Avaliação da Privacidade Diferencial Aplicada ao Aprendizado Federado Através de Ataques de Inversão de Modelo

Breno V. Manhães, João Pedro M. F. Ferreira, Guilherme A. Thomaz e Miguel Elias M. Campista

[1]Universidade Federal do Rio de Janeiro, GTA/DEL-Poli/PEE-COPPE

{bvalente,moretti,guiaraujo,miguel}@gta.ufrj.br

Abstract. *This work proposes a practical methodology for evaluating the effectiveness of Differential Privacy (DP) mechanisms in Federated Learning (FL) against model inversion attacks. To this end, an adversarial setting inspired by the red team/blue team (RT/BT) cybersecurity paradigm is adopted, where the blue team implements DP-based defenses and the red team attempts to reconstruct data from a specific client using the global model. Two types of attacks were implemented — a gradient-based attack and a naive baseline — and applied under varying levels of Gaussian noise. Experimental results show that even low noise levels are sufficient to significantly mitigate inversion attacks, as evidenced by both quantitative metrics (SSIM, PSNR, and MSE) and visual inspection of the reconstructed images. Meanwhile, model accuracy remains stable under moderate noise levels, with minimal impact on performance. These findings suggest that a favorable balance between privacy and utility can be achieved in practical FL scenarios.*

Resumo. *Este trabalho propõe uma metodologia prática para avaliar a eficácia de mecanismos de Privacidade Diferencial (DP) no Aprendizado Federado (FL) diante de ataques de inversão de modelo. Para isso, foi adotado um cenário adversarial baseado no paradigma de segurança cibernética equipe vermelha/equipe azul (RT/BT), onde a equipe azul implementa a proteção de DP e a equipe vermelha realiza ataques para recuperar dados de um cliente específico a partir do modelo global. Foram implementados dois tipos de ataques — um baseado em gradientes e outro ingênuo — aplicados em diferentes intensidades de ruído gaussiano. Os experimentos demonstraram que ruídos baixos já são suficientes para mitigar significativamente os ataques de inversão, conforme evidenciado tanto por métricas quantitativas (SSIM, PSNR e MSE) quanto pela análise visual das imagens reconstruídas. Por outro lado, observou-se que a acurácia do modelo global se mantém estável até níveis moderados de ruído, com impacto limitado no desempenho. Esses resultados indicam que é possível alcançar um bom equilíbrio entre privacidade e utilidade do modelo em cenários práticos.*

1. Introdução

O Aprendizado Federado (*Federated Learning* – FL) emergiu como uma abordagem poderosa para treinar modelos de aprendizado de máquina em larga escala, utilizando

dados distribuídos em múltiplos dispositivos ou silos, sem a necessidade de centralizar informações potencialmente sensíveis [Kairouz 2021]. Essa capacidade de aprendizado colaborativo, preservando a localização original dos dados, é particularmente promissora em domínios como saúde e finanças, onde a privacidade do indivíduo é uma preocupação primordial. No entanto, embora o FL evite a centralização direta dos dados brutos, o próprio processo de treinamento, envolvendo a troca de atualizações de modelo, como gradientes, ainda expõe os sistemas a riscos significativos de privacidade. Ataques como inferência de pertencimento (*membership inference*) [Shokri et al. 2017] e inversão de modelo (*Model Inversion*) [Fredrikson et al. 2015] podem potencialmente extrair informações sobre os dados de treinamento privados de um cliente participante.

Nesse contexto, a Privacidade Diferencial (*Differential Privacy* – DP) estabeleceu-se como uma referência consolidada para fornecer garantias formais e quantificáveis de privacidade em algoritmos de análise de dados, incluindo o FL [Dwork et al. 2006, Dwork and Roth 2014]. A DP funciona tipicamente pela adição de ruído cuidadosamente calibrado às informações compartilhadas, como os gradientes ou parâmetros do modelo, de modo a obscurecer a contribuição de qualquer indivíduo específico no resultado agregado. A força dessa proteção é geralmente parametrizada por *épsilon* (ϵ) e *delta* (δ), onde valores menores indicam maior privacidade. Contudo, a avaliação da privacidade em sistemas baseados em FL que operam por múltiplas rodadas de comunicação apresenta desafios complexos [Zhu et al. 2017]. Essa avaliação é crucial para entender o risco real ao qual os dados dos usuários estão expostos, já que garantias teóricas nem sempre se traduzem em segurança prática [Bagdasaryan et al. 2019].

A abordagem padrão de avaliação de privacidade consiste em utilizar teoremas de composição para rastrear o orçamento de privacidade acumulado (ϵ, δ totais) ao longo do treinamento. Embora teoricamente sólida, essa métrica agregada muitas vezes resulta em limites superiores conservadores [Truex et al. 2020]. Crucialmente, ela oferece pouca intuição sobre a vulnerabilidade prática de um cliente específico a ataques reais que podem ocorrer persistentemente ao longo do treinamento. Como um ruído demasiado prejudica a convergência do modelo, consequentemente limitando seu uso em aplicações que exigem maior acurácia, focar exclusivamente na otimização do ϵ acumulado pode prejudicar excessivamente a qualidade e usabilidade de um modelo de FL sem incorrer em ganhos igualmente significativos na privacidade de seus clientes. Em suma, existe uma lacuna notável entre a garantia teórica fornecida pelo (ϵ, δ) composto e a segurança operacional real experimentada pelos participantes.

Este trabalho propõe uma abordagem inspirada na metodologia *equipe vermelha/equipe azul* (RT/BT), um conceito bem estabelecido na área de segurança da informação [Hughes 2020], para preencher essa lacuna e proporcionar uma avaliação mais realista e orientada à prática da eficácia dos mecanismos de DP em FL. O cenário adversarial simulado vai além dos trabalhos que apenas consideram métricas teóricas de privacidade [Geyer et al. 2017]. Neste trabalho, o Time azul (*Blue Team*) assume o papel do defensor do FL. Sua responsabilidade é implementar, configurar e gerenciar os mecanismos de DP. Isso inclui a escolha do tipo de mecanismo central ou local, a calibração do nível de ruído que controla o orçamento de privacidade por rodada e, potencialmente, a adaptação da estratégia ao longo do tempo. Já o Time Vermelho (Red Team) atua como o adversário. Seu objetivo é comprometer a privacidade de um cliente específico dentro

do sistema de FL. Especificamente, a equipe vermelha se concentra em lançar Ataques de Inversão de Modelo (*Model Inversion Attacks*) de forma contínua, ao longo de múltiplas rodadas de treinamento, tentando reconstruir ou inferir características dos dados privados do cliente alvo com base nas atualizações do modelo que observa, mesmo que ruidosas.

