Menos é Mais: Avaliação do Impacto da Compressão de Modelos na Eficiência do Aprendizado Federado

Guilherme M. A. Libardi[1], **Bruno Y. L. Kimura**[1], **Joahannes B. D. da Costa**[1]

[1]Instituto de Ciência e Tecnologia (ICT)
Universidade Federal de São Paulo (UNIFESP) – São José dos Campos, SP, Brasil
{guilherme.libardi, bruno.kimura, joahannes.costa}@unifesp.br

Resumo. *O crescimento de dispositivos conectados a Internet tem gerado grande volume de dados e, cada vez mais, impulsionado o uso de Inteligência Artificial (IA). Porém, abordagens centralizadas de aprendizado levantam preocupações relativas à privacidade dos usuários. O Aprendizado Federado (FL) desponta como alternativa distribuída, pois evita o compartilhamento de dados brutos dos usuários e atende legislações de privacidade. No entanto, a comunicação frequente entre dispositivos e servidor no FL gera alto consumo de banda e energia. Para redes com recursos limitados, soluções que mitiguem esse problema são cruciais. Nesse contexto, considerando a possibilidade de redução do tráfego no FL, este trabalho investiga técnicas de compressão de modelos, visando equilibrar qualidade do modelo e custo de comunicação. Os resultados indicam que as técnicas de compressão reduzem efetivamente o volume de dados transmitidos, sem degradar o desempenho do modelo, mesmo em cenários com dados desbalanceados.*

Abstract. *The growing number of devices connected to the Internet has generated a large volume of data, increasingly driving the use of Artificial Intelligence (AI). However, centralized learning approaches raise concerns regarding user privacy. Federated Learning (FL) emerges as a distributed alternative, as it avoids sharing users' raw data and complies with privacy regulations. Nevertheless, frequent communication between devices and the server in FL leads to high bandwidth and energy consumption. In networks with limited resources, solutions to mitigate this problem are crucial. In this context, considering the possibility of reducing traffic in FL, we investigate different model compression techniques, aiming to balance model quality and communication cost. The results indicate that compression techniques effectively reduce the volume of transmitted data without degrading model performance, even in scenarios involving imbalanced data.*

1. Introdução

O número de dispositivos com capacidade de conexão com a Internet vem crescendo continuamente [Alves et al. 2024]. A maioria desses dispositivos são móveis e possuem capacidades avançadas de *hardware* que permitem o sensoriamento e coleta de inúmeros dados, tais como acelerômetro, luminosidade, posicionamento, giroscópio, dentre outros [Ficco et al. 2024]. Ou seja, uma grande quantidade de dados é gerada a cada instante, o que possibilita extrair padrões e direcionar aplicações baseadas nos perfis dos usuários, aumentando assim suas experiências de uso [Sabah et al. 2024].

Os avanços em Inteligência Artificial (IA) têm elevado de forma significativa sua relevância e aplicação no cotidiano. Esse crescimento colocou em evidência não apenas

a necessidade de técnicas de Aprendizado de Máquina, ou *Machine Learning (ML)* em inglês, mais robustas, mas também a importância de um controle rigoroso e adequado no tratamento de dados. Essa atenção é essencial para garantir tanto a qualidade dos modelos desenvolvidos para as mais diversas aplicações quanto a manutenção da privacidade dos dados dos usuários que estão envolvidos nesse processo [de Souza et al. 2024].

O controle efetivo dos dados dos usuários é uma tarefa desafiadora, sobretudo porque a maioria das abordagens em ML ainda se baseia em técnicas centralizadas de aprendizado [Thakur et al. 2025]. Esse paradigma pressupõe acesso irrestrito aos dados para treinamento dos modelos, o que, por sua vez, dificulta a implementação de medidas eficazes de privacidade. Nesse contexto, surge o Aprendizado Federado, ou *Federated Learning (FL)* em inglês, um paradigma de aprendizado em que o modelo é treinado de forma distribuída por diversos clientes e agregados por um servidor centralizado [McMahan et al. 2017]. Em FL, elimina-se a necessidade de compartilhamento dos dados brutos (*raw data*) que estão armazenados localmente nos dispositivos, compartilhando apenas os pesos dos modelos locais.

