Resiliência de NIDS Federados em SDN via Atestação Comportamental com Active Semantic Probing

Cassiano Darif Zago[1], Fábio Lúcio L. de Mendonça[1], Roger Immich[2]
Rodolfo I. Meneguette[3], Leandro A. Villas[4], Geraldo Pereira Rocha Filho[5]

[1]Universidade de Brasília (UnB)

[2]Universidade Federal do Rio Grande do Norte (UFRN)

[3]Universidade de São Paulo (USP)

[4]Universidade Estadual de Campinas (UNICAMP)

[5]Universidade Estadual do Sudoeste Bahia (UESB)

cassiano.zago@unb.br, fabio.mendonca@redes.unb.br, roger@imd.ufrn.br

meneguette@icmc.usp.br, lvillas@unicamp.br, geraldo.rocha@uesb.edu.br

Abstract. *In SDN environments, Federated Learning (FL)-based NIDS are compromised by non-IID heterogeneity and malicious clients evading parameter inspection. To address this, we propose Sentinel-Flow, a behavioral attestation framework that replaces parametric validation with an active challenge-response protocol. It integrates three components: (i) an Active Semantic Probing mechanism injecting out-of-band canary flows; (ii) attestation in Trusted Execution Environments (TEEs) to ensure measurement integrity; and (iii) a Trust Score-based governance model to dynamically filter clients before federated aggregation. Experiments demonstrate that while weight scaling attacks collapse the federated model (100% ASR, 53.27% accuracy), Sentinel-Flow successfully reduces ASR to 6.69% and restores overall accuracy to 94.29%.*

Resumo. *Em redes SDN, NIDS baseados em Aprendizado Federado (FL) são vulneráveis à heterogeneidade Não-IID e a clientes maliciosos que evadem a inspeção paramétrica. Para solucionar isso, propomos o Sentinel-Flow, um framework de atestação comportamental que substitui a validação de parâmetros por um protocolo ativo de desafio-resposta. A solução integra três componentes: (i) Active Semantic Probing com injeção out-of-band de canary flows; (ii) atestação em Ambientes de Execução Confiáveis (TEE) para garantir a integridade das medições; e (iii) um modelo de governança baseado em Trust Score para filtrar clientes antes da agregação federada. Resultados experimentais mostram que, enquanto ataques de escalonamento de pesos colapsam o modelo (ASR de 100% e acurácia de 53,27%), o Sentinel-Flow reduz o ASR para 6,69% e restaura a acurácia global para 94,29%.*

1. Introdução

A adoção de arquiteturas distribuídas (5G/6G e IoT) tornou os Sistemas de Detecção de Intrusão em Redes (*Network Intrusion Detection Systems* – NIDS) essenciais [de Oliveira et al. 2023]. Contudo, restrições de escalabilidade e processamento

na borda limitam abordagens centralizadas, impulsionando o aprendizado distribuído [Stallings and Brown 2012].

O Aprendizado Federado (*Federated Learning* – FL) permite o treinamento colaborativo preservando a privacidade e reduzindo custos. Em NIDS, o FL gera modelos globais capturando padrões locais [Stallings and Brown 2012, de Oliveira et al. 2023, Antonesi et al. 2025]. Entretanto, a abordagem introduz vulnerabilidades, como o comprometimento de clientes e a manipulação da agregação [Ferrag et al. 2025, Wang et al. 2023, Zhang et al. 2025].

Entre essas ameaças, destacam-se os ataques de envenenamento. O escalonamento de pesos (*weight scaling*) é especialmente crítico, pois amplifica gradientes maliciosos para dominar a agregação, camuflando-se na variabilidade natural de cenários Não-IID [Cinà et al. 2023, Enneifer et al. 2025, Arimanda et al. 2025]. Em NIDS federados, a heterogeneidade do tráfego dificulta ainda mais a distinção entre comportamentos legítimos e manipulações [Barbetta et al. 2010, Arimanda et al. 2025].

Atualmente, as defesas em FL focam em inspeções paramétricas e estatísticas no servidor [Yinusa and Faezipour 2025, Singh 2025, Kabir et al. 2024]. Em cenários NIDS reais, essa validação centralizada frequentemente penaliza nós honestos (devido ao Não-IID) e falha contra ameaças furtivas que não distorcem os pesos de forma evidente [Kabir et al. 2024, Kasyap and Tripathy 2024, Zhang et al. 2025]. Logo, é vital verificar a integridade funcional dos modelos sem depender exclusivamente da inspeção de parâmetros.

Para sanar essas limitações, propõe-se o *Sentinel-Flow*, que substitui a validação paramétrica passiva por uma atestação comportamental ativa. O *framework* orquestra um protocolo de desafio-resposta via SDN, injetando fluxos de teste (*canary flows*) em Ambientes de Execução Confiáveis (TEE) na borda. A aceitação das atualizações passa a depender do desempenho do modelo sob esses estímulos controlados [Kasyap and Tripathy 2024, Luo et al. 2025]. As principais contribuições deste trabalho são:

- uma análise experimental da vulnerabilidade de agregações clássicas (como o FedAvg) a ataques de escalonamento de pesos em cenários Não-IID;
- a proposta do *Sentinel-Flow*, mecanismo de atestação comportamental integrado ao plano de controle SDN;
- a formalização de um modelo de confiança dinâmico para seleção de clientes baseado em evidências de tempo de execução;
- uma avaliação empírica que comprova a redução da taxa de sucesso do ataque e a recuperação da acurácia global.

O restante do artigo está organizado da seguinte forma: A Seção 2 apresenta os trabalhos relacionados. A Seção 3 apresenta a solução proposta. A Seção 4 apresenta a metodologia experimental. Os resultados são analisados na Seção 5. Por fim, a Seção 6 apresenta as conclusões e os trabalhos futuros.