Os resultados obtidos através da interação simulada Equipe Vermelha/Equipe Azul revelam que o ajuste do nível de ruído (σ), baseado tanto na observação direta da eficácia do ataque de inversão quanto nas métricas quantitativas dos ataques, proporciona *insights* mais práticos do que a mera contabilidade teórica de ϵ comumente avaliada em outros trabalhos. A análise quantificou a troca entre privacidade e utilidade, demonstrando como níveis crescentes de ruído progressivamente degradam a qualidade das imagens reconstruídas (medida por SSIM, PSNR, MSE e visualmente) pela Equipe Vermelha, mas impactam a acurácia do modelo global. A avaliação foi além, analisando a eficácia dos ataques (baseado em gradiente e ingênuo) e a acurácia do modelo ao longo de 50 rodadas de treinamento para um cliente-alvo individual, permitindo identificar faixas de ruído que oferecem proteção substancial contra ataques de inversão sem comprometer excessivamente o desempenho do modelo.

Este trabalho está estruturado da seguinte forma. A Seção 2 discute trabalhos relacionados. A Seção 3 apresenta as insuficiências relacionadas às métricas existentes de privacidade, enquanto a Seção 4 apresenta os conceitos de Ataques de Inversão de Modelo relevantes para este estudo. A Seção 5 detalha a metodologia proposta de avaliação adversarial equipe vermelha/equipe azul e descreve o ambiente experimental. A Seção 6 apresenta e analisa os resultados obtidos nas simulações. Finalmente, a Seção 7 conclui este trabalho e sugere direções futuras.

2. Trabalhos Relacionados

A integração da privacidade diferencial no aprendizado federado tem sido explorada como uma solução para mitigar riscos de privacidade decorrentes da troca de atualizações de modelo entre clientes e servidor. Diversos estudos propuseram mecanismos para garantir DP em FL, abordando tanto a perspectiva centralizada no servidor quanto a local nos clientes.

Geyer *et al.* propuseram um algoritmo que preserva a privacidade diferencial no nível do cliente durante a otimização federada, equilibrando a perda de privacidade e o desempenho do modelo [Geyer et al. 2017]. Seus estudos empíricos sugerem que, com um número suficientemente grande de clientes participantes, é possível manter a privacidade diferencial no nível do cliente com apenas uma pequena perda de desempenho do modelo. Já *Naseri et al.* investigaram o uso de DP para proteger tanto a privacidade quanto a robustez em FL, avaliando técnicas de Privacidade Diferencial Local (LDP) e Central (CDP) [Naseri et al. 2022]. Seus experimentos mostraram que ambas as variantes de DP defendem contra ataques de *backdoor*, embora com diferentes níveis de compromisso entre proteção e utilidade.

Wei *et al.* propuseram o *framework noising before model aggregation FL*, no qual ruídos artificiais são adicionados aos parâmetros no lado dos clientes antes da agregação, garantindo DP sob diferentes níveis de proteção [Wei et al. 2019]. Eles também desenvolveram limites teóricos de convergência da função de perda do modelo treinado em FL, revelando propriedades-chave como o compromisso entre desempenho de con-

vergência e níveis de proteção de privacidade. No contexto brasileiro, Alves *et al.* apresentaram o PEGASUS, que utiliza garantias de DP para mitigar ataques adversários em FL [Alves et al. 2023]. Além disso, o PEGASUS emprega uma estratégia de seleção de clientes que adapta dinamicamente a quantidade de dispositivos que treinam o modelo, visando lidar com o acréscimo da perda de privacidade ao longo das rodadas de comunicação.

Apesar dessas contribuições, a maioria dos trabalhos concentra-se em métricas teóricas de privacidade, como o orçamento acumulado de ϵ, sem considerar a eficácia prática dos mecanismos de DP contra ataques persistentes. Assim, a principal contribuição deste trabalho para o estado da arte reside no desenvolvimento e aplicação de um *framework* de avaliação adversarial que permite uma análise empírica e contextualizada da real proteção oferecida pela Privacidade Diferencial no Aprendizado Federado a partir do uso de ataques de inversão de modelo, complementando as análises puramente teóricas de orçamentos de privacidade.

3. Métricas Teóricas de Privacidade: ϵ e δ

A privacidade diferencial é formalizada por meio dos parâmetros ϵ e δ, que quantificam o grau de proteção oferecido por um algoritmo. Intuitivamente, esses parâmetros garantem que a presença ou ausência de um único dado no conjunto não altere significativamente a distribuição das saídas do algoritmo, oferecendo uma proteção estatística à informação individual [Dwork and Roth 2014].

Essas métricas vêm sendo amplamente utilizadas para avaliar a privacidade ao longo do aprendizado federado. Nesse cenário, o treinamento é realizado de forma iterativa, com múltiplas rodadas de comunicação entre os dispositivos clientes e o servidor central. Como cada rodada utiliza mecanismos de privacidade diferencial, o risco de vazamento de informações tende a se acumular ao longo do tempo.

