FedSBS: Seleção de Participantes Baseado em Pontuação para Aprendizado Federado no Cenário de Detecção de Intrusão

Helio N. Cunha Neto, Natalia C. Fernandes, Diogo M. F. Mattos

[1]MídiaCom - PPGEET/TET/UFF
Universidade Federal Fluminense - UFF

Resumo. *Sistemas de Detecção de Intrusão baseados em Aprendizado Federado apresentam desafios para a segurança cibernética, incluindo a gestão de dados desbalanceados e interferência de participantes maliciosos. O Aprendizado Federado é uma abordagem colaborativa de aprendizado de máquina e permite que participantes treinem modelos com seus dados locais, preservando a privacidade. Os modelos locais são agregados em um modelo global. Contudo, participantes maliciosos podem comprometer o modelo global com dados aleatórios ou enviesados. Este artigo propõe o método FedSBS para a seleção de participantes. O FedSBS visa pontuar a contribuição de cada participante e, então, proceder com a seleção de participantes. O método busca minimizar os riscos representados por participantes maliciosas concomitantemente com a otimização do desempenho do modelo global. A proposta demonstra desempenho superior quando comparada a outros métodos de seleção de participantes, atingindo 80% da métrica F1, 90% de acurácia e 69% de precisão no conjunto de testes com presença de participantes maliciosos. FedSBS mantém o desempenho do modelo global mesmo no cenário com 60% de participantes maliciosos.*

Abstract. *Intrusion Detection Systems based on Federated Learning pose challenges for cybersecurity, including managing imbalanced data and interference from malicious participants. Federated Learning is a collaborative approach to machine learning that allows participants to train models with their local data while preserving privacy. A global model aggregates local models. However, malicious participants can compromise the global model with random or biased data. This article proposes the FedSBS method for participant selection. FedSBS aims to score each participant's contribution and then proceed with the selection process. The method seeks to minimize the risks posed by malicious participants while optimizing the performance of the global model. The proposal demonstrates superior performance compared to other participant selection methods, achieving 80% F1-score, 90% accuracy, and 69% precision in the test set with malicious participants. FedSBS maintains the performance of the global model even with up to 60% malicious participants.*

1. Introdução

Os ataques cibernéticos são uma ameaça constante que continua a crescer tanto na frequência quanto na complexidade [Andreoni Lopez et al., 2019], acentuando a necessidade de mecanismos avançados de segurança. Um Sistema de Detecção de Intrusão —

Este trabalho foi realizado com recursos do CNPq, CAPES, FAPERJ, RNP e Prefeitura de Niterói/FEC/UFF (Edital PDPA 2020) .

Intrusion Detection System (IDS) é uma linha de defesa crucial projetada para prevenir ameaças e ataques em sistemas de computadores, redes ou aplicações. O IDS monitora constantemente o tráfego da rede, analisa registros de sistema e examina outras fontes de dados pertinentes. A relevância de um IDS reside em sua abordagem proativa para detecção de ameaças, capacidade de se adaptar a novas ameaças e manutenção da integridade e segurança das infraestruturas digitais. Trabalhos recentes têm utilizado técnicas de Aprendizado de Máquina — *Machine Learning* (ML) para aumentar as capacidades de IDSes, conciliando capacidades de reconhecimento de padrões e previsão. Os algoritmos de ML demonstram potencial na detecção de novos métodos de intrusão, melhorando assim o desempenho e a adaptabilidade do IDS. Contudo, para se obter desempenho efetivo em Sistemas de Detecção de Intrusão baseado em Aprendizado de Máquina — *Machine Learning-based Intrusion Detection System* (ML-IDS), exige-se a coleta e a centralização de dados de rede para treinar os modelos de ML. Esse processo pode apresentar desafios consideráveis, especialmente em meio a regulamentações cada vez mais rigorosas de proteção de privacidade de dados pessoais, como a Lei Geral de Proteção de Dados (LGPD) [Cunha Neto et al., 2021]. Portanto, a importância do ML para o IDS reside na capacidade de permitir o desenvolvimento de uma nova geração de IDS que aprenda com o tráfego de rede de forma incremental. Essa abordagem pode melhorar a eficiência do IDS, mas requer planejamento cuidadoso para prevenir possíveis vazamentos de dados. O Aprendizado Federado — *Federated Learning* (FL) surge como uma estratégia eficaz para treinar coletivamente um modelo de aprendizado de máquina em diversas fontes de dados, preservando a privacidade e a segurança.

O FL é uma abordagem eficaz para melhorar o desempenho de IDSes [Agrawal et al., 2022]. Em um IDS baseado em FL, múltiplos IDSes treinam um modelo local de aprendizado de máquina com seus dados locais. Em seguida, os IDSes enviam os parâmetros do seu modelo para um servidor centralizado, o servidor de agregação. No FL, esses IDSes são os participantes da federação. O servidor de agregação seleciona aleatoriamente um subconjunto de participantes a cada rodada de agregação e agrega seus modelos locais. Uma rodada de agregação refere-se ao processo iterativo de atualização do modelo global com base nas atualizações dos modelos locais dos participantes selecionados. O modelo global atualizado, incorporando as contribuições dos participantes selecionados, é então distribuído aos participantes. O processo iterativo de seleção de participantes e agregação de modelo permite que o modelo global aprenda a partir de diversas fontes de dados, preservando a privacidade de dados locais. O processo de treinamento local permite que o modelo global aprenda de diversas fontes de dados sem a exposição dos dados locais na rede. No entanto, devido à ausência de acesso aos dados, o servidor de agregação não garante que os participantes selecionados são honestos. Participantes maliciosos podem distorcer o processo de treinamento de várias maneiras. Os participantes maliciosos podem fornecer dados falsos ou enviesados, introduzir *backdoors* no modelo ou deliberadamente enviar parâmetros de modelo corrompidos durante a fase de agregação [Cunha Neto et al., 2023]. em um ataque de *backdoor*, o participante malicioso treina seu modelo local para realizar a tarefa de aprendizado pretendida e também para responder de uma maneira específica a determinadas entradas. Isso pode levar a um modelo global impreciso, enviesado ou vulnerável a ataques. Tais ações comprometem a eficácia e a confiabilidade do modelo global, podendo resultar em falhas de classificação e em brechas de segurança

potencialmente graves. Portanto, identificar e mitigar a ação de participantes maliciosos é um desafio significativo para o aprendizado federado [Cunha Neto et al., 2023].