2. Trabalhos Relacionados

Nesta seção são apresentados e discutidos os principais trabalhos relacionados à segurança em FL aplicado a NIDS, com foco em ataques de envenenamento e mecanismos de de-

fesa. A análise busca identificar as estratégias adotadas na literatura, incluindo abordagens baseadas em agregação robusta, validação no lado do servidor e detecção semântica, bem como os modelos de avaliação e métricas utilizadas para mensurar a eficácia dessas soluções. A Tabela 1 apresenta essas contribuições, organizando os trabalhos de acordo com suas principais características e limitações.

Tabela 1. Comparativo de Trabalhos Relacionados e Lacunas Identificadas

Ref.	**Contexto/Foco**	**Contribuição Principal**	**Limitação / Relação com este Trabalho**
[Ferrag et al. 2025]	IA em Cibersegurança	Vulnerabilidades sistêmicas em modelos distribuídos.	Não propõe mecanismos de governança ativa via SDN.
[Antonesi et al. 2025]	Modelagem de Sequências	*Baseline* de alta capacidade na borda (Transformers).	Foca em capacidade paramétrica, negligenciando atestação comportamental.
[Lazzaro et al. 2025]	Defesas Clean-Label	Ataques persistentes sem troca explícita de rótulo.	Demonstra a falha inerente da detecção *White-Box* (inspeção de pesos).
[Kabir et al. 2024]	Defesa FL (Validação)	*Server-Side Validation* para filtrar clientes maliciosos.	Rejeita falsamente nós honestos em tráfego Não-IID; validação estática fora da borda.
[Enneifer et al. 2025]	Moderação de Borda	Envenenamento empírico "furtivo".	Reforça a ineficácia de filtros estatísticos em topologias descentralizadas.
[Kumar et al. 2025]	Ataques de Orçamento	Manipulação adversarial em $< 1\%$.	Suporta a vulnerabilidade com *Weight Scaling*.
[Wang et al. 2023]	Aprendizado Federado	Ataques adaptativos bizantinos em malhas dinâmicas.	Evidencia a necessidade de respostas dinâmicas baseadas em Desafio-Resposta.
[Yinusa and Faezipour 2025]	CNNs (Edge)	Vulnerabilidade a manipulações locais no plano de dados.	Justifica a necessidade de isolamento da atestação via hardware (TEE).
[Singh 2025]	FL Seguro	Autoencoders + LSTM para varredura.	Valida a arquitetura NIDS local, sem validação *Out-of-band*.
[Zhang et al. 2025]	Amostras Incertas em FL	Ataque de evasão via pontos de fronteira latentes.	Destaca as falhas de agregações puramente geométricas (Krum, Trimmed Mean).
[Kasyap and Tripathy 2024]	Computação na Borda	Envenenamento tratado como anomalia semântica (OOD).	Apoia a necessidade do uso de *Canary Flows* ancorados na realidade.
[Arimanda et al. 2025]	FL Intrusion Detection	Queda de desempenho sob regime de tráfego Non-IID.	Prova que o ruído natural Não-IID camufla a inserção do *backdoor* NIDS.
[Luo et al. 2025]	Remoção de Backdoor	Feedback guiado diretamente na inferência.	Base teórica do *probing*, apenas em *Server-Side*.
Solução proposta	**NIDS Federado em SDN**	***Behavioral Attestation* (*Black-Box*)**	**Desafio-Resposta dinâmico *Out-of-band*, ancorado em TEE na borda.**

A adoção de técnicas de Aprendizado de Máquina em redes críticas tem ampliado a superfície de ataque desses sistemas. al. [Ferrag et al. 2025] destacam que modelos distribuídos e IA generativa introduzem vulnerabilidades intrínsecas, nas quais manipulações locais podem comprometer o comportamento global de sistemas como NIDS. Nesse contexto, arquiteturas profundas têm sido utilizadas na borda para melhorar a capacidade de reconhecimento de padrões [Antonesi et al. 2025]. No entanto, permanece em aberto se essa capacidade paramétrica contribui para a robustez ou, ao contrário, facilita a incorporação de padrões maliciosos durante o treinamento.

Ataques recentes têm se tornado cada vez mais sofisticados e furtivos. Enneifer et al. [Enneifer et al. 2025] demonstram que estratégias modernas evitam gerar anomalias evidentes em métricas tradicionais, enquanto Lazzaro et al. [Lazzaro et al. 2025] evidenciam a eficácia de ataques *clean-label*. Em cenários de FL, adversários bizantinos e adaptativos exploram a heterogeneidade dos dados para mascarar atualizações maliciosas [Wang et al. 2023, Zhang et al. 2025]. Nesse sentido, Kasyap e Tripathy [Kasyap and Tripathy 2024] argumentam que abordagens baseadas apenas em propriedades geométricas são insuficientes, sendo necessário considerar aspectos semânticos, como a detecção de amostras fora da distribuição.

No campo das defesas, diferentes estratégias têm sido propostas para mitigar ataques de envenenamento em FL. Métodos baseados em sanitização estatística e agregação robusta buscam reduzir o impacto de contribuições adversariais por meio da análise das atualizações locais, utilizando métricas como norma, distância ou similaridade entre gradientes [Yinusa and Faezipour 2025, Singh 2025]. Essas abordagens incluem técnicas de filtragem, normalização e ponderação adaptativa, visando atenuar a influência de clientes potencialmente maliciosos.

Alternativamente, abordagens mais recentes adotam mecanismos de validação no lado do servidor (*server-side validation*), como o FLShield [Kabir et al. 2024], nos quais o agregador avalia as atualizações locais com base no desempenho de modelos candidatos em um *dataset* de validação centralizado. Essa estratégia introduz um critério adicional baseado no comportamento do modelo, indo além da inspeção puramente paramétrica. Contudo, tais métodos dependem fortemente da representatividade do *dataset* global, o que compromete sua eficácia em cenários com alta heterogeneidade de dados.