Vários desafios devem ser superados para que o FL seja efetivamente implantado, onde dois dos principais são destacados a seguir. *(i) Comunicação eficiente*: a comunicação frequente entre dispositivos e servidor central gera alto consumo de banda e energia. Técnicas como compressão, quantização e redução da frequência de sincronização ajudam a mitigar o problema. Isso melhora a escalabilidade e viabiliza o uso em redes com recursos limitados. *(ii) Qualidade do modelo*: distribuições de dados não balanceadas entre os dispositivos do treinamento federado dificultam a convergência e a generalização do modelo global. Ou seja, deve haver um equilíbrio entre esses fatores para a implementação efetiva de FL.

Considerando aspectos discutidos acima, este trabalho visa realizar uma avaliação exploratória das principais técnicas de compressão de modelos utilizadas no contexto de FL, considerando sua influência direta no equilíbrio entre qualidade do modelo e comunicação eficiente. Para isso, são realizados experimentos comparativos em ambientes de simulação, com o objetivo de coletar métricas que permitam avaliar e contrastar diferentes métodos do estado da arte. Especificamente, investigam-se os impactos das técnicas de compressão – poda, esparsificação e quantização – sobre dois conjuntos de dados amplamente utilizados na literatura: MNIST e Fashion-MNIST. Sendo assim, a principal contribuição deste trabalho está na demonstração empírica de que técnicas de compressão podem ser aplicadas em FL sem comprometer a acurácia do modelo, mesmo em cenários desbalanceados.

O restante do trabalho está organizado da seguinte forma. A Seção 2 apresenta o referencial teórico com os principais conceitos envolvidos no trabalho. A Seção 3 descreve a metodologia de avaliação e os resultados obtidos. Por fim, a Seção 5 apresenta as conclusões e direções para trabalhos futuros.

2. Referencial Teórico

Esta seção apresenta os principais conceitos e desafios em FL, bem como as técnicas de compressão utilizadas neste trabalho.

2.1. *Federated Learning (FL)*

O FL representa uma solução promissora para preservar a privacidade durante o treinamento colaborativo de modelos, evitando o compartilhamento de dados brutos entre

dispositivos. Essa solução combina conceitos de otimização distribuída, ML e proteção de privacidade, sendo aplicado em cenários diversos como saúde, finanças e Internet das Coisas (IoT) [Liu et al. 2024].

Entre os principais desafios enfrentados em FL estão a heterogeneidade e custo de comunicação. A presença de dados Não-Independentemente e Identicamente Distribuídos (Não-IID) impacta a convergência do modelo global [Lu et al. 2024], sendo causada pela diferença entre os dados dos clientes, denominada de *heterogeneidade estatística*. Dados Não-IID ocorrem quando as observações apresentam dependências ou seguem distribuições diferentes, violando a suposição de independência e igualdade de distribuição [Zhu et al. 2021]. Isso pode acontecer em situações com correlação temporal, espacial ou de contexto, tornando a análise estatística mais complexa.

Adicionalmente, os dispositivos dos clientes são heterogêneos, em termos de recursos e capacidades computacionais, representando a *heterogeneidade sistêmica* envolvida no processo de FL. Particularmente, o treinamento distribuído em si exige múltiplas trocas de atualizações de modelos, o que pode gerar aumento na sobrecarga de comunicação entre clientes e servidor durante o processo de treinamento [de Souza et al. 2024]. Assim, considerar a heterogeneidade e manter eficiência na utilização dos recursos de rede ainda são pontos de investigação em FL [Maciel et al. 2024].

2.2. Compressão de Modelos

A compressão de modelos considera o uso de técnicas que buscam reduzir a quantidade de dados transmitidos durante o treinamento, reduzindo a perda de acurácia global causada por essa compressão [Hoefler et al. 2021]. Dessa forma, a compressão de modelos é essencial para diminuir o custo de comunicação do treinamento, sendo um dos componentes importantes que definem a viabilidade do treinamento em FL.