Este artigo propõe o método Seleção Federada Baseada em Pontuação – *Federated Score-Based Selection* (FedSBS) que mantêm o desempenho do aprendizado federado na presença de participantes maliciosos. A solução proposta permite a agregação eficiente dos modelos locais, escolhendo os participantes com base em sua contribuição. A proposta incorpora a abordagem *epsilon greedy* para selecionar os participantes usando uma pontuação baseada na contribuição e um mecanismo para bloquear participantes selecionados repetidas vezes. O método de pontuação de participantes é baseado no ganho de informação e considera ambas as perdas locais e globais. Dessa forma, o método de pontuação mede a contribuição do participante individualmente no grupo de participantes selecionados. Os participantes com alta perda local e baixa perda global são recompensados e, o caso contrário, penalizados. O FedSBS possui um mecanismo de bloqueio de participantes para promover uma distribuição mais equilibrada da seleção e evitar a seleção frequente dos mesmos participantes. O mecanismo acompanha as vezes que cada participante é selecionado e determina a probabilidade de bloquear um participante da seleção. Como resultado, os participantes selecionados várias vezes têm probabilidade cada vez maior de serem bloqueados para uma nova seleção. O método FedSBS foi simulado e se mostrou mais eficaz que outras abordagens de seleção de participantes. O FedSBS alcançou 80% na métrica F1, 90% de acurácia e 69% de precisão no conjunto de testes, mantendo a variação baixa entre as avaliações. O método proposto manteve o desempenho mesmo em cenários com 60% dos participantes maliciosos.

O restante do artigo está organizado da seguinte forma. A Seção 2 fornece um contexto do Aprendizado Federado. A Seção 3 apresenta os trabalhos relacionados. A proposta é descrita na Seção 4. A Seção 5 descreve o modelo do atacante. A Seção 6 descreve a avaliação e resultados. Por fim, a Seção 7 conclui o artigo

2. Fundamentos de Aprendizado Federado

O aprendizado federado é um processo que treina um modelo estatístico global a partir de dados oriundos de dispositivos remotos. Durante esse procedimento, os participantes treinam colaborativamente um modelo global enquanto mantêm os dados em seus próprios dispositivos. O intuito do aprendizado federado é treinar o modelo global a partir do treinamento de modelos locais, utilizando os dados armazenado localmente nos dispositivos. Assim, os dispositivos transmitem as atualizações parciais a um servidor central em cada rodada de agregação. Esse servidor coleta e combina os modelos intermediários para, posteriormente, distribuir o modelo global consolidado a todos os participantes.

Cada participante, representado por $n \in N$, usa seu conjunto de dados D_n para treinar o modelo local $\mathbf{w}_n^t$, enviando apenas os parâmetros do modelo local ao servidor de aprendizado federado em intervalos regulares. Em cada rodada de agregação, um subconjunto $\mathbb{S}_t$, em que $\mathbb{S}_t \in N$, dos participantes é selecionado aleatoriamente para fornecer os parâmetros de seus modelos locais ao servidor. O treinamento com todos os participantes não é escalável, dado que o número de participantes é finito, porém não limitado, tornando o treinamento inviável.

Posteriormente, todos os parâmetros dos modelos locais selecionados são agregados para gerar um modelo global, denominado $\mathbf{w}_G^t$. Os modelos locais são atualizados

internamente por τ atualizações locais, antes do envio dos parâmetros dos modelos locais para o servidor realizar a agregação global [Cunha Neto et al., 2020].

O propósito de cada modelo local é minimizar sua função de perda local $L(w^t)$. Essa função de perda pode variar conforme o problema em questão, por exemplo, o erro médio quadrático (*Mean Squared Error* - MSE) é utilizado para problemas de regressão e a entropia cruzada é usada para problemas de classificação (*Log Loss*). A função de perda global deve contemplar a perda de todos os participantes envolvidos na rodada de agregação [Cunha Neto et al., 2020].

O método mais comum para agregação dos modelos locais é a média federada (*Federated Averaging* - FedAvg) [Brendan McMahan et al., 2017], um método de treinamento baseado no Gradiente Descendente Estocástico – *Stochastic Gradient Descent* (SGD). O FedAvg seleciona um subconjunto aleatório de participantes $\mathbb{S}_t \in N$ que realizam atualizações locais em seus modelos usando seus dados locais. Os parâmetros dos modelos locais são enviados ao servidor de agregação, que realiza a combinação dos parâmetros por meio de uma média ponderada. A média é ponderada pela relação entre o tamanho do conjunto de dados do participante e o tamanho do conjunto de dados de todos os participantes envolvidos na iteração de agregação. No entanto, o FedAvg não possui garantias de convergência e pode divergir em cenários reais quando os dados são heterogêneos [Cunha Neto et al., 2020].

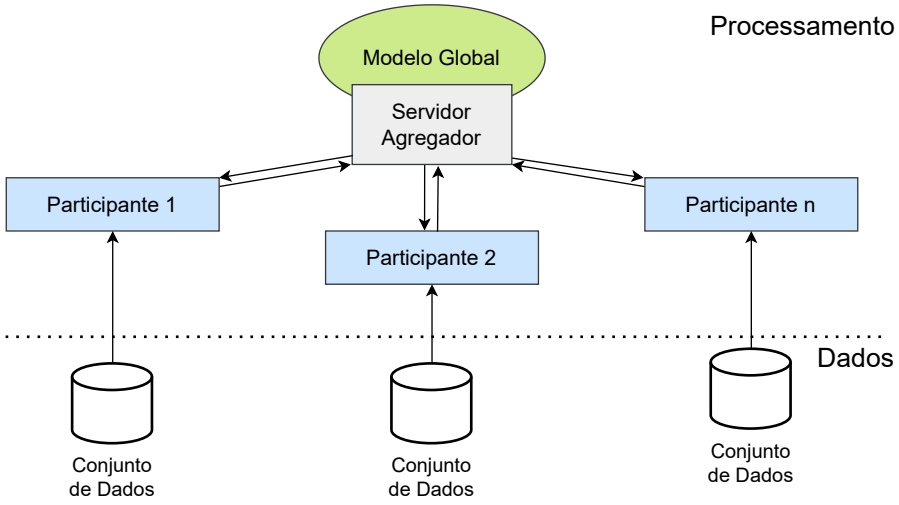


Figura 1. A arquitetura do aprendizado federado. Cada participante do aprendizado federado possui um conjunto de dados local e treina seus modelos com base nele. Apenas os parâmetros do modelo são enviados ao servidor de agregação, que une os modelos locais para criar um modelo global, consolidando o conhecimento dos participantes.