Entretanto, em ambientes NIDS, o tráfego de rede apresenta natureza intrinsecamente Não-IID, o que limita severamente a aplicabilidade de abordagens estatísticas e geométricas. Algoritmos de agregação bizantino-robustos tradicionais, como Krum, Median ou Trimmed Mean, baseiam-se em métricas de distância euclidiana para descartar atualizações anômalas. Em cenários Não-IID, características locais legítimas são frequentemente interpretadas como desvios estatísticos, resultando na rejeição de nós honestos (falsos positivos) e no descarte de assinaturas de ataque cruciais capturadas na borda. Além disso, métodos baseados em inspeção de parâmetros (*white-box*) não capturam comportamentos adversariais latentes, como *backdoors* ativados apenas em condições específicas.

Para contornar essas limitações de agregação, o *Sentinel-Flow* propõe um mecanismo de atestação comportamental *black-box* realizado na borda, ancorado em Ambientes de Execução Confiáveis (TEE, como Intel SGX ou ARM TrustZone). Diferentemente das abordagens existentes, a proposta elimina a dependência de distribuições estatísticas globais e avalia diretamente o comportamento dos modelos sob estímulos controlados, respeitando a variabilidade inerente a cenários Não-IID.

3. Solução Proposta

Esta seção apresenta o *Sentinel-Flow*, um solução de atestação comportamental para mitigar ataques de envenenamento em sistemas de FL aplicados a NIDS em cenários Não-IID. A proposta parte da limitação fundamental das abordagens tradicionais, nas quais a validação das contribuições locais é realizada por meio da inspeção de parâmetros no ser-

vidor agregador. Em ambientes heterogêneos, essa estratégia torna-se ineficaz, uma vez que atualizações maliciosas podem se camuflar dentro da variabilidade estatística natural. Diante disso, o *Sentinel-Flow* substitui a validação passiva baseada em parâmetros por um mecanismo ativo de atestação comportamental. Nesse modelo, a confiabilidade dos clientes é inferida a partir do comportamento do modelo durante a inferência, e não da análise direta de seus gradientes. A validação é deslocada para a borda e realizada por meio de um protocolo de desafio-resposta, orquestrado por um controlador SDN.

3.1. Visão Geral da Arquitetura

A Figura 1 apresenta a arquitetura do *Sentinel-Flow* e o fluxo de interação entre seus principais componentes. O controlador SDN atua como elemento central de orquestração, sendo responsável por conduzir o processo de validação comportamental de forma *out-of-band*, independente do ciclo de treinamento federado.

Inicialmente, o controlador injeta fluxos de teste (*canary flows*) diretamente nos Ambientes de Execução Confiáveis (TEE) dos nós de borda. Esses fluxos representam padrões conhecidos de tráfego malicioso e são processados de forma isolada, garantindo integridade e evitando interferência do sistema operacional do cliente. Em seguida, o comportamento do modelo local é observado a partir de sua capacidade de detectar corretamente esses fluxos. O resultado desse processo é convertido em um índice de confiança (*Trust Score*), que reflete a aderência do modelo ao comportamento esperado de um detector de intrusão. Por fim, durante a etapa de agregação, apenas os clientes que atingem o limiar de confiança têm suas atualizações (Δw_i) incorporadas ao modelo global. Clientes com baixo desempenho são automaticamente rejeitados pelo controlador SDN, impedindo que contribuições adversariais influenciem o processo de agregação.

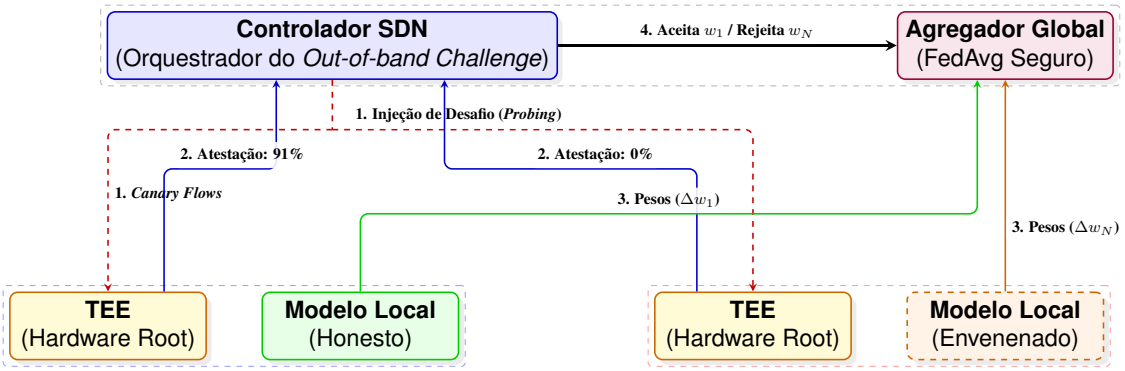


Figura 1. Arquitetura do *Sentinel-Flow*. O controlador SDN orquestra um protocolo de desafio-resposta , injetando *canary flows* nos TEEs da borda para avaliar o comportamento dos modelos locais. A agregação federada torna-se condicional ao *Trust Score*, permitindo filtrar clientes comprometidos antes da incorporação de suas atualizações.

3.2. Formalização do Modelo de Ameaça

Considere um sistema de FL composto por N clientes, cujo objetivo é minimizar a função de perda global $F(w) = \sum_{i=1}^{N} p_i F_i(w)$, onde $F_i(w)$ representa a função de perda local do cliente i e p_i sua contribuição relativa, tipicamente proporcional ao tamanho do conjunto de dados local.

