**MITIGAÇÃO DE ATAQUES POR ENVENENAMENTO EM APRENDIZADO FEDERADO: AVALIAÇÃO DE ABORDAGENS BASEADAS EM OUTLIERS**

**Resumo**

O aprendizado federado é um paradigma distribuído de aprendizado de máquina que preserva a privacidade ao permitir o treinamento colaborativo de modelos sem o compartilhamento de dados brutos. Apesar dos avanços em privacidade, esta abordagem ainda é vulnerável a ataques por envenenamento, que comprometem a integridade do modelo global por meio da inserção de dados corrompidos ou manipulação de pesos e gradientes compartilhados. Tendo em vista que clientes maliciosos tendem a gerar atualizações com comportamento estatístico discrepante, o uso de técnicas de detecção de *outliers* como mecanismos para mitigação de ataques por envenenamento têm se mostrado promissoras. Este estudo adota uma metodologia mista, composta por uma Revisão Integrativa da Literatura, a fim de identificar soluções de defesa e mitigação de ataques por envenenamento, e uma etapa experimental baseada em simulações computacionais, com objetivo de reproduzir cenários de ataque e validação de diferentes estratégias de mitigação. Os resultados esperados incluem a identificação de abordagens eficazes para detectar e reduzir o impacto de clientes maliciosos em sistemas de aprendizado federado, contribuindo para o fortalecimento da cibersegurança em aplicações sensíveis à privacidade e conformidade regulatória.

**1. Introdução**

O Aprendizado Federado (FL) emergiu como um paradigma revolucionário dentro do campo de Aprendizado de Máquina (ML). Nessa abordagem descentralizada, os dados permanecem localizados em dispositivos individuais ou servidores locais. Em vez de coletar e centralizar dados em um repositório para treinar um modelo, o FL permite que modelos sejam treinados diretamente nos dispositivos que geram ou possuem os dados. Após o treinamento local, as atualizações dos parâmetros do modelo ou gradientes – ao invés dos dados – são enviadas a um servidor central, onde são agregadas para criar um modelo global atualizado. Esse processo de treinamento e agregação locais é repetido iterativamente, aprimorando continuamente o modelo global sem expor os dados locais (LI *et al*., 2021). Essa abordagem é especialmente crítica em cenários que exigem alto grau de segurança e confidencialidade, como os regulados pela Lei Geral de Proteção de Dados (DE SOUZA *et al*., 2022) e *General Data Protection Regulation* (ANITHA; JEGATHEESAN, 2023).

Ainda que o FL ofereça benefícios expressivos em termos de privacidade e segurança em comparação a abordagens centralizadas de desenvolvimento de modelos de ML, a arquitetura distribuída do FL ainda apresenta vulnerabilidades, dentre os quais se destacam os ataques de envenenamento de dados e modelos (ZHANG *et al*., 2022). A literatura recente destaca os ataques de envenenamento como um dos principais desafios em implementações de FL. Ataques por envenenamento ocorrem por meio da alteração maliciosa dos gradientes ou coeficientes de modelos treinados localmente. Esta alteração se dá por um meio de um agente malicioso, que insere dados corrompidos (envenenamento de dados) ou alteração direta dos pesos ou gradientes (envenenamento de modelos). Em ambas as situações, o agente malicioso é capaz de comprometer a integridade do modelo global.

A detecção de *outliers* desponta como uma abordagem eficaz para mitigar os efeitos de ataques por envenenamento em FL. Estudos recentes identificam que, diante de um ambiente benigno, as atualizações locais dos pesos ou gradientes fornecidas pelos clientes tendem a seguir distribuições estatísticas relativamente consistentes. Já com a presença de agentes maliciosos, essas atualizações tendencialmente se comportam como *outliers* estatísticos (YAZDINEJAD *et al*., 2024). Neste sentido, a detecção de *outliers* e a adoção de estratégias que envolvam o desenvolvimento de modelos globais menos sensíveis a *outliers* se destacam como abordagens promissora para mitigar esse tipo de ataque em ambientes de FL (BHAGOJI *et al*., 2019; ZHANG *et al*., 2022).

Embora a associação de agentes maliciosos a *outliers* represente uma solução estratégica para o avanço da cibersegurança em aplicações de FL, ela deve ser continuamente aprimorada para lidar com ameaças cada vez mais avançadas (SIKANDAR *et al*. 2023). Blanchard *et al*. (2017) apontam desafios e oportunidades neste sentido, propondo a investigação do desempenho de métodos de agregação de modelos locais teoricamente menos sensíveis à *outliers* frente a ataques de envenenamento de modelos. Zhang *et al*. (2022) apontam para a necessidade de desenvolver soluções para ataques mais sofisticados, que expõem vulnerabilidades adicionais no âmbito do FL. Neste sentido, é necessário investigar limitações e oportunidades de abordagens baseadas em *outliers* como mecanismo de defesa contra ataques por envenenamento em aplicações de FL.

