Dynamic Trust-Aware Federated Aggregation for Privacy-Aware Multicentric Healthcare AI

Elarbi Badidi[1,*], Omar Elharrouss[1]

[1]Department of Computer Science and Software Engineering, College of Information Technology, UAE University, Al-Ain, UAE

Abstract

Multicentric healthcare AI must learn across hospitals, clinics, laboratories, and edge medical devices without pooling sensitive patient data. Federated learning (FL) supports this goal, but clinical deployments face non-IID site populations, unequal compute, intermittent participation, misconfigured clients, and poisoning attacks. Existing aggregation schemes often weight updates by sample count alone, while trust systems usually act as pre-selection filters rather than live aggregation signals. This paper proposes Dynamic Trust-Aware Federated Aggregation (DTAFA), which evaluates each participating site through gradient quality, behavioral consistency, and data integrity, then smooths the resulting trust score across rounds. Aggregation weights combine cohort size, smoothed trust, and a current-round quality gate, suppressing unreliable contributors without permanently excluding sites after transient failures. A NumPy evaluation on CIFAR-10 and Shakespeare, used as privacy-safe proxies for visual and sequential healthcare workloads, shows that DTAFA improves over FedAvg by up to 28.7 percentage points under sign-flip Byzantine attacks and remains competitive with coordinate-wise median aggregation at 1.55× FedAvg server runtime.

Keywords

Multicentric healthcare AI, Privacy-aware federated learning, Trust-aware aggregation, Non-IID data, Poisoning attack resilience

1. Introduction

Healthcare AI increasingly depends on evidence distributed across hospitals, clinics, laboratories, and edge devices that observe different populations, acquisition protocols, coding practices, and resource constraints. Federated learning (FL) is attractive because sites exchange model updates rather than raw patient records [1]. However, multicentric FL also combines non-IID case mixes, uneven compute and network capacity, intermittent participation, and the risk of misconfigured or malicious clients [2, 3].

Standard aggregation weights updates by local dataset size, so a large institution can dominate even when its distribution is shifted or its update is corrupted [1]. FedProx reduces drift through proximal regularization, but it does not decide whether a site's current contribution should be trusted. Trust systems based on fuzzy scoring, reputation, or blockchain auditing are often applied before training or after the fact, decoupling trust evaluation from the aggregation step [4].

DTAFA closes this gap by embedding adaptive trust directly into per-round aggregation. Its main contributions are: (i) a three-signal trust score over gradient quality, behavioral consistency, and data integrity; (ii) quality-gated, trust-scaled weights that reduce unreliable influence without permanent exclusion; (iii) trust-informed site selection with periodic re-evaluation; and (iv) a reproducible benchmark on non-clinical visual and sequential workloads.

The rest of this paper is organized as follows: Section II surveys related work. Section III defines the system model. Section IV details DTAFA. Section V presents experimental results. Section VI concludes the paper.

Joint Proceedings of the AIME 2026 Workshops: 1st International Workshop on Multicentric and Privacy-preserving Learning in Healthcare, Foundation Models for Public Health and Epidemiology: From Promise to Practice, and First International Workshop on Knowledge Graphs for Health, July 10, 2026, Ottawa, Canada

*Corresponding author.

†These authors contributed equally.

ebadidi@uaeu.ac.ae (E. Badidi); omar.elharrouss@uaeu.ac.ae (O. Elharrouss)

https://faculty.uaeu.ac.ae/ebadidi/ (E. Badidi); https://research.uaeu.ac.ae/en/persons/omar-el-harrouss/ (O. Elharrouss)

0000-0001-9121-8766 (E. Badidi); 0000-0002-5341-5440 (O. Elharrouss)

2. Related Work

Healthcare FL has progressed from feasibility studies to clinically motivated multicentric workflows. Multi-institutional brain-tumor segmentation showed that collaborative learning can be performed without sharing patient records, while later studies applied FL to rare cancer boundary detection and electronic-health-record mortality prediction [5, 6, 7]. These studies demonstrate the value of keeping data at local institutions, but they do not address how a federation should change the influence of sites whose updates become unreliable, shifted, or adversarial during training.