Neste trabalho, assume-se a existência de um subconjunto de clientes comprometidos $\mathcal{A} \subseteq \{1, \ldots, N\}$, capazes de manipular tanto seus dados locais quanto suas atualizações de modelo. Esses ataques incluem a inserção de amostras envenenadas x_{poison} (*data poisoning*) e a modificação das atualizações locais Δw_i. Em particular, ataques de *weight scaling* consistem na amplificação artificial das atualizações locais, podendo ser modelados como:

$$\tilde{\Delta w_i} = \alpha \cdot \Delta w_i, \quad \alpha > 1, \ i \in \mathcal{A}, \tag{1}$$

o que permite que clientes maliciosos influenciem desproporcionalmente o processo de agregação global.

Em ambientes Não-IID, nos quais as distribuições locais $P_i(X, Y)$ diferem significativamente entre os clientes, as atualizações Δw_i apresentam variabilidade intrínseca. Nesse contexto, as atualizações adversariais $\tilde{\Delta w_i}$ tornam-se estatisticamente indistinguíveis das contribuições legítimas, dificultando sua detecção por métodos baseados em inspeção de parâmetros (*white-box*). Como consequência, algoritmos de agregação baseados em média, como o FedAvg, podem incorporar contribuições adversariais sem evidências explícitas na distribuição dos pesos, resultando na degradação do modelo global ou na inserção silenciosa de *backdoors*.

3.3. Atestação Comportamental

Para superar as limitações das abordagens baseadas em inspeção paramétrica, o *Sentinel-Flow* adota um mecanismo de atestação comportamental, no qual a confiabilidade dos clientes é inferida a partir do seu comportamento em tempo de execução, em vez da análise direta de seus gradientes. Essa estratégia é particularmente adequada para cenários Não-IID, nos quais variações legítimas dificultam a identificação de atualizações maliciosas.

A proposta é implementada por meio de um protocolo de desafio-resposta (*challenge-response*) orquestrado por um controlador SDN, operando de forma *out-of-band* em relação ao processo de treinamento federado. O mecanismo consiste em três etapas principais:

1. O controlador SDN seleciona um conjunto de fluxos de teste (*canary flows*) representativos de padrões conhecidos de tráfego malicioso;
2. Esses fluxos são injetados diretamente em Ambientes de Execução Confiáveis (TEE) nos nós de borda, de forma transparente ao sistema operacional e indistinguível do tráfego legítimo;
3. O comportamento do modelo é avaliado com base na sua capacidade de identificar corretamente os fluxos de teste, sendo essa resposta utilizada para inferir sua confiabilidade e condicionar sua participação na agregação federada.

Diferentemente das abordagens tradicionais, nas quais a validação ocorre no servidor de agregação, o *Sentinel-Flow* desloca esse processo para a borda e adota uma perspectiva *black-box*, considerando apenas a saída do modelo diante de estímulos controlados. Dessa forma, elimina-se a dependência de suposições sobre a distribuição dos parâmetros e reduz-se a vulnerabilidade a ataques que exploram a variabilidade Não-IID.

3.4. Modelo de Confiança e Agregação Segura

A confiabilidade de cada cliente i na rodada t é quantificada por meio de um índice dinâmico, denominado *Trust Score*, definido como:

$$Trust(i)_t = \frac{\sum_{k=1}^{K} \mathbb{I}\big(M_i(\tilde{x}_{atk,k}) = \text{Ataque}\big)}{K}, \tag{2}$$

onde M_i representa o modelo local do cliente i, $\tilde{x}_{atk,k}$ corresponde às K amostras canário injetadas no TEE, e $\mathbb{I}(\cdot)$ é a função indicadora, que assume valor 1 para predições corretas e 0 caso contrário.

Com base nesse índice, define-se o subconjunto de clientes confiáveis:

$$S_t = \{i \in \{1, \ldots, N\} \mid Trust(i)_t \geq \tau_{safe}\}, \tag{3}$$

onde τ_{safe} é um limiar de confiança que controla a inclusão de clientes na etapa de agregação. O limiar τ_{safe} foi definido empiricamente em 60% ($\tau_{safe} = 0.6$) como um ponto de equilíbrio para cenários Não-IID. Um limiar excessivamente rigoroso (ex: 90%) poderia rejeitar nós honestos cujos modelos locais ainda estão convergindo ou que possuem baixa representatividade em suas amostras de treinamento originais. Por outro lado, o valor de 60% garante que o modelo demonstre uma capacidade preditiva mínima e intencional contra intrusões, filtrando efetivamente adversários que aplicam *clean-label poisoning* severo.

A atualização do modelo global é então realizada exclusivamente sobre esse subconjunto, por meio de uma agregação ponderada pelos tamanhos dos conjuntos de dados locais:

$$w_{t+1} = \sum_{i \in S_t} \frac{n_i}{\sum_{j \in S_t} n_j} w_i^{t+1}. \tag{4}$$

Esse mecanismo implementa um filtro comportamental pré-agregação, garantindo que apenas clientes que demonstram desempenho consistente na detecção dos fluxos canários contribuam para o modelo global, reduzindo a influência de atualizações adversariais.

3.5. Mecanismo de Atestação (Sentinel-Flow)

O processo de atestação comportamental no *Sentinel-Flow* é executado a cada rodada de treinamento federado e tem como objetivo avaliar a confiabilidade dos clientes antes da etapa de agregação. Diferentemente de abordagens tradicionais, que operam sobre as atualizações de modelo no servidor, o mecanismo proposto utiliza um protocolo de desafio-resposta para verificar o comportamento dos modelos em tempo de execução.

O Algoritmo 1 descreve o procedimento de auditoria orquestrado pelo controlador SDN. Para cada cliente participante, o sistema injeta um conjunto de fluxos de teste (*canary flows*) no Ambiente de Execução Confiável (TEE) e observa a resposta do modelo. A partir dessas respostas, é calculado um índice de confiança (*Trust Score*), utilizado para decidir sobre a participação do cliente na agregação. Clientes que não atingem o limiar mínimo de confiança τ_{safe} são considerados potencialmente comprometidos e têm suas atualizações descartadas. A agregação global é então realizada exclusivamente sobre o subconjunto de clientes aprovados, reduzindo a influência de contribuições adversariais.