A Figura 1 mostra que o aprendizado federado opera de forma descentralizada e colaborativa. Cada participante detém seu próprio conjunto de dados local e realiza o treinamento de modelos locais com base nesses dados. Importante ressaltar que apenas os parâmetros do modelo são compartilhados, sendo enviados ao servidor de agregação. O servidor de agregação, por sua vez, reúne os modelos locais de todos os participantes para formar um modelo global mais robusto, incorporando assim o conhecimento coletivo de todos os participantes na rede.

3. Trabalhos Relacionados

O aprendizado federado é uma tecnologia amplamente adotada para aprendizado de máquina colaborativo. No entanto, participantes mal-intencionados podem comprometer todo o treinamento. Uma maneira eficaz de minimizar o efeito de participantes maliciosos é a seleção de participantes [Lai et al., 2021, Song et al., 2021, Nishio e Yonetani, 2019, Lee, 2022]. Apesar do esforço de pesquisa para evitar participantes maliciosos, o desafio permanece.

O principal objetivo da seleção de participantes é encontrar os melhores participantes para a agregação dos modelos locais. Cada método de seleção de participantes tem uma métrica para definir os melhores participantes. Por exemplo, os participantes podem ser selecionados com base na qualidade de seus dados, poder computacional ou conexão. Wang *et al.* propuseram um método de pontuação de reputação para avaliar a contribuição de modelos locais no aprendizado federado. O método determina quais modelos locais contribuem para a melhoria do modelo global, permitindo a exclusão de modelos de baixo desempenho ou maliciosos que podem degradar o desempenho do modelo global [Wang e Kantarci, 2020]. No início de cada rodada de agregação, é atribuída uma pontuação de reputação a cada usuário, calculada usando uma fórmula que considera a acurácia do modelo local, a acurácia do conjunto de teste e a acurácia do modelo global na agregação anterior. Uma limitação notável desta estratégia está na sua dependência de agregar atualizações locais de todos os participantes, após a exclusão de participantes de baixo desempenho. Em contraste com essa metodologia, o método FedSBS proposto, emprega pontuação baseada em perdas locais e globais, em vez da acurácia. Vale ressaltar que para avaliação de modelos de aprendizado de máquina, é necessário outras métricas além da acurácia, particularmente em contextos com conjuntos de dados desbalanceados.

Song *et al.* propuseram um treinamento baseado em reputação para redes sem fio que empregam uma função de distribuição beta para pontuar a reputação dos participantes [Song et al., 2021]. Os autores também propuseram uma política de agendamento baseada em reputação que considera problemas de canal sem fio, como interferência e conexão degradada. Além de avaliar as contribuições dos participantes, o método de reputação também foi avaliado para detectar participantes maliciosos [Song et al., 2021]. Song *et al.* usam a distribuição beta, a partir das observações do participante e do servidor, para calcular a reputação do participante. Por outro lado, o FedSBS utiliza a perda local e global para determinar as pontuações e incorpora o mecanismo de bloqueio de participantes, mitigando o risco de seleção recorrente de participantes e a potencial influência de participantes maliciosos.

Lai *et al.* propuseram o Oort, uma seleção de participantes orientada ao FL [Lai et al., 2021]. Oort prioriza os participantes com alta perda local. Os autores modelaram a seleção de participantes como um problema de bandidos multi-armados, em que cada participante é um "braço" do bandido e a pontuação obtida é a recompensa [Lai et al., 2021]. Em seguida, o esquema de seleção de participantes pode explorar participantes não selecionados ou explorar os já selecionados. Lai *et al.* usam a perda local como critério para pontuar os participantes, complementando com um mecanismo de bloqueio projetado para reduzir o excesso de vezes que um mesmo participante é selecionado. A abordagem FedSBS proposta distingue-se pelo uso tanto da perda local quanto da global na pontuação dos participantes. A incorporação de perdas locais e globais em

um sistema de pontuação de aprendizado federado promove um ambiente de treinamento equilibrado.

Nishio *et al.* propuseram o método de seleção de participantes *Federated Learning with Client Selection* (FedCS) [Nishio e Yonetani, 2019]. FedCS é um arcabouço para o ambiente de Computação de Borda de Acesso Múltiplo — *Multi-access Edge Computing* (MEC). FedCS primeiro solicita algumas informações de recursos para um subconjunto de participantes. Com base nessas informações, FedCS seleciona o máximo de participantes possível que preenchem recursos mínimos pré-definidos. Os autores mais tarde estenderam seu trabalho para participantes com diferentes distribuições de dados [Yoshida et al., 2020]. FedCS, que depende de informações fornecidas pelos participantes, pode ser suscetível a informações falsas, especialmente quando lida com participantes maliciosos. Em contraste, o método de pontuação proposto neste artigo se baseia na perda local e global, uma métrica que pode ser calculada pelo servidor de agregação, aumentando assim a robustez na seleção de participantes.

