Learning Without Sharing: A Comparative Study of Federated Learning Models for Healthcare

Anja Campmans[1], Mina Alishahi[1] and Vahideh Moghtadaiee[2]

[1]*Department of Computer Science, Open Universiteit, Amsterdam, The Netherlands*

[2]*Cyberspace Research Institute, Shahid Beheshti University, Tehran, Iran*

Keywords: Federated Learning, Health Data, Privacy, Accuracy.

Abstract: Federated Learning (FL) has emerged as a powerful approach for training machine learning (ML) models on decentralized healthcare data while maintaining patient privacy. However, selecting the most suitable FL model remains a challenge due to inherent trade-offs between accuracy and privacy. This study presents a comparative analysis of multiple FL optimization strategies applied to two real-world tabular health datasets. We evaluate the performance of FL models in terms of predictive accuracy, and resilience to privacy threats.Our findings provide insights into the practical deployment of FL in healthcare, highlighting key trade-offs and offering recommendations for selecting suitable FL models based on specific privacy and accuracy requirements.

1 INTRODUCTION

The increasing digitization of healthcare data has opened new opportunities for developing machine learning (ML) models that improve diagnostics, treatment planning, and patient care. However, the sensitive nature of medical records, coupled with stringent privacy regulations such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA), restricts the direct sharing of health data (Gaballah et al., 2024). Traditional centralized ML approaches, where data from multiple sources is aggregated into a single repository for model training, pose significant privacy risks (Sheikhalishahi et al., 2022). These concerns have fueled interest in Federated Learning (FL) as a privacy-preserving alternative (Nguyen et al., 2021), (Tasbaz et al., 2024). FL enables multiple institutions or devices to collaboratively train models without exposing raw data, thereby maintaining data confidentiality while leveraging distributed learning.

Despite its promise, FL introduces several challenges, particularly in healthcare applications. One of the primary concerns is the trade-off between privacy and accuracy. Privacy-enhancing techniques, such as differential privacy (DP) and secure aggregation (SA) mitigate risks associated with data exposure but often come at the cost of reduced model performance. Additionally, the decentralized and non-independently and non-identically distributed (non-IID) nature of healthcare data creates significant obstacles related to model convergence, communication efficiency, and susceptibility to adversarial threats such as model poisoning attacks (Torki et al., 2025). Addressing these challenges is crucial for enabling the widespread adoption of FL in medical AI.

While previous studies have explored FL's applicability in healthcare (Ouadrhiri and Abdelhadi, 2022)(Zhang et al., 2023), existing works have primarily focused on either privacy-preserving mechanisms or performance optimization in isolation (Coelho et al., 2023)(Hernandez et al., 2022). To the best of our knowledge, there is no comprehensive comparative analysis that systematically evaluates multiple FL models across different privacy-accuracy trade-offs in the context of healthcare using tabular datasets. This study fills that gap by empirically assessing various FL optimization strategies on real-world health tabular datasets, providing a multi-faceted evaluation of their performance focusing on three key aspects:

- Model accuracy: Assessing the generalization capability of each FL model to unseen data in a decentralized healthcare setting.
- Privacy preservation: Evaluating the effectiveness

Campmans, A., Alishahi, M. and Moghtadaiee, V.
Learning Without Sharing: A Comparative Study of Federated Learning Models for Healthcare.
DOI: 10.5220/0013635000003979
In *Proceedings of the 22nd International Conference on Security and Cryptography (SECRYPT 2025)*, pages 735-740
ISBN: 978-989-758-760-3; ISSN: 2184-7711

of different privacy-enhancing mechanisms in safeguarding patient data and resilience to attacks.

- Trade-off: Analyzing how different FL models balance predictive performance with privacy protection, offering insights into the inherent compromises in federated healthcare applications.

By systematically analyzing these factors, this paper provides empirical evidence on the strengths and limitations of various FL models based on healthcare-specific requirements.

2 PRELIMINARIES

Federated Learning (FL) involves training a global model across multiple decentralized clients without sharing raw data. Various FL optimization algorithms have been proposed to enhance convergence speed, robustness, and privacy. Below are the definitions of the FL models in this study summarized in Table 1.