Caso nenhum cliente satisfaça o critério de confiança, o sistema realiza um *rollback* preventivo, preservando o modelo da rodada anterior.

Algoritmo 1: Protocolo de Atestação Comportamental Sentinel-Flow

Input: Modelos Locais $w_1^{t+1} \dots w_N^{t+1}$, Fluxos Canário C_{test}, Limiar τ_{safe}
Output: Modelo Global w_{t+1} Atualizado e Limpo

$S_t \leftarrow \emptyset$; `// Conjunto de nós com Atestação Positiva`
`// Auditoria SDN Out-of-Band (Black-Box Probing)`
foreach *nó* $i \in 1 \dots N$ **do**
  $SDN.injetar\_desafio(C_{test}, TEE\_porta\_i)$;
  $Alertas \leftarrow Ler\_Resposta\_TEE(i)$;
  $Trust(i) \leftarrow Alertas/|C_{test}|$;
  if $Trust(i) \geq \tau_{safe}$ **then**
    $S_t \leftarrow S_t \cup \{i\}$; `// Nó aprovado comportamentalmente`
  end
  else
    Log: *Bloqueio de nó comprometido (*i*)*;
  end
end
`// Agregação Segura e Re-normalizada`
if $|S_t| > 0$ **then**
  $w_{t+1} \leftarrow FedAvg(w_{i \in S_t}^{t+1})$;
end
else
  $w_{t+1} \leftarrow w_t$; `// Rollback preventivo se a rede colapsar`
end
return w_{t+1};

Cabe ressaltar que o processo de *probing* ocorre a cada rodada de treinamento federado, porém introduz um *overhead* computacional marginal. Como a validação baseia-se exclusivamente na inferência de um pequeno lote de dados (ex: 100 amostras) dentro do TEE, a latência adicionada é na ordem de milissegundos. Esse custo é ínfimo se comparado ao tempo de comunicação da rede e às épocas de treinamento retropropagado local, garantindo a viabilidade da orquestração SDN mesmo em redes de alta velocidade. A eficácia estrutural do *Sentinel-Flow* distancia-se das abordagens tradicionais pela natureza de seus estímulos e do isolamento em hardware. Em vez de depender de validações estatísticas que penalizam nós operando em ambientes Não-IID (falsos positivos), o conjunto canário (C_{test}) é composto por um subconjunto determinístico de fluxos adversariais empíricos (*ground-truth*) de tráfego intrusivo universal.

A injeção destas sondas através do TEE garante que o sistema operacional do nó de borda não diferencie a sonda de uma requisição de rede padrão. Se o modelo classificar este tráfego estritamente malicioso como benigno, atesta-se inequivocamente a presença de um *backdoor* semântico. Este método elimina a alta taxa de rejeição característica de *Server-Side Validations*, assegurando atestação determinística frente a adversários sofisticados.

4. Metodologia Experimental

Com o objetivo de avaliar a vulnerabilidade estrutural de sistemas de FL em cenários Não-IID, bem como validar a eficácia do *Sentinel-Flow*, foi desenvolvido um protocolo experimental controlado baseado em simulação, como apresentada na Figura 2. A metodologia foi concebida para reproduzir, de forma progressiva, três estados distintos do sistema: (i) operação nominal, (ii) degradação sob ataque adversarial e (iii) recuperação por meio de um mecanismo de atestação comportamental.

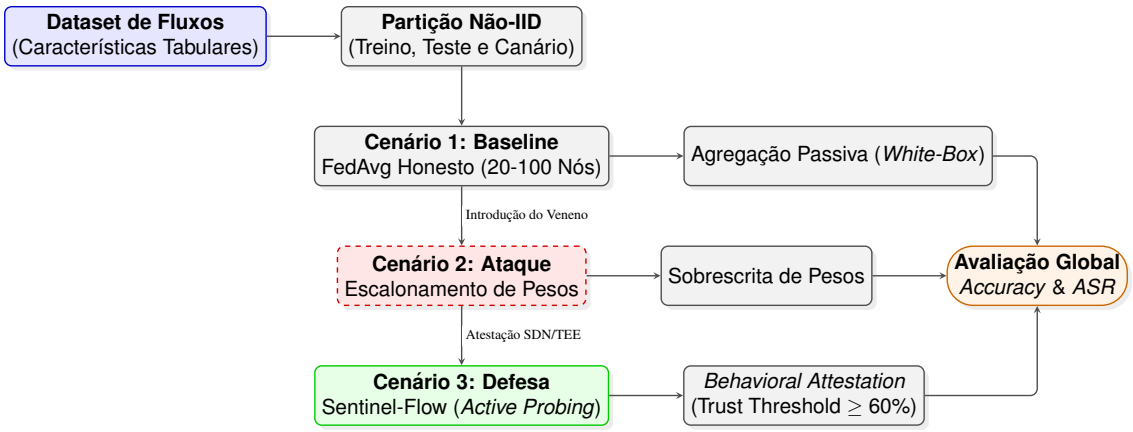


Figura 2. Fluxograma da metodologia experimental.

Os experimentos foram conduzidos utilizando dados derivados do conjunto *CIC-IDS2017*, representados por 78 atributos numéricos. A distribuição dos dados seguiu os princípios do FL em ambientes heterogêneos, nos quais 80% das amostras foram destinadas ao treinamento distribuído e 20% à avaliação global. O conjunto de treinamento foi particionado entre os clientes de forma não uniforme, simulando um cenário Não-IID. Adicionalmente, um subconjunto de 100 amostras maliciosas foi isolado para compor o conjunto de desafios semânticos (*canary set*), utilizado exclusivamente no processo de atestação comportamental.