A second line of work studies the optimization challenges created by heterogeneous clients and non-IID data. Broad FL surveys identify statistical heterogeneity, resource imbalance, privacy, and communication constraints as persistent barriers to deployment [8, 9]. FedProx addresses systems and statistical heterogeneity by adding a proximal term to local training [10], and asynchronous FL surveys describe how device variability affects round scheduling and convergence [2]. These approaches reduce drift or improve participation under heterogeneity, yet they do not evaluate whether a large healthcare site should be trusted in a particular aggregation round.

Robust FL has also examined poisoning and Byzantine behavior. Fang et al. show that local model poisoning can undermine Byzantine-robust aggregation rules [11], while FLTrust uses server-side trust bootstrapping to constrain malicious updates [12]. Median- and Krum-style methods, including Median-Krum variants, improve robustness through distance or statistical filtering [13]; however, such rules typically operate on the current update set and do not maintain institution-level trust histories. Ding et al. deploy a dual-layer anomaly detector for backdoor attacks [3], and graph attention or layer-wise aggregation can improve robustness or non-IID convergence [14, 15]. These methods are complementary to DTAFA, but they do not directly couple multi-signal trust to sample-size-aware aggregation weights.

Trust management for FL is closest to the present work. Cheng et al. score devices through multifactor fuzzy membership before training, but their trust rules are static and do not feed back into aggregation [4]. Hathout et al. adapt participation eligibility to data poisoning in MEC-FL [16], and Rjoub et al. use reinforcement learning to guide client selection from accumulated behavioral evidence [17]; once clients are selected, however, aggregation still loses much of the trust signal. Recent dynamic trust-scoring work explicitly targets unreliable or adversarial nodes [18], but does not frame trust as a healthcare-site aggregation weight combining gradient quality, behavioral consistency, data-integrity evidence, cohort size, and a quality gate in each round.

Privacy and auditability mechanisms further shape healthcare FL. Secure aggregation has been studied for real-world healthcare deployments [19], while blockchain-based FL architectures provide auditability guarantees at the cost of consensus latency [20, 21, 22]. DTAFA is designed to complement these mechanisms: it leaves raw data local and makes trust a first-class aggregation signal, updated continuously and applied directly each round.

3. System Model and Problem Formulation

Consider one coordinating server and heterogeneous healthcare clients $\mathcal{N} = \{1, \dots, N\}$. Client i may be a hospital, clinic, laboratory, or edge gateway; it holds local data $\mathcal{D}_i$ drawn from site-specific distribution $\mathcal{P}_i$ and has a resource profile (c_i, m_i, h_i, e_i) for compute, memory, channel quality, and availability [2]. Raw patient data remain local. At round t, the server selects $\mathcal{S}^t \subseteq \mathcal{N}$, broadcasts $\boldsymbol{w}^t$, receives updates $\{\Delta \boldsymbol{w}_i^t\}_{i \in \mathcal{S}^t}$, and aggregates them into $\boldsymbol{w}^{t+1}$.

Let $F(\boldsymbol{w}) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{N} |\mathcal{D}_i| F_i(\boldsymbol{w})$ be the global objective, where F_i is the local expected loss. Standard FedAvg minimizes F by weighting updates proportionally to $|\mathcal{D}_i|$; this is vulnerable when low-quality or malicious sites hold large cohorts [1]. DTAFA instead minimizes:

$$\min_{\boldsymbol{w}} \ \tilde{F}(\boldsymbol{w}) = \sum_{i \in \mathcal{S}^t} \lambda_i^t \, F_i(\boldsymbol{w}), \tag{1}$$

subject to $\sum_{i \in \mathcal{S}^t} \lambda_i^t = 1$, $\lambda_i^t \geq 0$, where λ_i^t is determined by cohort size and per-round trust $\tau_i^t \in [0, 1]$.