Assumindo que cada rodada é realizada com um mecanismo de privacidade que satisfaz (ϵ, δ)-privacidade, a aplicação direta do teorema de composição leva a uma estimativa conservadora da perda total de privacidade. Para T rodadas, uma composição avançada fornece

$$\epsilon_{\text{total}} = \epsilon\sqrt{2T\ln\left(\frac{1}{\delta}\right)} + T\epsilon\left(e^{\epsilon} - 1\right),$$

o que pode indicar uma perda significativa de privacidade conforme T aumenta. Essa abordagem teórica, embora rigorosa, nem sempre reflete de maneira precisa o impacto real, pois existem dois fatores principais que limitam sua exatidão. Primeiramente, a composição teórica tende a ser conservadora, muitas vezes superestimando a degradação da privacidade acumulada ao longo das rodadas, produzindo garantias excessivamente pessimistas [Truex et al. 2020]. Na prática, esse conservadorismo pode subestimar o real potencial de preservação da privacidade e comprometer a usabilidade de um modelo demasiado ruidoso. Em segundo lugar, em ambientes federados, os dados locais podem apresentar distribuições muito distintas. Essa heterogeneidade, combinada com a interação dinâmica entre múltiplas rodadas de atualização, gera efeitos complexos que não são capturados pelas composições teóricas tradicionais [Naseri et al. 2022]. Como consequência, os limites teóricos tornam-se ainda menos representativos do comportamento real do sistema.

Devido às duas limitações apontadas — o pessimismo das garantias teóricas e a dificuldade de capturar a heterogeneidade —, este trabalho propõe substituir a análise puramente teórica por uma abordagem baseada em experimentos práticos para avaliação da privacidade. Além disso, estudos recentes demonstram que a aplicação de privacidade diferencial local pode introduzir desafios adicionais, como aumento da variância nos dados e dificuldades na manutenção da acurácia do modelo [Sun et al. 2020]. Essas dificuldades evidenciam a necessidade de abordagens complementares para avaliar de forma mais realista o impacto da aplicação de privacidade diferencial em treinamentos multi-rodada.

4. Ataques de Inversão de Modelo

Os ataques de Inversão de Modelo representam uma classe de ataques que visam subverter a proteção de privacidade no FL. A ideia é inferir informações sensíveis sobre os dados de treinamento de um cliente específico a partir do modelo global compartilhado, com suas atualizações locais [Shi et al. 2024]. A premissa fundamental dos ataques de inversão no Aprendizado Federado é que o modelo treinado, mesmo que formado pela agregação de modelos de diversos clientes, ainda preserva características dos dados de treinamento de cada cliente individual. Ao explorar essa memorização, um atacante (a "equipe vermelha", neste contexto) tenta reconstruir ou inferir informações sobre os dados privados do cliente-alvo, utilizando o conhecimento que possui sobre o modelo global após uma sessão de treinamento local. É importante notar que este trabalho foca em ataques direcionados a um cliente específico, o que permite contornar a heterogeneidade dos dados entre os diferentes participantes do FL. A eficácia do ataque depende de diversos fatores, incluindo a arquitetura do modelo, a distribuição dos dados entre os clientes, a presença de mecanismos de proteção de privacidade e o conhecimento prévio do atacante sobre os dados do cliente.

Este trabalho implementa e avalia duas abordagens distintas para investigar a capacidade de um atacante realizar ataques de inversão direcionados a um cliente específico em um ambiente de FL. Essas estratégias representam diferentes filosofias de ataque: a primeira, denominada de **Inversão Ingênua (*Naive Inversion*)**, busca maximizar a confiança do modelo de forma heurística; a segunda, a **Inversão Baseada em Gradientes (*Gradient-Based Inversion*)**, utiliza um processo analítico de otimização baseado na minimização de uma função de perda. O funcionamento e as características de cada uma dessas abordagens, cujo objetivo final é avaliar a vulnerabilidade dos modelos em FL e quantificar a informação potencialmente extraída dos dados privados dos clientes, são apresentados a seguir:

Inversão Ingênua (*Naive Inversion*): esta abordagem representa uma tentativa intuitiva de gerar uma imagem que maximize a confiança do modelo na classe alvo. A *Naive Inversion* opera através de um processo iterativo onde uma imagem inicial, com ruído adicionado de forma localizada e aleatória, é gradualmente modificada de forma a aumentar a probabilidade de ser classificada como a classe alvo pelo modelo. O ataque da inversão ingênua inicia com uma imagem de entrada zerada, na qual um pequeno número de *pixels* selecionados aleatoriamente recebe valores iniciais uniformes. Iterativamente, o ataque gera perturbações localizadas e aleatórias, e as adiciona à imagem atual. A nova imagem é mantida apenas se aumentar a confiança (saída softmax do modelo local para o classe alvo). O processo continua até que a confiança atinja um limiar de 99% ou um número máximo de iterações.

Inversão Baseada em Gradientes (*Gradient-Based Inversion*): Em contraste, a *Gradient-Based Inversion* emprega uma abordagem baseada em otimização. O objetivo é encontrar uma imagem de entrada X^* que minimize uma função de perda $\mathcal{L}$, que mede a discrepância entre a classificação gerada na saída do modelo para uma entrada X, $f(X;\theta)$ (onde θ é o vetor de parâmetros do modelo), e uma classe alvo y. Matematicamente, busca-se resolver:

$$X^* = \arg\min_X \mathcal{L}(f(X;\theta), y).$$

A função de perda tipicamente utilizada é a entropia cruzada (*Cross-Entropy*), $\mathcal{L}_{CE}$. Neste caso, o ataque busca gerar uma imagem de entrada X^* que, ao ser processada pelo modelo $f(X;\theta)$, maximize a probabilidade que o modelo atribui à classe alvo desejada, y. Para fins de cálculo da perda, a classe alvo y é representada por um vetor *one-hot* correspondente, $\mathbf{y}_{\text{target}}$. A otimização, portanto, visa fazer com que a distribuição de probabilidades na saída da função softmax do modelo se alinhe o máximo possível com essa representação $\mathbf{y}_{\text{target}}$. Para alcançar isso, uma imagem inicial (por exemplo, ruído aleatório ou uma imagem zerada com perturbações) é ajustada iterativamente. Em cada passo, calcula-se o gradiente da função de perda $\mathcal{L}_{CE}$ (que compara a saída da softmax com $\mathbf{y}_{\text{target}}$) em relação aos pixels da imagem de entrada e atualiza-se a imagem na direção oposta ao gradiente (gradiente descendente). Esse processo torna a imagem progressivamente mais representativa dos padrões que o modelo, com seus parâmetros θ, aprendeu a associar à classe alvo y.