Algumas das principais técnicas de compressão são apresentadas a seguir. *(1)* ***Quantização***, que diminui o número de bits usados para representar os parâmetros do modelo, reduzindo a precisão numérica e, por consequência, o volume de dados transmitidos. *(2)* ***Esparsificação***, que seleciona apenas os gradientes ou pesos mais relevantes para envio, ignorando ou comprimindo o restante. Essa abordagem reduz drasticamente o custo de comunicação entre dispositivos e servidor, economiza largura de banda e energia, além de melhorar a escalabilidade do sistema [Hoefler et al. 2021]. *(3)* ***Poda de Modelos***, que é uma técnica de simplificação que remove conexões ou neurônios menos relevantes de um modelo, reduzindo sua complexidade e tamanho. Isso melhora a eficiência computacional, diminuindo uso de memória, tempo de inferência e, em alguns casos, aumentando a generalização [Jiang et al. 2022].

A compressão desempenha um papel fundamental na comunicação eficiente no FL, pois reduz o volume de dados transmitidos entre clientes e servidor, mitigando o gargalo de comunicação que pode inviabilizar o treinamento colaborativo. Em cenários com recursos limitados, essas técnicas se tornam determinantes para garantir tanto a escalabilidade quanto a acessibilidade do processo de treinamento.

3. Metodologia Experimental

Esta seção descreve os componentes importantes da metodologia empregada.

3.1. Conjuntos de Dados

Ressalta-se a importância de se considerar conjunto de dados[1] com complexidades distintas para analisar o impacto que heterogeneidade de dados pode trazer sobre as abordagens. Nesse sentido, os conjuntos de dados considerados foram o MNIST (menos complexo) e o Fashion-MNIST (complexidade mediana). O primeiro é um conjunto de imagens de dígitos manuscritos (0-9), amplamente utilizado como *benchmark* para tarefas de classificação de imagens. Consiste em 70.000 imagens em escala de cinza, com tamanho de 28x28 *pixels*. O segundo conjunto de dados, Fashion-MNIST, foi proposto como alternativa de maior complexidade ao MNIST. Tal conjunto consiste em 70.000 imagens em tons de cinza de 28x28 *pixels*, contudo, cada imagem representa um item de vestuário ou acessório de moda, como camisetas, calçados, bolsas, entre outros, distribuídos em 10 categorias. Para ambos os conjuntos, foram analisados dois cenários: IID, em que os dados estão uniformemente distribuídos entre os clientes; e não-IID, em que os dados estão desbalanceados entre os clientes, tanto em relação às classes quanto à quantidade de amostras por classe. A seção a seguir detalha a distribuição de dados entre os clientes.

3.2. Ambiente Experimental

As simulações de uma aplicação, onde 10 clientes colaboram com o treinamento federado, foram realizadas com o arcabouço Flower, versão 1.14.0. A rede neural utilizada foi uma *MultiLayer Perceptron* (MLP), com quatro camadas descritas a seguir. *Camada de Entrada*, que corresponde ao formato dos dados de entrada do modelo, variando de acordo com o *dataset*. *Camada Flatten* (Linearização), que transforma os dados de entrada em um vetor unidimensional. *Camada Oculta*, que contém 128 neurônios e utiliza a função de ativação *ReLU*. Por fim, uma *Camada de Saída*, que corresponde à quantidade de classes de saída dos dados de entrada e utiliza a função de ativação *Softmax* para classificação.

Dois cenários distintos de distribuição de dados foram considerados a partir da distribuição de Dirichlet [Yurochkin et al. 2019]: um com parâmetro $\alpha = 10$ no cenário IID, outro com $\alpha = 0.5$ no cenário não-IID. Na distribuição de Dirichlet, quanto menor o valor de α, mais desbalanceados tendem a ser os dados. Ressalta-se que o desbalanceamento dos dados ocorre tanto em relação ao número de amostras quanto ao número de classes distribuídas entre os clientes. Por fim, os experimentos foram realizados em um *desktop* com processador Intel(R) Core(R) i7-8700 6-core, 16GB de RAM, GPU GeForce GTX 1060 6GB e sistema operacional NixOS 25.05 x86_64.