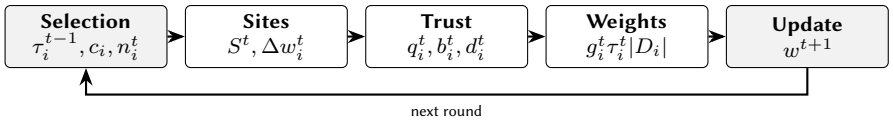

Figure 1: Compact DTAFA pipeline for privacy-aware multicentric healthcare learning.

4. The DTAFA Framework

4.1. Multi-Signal Trust Evaluation

DTAFA first computes an instantaneous trust observation $\hat{\tau}_i^t$ as a convex combination of three observable signals:

$$\hat{\tau}_i^t = \omega_q q_i^t + \omega_b b_i^t + \omega_d d_i^t, \tag{2}$$

where $\omega_q + \omega_b + \omega_d = 1$ and all three weights are nonnegative. The historical trust score is then updated by exponential smoothing:

$$\tau_i^t = \zeta\, \tau_i^{t-1} + (1-\zeta)\hat{\tau}_i^t, \tag{3}$$

with $\zeta \in (0,1)$.

Gradient quality begins with cosine alignment between client i's update and the coordinate-wise median of all received updates $\tilde{\boldsymbol{g}}^t$:

$$r_i^t = \frac{\langle \Delta \boldsymbol{w}_i^t,\, \tilde{\boldsymbol{g}}^t \rangle}{\|\Delta \boldsymbol{w}_i^t\| \cdot \|\tilde{\boldsymbol{g}}^t\|}. \tag{4}$$

The normalized score used in trust and aggregation is

$$q_i^t = \mathrm{clip}\left(\frac{r_i^t + 1}{2},\, 0,\, 1\right). \tag{5}$$

Values near 1 indicate alignment with the consensus direction; values near 0 flag adversarial or strongly drifted updates.

Behavioral consistency b_i^t tracks update-norm stability for received updates via an exponential moving average:

$$\bar{b}_i^t = \nu\, b_i^{t-1} + (1-\nu) \cdot \exp\!\left(-\frac{|\,\|\Delta \boldsymbol{w}_i^t\| - \bar{n}_i|}{\bar{n}_i + \epsilon}\right), \tag{6}$$

for $i \in \mathcal{S}^t$, where $\nu \in (0,1)$ is a behavioral memory factor and $\bar{n}_i$ is the running mean update norm. A light aging step sets $b_i^t = \chi \bar{b}_i^t$ for selected clients and $b_i^t = \chi b_i^{t-1}$ otherwise, so stale behavior scores do not remain fixed indefinitely.

Data integrity d_i^t penalizes clients whose reported label distribution $\hat{p}_i$ departs from the global reference $\bar{p}$:

$$d_i^t = 1 - \min(1,\ \kappa \cdot \mathrm{TV}(\hat{p}_i, \bar{p}))\,, \tag{7}$$

where $\kappa > 0$ controls sensitivity and $\mathrm{TV}(\cdot,\cdot) = \frac{1}{2}\sum_k |p_k - q_k|$. Clipping ensures that legitimate data heterogeneity degrades but does not zero out d_i^t. In a clinical deployment, $\hat{p}_i$ can be instantiated with approved aggregate metadata such as label prevalence, modality availability, or data-quality summaries rather than patient-level records.

4.2. Trust-Adaptive Aggregation

Aggregation weights scale dataset size by both smoothed trust and a current-round quality gate:

$$g_i^t = \epsilon_g + (1-\epsilon_g) q_i^t, \tag{8}$$

Algorithm 1 DTAFA Round

1: Select $\mathcal{S}^t$ using (12); compute median update $\tilde{\boldsymbol{g}}^t$.
2: **for** $i \in \mathcal{S}^t$ **do**
3: Evaluate q_i^t, b_i^t, and d_i^t using (5)–(7).
4: Update τ_i^t with (3); compute g_i^t and λ_i^t using (8)–(9).
5: Smooth λ_i^t into $\tilde{\lambda}_i^t$ using (10).
6: **end for**
7: Return $\boldsymbol{w}^{t+1}$ from (11) and updated trust logs.