5. Metodologia Aplicada

Para investigar a resiliência prática de mecanismos de DP em FL contra ataques, uma metodologia experimental inspirada no paradigma *equipe vermelha/equipe azul* (RT/BT) foi adotada. Este *framework* simula um confronto direto entre um defensor (equipe azul), responsável pela implementação e gerenciamento das proteções de privacidade, e um adversário (equipe vermelha), cujo objetivo é comprometer a privacidade de um participante específico através de ataques de inversão de modelo. O foco principal reside em avaliar como diferentes estratégias de DP, configuradas pela equipe azul, impactam a capacidade da equipe vermelha de extrair informações sobre os dados privados de um cliente-alvo individual ao longo de múltiplas rodadas de treinamento de FL.

5.1. Ambiente Experimental de Aprendizado Federado

O ambiente de FL foi simulado utilizando implementações customizadas em `Python` com a biblioteca `NumPy` para as operações centrais de rede neural e treinamento. O conjunto de dados utilizado foi o MNIST [Deng 2012], composto de imagens em escala de cinza (28×28 *pixels*) de dígitos manuscritos (0-9) usados em problemas de classificação. O conjunto de dados de treinamento foi particionado entre $N = 5$ clientes simulados. Atribuiu-se a cada cliente um conjunto de tamanho igual composto por amostras selecionadas aleatoriamente *sem reposição* do conjunto de dados original. Um cliente específico, selecionado aleatoriamente, foi designado como o **cliente-alvo** para os ataques da equipe vermelha.

Os experimentos usaram uma rede neural de duas camadas completamente conectadas (*fully-connected neural network*) implementada utilizando a biblioteca NumPy da linguagem Python. A arquitetura consiste em uma camada de entrada que recebe os *pixels*

das imagens de 28×28 em 784 neurônios, uma camada oculta com 10 neurônios e função de ativação ReLU e uma camada de saída com 10 neurônios, correspondentes aos dígitos 0-9, e função de ativação softmax. As matrizes de pesos (`W1`, `W2`) foram inicializadas com valores aleatórios pequenos obtidos de uma distribuição normal com média zero e desvio padrão 0,01 e os vetores de *biases* (`b1`, `b2`) com zeros. O treinamento federado ocorreu por R rodadas de comunicação, sendo que em cada rodada $r \leq R$ tem-se que:

1. o servidor distribui os parâmetros do modelo global atual ($W1_r, b1_r, W2_r, b2_r$) a todos os clientes participantes.
2. cada cliente k treina o modelo em seus dados locais por $E = 5$ épocas com uma taxa de aprendizado $\alpha = 0,1$.
3. os parâmetros atualizados do modelo local do cliente k ($W1_{r,k}, b1_{r,k}, W2_{r,k}, b2_{r,k}$) são retornados.
4. o servidor agrega os parâmetros recebidos de todos os clientes usando média simples (*Federated Averaging* – FedAvg), já que os clientes são iguais em número de imagens, para produzir o novo modelo global ($W1_{r+1}, b1_{r+1}, W2_{r+1}, b2_{r+1}$).
5. o procedimento retorna ao Passo 1 com $r = r + 1$, se $r \leq R$.

5.2. Configuração da equipe vermelha (Atacante)

A equipe vermelha atuou como um adversário com o objetivo de inferir informações sobre os dados privados do cliente-alvo. O modelo de atacante é capaz de acessar todos os modelos no servidor, incluindo as contribuições enviadas pelos clientes individuais e o modelo agregado. O acesso ocorre imediatamente após a conclusão do treinamento local desse cliente em uma dada rodada r, e após a aplicação de ruído, mas antes da agregação no servidor. O objetivo é reconstruir amostras de dados correspondentes a uma classe alvo, selecionada aleatoriamente a partir das classes presentes no conjunto de dados do cliente-alvo, utilizando apenas esses parâmetros do modelo local.

Os ataques (ingênuo e o baseado em gradientes) foram executados **em cada rodada** r de treinamento do FL, utilizando os parâmetros do modelo local do cliente-alvo específicos daquela rodada. Isso permite avaliar a evolução da capacidade de reconstrução da equipe vermelha conforme o modelo local do cliente é treinado. O sucesso da equipe vermelha é avaliado utilizando os seguintes métodos, calculados comparando a imagem reconstruída final com uma imagem real aleatoriamente selecionada do conjunto de dados privado do cliente-alvo:

- **Qualidade Visual:** Inspeção visual das imagens reconstruídas em diferentes marcos do ataque.
- **Métricas Quantitativas das Imagens:** o **Erro Quadrático Médio (*Mean Squared Error* – MSE)** calcula a média das diferenças quadráticas pixel a pixel entre a imagem original e a reconstruída; valores menores indicam menor erro numérico, mas não necessariamente melhor qualidade visual percebida. A **Razão Sinal-Ruído de Pico (*Peak Signal-to-Noise Ratio* – PSNR)**, expressa em decibéis (dB), mede a razão entre a máxima potência de sinal (valor máximo de pixel) e a potência do ruído (MSE); valores mais altos geralmente indicam reconstrução de maior qualidade, mas também podem não se correlacionar bem com a percepção humana. Por fim, o **Índice de Similaridade Estrutural (*Structural Similarity Index Measure* – SSIM)** foi empregado por ser projetado para medir a similaridade percebida, comparando informações de luminância, contraste e estrutura

local entre as imagens. O SSIM é amplamente adotado na avaliação de qualidade de imagem pois suas propriedades se alinham melhor com o sistema visual humano do que métricas baseadas puramente em erros de pixel. O SSIM varia entre -1 e 1, com valores mais próximos de 1 indicando maior similaridade estrutural.

- **Métricas Intrínsecas ao Ataque:** A confiança final alcançada (inversão ingênua) ou a perda final obtida (inversão baseada em gradientes) são registradas como indicadores da convergência do ataque.

5.3. Configuração da equipe azul (Defensor)

A equipe azul é responsável por implementar e gerenciar o mecanismo de DP, aplicado aos parâmetros do modelo de cada cliente após o treinamento local e antes de serem enviados ao servidor para agregação. Para cada tensor de parâmetro (`W1`, `b1`, `W2`, `b2`), adiciona-se ruído gaussiano amostrado de uma distribuição normal $\mathcal{N}(0, \sigma^2)$. O ruído é aplicado de forma independente a cada elemento dos tensores de parâmetros.