3.3. Técnicas de Compressão e Métricas de Avaliação

Para análise comparativa, as seguintes técnicas de compressão foram aplicadas: *(i) Poda*, poda de modelos baseada em magnitude; *(ii) Quantização*, quantização realizada pós-treinamento; *(iii) Esparsificação 15%*, onde a saída será 15% esparsa e é considerada uma esparsidade de nível baixo [Hoefler et al. 2021]; e *(iv) Quantização-QAT*, quantização ciente de treinamento (*Quantization-Aware Training*). As técnicas foram comparadas ao treinamento *Padrão*, em que não ocorre redução do modelo.

Para quantificar os resultados, foram utilizadas as seguintes métricas bem conhecidas da literatura. *(1) Acurácia*, que mede a proporção de predições corretas realizadas pelo modelo em relação ao total de exemplos avaliados, sendo crucial para avaliar o desempenho geral do modelo e sua capacidade de generalização. *(2) Perda*, que representa

[1] `https://flower.ai/docs/datasets/`

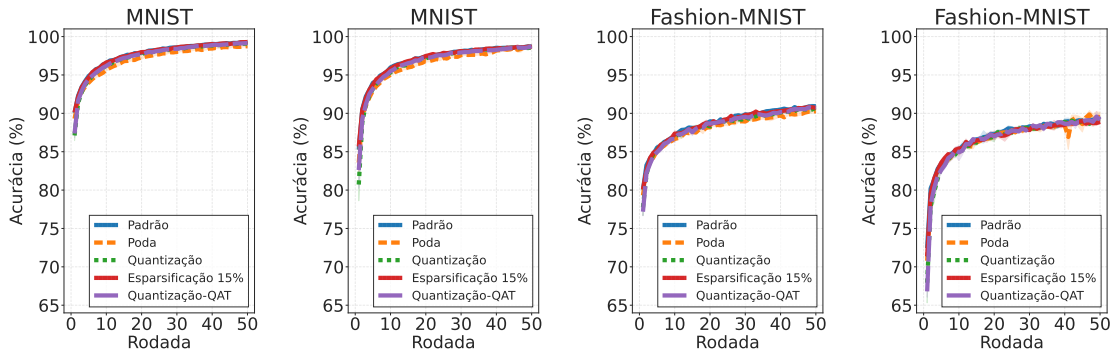
o erro cometido pelo modelo durante o treinamento e a validação, sendo obtida a partir de uma função de custo, que quantifica a discrepância entre as predições do modelo e os valores reais, portanto, auxiliando na avaliação da convergência e estabilidade do treinamento. *(3) Bytes transmitidos*, que indica a quantidade de dados transferidos (entre clientes e servidor) durante o treinamento, sendo relevante para avaliar o impacto das técnicas de compressão na redução do consumo de largura de banda.

4. Resultados Obtidos

Esta seção discute os resultados obtidos sobre a média de 10 execuções randomizadas.

4.1. Acurácia

A Figura 1 apresenta a evolução da acurácia do modelo global ao longo de 50 rodadas de comunicação para os diferentes cenários. A Figura 1(a) mostra os resultados do cenário IID e a Figura 1(b) os resultados do cenário não-IID. Para o conjunto MNIST, observa-se uma convergência rápida para altos níveis de acurácia (aproximadamente 97–99%) em ambos cenários. Todas as técnicas de compressão apresentaram desempenho muito próximo ao modelo Padrão (sem compressão). Embora a distribuição não-IID exiba uma variabilidade ligeiramente maior nas rodadas iniciais em comparação com a IID, a acurácia final converge para valores semelhantes, indicando robustez das abordagens à heterogeneidade dos dados. A Poda apresenta acurácia marginalmente inferior em alguns pontos, mas com mínima diferença.