2.1 Federated Learning Models

FedAvg (Federated Averaging): FedAvg, introduced by McMahan et al. (2016), aggregates local model updates by computing a weighted average of local model parameters. Given K clients, each with a dataset partition P_k of size n_k, the global objective is formulated as:

$$\min_{w\in\mathbb{R}^d} \quad f(w)=\sum_{k=1}^{K}\frac{n_k}{n}F_k(w), \tag{1}$$

where w represents the model parameters, $f(w)$ is the loss function of the global model, and $F_k(w)=\frac{1}{n_k}\sum_{i\in P_k} f_i(w)$ is the local loss function for client k. The total number of data samples across all clients is given by $n=\sum_{k=1}^{K} n_k$. Each client trains a local model and sends the updated parameters back to the server, which performs averaging to update the global model.

FedOpt (Federated Optimization) (Reddi et al., 2020): FedOpt extends FedAvg by allowing the use of adaptive optimizers like Adam, Yogi, and Adagrad at the server level. This flexibility helps improve convergence and performance, especially with non-IID data. Unlike FedAvg, which uses simple averaging, FedOpt enables more update strategies that adjust learning rates based on gradient history or gradient magnitude:

$$\Delta w^{(t)}=\sum_{k=1}^{K}\frac{n_k}{N}(w_k^{(t)}-w^{(t)}), \tag{2}$$

$$w^{(t+1)}=w^{(t)}-\eta G(\Delta w^{(t)}), \tag{3}$$

where $G(\cdot)$ represents an adaptive optimization function such as Adam or Adagrad. It is effective in scenarios with noisy updates or uneven data distributions.

FedAdagrad (Duchi et al., 2011): An adaptation of FedAvg that incorporates the Adagrad optimization technique to adjust learning rates based on the accumulation of past squared gradients. In FedAdagrad, local updates are computed similarly to FedAvg, but each parameter's learning rate is adjusted dynamically during the training process, ensuring that rare features are updated more aggressively while frequently updated parameters are adjusted more cautiously. This makes it effective in handling heterogeneous or imbalanced data distributions in FL.

$$w^{t+1}=w^t-\frac{\eta}{\sqrt{G^t+\varepsilon}}\cdot g^t, \tag{4}$$

where $G^t=\sum_{\tau=1}^{t}(g^\tau)^2$ accumulates past squared gradients, and ε prevents division by zero.

FedAdam (Kingma and Ba, 2017): FedAdam is a variant of FedAvg that incorporates the Adam optimizer, which adapts learning rates for each parameter based on both first-order and second-order moments of the gradient. In traditional Adam, the first moment estimate tracks the average gradient (momentum), while the second moment tracks the uncentered variance (squared gradient). It provides better convergence by taking into account the gradient's history, allowing for faster and more robust convergence, especially with non-IID data:

$$m^{t+1}=\beta_1 m^t+(1-\beta_1)g^t, \tag{5}$$

$$v^{t+1}=\beta_2 v^t+(1-\beta_2)(g^t)^2, \tag{6}$$

$$\hat{m}^{t+1}=\frac{m^{t+1}}{1-\beta_1^{t+1}}, \quad \hat{v}^{t+1}=\frac{v^{t+1}}{1-\beta_2^{t+1}}, \tag{7}$$

$$w^{t+1}=w^t-\frac{\eta}{\sqrt{\hat{v}^{t+1}}+\varepsilon}\hat{m}^{t+1}. \tag{8}$$

FedYogi (Zaheer et al., 2018): Similar to FedAdam, FedYogi also adapts learning rates based on first-order (momentum) and second-order (variance) gradient estimates, but with a key difference in how updates are performed. In Adam, large gradients can lead to aggressive updates, which may cause instability, especially in FL environments with non-IID data. FedYogi mitigates this by using a Yogi-style update rule, which reduces the risk of large, destabilizing updates when gradients are large. Specifically, FedYogi maintains more conservative updates by modifying how the second moment is updated, ensuring stability across varying gradient scales for noisy or highly heterogeneous data.

FedAvgM (FedAvg with Momentum) (Hsu et al., 2019): An extension of FedAvg that incorporates momentum into the update rule, inspired by the traditional stochastic gradient descent (SGD) with momentum. In FedAvgM, the server maintains a velocity term that accumulates the gradients over time, allowing it to smooth out fluctuations in the updates. This is particularly useful in FL, where local client updates may be noisy or vary due to non-IID data. By using momentum, FedAvgM accelerates convergence by moving faster in directions with consistent gradients and damping oscillations in directions with conflicting updates. This helps reduce the number of communication rounds needed for the global model to converge.

