela sua capacidade de ser altamente eficiente, preciso e rápido na resolução de problemas de classificação e regressão. Adicionalmente, é amplamente utilizado devido a sua habilidade em lidar com volumes massivos de dados, parâmetros esparsos e problemas complexos, além de também ser robusto na detecção de padrões não lineares.

A Tabela IV exibe os hiperparâmetros que foram utilizados para o treinamento do modelo de *Machine Learning*, destacando o nome do hiperparâmetro e a sua respectiva descrição.

TABELA IV
HIPERPARÂMETROS DO XGBOOST UTILIZADOS NO PROJETO

| **Hiperparâmetro** | **Descrição** |
|---|---|
| eta | Taxa de aprendizado |
| max_depth | Profundidade máxima da árvore |
| subsample | Amostragem de linhas por árvore |
| colsample_bytree | Amostragem de colunas por árvore |
| lambda | Regularização L2 |
| alpha | Regularização L1 |

#### *c.2 Optuna*

A seleção adequada de cada valor dos hiperparâmetros apresentados na subseção anterior representa um desafio relevante, uma vez que diferentes combinações podem resultar em desempenhos distintos para o mesmo modelo e conjunto de dados. Para identificar os valores que proporcionassem o melhor desempenho, empregou-se o *framework* de otimização de hiperparâmetros *Optuna*, permitindo automatizar o processo de busca e avaliação.

A seguir, apresenta-se a Tabela V contendo os hiperparâmetros utilizados, acompanhados dos intervalos de valores explorados pelo *Optuna* para determinar a combinação mais adequada ao modelo.

TABELA V
FAIXA DE VALORES DOS HIPERPARÂMETROS DO XGBOOST

| **Hiperparâmetro** | **Faixa de Valores** |
|---|---|
| eta | 0.01 – 0.3 |
| max_depth | 6 – 12 |
| subsample | 0.6 – 1.0 |
| colsample_bytree | 0.6 – 1.0 |
| lambda | $10^{-8}$ – 1.0 |
| alpha | $10^{-8}$ – 1.0 |

Durante a otimização, o *Optuna* registra internamente todas as combinações testadas, métricas avaliadas e eventuais interrupções por *pruning*. Ao término da busca, o *Optuna* apresenta o menor valor da função de perda *multiclass log loss* (*mlogloss*) encontrado e o conjunto de hiperparâmetros responsáveis pelo melhor desempenho, permitindo sua aplicação direta no modelo federado.

TABELA VI
MELHORES VALORES DE HIPERPARÂMETROS ENCONTRADOS PELO OPTUNA

| **Hiperparâmetro** | **Valor** |
|---|---|
| eta | 0.1673 |
| max_depth | 9 |
| subsample | 0.83 |
| colsample_bytree | 0.72 |
| lambda | 0.0124 |
| alpha | 0.0019 |

Ao final do treinamento, realizado com os melhores valores de hiperparãmetros, o valor de mlogloss obtido foi de 0.8032.

### D. Aprendizado Federado

Devido ao grande volume de dados gerado por cada dispositivo e à necessidade de realizar treinamentos mais eficientes e robustos, aplicou-se o Aprendizado Federado. Essa abordagem permite o treinamento colaborativo de modelos de aprendizado de máquina diretamente nos dispositivos, utilizando dados descentralizados sem a necessidade de compartilhamento ou centralização das informações. Reduzindo a sobrecarga de comunicação e possibilitando atualizações contínuas do modelo, incorporando o conhecimento distribuído de maneira escalável e segura.

A tabela VII, a seguir, mostra as principais configurações prédefinidas do ambiente de Aprendizado Federado.