É importante frisar que a eficácia do *probing* depende da *finesse* dos *canary flows*. Para mitigar adversários adaptativos e ataques de duas caras (*two-faced attacks*) onde o nó malicioso identifica tratar-se de um teste e responde corretamente apenas para burlar a auditoria, as sondas devem ser estatisticamente similares ao tráfego de produção. No contexto deste trabalho, as amostras do *canary set* mantêm a mesma dimensionalidade e distribuição de *features* do CIC-IDS2017. Embora o dataset atue como um referencial arquitetural para validar a mecânica do ataque, o *framework* é inerentemente agnóstico aos dados, sendo diretamente aplicável a datasets mais recentes. Em implementações dinâmicas, o controlador SDN pode rotacionar as amostras canário a cada rodada, impedindo o sobreajuste (*overfitting*) adversarial.

A avaliação foi estruturada em três cenários consecutivos. No primeiro cenário (*baseline*), estabelece-se o limite superior de desempenho do sistema em condições não adversariais, no qual clientes, variando entre 10, 20 e 40 nós de borda, realizam treinamento local sobre seus dados, enquanto o servidor executa a agregação global por meio do algoritmo *FedAvg* (Baseline). No segundo cenário, introduz-se um ataque de envenenamento que combina *clean-label poisoning* com escalonamento de pesos (*weight*

scaling), em que cinco clientes maliciosos manipulam seus dados ao rotular amostras intrusivas como benignas e amplificam suas atualizações para maximizar a influência na agregação federada. No terceiro cenário, aplica-se o mecanismo de atestação comportamental (*Sentinel-Flow*), no qual, antes da agregação, um controlador SDN injeta fluxos de teste (*canary flows*) nos Ambientes de Execução Confiáveis (TEE) de cada cliente, condicionando a participação ao desempenho do modelo na identificação dessas amostras.

A avaliação do sistema baseia-se em duas métricas complementares. A primeira é a acurácia global, definida como a proporção de classificações corretas sobre o conjunto de teste legítimo, refletindo a capacidade geral de detecção do modelo. A segunda é a *Attack Success Rate* (ASR), que corresponde ao percentual de amostras maliciosas classificadas incorretamente como benignas, representando a taxa de sucesso do atacante na inserção de *backdoors*. A Tabela 2 apresenta o conjunto de parâmetros que foi utilizados para gerar os resultados.

Tabela 2. Configuração dos Experimentos

Categoria	Configuração
Distribuição dos Dados	Não-IID entre clientes
Canary Set	100 amostras maliciosas
Número de Clientes	10, 20 e 40
Clientes Maliciosos	5 (fixo)
Modelo Local	MLP: Linear(78,64) $\rightarrow$ ReLU $\rightarrow$ Linear(64,1) $\rightarrow$ Sigmoid
Rodadas Federadas	200
Tipo de Ataque	Clean-label poisoning + Weight Scaling
Estratégia do Ataque	Inversão de rótulos + amplificação de gradientes
Defesa	Sentinel-Flow (Behavioral Attestation)
Validação	Canary Flows via SDN + TEE
Trust Threshold	$\geq 60\%$

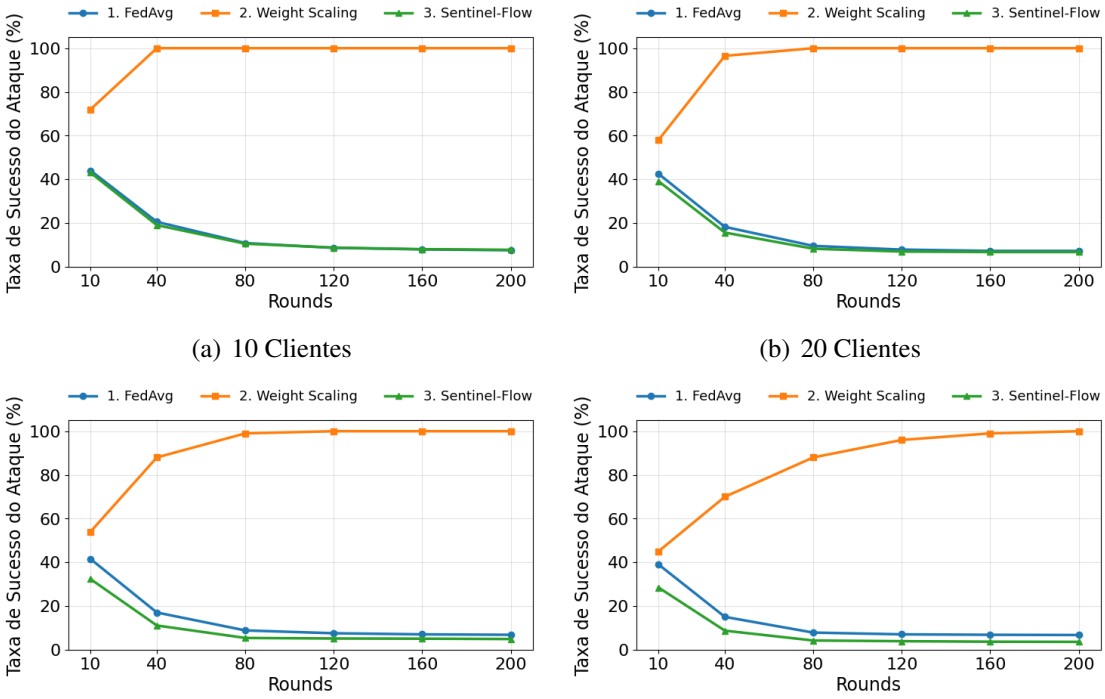
5. Resultados

A Figura 3 apresenta a evolução da Taxa de Sucesso do Ataque (ASR) ao longo de 200 rodadas de treinamento federado, considerando diferentes tamanhos de federação. Observa-se que o *Sentinel-Flow* apresenta convergência estável em todos os cenários, com redução consistente do ASR para valores próximos de 5%, mesmo sob partição Não-IID. Na Figura 3a, com 10 clientes, a presença do ataque de *weight scaling* resulta em degradação do sistema, elevando o ASR de 72,0% para 100,0% já nas primeiras rodadas e mantendo esse patamar até o final do teste, caracterizando o colapso completo da agregação global. Esse efeito é intensificado pela elevada proporção de clientes maliciosos, que representam metade da federação. Em contraste, o *Sentinel-Flow* mantém o ASR próximo ao *baseline* ao longo de todo o processo, evidenciando sua capacidade de filtrar contribuições adversariais mesmo em cenários altamente desfavoráveis.