where $\epsilon_g > 0$ prevents any selected site from being permanently muted by a single noisy round. The unsmoothed aggregation weight is

$$\lambda_i^t = \frac{g_i^t \cdot \tau_i^t \cdot |\mathcal{D}_i|}{\sum_{j \in \mathcal{S}^t} g_j^t \cdot \tau_j^t \cdot |\mathcal{D}_j|}. \tag{9}$$

When trust and quality scores are equal, (9) reduces exactly to FedAvg. The multiplicative coupling prevents large-cohort sites from dominating when their historical trust or current update quality is low, addressing a known vulnerability under non-IID poisoning [16]. A smoothing step prevents abrupt weight shifts:

$$\tilde{\lambda}_i^t = \eta_\lambda \, \lambda_i^t + (1 - \eta_\lambda) \, \tilde{\lambda}_i^{t-1}, \tag{10}$$

and the global update applies these smoothed weights directly:

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t + \sum_{i \in \mathcal{S}^t} \tilde{\lambda}_i^t \, \Delta \boldsymbol{w}_i^t. \tag{11}$$

Server-side overhead is dominated by computing $\tilde{\boldsymbol{g}}^t$ and cosine scores over the K selected updates, which scales as $\mathcal{O}(K \cdot d)$ in the reported implementation. A random projection to $d' < d$ can reduce this to $\mathcal{O}(K \cdot d')$ when model dimension is large, while preserving directional sensitivity [23].

4.3. Site Selection Policy

At the start of each round the server assigns each site a sampling score:

$$s_i^t = (1 - \xi)\left(\rho \, \tau_i^{t-1} + (1 - \rho) \cdot \frac{c_i}{\max_j c_j}\right) + \xi \cdot \frac{1}{1 + n_i^t}, \tag{12}$$

balancing trust history, normalized compute capacity, and an exploration term based on prior participation count n_i^t. The server samples K sites without replacement with probability proportional to s_i^t, giving low-trust or infrequently selected sites periodic re-evaluation opportunities and preventing permanent exclusion after transient failures [17].

4.4. Healthcare Deployment Considerations

DTAFA complements privacy-preserving FL rather than replacing it: secure aggregation or differential privacy can protect updates, while trust uses update-level statistics and approved aggregate metadata. Trust logs also provide an audit trail of which sites influenced each round, supporting robustness monitoring and post-deployment review without permanently excluding sites after short-lived failures.

5. Experimental Evaluation

5.1. Setup

We evaluate DTAFA on CIFAR-10 image classification and Shakespeare next-character prediction, used as privacy-safe proxies for visual and sequential healthcare workloads rather than as clinical records. To keep the repository executable without a deep-learning runtime, both tasks use a NumPy multinomial

Table 1
Final Test Accuracy (%) Under Varying Byzantine Fractions

	CIFAR-10				Shakespeare			
Method	0	0.1	0.2	0.3	0	0.1	0.2	0.3
FedAvg	38.9	24.5	7.3	5.4	27.4	18.9	4.9	8.9
FedProx	38.9	24.5	7.3	5.4	27.4	18.9	4.9	8.9
TrustFed	**39.0**	24.6	8.4	5.4	**27.5**	18.9	4.7	8.9
FedMed	38.6	36.3	35.1	30.0	24.3	24.0	**23.7**	**21.6**
DTAFA	38.9	**38.6**	**36.0**	**33.4**	27.2	**26.1**	23.4	20.7