FedMedian (Yin et al., 2018): A robust version of FedAvg that replaces the average of local model updates with the median of updates. In FL, client updates may contain outliers or even be malicious (in the case of adversarial clients). FedMedian mitigates this by taking the element-wise median of the local model updates instead of the average, which reduces the influence of extreme values. This provides robustness against outlier updates, making FedMedian well-suited for environments where some clients may provide noisy or untrustworthy updates.

FedProx (Sahu et al., 2018): FedProx extends FedAvg by introducing a proximal term in the local optimization objective. In FL, client devices may have differing datasets, leading to diversity in local updates. This can destabilize the training process, particularly when client models diverge significantly from the global model. FedProx stabilizes training by adding a proximal term that penalizes large deviations between the client's local model and the global model, effectively regularizing the local updates:

$$F_k^{\text{prox}}(w) = F_k(w) + \frac{\mu}{2}\|w - w^t\|^2, \tag{9}$$

where μ is the regularization parameter that controls the strength of the penalty. By tuning μ, FedProx can handle client heterogeneity more effectively, ensuring that local updates do not diverge too far from the global model, which improves convergence in non-IID settings.

FedTrimmedAvg (Yin et al., 2018): Similar to FedMedian, this method enhances robustness by performing a trimmed mean of the local updates, excluding extreme values before averaging. In FedTrimmedAvg, a portion of the smallest and largest values of the client updates are discarded, and the remaining updates are averaged. This is useful in situations where outlier updates can disrupt the training process, as it ensures that adversarial or noisy client contributions are removed from the aggregation process, resulting in a more stable global model.

Table 1: FL models summary.

Algorithm	Key Feature	Strength
FedAvg	Averaging local updates	Simple and efficient
FedOpt	Adaptive optimizers	Better convergence
FedAdagrad	Learning rate adjustment	Handles sparse data
FedAdam	Momentum + adaptive rates	Fast convergence
FedYogi	Controlled variance	More stable updates
FedAvgM	Momentum-based averaging	Smoother updates
FedMedian	Median aggregation	Robust to outliers
FedProx	Proximal term	Handles client heterogeneity
FedTrimmedAvg	Trimmed mean aggregation	Adversarial robustness

3 METHODOLOGY

To evaluate the performance of different FL approaches in a healthcare setting, we experiment with multiple FL optimization methods: FedAdagrad, FedAdam, FedYogi, FedAvg, FedOpt, FedAvgM, FedMedian, FedProx, and FedTrimmedAvg. These models vary in their optimization strategies, ranging from adaptive gradient-based methods (FedAdagrad, FedAdam, FedYogi) to classical aggregation techniques (FedAvg, FedAvgM) and robust aggregation methods designed to mitigate the impact of noisy or adversarial updates (FedMedian, FedTrimmedAvg). Additionally, FedProx introduces a regularization term to improve convergence in heterogeneous data settings, while FedOpt provides a generalized optimization framework for FL.

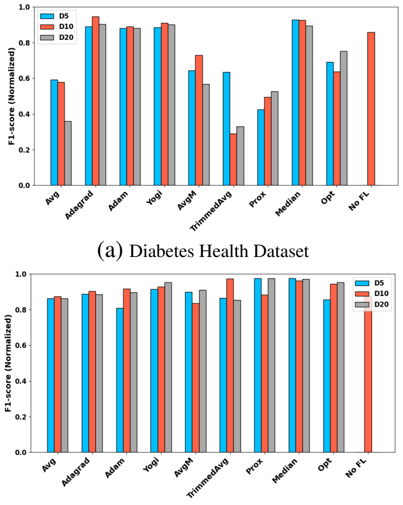
We conduct experiments on two tabular health datasets, where features include patient demographics, medical history, and clinical outcomes. To simulate realistic FL scenarios, data is distributed non-IID among agents, reflecting variations in patient populations across institutions. We evaluate the models under three different settings: 5, 10, and 20 federated agents, representing increasing levels of decentralization and data fragmentation. To assess the effectiveness of each FL model, we consider both accuracy and privacy trade-offs (Sheikhalishahi and Martinelli, 2018).

Accuracy Metric: We use the F1-score, which balances precision and recall, making it well-suited for imbalanced medical datasets where false positives and false negatives have significant implications.