TABELA VII
CONFIGURAÇÕES DO AMBIENTE DE APRENDIZADO FEDERADO

| **Parâmetro** | **Valor** | **Descrição** |
|---|---|---|
| num_clients | 4 | Número de clientes participantes |
| num_boost_round | 50 | Total de rounds de boosting do XGBoost |
| rounds_per_client | 5 | Rounds locais executados por cliente |

### E. Explicação com SHAP

A interpretação de modelos de aprendizado de máquina é essencial para garantir transparência, confiabilidade e compreensão dos fatores que influenciam as decisões algorítmicas. Entre os métodos de explicação mais reconhecidos está o SHAP (SHapley Additive exPlanations), fundamentado na Teoria dos Jogos Cooperativos. Esse método quantifica a contribuição de cada feature para a predição, indicando se ela aumenta ou reduz a saída do modelo em relação ao valor de referência (base value) [6]. Em problemas multiclasse, cada instância apresenta um vetor de contribuições, permitindo analisar separadamente o impacto das variáveis em cada classe prevista. Dentre as visualizações fornecidas estão os summary plots, que sintetizam quais features exercem maior influência nas decisões de cada classe do modelo, revelando não apenas a importância relativa das variáveis, mas também a direção e a distribuição de seus efeitos.

## IV. RESULTADOS E DISCUSSÃO

A avaliação do modelo foi conduzida em dois cenários: (i) classificação multiclasse, contemplando as categorias Be-

nign, DDoS, DoS e Recon; e (ii) classificação binária, distinguindo apenas entre Benign e Attack. Os resultados evidenciam diferenças significativas entre os dois cenários, tanto em termos de acurácia quanto de capacidade discriminativa entre as diferentes classes de tráfego.

É importante destacar que, em sistemas de detecção de ameaças, é fundamental que o modelo apresente desempenho equilibrado em todas as métricas, pois tanto falsos negativos quanto falsos positivos trazem riscos significativos, embora de naturezas diferentes. Falsos negativos representam ataques que não foram detectados, permitindo que atividades maliciosas passem despercebidas e comprometam a rede ou o dispositivo. Por outro lado, falsos positivos também são prejudiciais: quando tráfego benigno é erroneamente classificado como ataque, o sistema pode acionar alarmes indevidos ou executar ações de bloqueio, resultando em prejuízos operacionais.

A seguir são apresentados os resultados para ambos os cenários, além da análise interpretativa das predições do modelo multiclasses.

### A. *Classificação multiclasses*

No cenário multiclasses, o modelo demonstrou desempenho geral razoável, com acurácia de 85% e F1-score de 83%. Ainda assim, há limitações importantes relacionadas à discriminação entre classes com padrões de tráfego semelhantes. A tabela VIII e a figura 4 exibem as métricas de desempenho e a matriz de confusão, respectivamente, desse cenário.

TABELA VIII
MÉTRICAS DE CLASSIFICAÇÃO PARA O CENÁRIO MULTICLASSES

| Classe | Precisão | Recall | F1-score | Suporte |
|---|---|---|---|---|
| 0 (Benign) | 0.84 | 0.87 | 0.86 | 2486 |
| 1 (DDoS) | 0.86 | 0.97 | 0.91 | 77267 |
| 2 (DoS) | 0.72 | 0.35 | 0.47 | 18374 |
| 3 (Recon) | 0.79 | 0.74 | 0.77 | 1585 |
| **Acurácia** | | | 0.85 | |
| **Macro média** | 0.80 | 0.73 | 0.75 | 99712 |
| **Média ponderada** | 0.834 | 0.85 | 0.83 | 99712 |

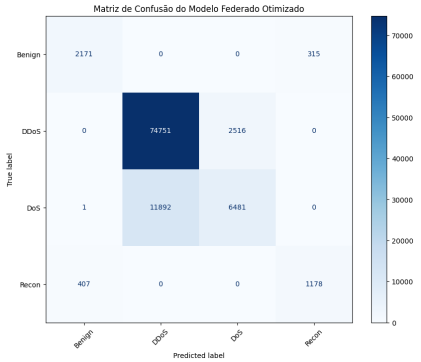


Fig. 4. Matriz de confusão para o cenário multiclasses

Os resultados indicam que o modelo apresenta boa capacidade de identificar tráfego benigno, com precisão de 84% e recall de 87%, embora ainda existam falsos positivos e falsos negativos. A classe Recon apresenta desempenho ligeiramente inferior, com precisão de 79% e recall de 74%, refletindo maior incidência de erros. A matriz de confusão revela que esses equívocos ocorrem especificamente entre as classes Benign e Recon, evidenciando que o modelo tende a confundir tráfego legítimo com atividades de reconhecimento e vice-versa.