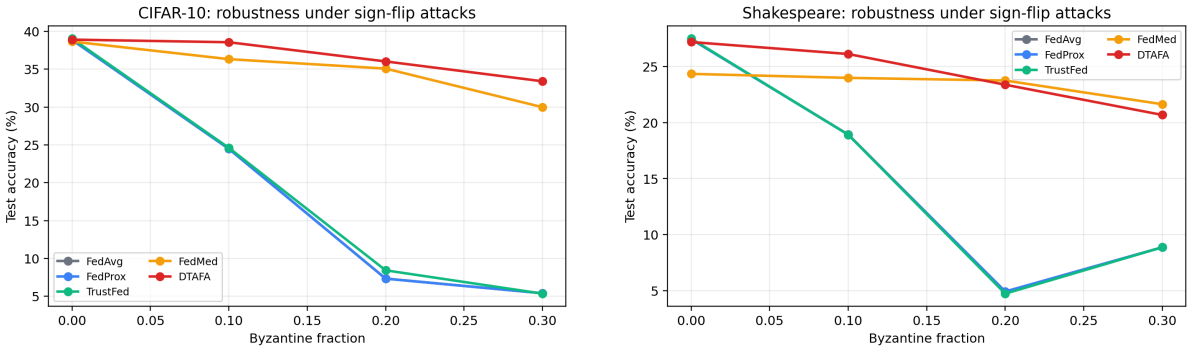

Figure 2: Final test accuracy under sign-flip Byzantine attacks on CIFAR-10 and Shakespeare.

softmax model. CIFAR-10 uses 8×8 average-pooled RGB features plus channel statistics; Shakespeare uses an eight-character context over the 36 most frequent characters. The splits contain 20,000/5,000 CIFAR-10 train/test examples and 50,000/10,000 Shakespeare train/test contexts.

Both tasks use $N = 100$ virtual sites partitioned by a Dirichlet distribution ($\mu = 0.5$) to emulate non-IID patient mix and site size [4]. Compute scores are sampled from $[0.2, 1.0]$, $K = 10$ sites participate per round, and Byzantine fractions $\beta \in \{0, 0.1, 0.2, 0.3\}$ submit sign-flipped updates scaled by 5. All methods use three local SGD steps, batch size $B = 64$, and $T = 60$ rounds averaged over seeds 7, 11, and 19. DTAFA uses $(\omega_q, \omega_b, \omega_d) = (0.65, 0.20, 0.15)$, $\zeta = 0.5$, $\nu = 0.8$, $\chi = 0.999$, $\kappa = 1.15$, $\epsilon_g = 0.10$, $\eta_\lambda = 0.7$, $\rho = 0.6$, and $\xi = 0.15$; code and CSV outputs are in `experiments/`.

5.2. Baselines

DTAFA is compared with FedAvg sample-count weighting [1], FedProx proximal regularization [2], TrustFed static data-integrity weighting [17], and FedMed coordinate-wise median aggregation [3].

5.3. Results

Robustness accuracy. Table 1 reports final test accuracy at $T = 60$, with the same values plotted in Fig. 2. On CIFAR-10, DTAFA is essentially tied in the benign case and becomes strongest as Byzantine pressure increases, retaining 33.4% accuracy at $\beta = 0.3$ versus 30.0% for FedMed and 5.4% for FedAvg/FedProx. On Shakespeare, DTAFA is best at $\beta = 0.1$ and remains close to FedMed at higher attack rates.

Convergence and trust behavior. Figure 3 shows trajectories at $\beta = 0.2$. DTAFA reaches 90% of final CIFAR-10 accuracy in 13.3 rounds, matching FedMed while avoiding FedAvg/FedProx/TrustFed collapse. On Shakespeare it reaches the same threshold in 25.0 rounds, ahead of FedMed's 36.7 rounds. At $\beta = 0.3$, final DTAFA trust separates honest and Byzantine sites: 0.64 versus 0.54 on CIFAR-10 and 0.62 versus 0.56 on Shakespeare.