Pesquisas anteriores sobre métodos de seleção de participantes dependem de métricas como acurácia [Wang e Kantarci, 2020], perda local [Lai et al., 2021] ou informações fornecidas pelos participantes [Nishio e Yonetani, 2019]. No entanto, essas métricas não fornecem uma avaliação precisa das contribuições dos participantes. Em contraste, o FedSBS incorpora um novo método de pontuação de participantes que considera a perda local e global, aprimorando a seleção de participantes. Além disso, o FedSBS possui um mecanismo de bloqueio de participantes baseado na distribuição de Boltzmann para evitar a seleção repetidas vezes dos mesmos participantes, promovendo uma seleção equilibrada. Ao utilizar essas estratégias, o método proposto melhorara significativamente a eficiência do treinamento de aprendizado federado com participantes maliciosos.

4. A Proposta de Seleção Federada Baseada em Pontuação – FedSBS

Este artigo aborda a seleção eficiente de participantes no cenário de aprendizado federado, de maneira a otimizar a aprendizagem do modelo global e minimizar o impacto potencial de participantes maliciosos ou de baixo desempenho. Para isso, é proposta a seleção de participantes baseada na pontuação que cada participante recebe de acordo com um método de pontuação baseado no ganho de informação calculado sobre suas perdas local e global. A proposta é composta por um Método de Pontuação de Participantes e por um Método de Seleção de participantes, descritos a seguir.

4.1. Método de Pontuação de Participantes

O método de pontuação de participantes proposto se inspira no Ganho de Informação – *Information Gain* (IG). O IG quantifica a informação que um subconjunto contribui para fazer previsões de classe precisas [Azhagusundari et al., 2013]. O principal uso de IG é para seleção de características e criação de árvores de decisão. O IG é definido por

$$I = Entropia(\mathbb{P}) - \sum_{n=1}^{N} \varphi_n Entropia(\mathbb{S}_t), \tag{1}$$

em que N é o número de subconjuntos e φ_n pondera a entropia do subconjunto $\mathbb{S}_t$. Normalmente, φ é definido como o número de amostras no subconjunto $\mathbb{S}_t$ dividido pelo

número total de amostras no conjunto de dados completo $\mathbb{P}$, $\forall\ \mathbb{S}_t \in \mathbb{P}$. Portanto, o IG mede a redução na entropia de um determinado subconjunto, isto é, a incerteza associada a um subconjunto de amostras selecionadas aleatoriamente [Momma e Bennett, 2002].

O principal objetivo do cenário de seleção de participantes para problemas de classificação é reduzir a entropia cruzada. Consequentemente, pode-se adaptar o IG como

$$I = L(\mathbf{w}_G^t) - \varphi_n L_n(\mathbf{w}_n), \tag{2}$$

em que $L(\mathbf{w}_G^t)$ é a função de perda global, e $L_n(\mathbf{w}_n)$ é a função de perda local do participante n. O somatório da Equação 1 é removido para determinar o IG de cada participante individualmente. É importante destacar as diferenças entre o IG e a proposta em relação aos seus valores retornados. O IG, que é baseado no cálculo da entropia, retorna um valor dentro do intervalo de zero a um. Por outro lado, o método de pontuação proposto, que utiliza a entropia cruzada, pode produzir qualquer valor positivo ou negativo. Essa divergência surge da diferença inerente entre a entropia e a entropia cruzada. Enquanto a entropia é limitada entre zero e um, a entropia cruzada não tem um limite superior.

A Equação 2 não aborda as propriedades necessárias para o cenário de seleção de participantes. A seleção considera participantes com valores baixos de $L(\mathbf{w}_G^t)$ e altos de $L_n(\mathbf{w}_n)$, isto é, baixa perda global e alta perda local. O objetivo principal do treinamento em FL é minimizar a função de perda global, denotada por $L(\mathbf{w}_G^t)$. Isso significa que o esforço é para alcançar baixa perda global para otimizar o desempenho geral do modelo. Ao mesmo tempo, a seleção de participantes dentro do FL opera com um objetivo contrastante. O objetivo é identificar participantes associados a um valor de perda local alto e trabalhar para minimizá-lo. Enquanto o treinamento global do FL visa a redução da perda global, o processo de seleção de participantes se concentra na minimização das perdas locais. Como resultado, um método de pontuação eficaz para a seleção de participantes deve encontrar um equilíbrio entre a menor perda global e a maior perda local. Para isso, utiliza-se o logaritmo negativo de ambas as funções de perda local e global para alcançar essa propriedade. Portanto, a definição do método de pontuação é dada por

$$\begin{aligned} I = & -\ln(L(\mathbf{w}_G^t)) - \varphi_n \times -\ln(L_n(\mathbf{w}_n)) \\ = & -\ln(L(\mathbf{w}_G^t)) + \varphi_n \ln(L_n(\mathbf{w}_n)). \end{aligned} \tag{3}$$

Assim como o IG, o método de pontuação de participantes proposto retorna um valor maior para participantes que podem contribuir para o treinamento e um valor baixo para aqueles que pouco contribuem ou não contribuem. Com base na Equação 3, os participantes com perda local menor do que a perda global são pontuados com um valor negativo. Por outro lado, participantes com perda local maior do que a perda global têm um valor de pontuação positiva. É essencial destacar que o método de pontuação de participantes não pode substituir o IG. A proposta é uma equação de pontuação baseada em IG para o cenário de seleção de participantes em FL e não uma medida alternativa.

O método utiliza a entropia de cada conjunto de dados do participante para determinar a ponderação de sua perda local individual. A proposta usa a entropia do conjunto de dados para punir os participantes com um conjunto de dados desequilibrado. Todo conjunto de dados que a classe alvo possui uma distribuição desigual é chamado de conjunto

de dados desequilibrado. Nesse caso, usar a entropia para ponderar $\ln(L_n(\mathbf{w}_n))$ pode ter um efeito diferente se $\ln(L_n(\mathbf{w}_n))$ for positivo ou negativo. Além disso, se $\ln(L_n(\mathbf{w}_n))$ for positivo, então $\varphi_n = entropia$; caso contrário, $\varphi_n = 1 - entropia$.

$$\varphi_n = \begin{cases} 1 + \sum p_c \log_2(p_c), & \textbf{se } \ln(L_n(\mathbf{w}_n)) < 0 \\ -\sum p_c \log_2(p_c), & \textbf{se } \ln(L_n(\mathbf{w}_n)) \geqslant 0 \end{cases} \tag{4}$$

A Equação 4 define φ_n, em que p_c é a proporção da classe c no conjunto de dados do participante n.