A classe DDoS apresentou o melhor desempenho entre os ataques, alcançando precisão de 86% e recall de 97%, o que indica a ocorrência de alguns falsos positivos, mas uma taxa muito baixa de falsos negativos. Em contraste, a classe DoS revelou-se o principal ponto fraco do modelo, com precisão de 72% e recall de apenas 35%, evidenciando grande dificuldade em detectar corretamente esse tipo de ataque. A matriz de confusão mostra que a maior parte desses erros decorre da confusão entre as classes DoS e DDoS, com o modelo classificando frequentemente tráfego DoS como DDoS, o que reforça a semelhança entre os padrões dessas categorias no dataset utilizado.

### B. *Classificação binária*

No cenário binário, o modelo apresentou desempenho ótimo, alcançando acurácia e F1-score de 99%. Esse resultado superior decorre da agregação das diferentes categorias de ataque em uma única classe maliciosa, o que elimina os erros de confusão entre os tipos específicos de ataques no cenário multiclasses, levando a uma classificação significativamente mais consistente. A tabela IX e a figura 5 exibem as métricas de desempenho e a matriz de confusão, respectivamente, desse cenário.

TABELA IX
MÉTRICAS DA CLASSIFICAÇÃO BINÁRIA DO MODELO FEDERADO OTIMIZADO

| Classe | Precisão | Recall | F1-score | Suporte |
|---|---|---|---|---|
| 0 (Benign) | 0.85 | 0.88 | 0.86 | 2486 |
| 1 (Attack) | 1.00 | 1.00 | 1.00 | 97226 |
| **Acurácia** | | | 0.99 | |
| **Macro média** | 0.93 | 0.94 | 0.93 | 99712 |
| **Média ponderada** | 0.99 | 0.99 | 0.99 | 99712 |

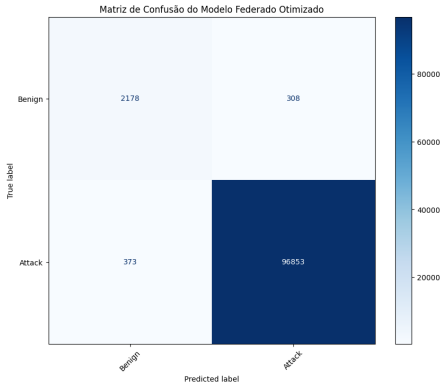


Fig. 5. Matriz de confusão binária do modelo federado otimizado

Os resultados demonstram um desempenho quase perfeito para a classe maliciosa, com acurácia, precisão e F1-score de 100%. A matriz de confusão confirma a baixíssima incidência de falsos positivos e falsos negativos para essa classe. Já a classe benigna apresentou desempenho razoável, com precisão de 85%, recall de 88% e F1-score de 86%. Esse comportamento é amplamente esperado em cenários práticos, onde a prioridade é a distinção entre tráfego benigno e malicioso, independentemente da tipologia específica do ataque.

### C. *Explicação com SHAP*

Os resultados globais obtidos por meio de SHAP permitem compreender influência de cada feature nas decisões do modelo. A figura 6 apresenta as features que tiveram maior impacto nas classificações em geral.