Privacy Metric: We evaluate membership inference attack (MIA) vulnerability, which quantifies how well an adversary can determine whether a given sample was part of the training set. A higher MI success rate indicates a greater privacy risk.

$$\mathbb{A} = \left|\Pr[\hat{M} = 1 \mid M = 1] - \Pr[\hat{M} = 1 \mid M = 0]\right| \tag{10}$$

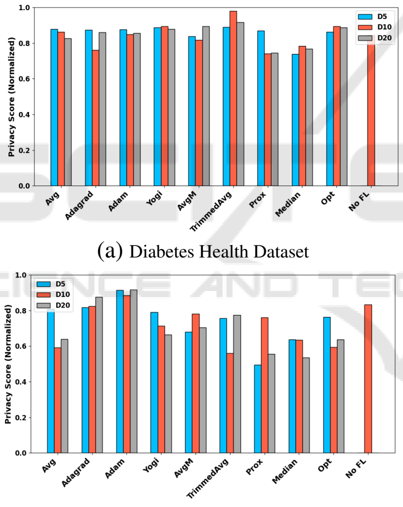
where $M = 1$ means the sample was in the training set, and $M = 0$ means it was not. $\hat{M}$ is the attacker's



(a) Diabetes Health Dataset

(b) Covid19 Dataset

Figure 1: Accuracy (F1 score) results on Diabetes and Covid19 datasets when data is distributed among 5, 10, and 20 parties.



(a) Diabetes Health Dataset

(b) Covid19 Dataset

Figure 2: Privacy (MIA) results on Diabetes and Covid19 datasets when data is distributed among 5, 10, and 20 parties.

guess (1 for member, 0 for non-member). The probabilities represent the attack model's confidence in distinguishing members from non-members.

Each FL model is trained under the different agent configurations (5, 10, and 20 agents denoted as D5, D10, and D20), ensuring consistent hyperparameter settings across experiments. The evaluation framework measures the trade-off between F1-score and privacy leakage for each optimization method, highlighting the impact of FL aggregation strategies on both predictive performance and patient data protection.

4 EXPERIMENTAL RESULTS

In this section, we present the datasets used in this study, and the experimental results including the accuracy, privacy, and trade-off results (codes here[1]).

Two well-known tabular medical related datasets have been used in this study: **1) Diabetes Health Indicators dataset[2]:** The first dataset is the Diabetes Health Indicators dataset. This dataset contains 253680 patient records of 21 input variables, and a binary output label. These 21 input variables contain 14 binary variables, and 7 numerical variables. **2) COVID-19 Dataset[3]:** The second dataset is the COVID-19 dataset, which contains 263007 patient records, each containing 18 binary variables, 5 numerical variables and a binary output variable.

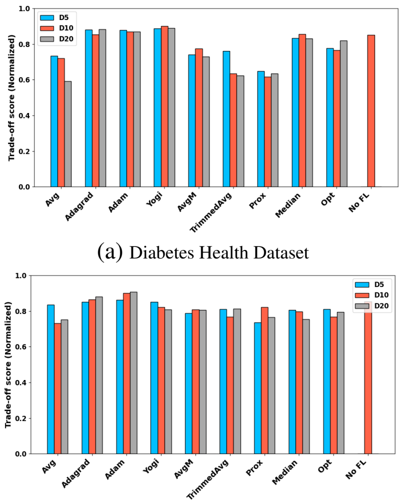
4.1 Accuracy Results

Figure 1 presents the F1-score comparison for the Diabetes Health Indicators and COVID-19 datasets, revealing distinct performance patterns across FL models. Models perform significantly better on the COVID-19 dataset, maintaining higher and more stable F1-scores across D5, D10, and D20, likely due to more uniform data distributions across federated clients, making it less susceptible to performance degradation as decentralization increases. In contrast, the Diabetes dataset experiences a sharp accuracy drop at D20, particularly for Avg and TrimmedAvg, indicating greater learning challenges from non-IID distributions. The best-performing models vary by dataset. For Diabetes, Median, Adagrad, and Yogi achieve the highest F1-scores at lower decentralization levels, while Prox struggles, suggesting its regularization is less effective in this setting. For COVID-19, Prox, Median, and Yogi consistently perform well, with Prox maintaining an F1-score near 0.97, benefiting from its stabilization technique. Therefore, Median aggregation proves robust across both datasets, while adaptive optimizers (Adagrad, Adam, Yogi) show dataset-dependent variations. Prox and Yogi excel in COVID-19, while Adagrad is better suited for Diabetes, emphasizing the need for dataset-specific FL model selection in healthcare applications.