**2. Objetivos do Projeto**

O objetivo do projeto é investigar estratégias de prevenção/mitigação de ataques de envenenamento de dados e de modelos em aplicações de FL. Para alcançar este objetivo geral, foram estabelecidos os seguintes objetivos específicos:

* Analisar as vulnerabilidades do FL, com ênfase nos ataques por envenenamento de dados e modelos, destacando sua influência na integridade do modelo global;
* Investigar os métodos de prevenção/mitigação de ataques de envenenamento baseados em *outliers*, avaliando seus benefícios e limitações;
* Reproduzir, por meio de simulação, cenários de ataques por envenenamento de dados e de modelos e realizar análise comparativa do efeito das estratégias identificadas na detecção de clientes maliciosos e na qualidade do modelo global; e
* Validar abordagens de defesa/mitigação de ataques a fim de aprimorar a detecção de clientes maliciosos em sistemas de FL.

Com isso, pretende-secontribuir para a evolução da segurança das aplicações de FL, garantindo sua aplicabilidade em cenários com alta exigência de privacidade, robustez e conformidade regulatória.

**3. Procedimentos Metodológicos**

Esta pesquisa adotará uma abordagem metodológica mista, combinando métodos quantitativos e qualitativos, com foco na análise de ameaças e avaliação de soluções voltadas à segurança em FL. Como primeiro passo, será realizada uma Revisão Integrativa da Literatura, baseada no protocolo proposto por Whittemore e Knafl (2005), com o objetivo de mapear ataques de envenenamento e as principais técnicas de mitigação destes ataques discutidas na literatura atual. Em seguida, no segundo estágio, será desenvolvida uma Pesquisa Experimental complementar, seguindo a abordagem de Estudo de Caso proposta por Runeson e Höst (2009). Essa etapa consistirá na simulação de cenários adversos em ambientes federados a fim de reproduzir condições realistas de ataques. As soluções defensivas aplicadas serão avaliadas com base em métricas quantitativas, conforme descrito em Yang *et al*. (2021). Essa análise visa mensurar a eficácia prática das estratégias de mitigação frente aos desafios mapeados anteriormente. A proposta de abordagem mista, que associa o levantamento teórico com avaliação experimental por meio de simulação, permite alinhar embasamento científico com validação prática sobre soluções para aumento da segurança em aplicações de FL.

No que se refere à revisão Integrativa da Literatura, a aplicação seguirá as seguintes etapas: inicialmente, será feita a identificação do problema de pesquisa, delimitando o foco nos desafios e soluções em aprendizagem federada e segurança de dados. Em seguida, será realizada a busca na literatura, utilizando bases acadêmicas relevantes para coletar artigos e estudos recentes. Após a busca, será feita a avaliação dos dados, verificando a qualidade e a pertinência dos materiais encontrados. Os dados selecionados passarão então por uma análise criteriosa, com o objetivo de sintetizar informações-chave e identificar lacunas existentes. Por fim, a apresentação dos resultados será realizada de forma estruturada, organizando os achados em categorias temáticas.

Para o estudo de caso, as etapas serão aplicadas da seguinte forma: primeiro, será realizado o projeto do estudo de caso, definindo o contexto e os parâmetros de análise, com foco em um ambiente de simulação de ataques e defesas em aprendizado federado. A seguir, ocorrerá a preparação para a coleta de dados, envolvendo a definição de métricas de avaliação e procedimentos de registro de eventos. A coleta de evidências será feita a partir da execução de experimentos controlados. Os dados obtidos serão então submetidos à análise, com o objetivo de interpretar o comportamento dos modelos e os efeitos das técnicas de mitigação aplicadas. Finalmente, será elaborado um relatório detalhado, apresentando os resultados e discutindo as implicações para a área de segurança em aprendizado federado.

**4. Cronograma**

Os procedimentos metodológicos descritos na Seção anterior compreendem um conjunto de etapas bem definidas, propostas por Whittemore e Knafl (2005) – para revisão integrativa de literatura – e Runeson e Höst (2009) – para o estudo de caso – que orientam o desenvolvimento da pesquisa. Para cada etapa, foram em atividades práticas a serem cumpridas ao longo dos doze meses de prazo deste projeto. O cronograma apresentado na Tabela 1 descreve as atividades metodológicas a serem executadas, bem como a duração prevista para sua realização. A adoção dessas abordagens metodológicas confere rigor e coerência à condução do projeto, permitindo que a pesquisa avance de forma estruturada. Considerando, que a pesquisa é um processo ao qual a descobertas são um pressuposto, redirecionamentos e aprofundamentos são previstos, de modo que os prazos foram definidos com margem para flexibilidade para acomodar investigações adicionais que venham a surgir no decorrer do projeto, sem comprometer a qualidade ou os objetivos propostos.