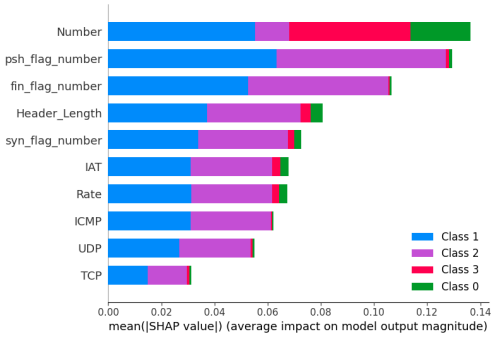


Fig. 6. Importância global das features segundo SHAP

É possível perceber que as características de fluxo com maior peso nas classificações são o número de pacotes, valores das flags PSH e FIN, seguidos pelo comprimento de cabeçalho, valor da flag SYN, intervalo entre a chegada de pacotes e taxa de transmissão. Isso demonstra coerência com o comportamento típico de tráfego malicioso, especialmente em ataques de negação de serviço (DoS e DDoS), que tendem a gerar volumes altos de pacotes, intervalos curtos entre transmissões e uso recorrente de flags como SYN, PSH e FIN para sobrecarregar ou manipular o estado da conexão.

Obervando os resultados globais por classe, pode-se analisar não apenas a magnitude do impacto das features, mas também a direção com que contribuem para as decisões do modelo, indicando se favorecem ou reduzem a probabilidade de pertencimento a uma determinada classe. As figuras 7, 8, 9 e 10 apresentam as distribuições dos valores de Shapley para cada feature em cada classe.

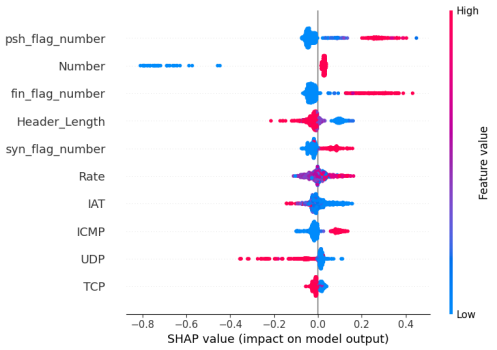


Fig. 7. Importância global das features para classe DDoS

Os resultados para a classe DDoS demonstram que as variáveis com maior contribuição positiva incluem as flags PSH, FIN e SYN. A ativação dessas flags pode indicar padrões de manipulação e sinalização atípica no fluxo TCP, associados a comportamentos anômalos característicos, como interrupções artificiais, renegociações incompletas e abertura massiva de conexões simuladas.

Além disso, os atributos de taxa de transmissão e intervalo entre a chegada de pacotes também possuem contribuições positivas, com valores altos do primeiro e valores baixos do segundo favorecendo a predição por essa classe. Isso é compatível com o perfil de saturação de ataques de DDoS.

Em contraste, A variável com maior contribuição negativa é o número de pacotes no fluxo: quanto menor seu valor, menos provável será a classificação de uma instância como DDoS, o que também remete ao perfil de saturação dessa classe.

Cabe destacar que a predominância da classe DDoS no dataset favorece o aprendizado detalhado de seus padrões, reforçando a associação positiva dessas features com essa classe e ampliando seu papel como elementos distintivos em relação às demais categorias.

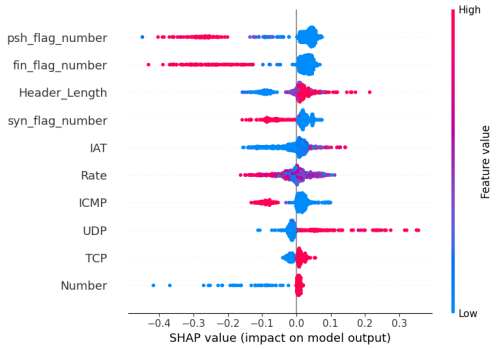


Fig. 8. Importância global das features para classe DoS

Para a classe DoS, o comportamento do modelo apresenta diferenças marcantes em relação ao DDoS. As flags PSH e FIN exibem contribuições negativas significativas, com valores altos reduzindo a probabilidade de classificação de uma instância como DoS. Esse comportamento é coerente com a composição dos ataques constantes no dataset para essa categoria, que tipicamente não fazem uso dessas flags.