[1] https://github.com/Anja-Hobby/evalFLonTabularDatasets

[2] https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

[3] https://github.com/marianarf/covid19_mexico_data



(a) Diabetes Health Dataset

(b) Covid19 Dataset

Figure 3: Trade-off (privacy-accuracy) results on Diabetes and Covid19 datasets when data is distributed among 5, 10, and 20 parties.

4.2 Privacy Results

Figure 2 compares privacy scores for the Diabetes and COVID-19 datasets across different decentralization levels (D5, D10, D20). FL models exhibit higher privacy risks in the Diabetes dataset, where most models have elevated privacy scores, indicating greater vulnerability to privacy leakage. In contrast, the COVID-19 dataset shows lower privacy scores, suggesting stronger resistance to privacy attacks, likely due to its data distribution characteristics. For Diabetes, TrimmedAvg (0.9799 at D10) and Avg (0.8774 at D5) are the most privacy-vulnerable, whereas Median (0.7389 at D5, 0.7688 at D20) and Prox (0.7405 at D10, 0.7441 at D20) offer better privacy protection, though Prox trades off some accuracy. In the COVID-19 dataset, Prox and Median achieve the lowest privacy scores (0.5–0.64), while Avg and Adam (0.91) remain more privacy-exposed. TrimmedAvg shows the largest fluctuations in privacy risks, particularly at D10. Overall, FL models face greater privacy risks in the Diabetes dataset, while Prox and Median consistently offer stronger privacy protection. These findings emphasize the importance of selecting FL models that balance privacy and accuracy based on dataset characteristics.

4.3 Trade-off Results

Table 2: Guidelines for selecting FL models in healthcare.

Objective	Recommended Models
High Model Accuracy	FedAdam, FedYogi
Strong Privacy Preservation	FedProx, FedMedian
Privacy-Accuracy Trade-off	FedAvg + DP, FedAdagrad
Handling Non-IID Data	FedProx, FedYogi
Communication Efficiency	FedAvg, FedMedian
Adversarial Robustness	FedMedian, FedYogi

The trade-off score results for the Diabetes and COVID-19 datasets, shown in Figure 3, highlighting how FL models balance accuracy and privacy. For the Diabetes dataset, trade-off scores generally decline as the number of nodes increases (Avg: 0.7345 at D5 to 0.5926 at D20), consistent with previous findings. However, Yogi (0.8864 → 0.8898) and Adagrad (0.8816 → 0.8818) maintain stable scores, indicating effective balance. Median and Adam also perform well, with scores above 0.83. In the COVID-19 dataset, Adam excels (0.8618 → 0.9071), while Adagrad and Yogi perform well initially, but Yogi's score drops at D20 (0.8088). TrimmedAvg and Prox, which struggled with Diabetes, improve here, with Prox peaking at 0.8217 at D10. However, Median's score declines (0.8059 → 0.7536), making it less favorable for COVID-19. Hence, Adam, Adagrad, and Yogi show the best trade-offs, with Adam excelling in COVID-19 and Yogi remaining stable for Diabetes. Median is strong for Diabetes but weaker for COVID-19, while Prox improves in COVID-19 but remains less effective in Diabetes. These results emphasize the need for dataset-specific FL model selection.

4.4 Guidelines

Choosing the right FL model depends on trade-offs between accuracy, privacy, communication efficiency, and robustness to data heterogeneity. Table 2 outlines selection guidelines for federated healthcare applications.

- For high accuracy, FedAdam and FedYogi are preferred for their adaptive learning rates, enhancing convergence in non-IID settings, ideal for tasks like disease diagnosis and clinical decision support.
- When privacy is key, FedProx and FedMedian offer stronger protection, and FedProx ensures stable training in diverse settings, while FedMedian resists adversarial attacks, making them ideal for sensitive tasks like personalized healthcare.
- To balance privacy and accuracy, FedAvg with DP and FedAdagrad provide practical solutions—FedAvg with DP protects confidentiality with solid performance, while FedAdagrad ensures stability across diverse

clients, making them suitable for federated health monitoring.