(a) $\alpha = 10$ (b) $\alpha = 0.5$ (c) $\alpha = 10$ (d) $\alpha = 0.5$

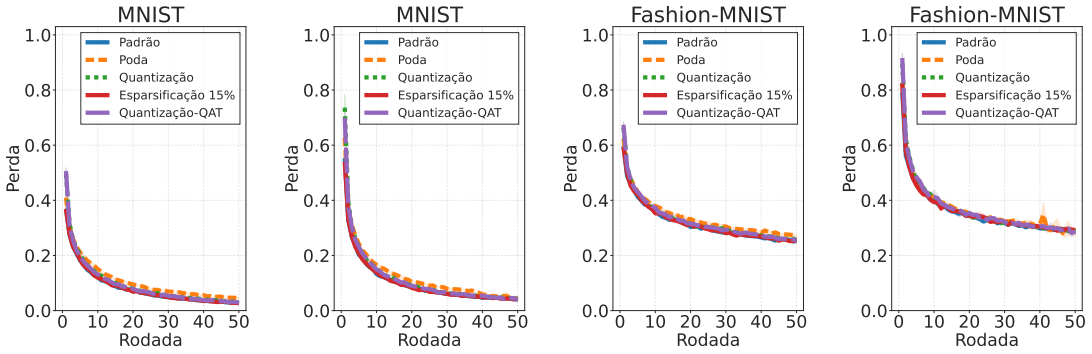
Figura 1. Acurácia do modelo com diferentes técnicas de compressão.

No conjunto Fashion-MNIST, a acurácia final atingida é inferior à de MNIST (cerca de 90–91%), como esperado considerando sua maior complexidade. A convergência também é ligeiramente mais lenta. Novamente, as diferentes técnicas de compressão acompanham o desempenho do modelo Padrão, tanto em cenários IID quanto não-IID, mostrados nas Figuras 1(c) e 1(d), respectivamente. A distribuição Não-IID parece introduzir uma leve instabilidade adicional na curva de aprendizado em comparação com a IID, mas sem impactar significativamente a acurácia final. Em suma, os resultados de acurácia sugerem que as técnicas de compressão avaliadas preservam o desempenho do modelo para ambos conjuntos e distribuições de dados, indicando potencial promissor para serem aplicadas em tarefas de maior complexidade.

4.2. Perda

A Figura 2 apresenta a evolução da perda do modelo global, complementando a análise de acurácia. Como esperado, as curvas de perda exibem um comportamento inverso ao da

acurácia, com uma redução acentuada nas rodadas iniciais seguida por uma estabilização. Para o MNIST a perda converge para valores muito baixos (inferiores a 0.1), refletindo a alta acurácia alcançada. As curvas para todas as técnicas de compressão são visualmente quase indistinguíveis da curva Padrão, tanto no cenário IID quanto no não-IID, mostrados nas Figuras 2(a) e 2(b), respectivamente. A ligeira variabilidade observada na acurácia não-IID também se reflete aqui, com uma convergência um pouco menos suave, mas atingindo níveis de perda finais similares aos do cenário IID.



(a) $\alpha = 10$ (b) $\alpha = 0.5$ (c) $\alpha = 10$ (d) $\alpha = 0.5$

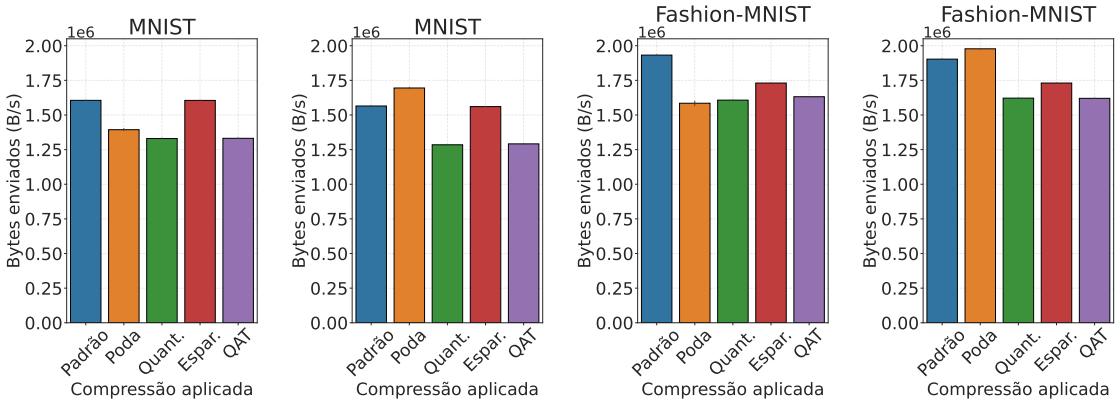
Figura 2. Perda do modelo com diferentes técnicas de compressão.