A equipe azul controla o nível de privacidade ajustando o parâmetro σ. Múltiplos cenários experimentais foram testados, variando sistematicamente σ de 0 (caso base do FedAvg, sem DP) até 0,1, adicionando 0,01 a cada simulação. Embora a composição formal do orçamento de privacidade (ϵ, δ) não tenha sido calculada, a avaliação se concentrou no impacto da adição de ruído tanto na resistência a ataques de inversão de modelo quanto na utilidade do modelo.

Por fim, a eficácia da equipe azul em preservar a utilidade do modelo é medida pela acurácia média do modelo global final (`W1_R`, `b1_R`, `W2_R`, `b2_R`), agregado a partir dos modelos locais de *todos* os clientes.

5.4. Procedimento Experimental

O procedimento experimental completo seguido neste trabalho está ilustrado na Figura 1, que representa a interação cíclica entre a equipe azul (defensora) e a equipe vermelha (adversária) ao longo do treinamento federado. As etapas numeradas na figura correspondem diretamente às ações descritas abaixo, permitindo mapear visualmente cada fase do experimento. Além disso, a figura explicita os dois ciclos principais: o loop interno de rodadas de aprendizado federado e o loop externo que reinicia o processo para diferentes configurações de ruído.

Para cada configuração de privacidade definida pela equipe azul:

1. o FL é inicializado. Um cliente e uma classe alvo de seu conjunto de dados são selecionados para o ataque.
2. em cada rodada r, após o treinamento local do cliente-alvo e antes da agregação, a equipe azul adiciona ruído ao modelo treinado e então o envia para a equipe vermelha, que obtém os parâmetros ruidosos do modelo local do cliente-alvo.
3. a equipe vermelha executa ambos os ataques (ingênuo e baseada em gradientes) usando os parâmetros ruidosos do modelo do cliente-alvo para tentar reconstruir uma imagem da classe alvo.
4. as métricas de sucesso da equipe vermelha (MSE, PSNR, SSIM, inspeção visual, confiança/perda final) e a métrica de utilidade do modelo da equipe azul (acurácia média do modelo global nos dados dos clientes após a agregação da rodada) foram registradas.

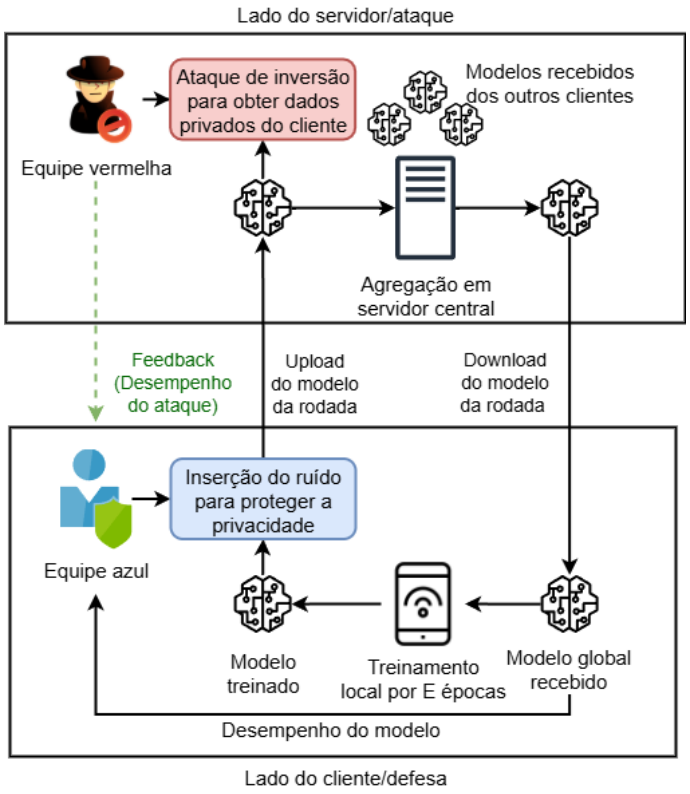


Figura 1. Metodologia de Equipe Azul contra Equipe Vermelha aplicada. A rodada r começa no download do modelo da rodada e termina ao fim da agregação no servidor.

5. o modelo global é agregado e uma nova rodada r é reiniciada se $r \leq R$. O treinamento do FL é executado por $R = 50$ rodadas, com $E = 5$ iterações locais. Portanto, caso $r = R$, o processo retorna ao início para um diferente valor de ruído.

A análise final compara o sucesso dos ataques da equipe vermelha (qualidade da reconstrução) sob diferentes níveis de proteção de DP implementados pela equipe azul, correlacionando-o com a perda de utilidade (acurácia) do modelo global. Isso permite uma avaliação empírica do compromisso entre a privacidade (medida pela resistência aos ataques específicos implementados) e a utilidade do modelo no cenário FL construído.