Tabela 1 – Cronograma do projeto.

| **Atividades** | **Meses** | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| **1. Revisão Integrativa da Literatura** |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.1. Identificação do problema e formulação da questão de pesquisa |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.2. Busca sistemática em bases científicas sobre ataques e defesas em FL |  |  |  |  |  |  |  |  |  |  |  |  |
| 1.3. Triagem, avaliação e síntese crítica da literatura selecionada |  |  |  |  |  |  |  |  |  |  |  |  |
| **2. Design e Preparação do Estudo Experimental** |  |  |  |  |  |  |  |  |  |  |  |  |
| 2.1. Definição dos cenários de ataque em ambientes federados |  |  |  |  |  |  |  |  |  |  |  |  |
| 2.2. Configuração do ambiente de simulação |  |  |  |  |  |  |  |  |  |  |  |  |
| 2.3. Modelagem e preparação dos protocolos de mitigação a serem testados |  |  |  |  |  |  |  |  |  |  |  |  |
| **3. Execução de Simulações e Coleta de Dados** |  |  |  |  |  |  |  |  |  |  |  |  |
| 3.1. Execução de ataques controlados |  |  |  |  |  |  |  |  |  |  |  |  |
| 3.2. Aplicação de estratégias defensivas e métricas de avaliação |  |  |  |  |  |  |  |  |  |  |  |  |
| 3.3. Coleta de evidências quantitativas |  |  |  |  |  |  |  |  |  |  |  |  |
| **4. Análise e Validação** |  |  |  |  |  |  |  |  |  |  |  |  |
| 4.1. Análise estatística dos dados |  |  |  |  |  |  |  |  |  |  |  |  |
| 4.2. Validação das estratégias defensivas aplicadas |  |  |  |  |  |  |  |  |  |  |  |  |
| 4.3. Ajustes e reexecução de cenários críticos com diferentes parâmetros |  |  |  |  |  |  |  |  |  |  |  |  |
| **5. Redação e Submissão** |  |  |  |  |  |  |  |  |  |  |  |  |
| 5.1. Escrita do artigo científico e relatório |  |  |  |  |  |  |  |  |  |  |  |  |
| 5.2. Revisão e ajustes estruturais |  |  |  |  |  |  |  |  |  |  |  |  |
| 5.3. Preparação para apresentação e submissão final |  |  |  |  |  |  |  |  |  |  |  |  |

**Referências Bibliográficas**

ANITHA, G.; JEGATHEESAN, A. Navigating Privacy Concerns in Federated Learning A GDPR-Focused Analysis. In: **Handbook on Federated Learning**. CRC Press, 2023. p. 106-130.

BHAGOJI, A. N. et al. Analyzing federated learning through an adversarial lens. In: **International conference on machine learning**. PMLR, 2019. p. 634-643.

BLANCHARD, P. et al. Machine learning with adversaries: Byzantine tolerant gradient descent. **Advances in neural information processing systems**, v. 30, 2017.

DE SOUZA, L. A. C. et al. Aprendizado federado com agrupamento hierárquico de clientes para aumento da acurácia. In: **Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)**. SBC, 2022. p. 545-558.

LI, Q. et al. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. **IEEE Transactions on Knowledge and Data Engineering**, v. 35, n. 4, p. 3347-3366, 2021.

RUNESON, P.; HÖST, M. Guidelines for conducting and reporting case study research in software engineering. **Empirical software engineering**, v. 14, p. 131-164, 2009.

SIKANDAR, H. S. et al. A detailed survey on federated learning attacks and defenses. **Electronics**, v. 12, n. 2, p. 260, 2023.

WHITTEMORE, R.; KNAFL, K. The integrative review: updated methodology. **Journal of advanced nursing**, v. 52, n. 5, p. 546-553, 2005.

YANG, Q. et al. Federated machine learning: Concept and applications. **ACM Transactions on Intelligent Systems and Technology (TIST)**, v. 10, n. 2, p. 1-19, 2019.

YAZDINEJAD, A et al. A robust privacy-preserving federated learning model against model poisoning attacks. **IEEE Transactions on Information Forensics and Security**, 2024.

ZHANG, J. et al. Security and privacy threats to federated learning: Issues, methods, and challenges. **Security and Communication Networks**, v. 2022, n. 1, p. 2886795, 2022.