Para Fashion-MNIST, a perda estabiliza em um patamar mais elevado (aproximadamente 0.25 – 0.30), consistente com a menor acurácia obtida neste conjunto. Assim como no MNIST, as técnicas de compressão demonstram um comportamento de perda significativamente próxima ao do modelo Padrão, indicando que a compressão não introduziu um viés significativo no processo de otimização. A comparação entre IID novamente sugere que a heterogeneidade dos dados impacta mais a dinâmica da convergência do que o valor final da perda, conforme mostrado nas Figuras 2(c) e 2(d), respectivamente. Tais resultados reforçam que as técnicas de compressão mantêm a capacidade de aprendizado do modelo em níveis comparáveis ao cenário sem compressão, mesmo sob diferentes condições de distribuição de dados.

4.3. Bytes Transmitidos

A Figura 3 apresenta a quantidade de bytes enviados durante o treinamento para as diferentes técnicas de compressão. Focando inicialmente no conjunto MNIST com dados IID, Figura 3(a), é importante observar que a quantidade de bytes enviados na simulação Padrão (1.6×10^6 bytes) indica que o treinamento já apresenta intrinsecamente uma quantidade de bytes inferior. A baixa complexidade de MNIST exige menos conexões ativas para o modelo aprender suas características, tornando as atualizações inerentemente mais esparsas. Consequentemente, métodos como a Esparsificação 15%, que reduzem menos parâmetros por conta do baixo nível de esparsidade, mostram um impacto reduzido na comunicação para o MNIST, principalmente no cenário não-IID.

Em contraste com a esparsificação, as técnicas de Quantização e Quantização-QAT demonstram ser mais eficientes na redução de bytes para o MNIST (conforme mostrado nas Figuras 3(a) e 3(b)), alcançando $\approx 1.35 \times 10^6$ bytes. Isso ocorre porque seu mecanismo, baseado na redução da precisão numérica, independe da esparsidade estrutural do modelo. Já no Fashion-MNIST, o volume de transmissão Padrão é significativamente maior ($\approx 1.9 - 1.95 \times 10^6$ bytes), refletindo uma esparsidade natural muito menor devido



(a) $\alpha = 10$ (b) $\alpha = 0.5$ (c) $\alpha = 10$ (d) $\alpha = 0.5$

Figura 3. Bytes Transmitidos com diferentes técnicas de compressão.

à maior complexidade do *dataset*, conforme mostrado nas Figuras 3(c) e 3(d), respectivamente. Nesse cenário, a Esparsificação 15% obtém uma redução mais significativa (para $\approx 1.7 \times 10^6$ bytes), pois há mais parâmetros não nulos sobre os quais a técnica pode efetivamente operar. As abordagens de Quantização mantêm sua alta e consistente eficiência também no Fashion-MNIST ($\approx 1.6 \times 10^6$ bytes).

Por fim, a análise revela um comportamento peculiar e sensível à distribuição dos dados para a técnica de Poda. Por um lado, a Poda reduziu eficientemente a comunicação em cenários IID para ambos os conjuntos: na Figura 3(a), $\approx 1.4 \times 10^6$ sobre MNIST; na Figura 3(c) $\approx 1.55 \times 10^6$ sobre Fashion-MNIST. Por outro lado, a técnica falha consideravelmente em ambientes não-IID, Figura 3(b) e Figura 3(d), onde a quantidade de bytes transmitidos se iguala ou até supera a do modelo Padrão ($\approx 1.95 \times 10^6$). Uma hipótese plausível é que a heterogeneidade dos dados leva os clientes a podarem conjuntos diferentes de pesos, e a agregação subsequente dessas estruturas esparsas não coincidentes resulta em uma representação densa ou ineficiente no servidor, anulando o benefício da compressão local. As demais técnicas mostram-se robustas nesse aspecto, com pouca variação nos bytes transmitidos entre IID e não-IID.