6. Resultados

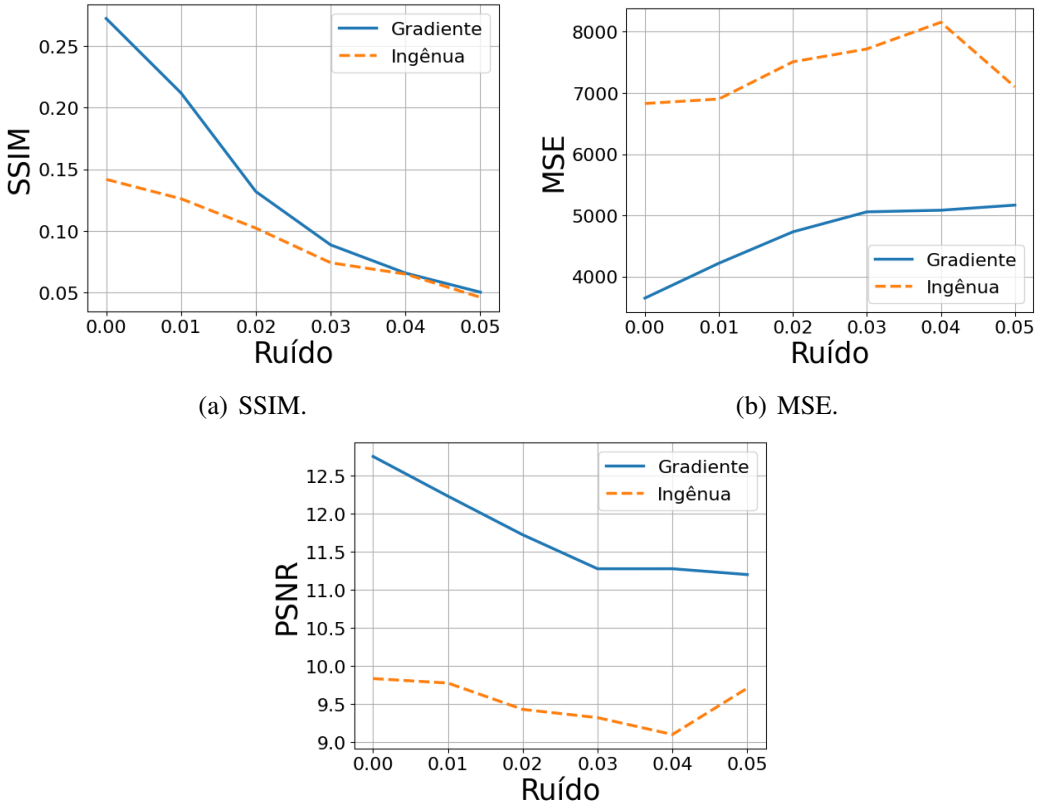
Esta seção apresenta os resultados das avaliações empírica e quantitativa, dividida em análises que examinam a eficácia comparativa dos ataques de inversão, a vulnerabilidade por classe de dados, o impacto do ruído na acurácia do modelo global, cujo caso sem ruído (ruído = 0) representa a linha de base do FedAvg sem DP, e, por fim, uma análise visual dos algarismos reconstruídos.

6.1. Análise inicial

Esta seção compara a eficácia dos dois tipos de ataques utilizando as métricas SSIM, MSE e PSNR para os ruídos de 0 a 0,05. As Figuras 2(a), 2(b) e 2(c) ilustram esses

resultados. Para cada tentativa de reconstrução de uma imagem de uma classe alvo, estas métricas foram calculadas comparando a imagem reconstruída com todas as imagens reais da respectiva classe presentes no conjunto de dados privado do cliente-alvo, sendo o valor utilizado aquele que representou a maior similaridade (ou seja, maior SSIM/PSNR ou menor MSE) entre a reconstrução e qualquer uma das instâncias reais.

Observa-se que o ataque baseado em gradientes apresenta métricas de similaridade significativamente superiores em comparação ao ataque de inversão ingênua, indicando imagens reconstruídas de melhor qualidade. A inclusão do ataque ingênuo, apesar de seu menor desempenho de acordo com as métricas exploradas, serve como um importante ponto de referência devido à sua simplicidade conceitual e menor custo computacional para implementação e execução. Ele representa uma abordagem de ataque mais direta e menos sofisticada, permitindo contrastar o ganho de eficácia obtido com o método mais complexo baseado em gradientes. Além disso, para o ataque baseado em gradientes, é possível notar que todas as métricas apresentam piora significativa quando o modelo é exposto a níveis crescentes de ruído. Isso demonstra que a equipe azul foi bem-sucedida em reduzir a eficácia dos ataques realizados pela equipe vermelha à medida que o ruído aumentava. Além disso, os resultados fornecem uma métrica prática e aplicável para avaliar a efetividade dos mecanismos de privacidade diferencial em cenários de uso real.



(a) SSIM.

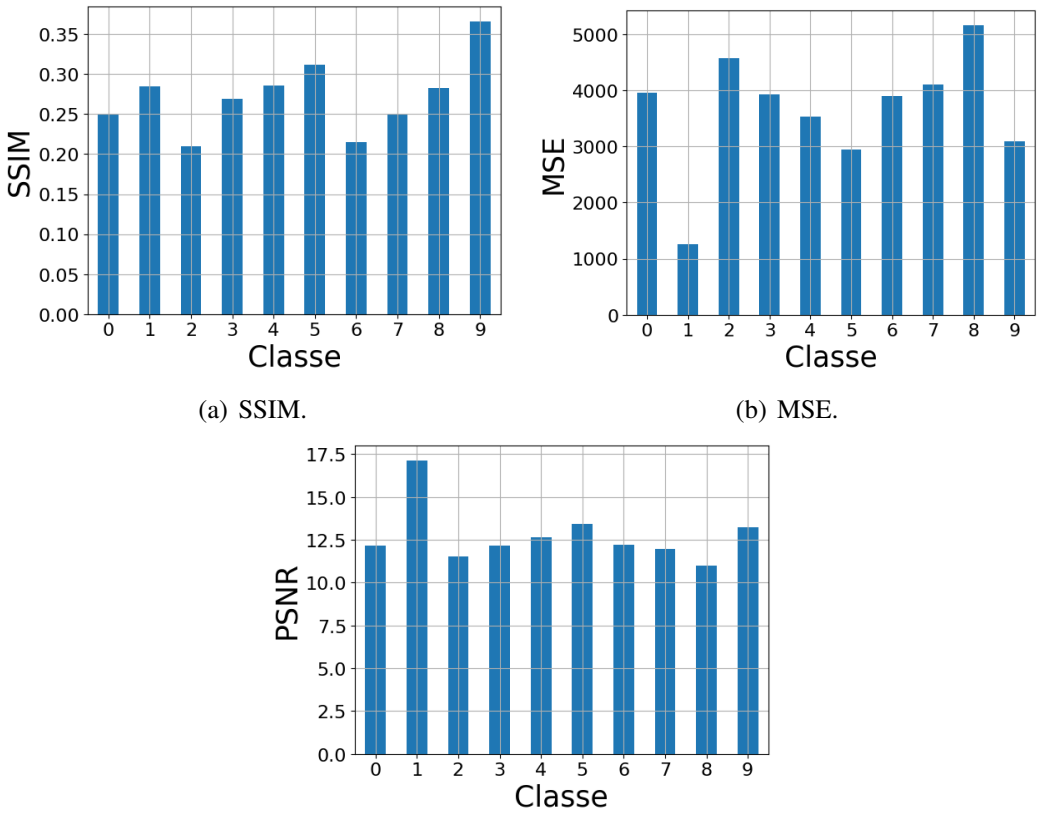
(b) MSE.