Na Figura 3b, com 20 clientes, observa-se comportamento semelhante, porém com progressão ligeiramente menos abrupta do ataque. O ASR cresce rapidamente de 58,0% para 100,0%, atingindo o regime de colapso em poucas dezenas de rodadas. A redução da fração relativa de clientes maliciosos contribui para esse atraso inicial, embora não seja

suficiente para evitar o comprometimento do modelo global. Por outro lado, o *Sentinel-Flow* mantém desempenho consistente, com valores de ASR alinhados ao baseline

Nas Figuras 3c e d), com 30 e 40 clientes, observa-se uma tendência clara de amortecimento da progressão do ataque. Embora o *weight scaling* ainda seja capaz de levar o ASR a 100,0%, o crescimento ocorre de forma mais gradual, exigindo maior número de rodadas para comprometer a agregação global. Esse comportamento decorre da menor influência relativa dos clientes maliciosos à medida que o tamanho da federação aumenta. Em contrapartida, o *Sentinel-Flow* não apenas mantém o ASR próximo ao baseline, como, em alguns casos, apresenta valores ainda inferiores, indicando maior estabilidade do modelo sob defesa. Essa redução é de aproximadamente 20% quando comparado com o baseline. Esses resultados evidenciam que a abordagem proposta é eficaz de forma consistente em diferentes escalas, preservando a integridade do teste federado mesmo diante de ataques adversariais persistentes.



(a) 10 Clientes (b) 20 Clientes

(c) 30 Clientes (d) 40 Clientes

Figura 3. Evolução da Taxa de Sucesso do Ataque (ASR) ao longo de 200 rodadas de treinamento federado para diferentes tamanhos de federação.

A Figura 4 apresenta a comparação entre os cenários de *baseline*, ataque de escalonamento e aplicação do *Sentinel-Flow*, considerando simultaneamente a acurácia global e a Taxa de Sucesso do Ataque (ASR). Em regime benigno, o modelo federado atinge acurácia de 86,6% e ASR de 7,2%, indicando desempenho consistente na detecção de intrusões. Entretanto, sob o ataque de *weight scaling*, observa-se degradação significativa do sistema, com a acurácia reduzida para 53,27% e o ASR elevado a 100%, caracterizando o colapso completo da capacidade de detecção do modelo global. Em contraste, a aplicação do *Sentinel-Flow* não apenas restaura o desempenho do sistema, como o supera, alcançando acurácia de 94,29% e reduzindo o ASR para 6,69%. Esse resultado evidencia

a eficácia do mecanismo de atestação comportamental em mitigar contribuições adversariais, preservando a integridade do aprendizado federado mesmo na presença de ataques severos.

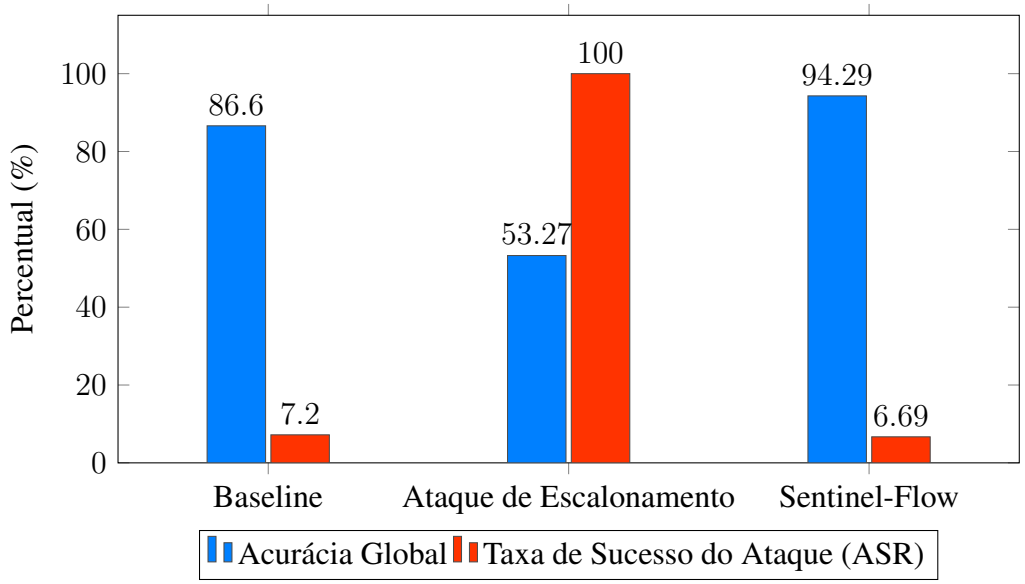


Figura 4. Análise do colapso preditivo induzido pelo ataque de *weight scaling*.

6. Conclusão e Trabalhos Futuros

Este trabalho propôs o *Sentinel-Flow*, um framework de atestação comportamental para aprendizado federado aplicado à detecção de intrusões em ambientes Não-IID, fundamentado na validação dos modelos locais por meio de sinais observáveis de inferência. A abordagem combina *active probing* via SDN com execução em ambientes TEE, permitindo avaliar a confiabilidade dos clientes a partir de sua resposta a fluxos de ataque conhecidos, previamente isolados do processo de treinamento. Em cenários experimentais controlados, os resultados demonstraram que o ataque de *weight scaling* é capaz de comprometer completamente a agregação global, elevando o ASR a 100% e reduzindo significativamente a acurácia do modelo. Em contraste, o *Sentinel-Flow* foi capaz de restaurar o desempenho do sistema, reduzindo o ASR para níveis próximos ao *baseline* e, em alguns casos, superando a acurácia observada em regime benigno, evidenciando sua eficácia na mitigação de contribuições adversariais.