5. Conclusão

Este trabalho apresenta uma avaliação exploratória focada na aplicação de técnicas de compressão com o objetivo de maximizar a eficiência da comunicação no contexto de FL. Os resultados mostram que a redução no volume de dados transmitidos contribui para uma maior eficiência do sistema, sem acarretar perdas significativas de desempenho.

Como trabalhos futuros, pretende-se propor uma técnica de compressão de modelos adaptativa, capaz de ajustar dinamicamente o nível de compressão com base nas condições da rede, nas características dos dispositivos participantes e na sensibilidade das atualizações do modelo, visando otimizar o *trade-off* entre eficiência de comunicação e desempenho do modelo. Propõe-se ainda a utilização de conjuntos de dados mais complexos, como CIFAR-10 e ImageNet. Por fim, pretende-se explorar outras abordagens de compressão, como fatoração *low-rank*, destilação de conhecimento e codificação baseada em entropia, além da aplicação de níveis mais altos de esparsificação.

Agradecimentos

Este projeto foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei nº 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado Arquitetura Cognitiva (Fase 3), DOU 01245.003479/2024-10.

Referências

[Alves et al. 2024] Alves, V. R. M., da Costa, J. B. D., Gonzalez, L., de Souza, A. M., and Villas, L. (2024). Seleção de clientes adaptativa baseada em privacidade diferencial para aprendizado federado. In *Anais Estendidos do XLII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 225–232. SBC.

[de Souza et al. 2024] de Souza, A. M., Maciel, F., da Costa, J. B. D., Bittencourt, L. F., Cerqueira, E., Loureiro, A. A., and Villas, L. A. (2024). Adaptive client selection with personalization for communication efficient federated learning. *Ad Hoc Networks*, 157:103462.

[Ficco et al. 2024] Ficco, M., Guerriero, A., Milite, E., Palmieri, F., Pietrantuono, R., and Russo, S. (2024). Federated learning for iot devices: Enhancing tinyml with on-board training. *Information Fusion*, 104:102189.

[Hoefler et al. 2021] Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., and Peste, A. (2021). Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124.

[Jiang et al. 2022] Jiang, Y., Wang, S., Valls, V., Ko, B. J., Lee, W.-H., Leung, K. K., and Tassiulas, L. (2022). Model pruning enables efficient federated learning on edge devices. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10374–10386.

[Liu et al. 2024] Liu, B., Lv, N., Guo, Y., and Li, Y. (2024). Recent advances on federated learning: A systematic survey. *Neurocomputing*, page 128019.

[Lu et al. 2024] Lu, Z., Pan, H., Dai, Y., Si, X., and Zhang, Y. (2024). Federated learning with non-iid data: A survey. *IEEE Internet of Things Journal*.

[Maciel et al. 2024] Maciel, F., de Souza, A. M., Bittencourt, L. F., Villas, L. A., and Braun, T. (2024). Federated learning energy saving through client selection. *Pervasive and Mobile Computing*, 103:101948.

[McMahan et al. 2017] McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

[Sabah et al. 2024] Sabah, F., Chen, Y., Yang, Z., Azam, M., Ahmad, N., and Sarwar, R. (2024). Model optimization techniques in personalized federated learning: A survey. *Expert Systems with Applications*, 243:122874.

[Thakur et al. 2025] Thakur, D., Guzzo, A., Fortino, G., and Piccialli, F. (2025). Green federated learning: A new era of green aware ai. *ACM Computing Surveys*, 57(8).

[Yurochkin et al. 2019] Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., and Khazaeni, Y. (2019). Bayesian nonparametric federated learning of neural networks. In *International conference on machine learning*, pages 7252–7261. PMLR.

[Zhu et al. 2021] Zhu, H., Xu, J., Liu, S., and Jin, Y. (2021). Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390.