(c) PSNR.

Figura 2. Evolução das métricas ao longo das rodadas para cada tipo de ataque.

6.2. Análise dos ataques por classe

A seguir, foi analisado o desempenho do ataque baseado em gradientes em cada classe (0 a 9) quando expostas a nenhum ruído, como mostrado nas Figuras 3(a), 3(b) e 3(c). A análise dos gráficos mostra como a classe 1 apresenta o melhor resultado para as métricas de MSE e PSNR, enquanto a classe 9 mostra o melhor resultado para SSIM. Pode-se concluir, portanto, que determinadas classes estão mais vulneráveis aos Ataques de Inversão que outras.



(a) SSIM. (b) MSE.

(c) PSNR.

Figura 3. Métricas SSIM, MSE e PSNR obtidas para o ataque baseado em gradientes, separadas por classe (0 a 9).

6.3. Análise da acurácia em função do nível de ruído

A Figura 4 ilustra como a acurácia do modelo global se comporta sob diferentes intensidades de ruído. Para aprofundar a análise, o limite superior de ruído foi estendido para 0,1. Os resultados apresentados são fruto de 30 simulações do ambiente de treinamento, cada uma com uma nova distribuição aleatória de amostras entre os clientes para garantir a robustez da análise. Os pontos no gráfico representam a acurácia média para cada nível de ruído, enquanto as linhas verticais indicam o intervalo de confiança de 95% (IC 95%), que mede a incerteza estatística das estimativas.

Confirma-se a tendência geral de queda na acurácia à medida que o ruído aumenta, um comportamento esperado e consistente com a teoria da privacidade diferencial, onde a adição de ruído visa proteger a privacidade, mas pode degradar o desempenho do modelo.

No entanto, observa-se um comportamento não-monótono para níveis de ruído muito baixos (entre 0.00 e 0.01), com um sutil aumento na acurácia média. Este fenômeno pode ser atribuído ao ruído atuando como uma forma de regularização, que mitiga o sobreajuste (*overfitting*) do modelo. Em regimes de baixo ruído, sua adição moderada pode, paradoxalmente, melhorar a capacidade de generalização do modelo, impedindo que ele memorize o conjunto de treino.

Para avaliar formalmente o impacto do ruído, foi realizado o teste de Mann-Whitney, comparando a acurácia em cada nível de ruído com a do cenário sem ruído (ruído 0). Os resultados indicam que o primeiro nível de ruído a apresentar uma diferença estatisticamente significativa na acurácia é 0.07. Isso sugere que o modelo é resiliente a níveis moderados de ruído. Com base nessa análise, um nível de ruído próximo de 0.05 ou 0.06 pode ser considerado ideal, pois maximiza a proteção à privacidade sem comprometer a acurácia do modelo com significância estatistíca.

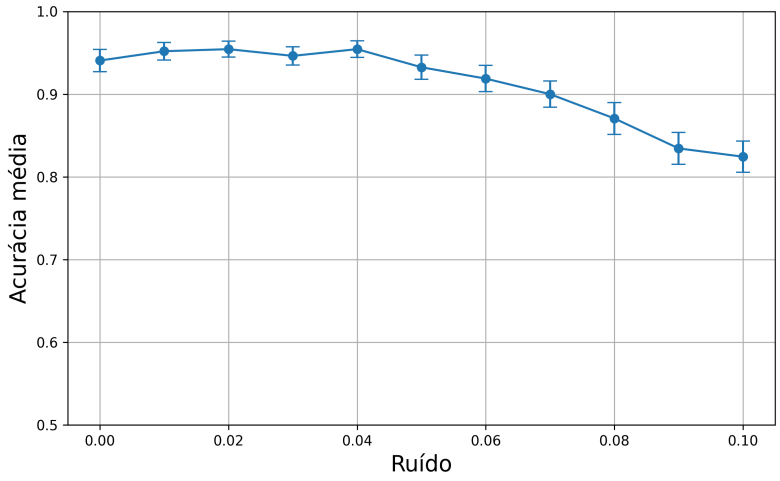


Figura 4. Acurácia do modelo global em função do nível de ruído.

6.4. Análise visual

Considerando a identificada maior vulnerabilidade da Classe 9 aos ataques de inversão (conforme discutido na Seção 6.2) e a importância de avaliar visualmente o impacto da Privacidade Diferencial, esta seção apresenta uma comparação das imagens reconstruídas para a Classe 9. A Figura 5 exibe os resultados obtidos pelos ataques Baseado em Gradientes e Ingênuo sob uma faixa estendida de níveis de ruído (σ), variando de 0,00 (sem ruído) até 0,05. Esta análise visual direta permite observar a degradação progressiva na qualidade e clareza das reconstruções à medida que a intensidade do ruído aumenta, e comparar a resiliência relativa de cada método de ataque.

A inspeção visual das imagens na Figura 5 e a comparação com as imagens reais do cliente atacado, na Figura 6, confirma as tendências observadas nas métricas quantitativas (SSIM, MSE, PSNR) apresentadas anteriormente. Nota-se que as reconstruções obtidas pelo ataque Baseado em Gradientes são visualmente mais detalhadas e reconhecíveis do que as do ataque de Inversão Ingênua, especialmente sob níveis de ruído baixos ou nulos ($\sigma \leq 0.02$). Conforme o nível de ruído (σ) aumenta, a qualidade de ambas as

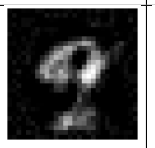
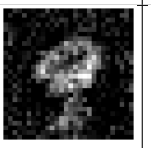
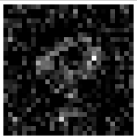
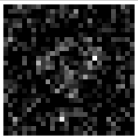
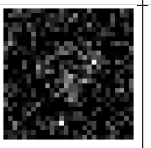
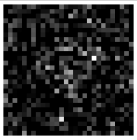
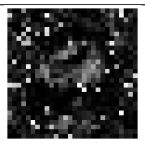
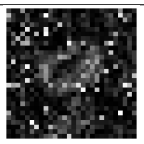
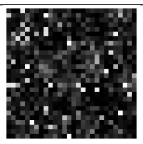
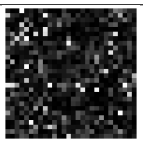
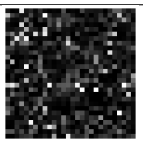
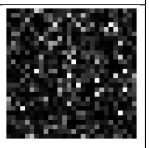
Ataque\Ruído (σ)	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**
Gradientes						
Ingênuo						

Figura 5. Comparativo visual das imagens da Classe 9 reconstruídas pelos ataques Baseado em Gradientes e Ingênuo, sob diferentes níveis de ruído (σ).