Por outro lado, o valor da flag SYN, a taxa de transmissão e o intervalo entre a chegada de pacotes apresentam contribuições negativas para a predição dessa classe. Valores baixos para a última e valores altos para as demais reduzem a probabilidade de classificação. Isso reflete um padrão impreciso para DoS aprendido pelo modelo, que pode explicar a alta taxa de falsos negativos dessa classe, pois muitas instâncias de DoS acabam sendo absorvidas pela classe DDoS devido à semelhança dos padrões e ao forte desequilíbrio entre as classes. Como consequência, o modelo depende mais da interação entre múltiplas features para identificar corretamente essa classe e distingui-la de DDoS.

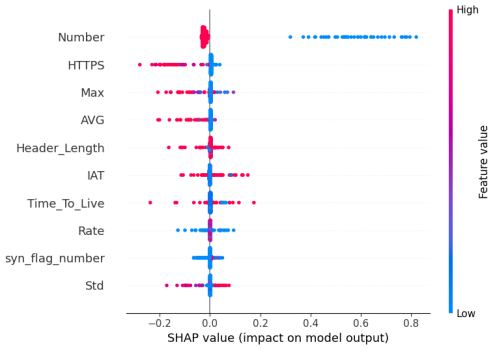


Fig. 9. Importância global das features para classe Recon

Os resultados para a classe Reconhecimento (Recon) evidenciam um padrão de contribuições claramente distinto dos demais ataques, refletindo a natureza exploratória, esparsa e de baixo volume típica desse tipo de tráfego. A feature com maior impacto positivo é o número de pacotes no fluxo, com valores menores aumentando a probabilidade de classificação como ataque de reconhecimento. Já as features com maior impacto negativo são protocolo HTTPS e os tamanhos máximo e médio dos pacotes, com valores elevados reduzindo a probabilidade de escolha pela classe.

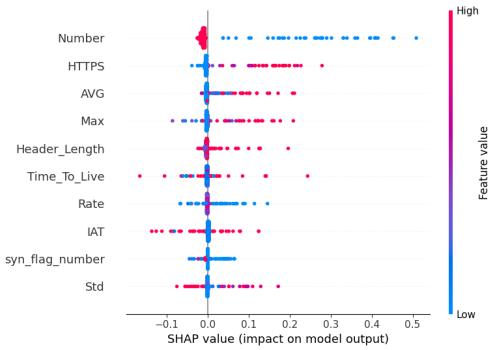


Fig. 10. Importância global das features para classe Benign

Para a classe Benign, o modelo identifica um padrão compatível com o comportamento típico do tráfego legítimo em ambientes IoT, caracterizado por fluxos curtos e de baixa volumetria. O número de pacotes no fluxo apresenta o impacto positivo mais expressivo, indicando que o modelo tende a associar fluxos com poucos pacotes a tráfego benigno. De forma semelhante, as features HTTPS e tamanhos médio e máximo de pacotes exercem contribuição positiva marcante, reforçando essas características como indicadores de normalidade no contexto analisado.

## V. Conclusão

Este trabalho apresentou uma aplicação de aprendizado federado para a tarefa de detecção de ameaças no contexto de IoT, utilizando o modelo XGBoost e o dataset CICIoT2023. A análise contemplou tanto o cenário de classificação multiclasses quanto o binário, além da aplicação de métodos de interpretabilidade baseados em SHAP para compreender o comportamento interno do modelo.

Os resultados demonstram que, embora o modelo apresente desempenho razoável no cenário multiclasses, ainda existem limitações relevantes na discriminação entre tipos específicos de ataques, devido a semelhança entre os padrões de tráfego e ao forte desbalanceamento presente no dataset.