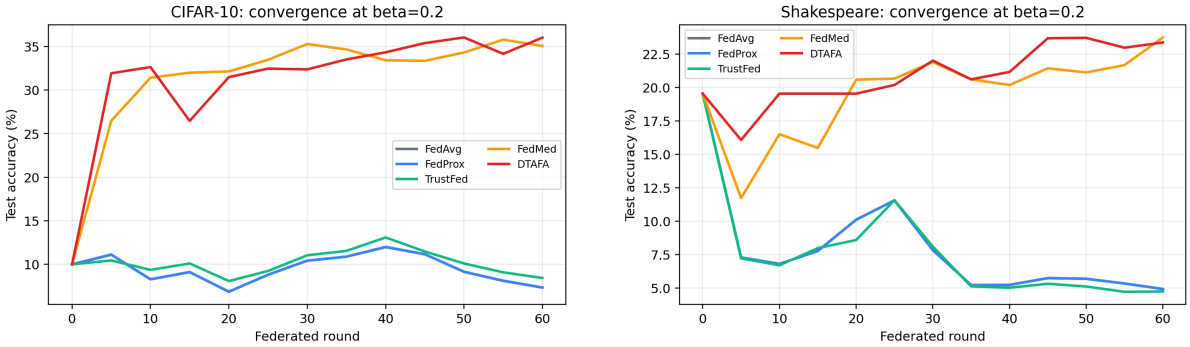

Figure 3: Convergence at $\beta = 0.2$ for CIFAR-10 and Shakespeare.

Table 2
Server Runtime and CIFAR-10 DTAFA Component Sensitivity

Method	Time
FedAvg	1.00×
FedProx	1.03×
TrustFed	1.19×
FedMed	1.60×
DTAFA	1.55×

Configuration	Acc. (%)	90% Round
Full DTAFA	36.0	13.3
No temporal smoothing	35.6	15.0
No deviation score	37.4	28.3
No site selection policy	34.9	13.3
No gradient clipping	29.9	35.0

Server overhead and sensitivity. Table 2 shows that DTAFA costs 1.55× FedAvg, comparable to FedMed's 1.60×, because it computes cosine agreement, trust updates, clipping, and trust-scaled aggregation. The CIFAR-10 ablation at $\beta = 0.2$ shows gradient clipping as the most important robustness component; removing it lowers accuracy from 36.0% to 29.9% and delays the 90%-of-final threshold from 13.3 to 35.0 rounds.

6. Conclusion

This paper presented DTAFA, a trust-aware aggregation framework for privacy-aware multicentric healthcare AI. By combining three-signal trust, exponential smoothing, a current-round quality gate, and trust-informed site selection, DTAFA reduces poisoned or unreliable influence while preserving periodic re-evaluation. On privacy-safe CIFAR-10 and Shakespeare proxies, DTAFA substantially improves over sample-count aggregation under sign-flip Byzantine attacks: at $\beta = 0.3$, it reaches 33.4% CIFAR-10 accuracy versus 5.4% for FedAvg/FedProx and 30.0% for FedMed, and 20.7% Shakespeare accuracy versus FedAvg's 8.9%. Future work will evaluate approved clinical cohorts, integrate secure aggregation and differential privacy, and extend DTAFA to asynchronous settings with stale-gradient discounting [15].

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] J. Chen, W. Li, G. Yang, E. al., S. Guo, Federated learning meets edge computing: A hierarchical aggregation mechanism for mobile devices, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 13736, Springer, 2022, pp. 183–197. URL: http://dx.doi.org/10.1007/978-3-031-19211-1_38. doi:10.1007/978-3-031-19211-1_38.

[2] C. Xu, Y. Qu, Y. Xiang, L. Gao, Asynchronous federated learning on heterogeneous devices: A survey, Comput. Sci. Rev. 50 (2023) 100595. URL: http://dx.doi.org/10.1016/j.cosrev.2023.100595. doi:10.1016/j.cosrev.2023.100595.

[3] B. Ding, P. Yang, S.-J. Huang, FedDLAD: A federated learning dual-layer anomaly detection framework for enhancing resilience against backdoor attacks, International Joint Conferences on Artificial Intelligence Organization, California, 2025, pp. 5021–5029. URL: https://www.ijcai.org/proceedings/2025. doi:10.24963/ijcai.2025.

[4] M. Cheng, W. Chen, W. Fang, Z. Wang, T. Xu, J. Liu, V. C. M. Leung, MFTE: Multifactor and fuzzy trust evaluation for federated learning in mobile edge computing, Comput. Netw. 265 (2025) 111340. URL: https://linkinghub.elsevier.com/retrieve/pii/S138912862500307X. doi:10.1016/j.comnet.2025.111340.