- In heterogeneous data settings, FedProx and FedYogi are the best choices. FedProx prevents divergence by regularizing updates, while FedYogi adapts learning rates for client variability, making them ideal for multi-institutional studies and mobile health.
- For communication efficiency, FedAvg and FedMedian minimize overhead. FedAvg reduces update frequency, and FedMedian boosts robustness, making them suitable for remote healthcare and edge-based AI.
- For adversarial robustness, FedMedian and FedYogi resist malicious updates. FedMedian mitigates adversarial impact, while FedYogi stabilizes training, ideal for secure wearable health and remote diagnosis.

By aligning FL model selection with these criteria, healthcare practitioners and researchers can ensure optimal performance while maintaining privacy and efficiency in federated medical AI systems.

5 CONCLUSION

This study evaluated various federated learning (FL) optimization strategies on tabular health datasets, analyzing accuracy, privacy, and trade-offs. Adaptive optimizers like *FedYogi* and *FedAdam* showed improved performance in non-IID settings, while robust aggregation methods like *FedMedian* and *FedTrimmedAvg* enhanced resilience against noisy updates. The findings emphasize that no single FL model is universally optimal; instead, selection should be guided by specific healthcare requirements, such as the need for high accuracy or privacy. Future work should explore stricter privacy mechanisms, such as differential privacy, and assess scalability in real-world deployments. We also plan to explore the performance of FL models, when data is distributed vertically rather than horizontally.

REFERENCES

Coelho, K. K., Nogueira, M., Vieira, A. B., Silva, E. F., and Nacif, J. A. M. (2023). A survey on federated learning for security and privacy in healthcare applications. *Computer Communications*, 207:113–127.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159.

Gaballah, S. A., Abdullah, L., Alishahi, M., Nguyen, T. H. L., Zimmer, E., Mühlhäuser, M., and Marky, K. (2024). Anonify: Decentralized dual-level anonymity for medical data donation. *Proc. Priv. Enhancing Technol.*, 2024(3):94–108.

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., and Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45.

Hsu, T. H., Qi, H., and Brown, M. (2019). Measuring the effects of non-identical data distribution for federated visual classification. *CoRR*, abs/1909.06335.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Nguyen, D. C., Pham, Q.-V., Pathirana, P. N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O. A., and Hwang, W.-J. (2021). Federated learning for smart healthcare: A survey.

Ouadrhiri, A. E. and Abdelhadi, A. (2022). Differential privacy for deep and federated learning: A survey. *IEEE Access*, 10:22359–22380.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan, H. B. (2020). Adaptive federated optimization. *CoRR*, abs/2003.00295.

Sahu, A. K., Li, T., Sanjabi, M., Zaheer, M., Talwalkar, A., and Smith, V. (2018). On the convergence of federated optimization in heterogeneous networks. *CoRR*, abs/1812.06127.

Sheikhalishahi, M. and Martinelli, F. (2018). Privacy-utility feature selection as a tool in private data classification. In *Distributed Computing and Artificial Intelligence, 14th International Conference*, pages 254–261. Springer International Publishing.

Sheikhalishahi, M., Saracino, A., Martinelli, F., and Marra, A. L. (2022). Privacy preserving data sharing and analysis for edge-based architectures. *Int. J. Inf. Sec.*, 21(1):79–101.

Tasbaz, O., Moghtadaiee, V., and Farahani, B. (2024). Feature fusion federated learning for privacy-aware indoor localization. *Peer-to-Peer Networking and Applications*, 17:2781–2795.

Torki, O., Ashouri-Talouki, M., and Alishahi, M. (2025). Fed-gwas: Privacy-preserving individualized incentive-based cross-device federated gwas learning. *Journal of Information Security and Applications*, 89:104002.

Yin, D., Chen, Y., Ramchandran, K., and Bartlett, P. L. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. *CoRR*, abs/1803.01498.

Zaheer, M., Reddi, S., Sachan, D., Kale, S., and Kumar, S. (2018). Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Zhang, F., Kreuter, D., Chen, Y., Dittmer, S., Tull, S., Shadbahr, T., Collaboration, B., Preller, J., Rudd, J. H. F., Aston, J. A. D., Schönlieb, C.-B., Gleadall, N., and Roberts, M. (2023). Recent methodological advances in federated learning for healthcare.