Por outro lado, ao considerar apenas a distinção entre tráfego benigno e malicioso, o modelo alcança desempenho quase perfeito. A agregação das classes de ataque reduz drasticamente os erros de confusão observados no cenário multiclasses, reforçando que, para aplicações práticas em ambientes IoT, a classificação binária pode ser uma abordagem adequada quando o objetivo é identificar comportamentos anômalos de maneira rápida e confiável.

A análise interpretável via SHAP complementou a avaliação quantitativa, evidenciando que atributos como número de pacotes, flags PSH/FIN/SYN, intervalo de chegada entre pacotes e taxa de transmissão desempenham papéis centrais na diferenciação entre classes. Observou-se que o modelo aprendeu padrões coerentes com a semântica dos tipos de ataque presentes no CICIoT2023, validando não apenas sua acurácia, mas também a plausibilidade de suas decisões. Além disso, também foi possível elucidar as causas da principal limitação observada na classificação multiclasses — a sobreposição de características entre DDoS e DoS.

Em síntese, os resultados destacam tanto a eficácia quanto os desafios da utilização de aprendizado de máquina para detecção de ameaças, ao mesmo tempo em que demonstram o valor da explicabilidade para compreender e aprimorar soluções baseadas em IA em cibersegurança. Tais sistemas podem alcançar excelente desempenho para detecção binária de ameaças em IoT. Entretanto, a classificação multiclasses ainda demanda aprimoramentos, seja por meio de técnicas de balanceamento ou engenharia de features mais especializadas. Assim, o sistema proposto é capaz de conciliar alto desempenho, preservação de dados locais e confiabilidade nas decisões.

Como trabalhos futuros, propõe-se expandir a avaliação para outros modelos de aprendizado de máquina, incluindo arquiteturas de Deep Learning capazes de capturar padrões mais complexos do tráfego. Além disso, recomenda-se testar a abordagem em diferentes datasets de ameaças cibernéticas, que apresentem conjuntos de features distintos e cenários mais diversificados. A aplicação de métodos de XAI nesses novos contextos permitirá comparar modelos e bases de dados não apenas do ponto de vista quantitativo, mas também qualitativo, oferecendo uma compreensão mais profunda sobre como cada modelo distingue as classes de tráfego e revelando quais atributos são mais relevantes para a caracterizar diferentes tipos de ataques.

## Referências

[1] Samir Lemes. “The Role of Software Engineering in Industry 4.0”. Em: *ResearchGate* (2023). Disponível em ResearchGate – ID de publicação 374532017.

[2] Vitor Fantin. *Cibersegurança em Alta: Tendências e Investimentos até 2028*. URL: https://www.linkedin.com/pulse/ciberseguran%C3%A7a-em-alta-tend%C3%AAncias-e-investimentos-at%C3%A9-ecake/ (acesso em 20/08/2025).

[3] R. Kalakoti, H. Bahsi e S. Nõmm. “Improving IoT Security With Explainable AI: Quantitative Evaluation of Explainability for IoT Botnet Detection”. Em: *IEEE Internet of Things Journal* 11.10 (2024), pp. 18237–18254. DOI: 10.1109/JIOT.2024.3360626.

[4] K. Fatema et al. “Federated XAI IDS: An Explainable and Safeguarding Privacy Approach to Detect Intrusion Combining Federated Learning and SHAP”. Em: *Future Internet* 17.6 (2025), p. 234. DOI: 10.3390/fi17060234.

[5] Euclides Carlos Pinto Neto, Sajjad Dadkhah, Raphael Ferreira, Alireza Zohourian, Rongxing Lu e Ali A. Ghorbani. “CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment”. Em: *Sensors* 23.13 (2023), p. 5941. DOI: 10.3390/s23135941. URL: https://www.mdpi.com/1424-8220/23/13/5941.

[6] W. Hsieh, Z. Bi e C. et al. Jiang. “A Comprehensive Guide to Explainable AI: From Classical Models to LLMs”. Em: *arXiv preprint* (2024). arXiv: 2412.00800 [cs.LG].