4.2. Método de Seleção de Participante

Junto à utilização do método de pontuação do participante, o FedSBS aborda elementos críticos para melhorar a robustez e eficiência da seleção de participantes, como balancear a relação entre *exploration* e *exploitation* usando *epsilon* ganancioso, e garantir a diversidade implementando um mecanismo de bloqueio. *Exploration* e *exploitation* são conceitos centrais na tomada de decisões sob incerteza. *Exploration* refere-se ao processo de buscar novas informações para aprender mais sobre o ambiente. Já *Exploitation* é o processo pelo qual utiliza-se o conhecimento adquirido para tomar decisões que acredita-se serem as melhores.

FedSBS usa a abordagem *epsilon* ganancioso para balancear entre a *exploration* de selecionar novos participantes e a *exploitation* do ganho já conhecido de selecionar participantes. A política *epsilon* ganancioso seleciona o atual melhor participante com a probabilidade de $1 - \epsilon$ e uma seleção aleatória de participantes com a probabilidade de ϵ, em que $0 < \epsilon < 1$. O valor de ϵ é reduzido por um fator de decaimento η a cada rodada de agregação. A variável η deve assumir um valor distinto dependendo do número de rodadas de agregação. Para que ϵ comece igual a 1, e no final do treinamento, ϵ seja igual a um valor desejável b, então $\eta = \sqrt[t]{b}$, em que t é o número de rodadas de agregação, e b é o último valor que ϵ deve assumir.

O método de seleção de participantes proposto incorpora um mecanismo de bloqueio de participantes para prevenir a seleção frequente dos mesmos participantes. Este mecanismo observa o número de vezes que cada participante é selecionado. Ele aplica uma distribuição de Boltzmann para determinar a probabilidade de um participante ser bloqueado para uma nova seleção após sua seleção inicial. Em essência, se um participante foi escolhido uma vez, sua probabilidade de ser bloqueado para uma nova seleção aumenta a cada seleção subsequente, promovendo uma distribuição mais equilibrada de seleção de participantes e reduzindo a probabilidade de seleção frequente em um pequeno subconjunto de participantes. Portanto, a seleção de um participante que já foi escolhido é determinada por

$$P(\Omega_n, T) = \exp\left(-\frac{\Omega_n}{T}\right), \tag{5}$$

em que Ω_n denota o número de vezes que o participante n foi selecionado anteriormente, enquanto T representa o hiperparâmetro de "temperatura", representando a aleatoriedade do sistema. Especificamente, a probabilidade de seleção de um participante é maior durante as seleções iniciais e quando o valor da temperatura, aleatoriedade, é alto. A proba-

bilidade diminui para seleções subsequentes, promovendo uma distribuição mais equitativa da seleção de participantes ao longo de diversas rodadas, com baixa aleatoriedade.

5. Modelo do Atacante

O atacante realiza um ataque bizantino, usando pontos de dados aleatórios para prejudicar o treinamento e tentando interromper o processo de treinamento. Ataques bizantinos em aprendizado federado ocorrem quando participantes mal-intencionados, ou nós bizantinos, injetam dados aleatórios no sistema para causar o colapso do modelo global. Esses ataques podem ocorrer quando os nós maliciosos fazem parte do sistema de aprendizado. Esses ataques podem impactar significativamente a integridade e o desempenho do modelo global. O atacante é um participante do ambiente de aprendizado federado e usa esse acesso para injetar pontos de dados falsos no conjunto de dados local. Esses pontos de dados falsos podem ser elaborados de tal forma a influenciar os resultados do treinamento e causar resultados imprecisos ou indesejáveis. São considerados três tipos distintos de participantes maliciosos: I) participante malicioso constante, II) participante p-malicioso, e III) participante malicioso de agregação k. O participante malicioso constante é um tipo de participante malicioso que se comporta maliciosamente durante todo o processo de treinamento. O participante p-malicioso utiliza seus dados reais, mas também pode usar dados falsos com uma probabilidade p que segue uma distribuição de Boltzmann. Finalmente, o participante malicioso de agregação k mantém um comportamento honesto até que atinja a rodada de agregação k, momento em que começa a usar dados falsos. Assim, o cenário de aprendizado federado apresenta desafios significativos, principalmente com a possibilidade de ataques bizantinos realizados por participantes maliciosos. Apesar das ameaças potenciais apresentadas por diferentes tipos de participantes maliciosos, o FedSBS fornece uma abordagem eficaz para manter a integridade do modelo global do aprendizado federado.

6. Avaliação da Proposta e Resultados Experimentais

A avaliação da proposta FedSBS considera a comparação com dois métodos de referência, baseados em trabalhos relacionados — Oort [Lai et al., 2021] e Wang *et al.* [Wang e Kantarci, 2020] — na presença de participantes maliciosos. O modelo de ataque é o apresentado na Seção 5. As principais métricas de desempenho examinadas nesta avaliação incluem acurácia, a métrica F1, precisão, sensibilidade e especificidade. Para avaliar a resiliência dos métodos em diferentes proporções de presença adversária, consideram-se dois cenários distintos com diferentes proporções de participantes maliciosos, 20% e 60%. Ao analisar o desempenho da proposta e dos trabalhos relacionados nessas condições, o objetivo é apresentar a eficácia e a robustez da proposta ao lidar com diferentes graus de participação adversária no sistema. A avaliação fornece a compreensão sobre cada proposta em relação à escolha do método mais apropriado para garantir uma seleção de participantes confiável e segura na presença de participantes maliciosos.