Os resultados também evidenciaram que a análise baseada exclusivamente em parâmetros apresenta limitações intrínsecas em cenários heterogêneos, nos quais variações legítimas decorrentes da distribuição Não-IID podem se sobrepor a padrões adversariais. Nesse contexto, a validação comportamental proposta mostrou-se não redundante em relação às abordagens tradicionais, uma vez que captura diretamente o impacto funcional das atualizações locais sobre a capacidade de detecção do modelo. A consistência observada ao longo de diferentes tamanhos de federação indica que, embora o aumento do número de clientes reduza a influência relativa dos participantes maliciosos, a vulnerabilidade estrutural do aprendizado federado persiste na ausência de mecanismos ativos de validação. Como limitação, o modelo de atestação considera um conjunto fixo

de fluxos canários, o que pode reduzir sua capacidade de generalização frente a ataques altamente adaptativos ou distribuídos.

Como trabalhos futuros, pretende-se ampliar o mecanismo de atestação com estratégias adaptativas de geração de *canary flows*, rotacionando as amostras dinamicamente para neutralizar adversários cientes do protocolo. Além disso, será conduzida uma análise rigorosa de sensibilidade do *overhead* e do custo computacional do TEE (latência, tempo de processamento e impacto na rede) em topologias reais de IoT e dispositivos de borda com restrição de hardware. Será investigada também a integração com técnicas complementares de robustez, buscando uma abordagem híbrida que avalie o desempenho do Sentinel-Flow em comparação com métodos bizantinos estabelecidos (como Krum e Median) em dados com extrema divergência.

Referências

Antonesi, G., Cioara, T., Anghel, I., Michalakopoulos, V., Sarmas, E., and Toderean, L. (2025). A systematic review of transformers and large language models in the energy sector: towards agentic digital twins. *Applied Energy*, 401:126670.

Arimanda, N., Radhakrishnan, R. V., and Padmavathi, U. (2025). Fl-ids++: A dynamic federated learning framework for intrusion detection with personalized non-iid data, adversarial resilience and energy-efficient lightweight models. *Future Generation Computer Systems*, 177:108234.

Barbetta, P. A., Bornia, A. C., and Reis, M. M. (2010). *Estatística para Cursos de Engenharia e Informática*. Atlas, São Paulo, 3ª edição.

Cinà, A. E., Grosse, K., Demontis, A., Vascon, S., Zellinger, W., Moser, B. A., Oprea, A., Biggio, B., Pelillo, M., and Roli, F. (2023). Wild patterns reloaded: A survey of machine learning security against training data poisoning. *ACM Computing Surveys*, 55:294–333.

de Oliveira, J. A., Gonçalves, V. P., Meneguette, R. I., de Sousa Jr, R. T., Guidoni, D. L., Oliveira, J. C., and Rocha Filho, G. P. (2023). F-nids—a network intrusion detection system based on federated learning. *Computer Networks*, 236:110010.

Enneifer, S., Baccini, F., Siciliano, F., Amerini, I., and Silvestri, F. (2025). The perils of stealthy data poisoning attacks in misogynistic content moderation. *Online Social Networks and Media*, 50:100334.

Ferrag, M. A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., Tihanyi, N., Bisztray, T., and Debbah, M. (2025). Generative ai in cybersecurity: A comprehensive review of llm applications and vulnerabilities. *Internet of Things and Cyber-Physical Systems*, 5:1–46.

Kabir, E., Song, Z., Rashid, R. U., and Mehnaz, S. (2024). Flshield: A validation based federated learning framework to defend against poisoning attacks. *2024 IEEE Symposium on Security and Privacy*, 1:2572–2590.

Kasyap, H. and Tripathy, S. (2024). Beyond data poisoning in federated learning. *Expert Systems With Applications*, 235:121192.

Kumar, K. N., Mohan, C. K., Cenkeramaddi, L. R., and Awasthi, N. (2025). Minimal data poisoning attack in federated learning for medical image classification: An attacker perspective. *Artificial Intelligence In Medicine*, 159:103024.

Lazzaro, D., Mura, R., Cinà, A. E., Laurita, G., Vercelli, G., Oneto, L., Biggio, B., and Roli, F. (2025). Poison once, fool many: Practical poisoning attacks against text-to-image retrieval systems. *Knowledge-Based Systems*, 334:115090.

Luo, T., Peng, H., Fu, A., Yang, W., Pang, L., Al-Sarawi, S. F., Abbott, D., and Gao, Y. (2025). Just a little human intelligence feedback! unsupervised learning assisted supervised learning data poisoning based backdoor removal. *Computer Communications*, 233:108052.

Singh, P. (2025). A secure federated learning framework based on autoencoder and long short-term memory with generalized robust loss function for detection and prevention of data poisoning attacks. *Biomedical Signal Processing and Control*, 102:107320.

Stallings, W. and Brown, L. (2012). *Computer security : principles and practice*. Pearson, Boston.

Wang, S., Li, Q., Cui, Z., Hou, J., and Huang, C. (2023). Bandit-based data poisoning attack against federated learning for autonomous driving models. *Expert Systems With Applications*, 227:120295.

Yinusa, A. and Faezipour, M. (2025). Enhancing the robustness of cnn-based lung cancer detection models against label-flipping poison attacks using defensive distillation. *Array*, 29:100637.

Zhang, H.-R., Wang, K.-X., Liang, X.-Y., and Yu, Y.-F. (2025). Dups: Data poisoning attacks with uncertain sample selection for federated learning. *Computer Networks*, 256:110909.