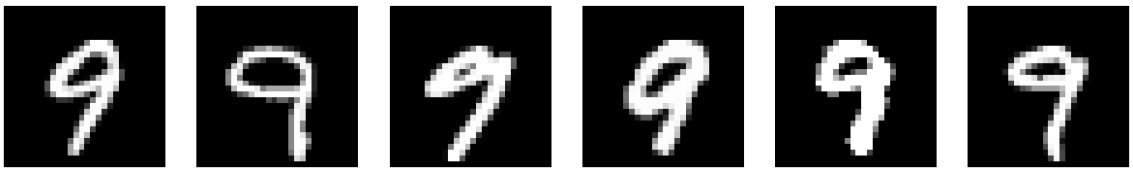


Figura 6. Imagens do conjunto de dados do cliente alvo.

reconstruções se deteriora, tornando os dígitos progressivamente mais indistintos e fragmentados. Essa degradação visual é consistente com a queda nos valores de SSIM e PSNR e o aumento no MSE reportados nas seções anteriores, reforçando a eficácia do ruído em dificultar a extração de informação visualmente útil.

7. Conclusões

Este trabalho apresentou uma metodologia baseada na dinâmica equipe vermelha vs. equipe azul como uma forma prática e eficaz de avaliar empiricamente o impacto de mecanismos de privacidade diferencial no contexto de aprendizado federado. A estrutura proposta permite simular, de maneira controlada, tanto a perspectiva do defensor quanto a do atacante, possibilitando uma análise concreta da eficácia dos mecanismos de proteção frente a ataques de reconstrução. Os resultados obtidos demonstram que essa abordagem é capaz de fornecer informações claras e relevantes sobre os níveis de ruído necessários para atingir um equilíbrio entre privacidade e desempenho do modelo.

Além de se mostrar eficiente na identificação do ponto de equilíbrio desejado, a metodologia proposta também se destaca por sua aplicabilidade em diferentes cenários experimentais. Sua estrutura modular e empírica facilita a adaptação a distintas configurações de ataque, arquiteturas de modelo e estratégias de privacidade, permitindo sua reutilização em estudos futuros. Dessa forma, contribui para o desenvolvimento de abordagens experimentais mais conectadas à prática do aprendizado federado, complementando análises teóricas existentes. Como desdobramento deste trabalho, futuras investigações podem considerar a aplicação de outros mecanismos de ruído, como o ruído Laplaciano, e a experimentação com arquiteturas de redes neurais mais complexas. Além disso, a utilização de conjuntos de dados com maior diversidade e complexidade, como CIFAR-10 ou TinyImageNet, pode ampliar a validade dos resultados. A avaliação da

metodologia frente a ataques mais avançados, como ataques inferenciais ou baseados em otimização adversarial, também representa uma direção promissora. Por fim, a aplicação da metodologia em cenários com múltiplos adversários ou com participantes maliciosos inseridos no processo federado pode abrir novas possibilidades de avaliação sob perspectivas de ameaça mais realistas e distribuídas.

Agradecimentos

O presente trabalho foi realizado com apoio do CNPq, Processo 405940/2022-0; da CAPES (Código de Financiamento 001 e 88887.954253/2024-00); da FAPERJ; e da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Processos 2023/00673-7 e 2023/00811-0. Os autores agradecem também à Fundação de Desenvolvimento da Pesquisa (Fundep), no âmbito do programa Rota 2030, e às empresas Stellantis e Mobway pelo apoio e colaboração que tornaram esta pesquisa possível.

Referências

Alves, T., Silva, J., Pereira, L., and Souza, M. (2023). Pegasus: Garantias de privacidade diferencial para mitigar ataques adversários em aprendizagem federada. In *Anais Estendidos do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, volume 32, pages 15453–15462.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference (TCC)*, pages 265–284. Springer.

Dwork, C. and Roth, A. (2014). *The Algorithmic Foundations of Differential Privacy*, volume 9. Foundations and Trends in Theoretical Computer Science.

Fredrikson, M., Jha, S., and Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333. ACM.

Geyer, R. C., Klein, T., and Nabi, M. (2017). Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*.

Hughes, M. Z. (2020). *Red Team: How to Succeed by Thinking Like the Enemy*. Basic Books.

Kairouz, P. e. a. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210.

Naseri, M., Hayes, J., and De Cristofaro, E. (2022). Local and central differential privacy for robustness and privacy in federated learning. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*.

Shi, Y., Kotevska, O., Reshniak, V., Singh, A., and Raskar, R. (2024). Dealing doubt: Unveiling threat models in gradient inversion attacks under federated learning, a survey and taxonomy. *arXiv preprint arXiv:2405.10376*.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE.

Sun, L., Qian, J., and Chen, X. (2020). Ldp-fl: Practical private aggregation in federated learning with local differential privacy. *arXiv preprint arXiv:2007.15789*.

Truex, S., Liu, L., Chow, K.-H., Gursoy, M. E., and Wei, W. (2020). Ldp-fed: Federated learning with local differential privacy. *arXiv preprint arXiv:2006.03637*.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H., Jin, S., Jin, Y., and Zhang, J. (2019). Federated learning with differential privacy: Algorithms and performance analysis. *arXiv preprint arXiv:1911.00222*.

Zhu, T., Li, G., Zhou, W., and Luo, C. (2017). *Differential Privacy and Applications*. Springer.