Para alcançar o objetivo, foi desenvolvido um simulador[1] baseado em Python usando o conjunto de dados CICIDS2017 [Sharafaldin et al., 2019]. Para implementação dos modelos utiliza-se a biblioteca PyTorch[2], tanto para a proposta quanto para os métodos de referência. O modelo de aprendizado de máquina utilizado consiste em um

[1]Disponível em https://github.com/helioncneto/FederatedLearningSimulator

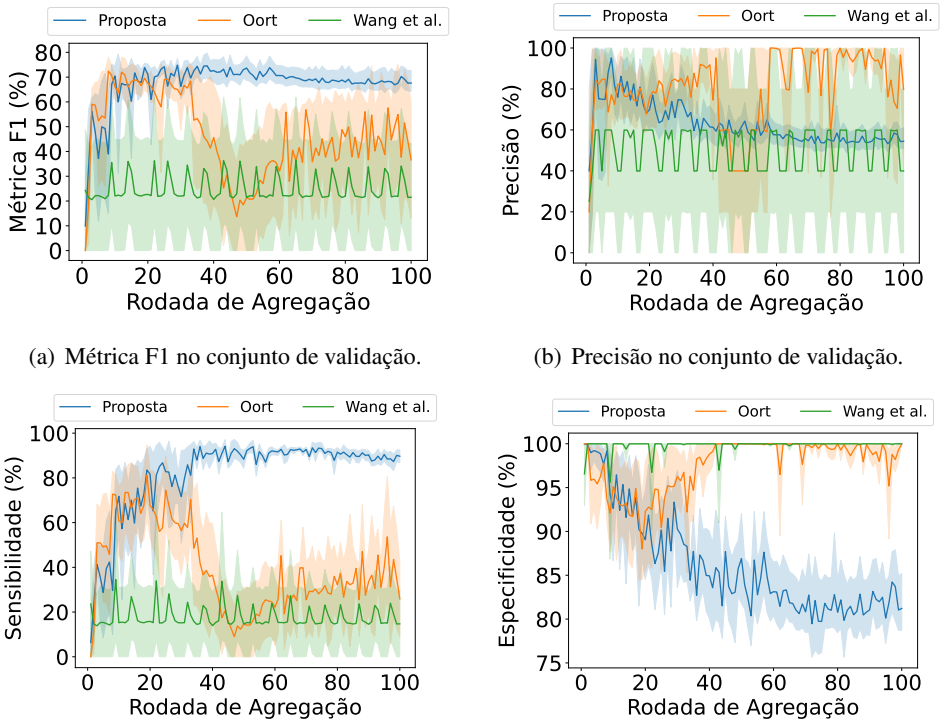
[2]Disponível em https://pytorch.org/, acessado em 10/06/2023.

perceptron multicamadas com duas camadas ocultas. A primeira camada oculta é composta por 50 neurônios, enquanto a segunda contém 100 neurônios. O simulador garante que o conjunto de dados seja dividido em três partes com um equilíbrio igual de classes - 80% de rótulos de tráfego de ataque e 30% de rótulos de tráfego normal. A primeira parte do conjunto de dados é usada para treinamento, a segunda para validação e a terceira para teste. Os conjuntos de treinamento, validação e teste são compostos por 90%, 5% e 5% de amostras, respectivamente. O simulador aplica a distribuição de Dirichlet para dividir o conjunto de treinamento para cada participante, simulando dados Non Independent and Identically Distributed (Non-IID). Essa abordagem de distribuição de dados é baseada na distribuição realizada por Kim *et al.* [Kim et al., 2022]. Para as avaliações, foi utilizado um computador equipado com processador Intel(R) Core(TM) i9-12900KF 12th Gen 3.2GHz, 126GB de RAM e placa de vídeo GeForce RTX 3060 com 12GB de memória Graphics Double Data Rate (GDDR) RAM.

A distribuição Dirichlet é utilizada nesse simulador para simular dados Non-IID. A distribuição Dirichlet é uma distribuição multivariada que gera uma distribuição de probabilidade sobre um conjunto de k categorias, em que k é um número inteiro positivo. Cada amostragem da distribuição Dirichlet produz um vetor de probabilidade, um vetor de k probabilidades que somam 1, em que os valores no vetor representam as probabilidades de cada categoria. Várias amostras são geradas independentes da distribuição Dirichlet, com diferentes conjuntos de parâmetros, representando diferentes distribuições subjacentes para gerar dados Non-IID. Os vetores de probabilidade resultantes são diferentes para cada observação, refletindo as diferentes distribuições subjacentes.

Nas simulações, cada participante detém uma quantidade diferente de dados e o simulador seleciona aleatoriamente a quantidade para cada participante. É importante destacar que cada avaliação foi executada por 100 rodadas de agregação, porque o desempenho do modelo global permanece relativamente estável após 90 rodadas de agregação. Para garantir relevância estatística, são realizadas cinco execuções para cada avaliação, cada uma empregando uma distribuição de dados única entre os participantes. Para garantir a relevância estatística, os resultados são apresentados como médias de rodadas de experimento associadas a um intervalo de confiança de 95%. Para cada rodada de agregação, apenas 30% dos participantes são selecionados para a agregação em cada avaliação. Em um trabalho anterior, verificou-se que selecionar mais participantes ou menos não impacta significativamente o desempenho do modelo global [Cunha Neto et al., 2021].

As Figuras 2 e 3 comparam a proposta FedSBS com os métodos de referência Oort e Wang *et al.*, avaliando seu desempenho em quatro métricas críticas: métrica F1, precisão, sensibilidade e especificidade. Uma característica chave do FedSBS é o bloqueador de participantes, que mitiga o impacto dos participantes maliciosos no processo de treinamento colaborativo. O bloqueador de participantes contribui para o treinamento estável e a eficácia superior do FedSBS, mesmo em cenários com uma presença significativa de participantes maliciosos (20% na Figura 2 e 60% na Figura 3). Nos testes conduzidos, cada uma dos três comportamentos distintos de participantes maliciosos representou um terço do total de participantes maliciosos. Um terço apresentou comportamento p-malicioso, outro se mostrou consistentemente malicioso, enquanto o último se tornou malicioso a partir da rodada de agregação k. Todos conduziram ataques bizantinos, gerando conjuntos de dados e classificações aleatórias com o objetivo de sabotar o



(a) Métrica F1 no conjunto de validação.

(b) Precisão no conjunto de validação.

(c) Sensibilidade no conjunto de validação.

(d) Especificidade no conjunto de validação.

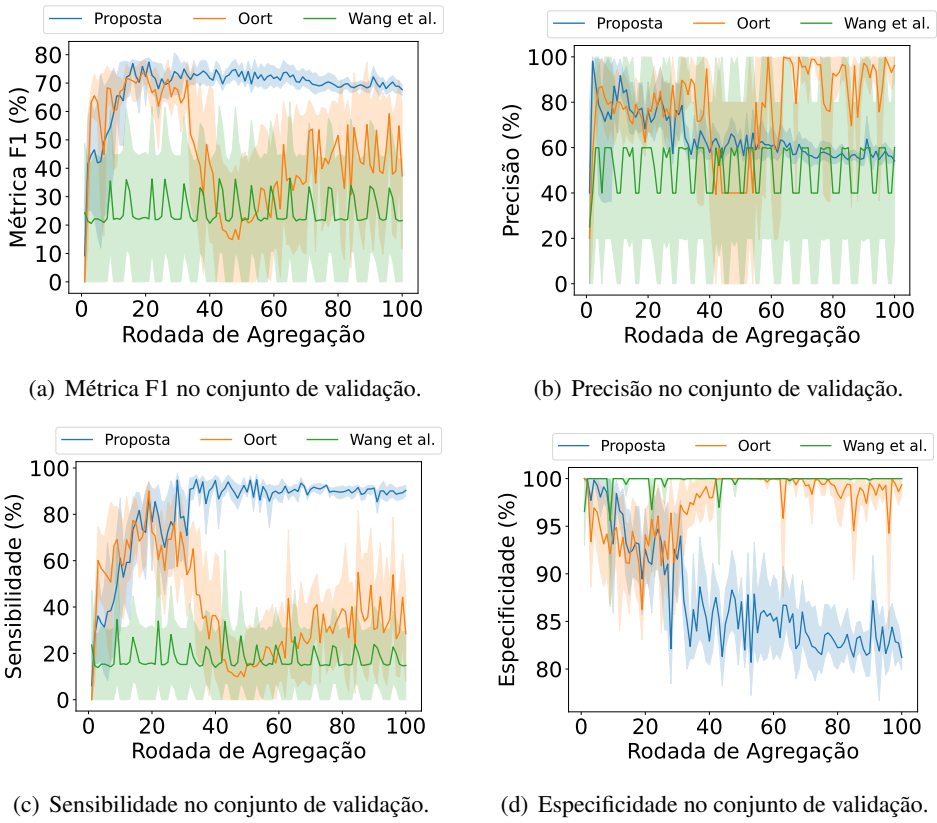
Figura 2. Comparação do desempenho do FedSBS, Oort e Wang *et al.* em quatro métricas de avaliação: métrica F1, precisão, sensibilidade e especificidade em um cenário com 20% de participantes maliciosos. A proposta FedSBS superou os outros métodos de seleção de participantes na métrica F1 e sensibilidade, com menor variação, enquanto os outros métodos se destacaram em especificidade ao classificar amostras predominantemente como tráfego normal.

Tabela 1. Comparação das métricas de desempenho do conjunto de testes (Média e Desvio padrão) para a Proposta, Oort [Lai et al., 2021] e Wang *et al.* [Wang e Kantarci, 2020] sob proporções variadas de participantes maliciosos (20% e 60%). A tabela destaca o desempenho superior do método proposto em todas as métricas, demonstrando sua resiliência mesmo na presença de participantes maliciosos. O valor μ é a média e σ é o desvio padrão.

		Acur.		F1		Prec.		Sens.		Espec.	
Taxa de Mal.	Método	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
20%	FedSBS	90.4	1.7	80.1	2.5	69.3	5.1	95.6	2.5	89.1	2.7
	Oort	81.2	1.4	10.5	12.4	60.0	48.9	6.0	7.1	100.0	0.0
	Wang *et al.*	82.8	3.5	21.0	25.9	39.9	48.9	14.4	17.7	100.0	0.0
60%	FedSBS	90.2	1.9	79.9	2.9	68.9	5.6	95.0	2.3	88.9	3.0
	Oort	82.3	3.0	17.7	22.6	59.9	48.9	11.5	15.1	100.0	0.0
	Wang *et al.*	82.8	3.5	21.0	25.9	39.9	48.9	14.4	17.6	100.0	0.0

processo de treinamento.

FedSBS proporciona um equilíbrio entre precisão e sensibilidade, como eviden-



(a) Métrica F1 no conjunto de validação.

(b) Precisão no conjunto de validação.

(c) Sensibilidade no conjunto de validação.

(d) Especificidade no conjunto de validação.

Figura 3. Gráficos de linha comparando o desempenho das proposta FedSBS, Oort e Wang *et al.* em quatro métricas de avaliação (métrica F1, precisão, sensibilidade e especificidade) em um cenário com 60% de participantes maliciosos. A proposta FedSBS demonstra eficácia e adaptabilidade superiores, superando outros métodos de seleção de participantes em métricas cruciais para dados desbalanceados e mantendo menor variação. O FedSBS superou as referências possuindo a melhor métrica F1 (a) e sensibilidade (c), indicando melhores resultados na identificação de amostras de ataques, mesmo quando estas são escassas no conjunto de dados. Em contrapartida, os modelos de referência apresentaram melhor precisão (b) e especificidade (d), tendendo a classificar as amostras pela classe mais frequente do conjunto de dados.

ciado pelo seu desempenho na métrica F1 superior aos métodos de referência. Em contraste, Oort e Wang *et al.* exibem alta variação, em algumas métricas. Os métodos de referência se destacam em especificidade, mas enfrentam dificuldades em sensibilidade, demonstrando dificuldade na detecção de ataques. Isso acontece, pois, o conjunto de dados possui menos amostras da classe ataque do que a classe tráfego normal. A resistência do FedSBS ao modelo de atacante, detalhado na Seção 5, destaca ainda sua adaptabilidade e robustez em cenários desafiadores. Ao incorporar o bloqueador de participantes e o método de seleção de participantes proposto, o método efetivamente evita participantes maliciosos, resultando em desempenho superior aos outros métodos e estabilidade.

A Tabela 1 compara as métricas de desempenho do conjunto de testes para a proposta FedSBS, Oort, e Wang *et al.*, avaliadas sob diferentes proporções de participantes maliciosos (20% e 60%). Notavelmente, o método FedSBS apresenta desempenho superior nas principais métricas, especialmente na métrica F1, Precisão e Sensibilidade. Essas métricas são críticas ao lidar com conjuntos de dados desbalanceados, pois proporcionam uma compreensão mais abrangente do desempenho do modelo na identificação de ambas

as classes, tráfego de ataques e tráfego normal.

Os resultados do conjunto de testes demonstram que o método de seleção proposto supera os metodos de referência em desempenho, mesmo quando avaliado em dados nunca antes vistos. Ao comparar o FedSBS com as referências, é evidenciado que o método proposto demonstra desempenho superior no conjunto de testes. O desempenho observado destaca a eficácia do FedSBS em lidar com os desafios apresentados por conjuntos de dados desbalanceados e participação adversária, tornando-o uma solução promissora para garantir uma seleção de participantes confiável e segura em tais cenários.

7. Conclusão

Esse artigo propos um método de treinamento para FL que inclui o método de seleção de participantes FedSBS. O objetivo principal é enfrentar o desafio em que qualquer cenário em aprendizado federado está iminente, a possibilidade de participantes maliciosos. O FedSBS emprega um método de pontuação de participantes baseado em ganho de informação para pontuar participantes e utiliza uma estratégia *epsilon* ganancioso para seleção. Para melhorar a robustez do método à repetição do mesmo conjunto de participantes, é introduzido um mecanismo bloqueador de participantes que impede que participantes superselecionados dominem o processo de aprendizado. Os resultados da avaliação da proposta demonstraram a eficácia da proposta em cenários incluindo participantes maliciosos. Em cenários com 20% e 60% de participantes maliciosos, o método proposto alcançou mais de 80% na métrica F1, 90% de acurácia e 69% de precisão no conjunto de testes, mostrando a resiliência na presença de participantes maliciosos. O método proposto também mostrou menor variação durante o treinamento, tendo um desvio padrão menor que os métodos comparados.

O método proposto de seleção de participantes para FL aborda efetivamente os desafios estatísticos inerentes, mesmo em ambientes com participantes maliciosos. Sua robustez, escalabilidade e desempenho superiores fazem do FedSBS uma solução promissora para promover sistemas de aprendizado federado seguros e eficientes. Pesquisas futuras visam investigar a integração de técnicas avançadas de preservação de privacidade, como o emprego de criptografia homomórfica ou a utilização de estratégias de mascaramento de parâmetros, para aumentar a segurança e confidencialidade do processo de treinamento de FL.

Referências

Agrawal, S., Sarkar, S., Aouedi, O., Yenduri, G., Piamrat, K., Alazab, M., Bhattacharya, S., Maddikunta, P. K. R. e Gadekallu, T. R. (2022). Federated learning for intrusion detection system: Concepts, challenges and future directions. *Computer Communications*, 195:346–361.

Andreoni Lopez, M., Mattos, D. M., Duarte, O. C. M. e Pujolle, G. (2019). Toward a monitoring and threat detection system based on stream processing as a virtual network function for big data. *Concurrency and Computation: Practice and Experience*, 31(20):e5344.

Azhagusundari, B., Thanamani, A. S. et al. (2013). Feature selection based on information gain. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 2(2):18–21.

Brendan McMahan, H., Moore, E., Ramage, D., Hampson, S. e Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. Em *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, volume 54.

Cunha Neto, H. N., Hribar, J., Dusparic, I., Mattos, D. M. F. e Fernandes, N. C. (2023). A survey on securing federated learning: Analysis of applications, attacks, challenges, and trends. *IEEE Access*, 11:41928–41953.

Cunha Neto, H. N., Mattos, D. M. e Fernandes, N. C. (2021). Fedsa: Arrefecimento simulado federado para a aceleraçao da detecçao de intrusao em ambientes colaborativos. Em *Anais do XXXIX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, p. 280–293. SBC.

Cunha Neto, H. N., Mattos, D. M. F. e Fernandes, N. C. (2020). Privacidade do usuário em aprendizado colaborativo: Federated learning, da teoria à prática. *Minicursos do Simpósio Brasileiro de Segurança de Informação e de Sistemas Computacionais - SBSeg*, 20:142–195.

Kim, G., Kim, J. e Han, B. (2022). Communication-efficient federated learning with acceleration of global momentum. *arXiv preprint arXiv:2201.03172*.

Lai, F., Zhu, X., Madhyastha, H. V. e Chowdhury, M. (2021). Oort: Efficient federated learning via guided participant selection. Em *15th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 21)*, p. 19–35.

Lee, W. (2022). Reward-based participant selection for improving federated reinforcement learning. *ICT Express*.

Momma, M. e Bennett, K. P. (2002). A pattern search method for model selection of support vector regression. Em *Proceedings of the 2002 SIAM International Conference on Data Mining*, p. 261–274. SIAM.

Nishio, T. e Yonetani, R. (2019). Client selection for federated learning with heterogeneous resources in mobile edge. Em *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, p. 1–7.

Sharafaldin, I., Lashkari, A. H., Hakak, S. e Ghorbani, A. A. (2019). Developing realistic distributed denial of service (ddos) attack dataset and taxonomy. Em *2019 International Carnahan Conference on Security Technology (ICCST)*, p. 1–8.

Song, Z., Sun, H., Yang, H. H., Wang, X., Zhang, Y. e Quek, T. Q. (2021). Reputation-based federated learning for secure wireless networks. *IEEE Internet of Things Journal*, 9(2):1212–1226.

Wang, Y. e Kantarci, B. (2020). A novel reputation-aware client selection scheme for federated learning within mobile environments. Em *2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, p. 1–6. IEEE.

Yoshida, N., Nishio, T., Morikura, M., Yamamoto, K. e Yonetani, R. (2020). Hybrid-fl for wireless networks: Cooperative learning mechanism using non-iid data. Em *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, p. 1–7. IEEE.