[5] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, S. Bakas, Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation, Lecture Notes in Computer Science 11383 (2019) 92–104. URL: http://dx.doi.org/10.1007/978-3-030-11723-8_9. doi:10.1007/978-3-030-11723-8_9.

[6] S. e. a. Pati, Federated learning enables big data for rare cancer boundary detection, Nat. Commun. 13 (2022) 7346. URL: https://www.nature.com/articles/s41467-022-33407-5. doi:10.1038/s41467-022-33407-5.

[7] A. Vaid, S. K. Jaladanki, J. Xu, S. Teng, A. Kumar, S. Lee, S. Somani, I. Paranjpe, J. K. De Freitas, T. Wanyan, K. W. Johnson, M. Bicak, E. Klang, Y. J. Kwon, A. Costa, S. Zhao, R. Miotto, A. W. Charney, E. Böttinger, Z. A. Fayad, G. N. Nadkarni, F. Wang, B. S. Glicksberg, Federated learning of electronic health records to improve mortality prediction in hospitalized patients with covid-19: Machine learning approach, JMIR Med. Inform. 9 (2021) e24207. URL: http://medinform.jmir.org/2021/1/e24207/. doi:10.2196/24207.

[8] T. Li, A. K. Sahu, A. Talwalkar, V. Smith, Federated learning: Challenges, methods, and future directions, IEEE Signal Process. Mag. 37 (2020) 50–60. URL: https://ieeexplore.ieee.org/document/9084352/. doi:10.1109/msp.2020.2975749.

[9] D. M. Jimenez G., D. Solans, M. Heikkila, A. Vitaletti, N. Kourtellis, A. Anagnostopoulos, I. Chatzigiannakis, Non-iid data in federated learning: A survey with taxonomy, metrics, methods, frameworks and future directions, arXiv [cs.LG] (2024). URL: http://dx.doi.org/10.48550/arXiv.2411.12377. doi:10.48550/arXiv.2411.12377. arXiv:2411.12377.

[10] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, arXiv [cs.LG] (2020). URL: https://arxiv.org/abs/1812.06127. arXiv:1812.06127.

[11] M. Fang, X. Cao, J. Jia, N. Gong, Local model poisoning attacks to byzantine-robust federated learning, in: Proceedings of the 29th USENIX Security Symposium, USENIX Association, 2020, pp. 1605–1622. URL: https://www.usenix.org/conference/usenixsecurity20/presentation/fang.

[12] X. Cao, M. Fang, J. Liu, N. Z. Gong, Fltrust: Byzantine-robust federated learning via trust bootstrapping, arXiv [cs.CR] (2020). doi:10.48550/arXiv.2012.13995. arXiv:2012.13995.

[13] F. Colosimo, F. De Rango, Median-krum: A joint distance-statistical based byzantine-robust algorithm in federated learning, in: Proceedings of the Int'l ACM Symposium on Mobility Management and Wireless Access, ACM, New York, NY, USA, 2023. URL: https://dl.acm.org/doi/10.1145/3616390.3618283. doi:10.1145/3616390.3618283.

[14] Y. Sanjalawe, S. Al-E'mari, T. Alqurashi, Z. H. Alharbi, S. N. Makhadmeh, M. Alsharaiah, Adaptive graph attention-based federated learning for IoT intrusion detection: mitigating poisoning attacks, PeerJ Comput. Sci. 11 (2025) e3281. URL: http://dx.doi.org/10.7717/peerj-cs.3281. doi:10.7717/peerj-cs.3281.

[15] Y. Xu, Y. Zhu, Z. Wang, H. Xu, Y. Liao, Enhancing federated learning through layer-wise aggregation over non-IID data, IEEE Trans. Serv. Comput. 18 (2025) 798–811. URL: https://ieeexplore.ieee.org/document/10857658/. doi:10.1109/tsc.2025.3536309.

[16] B. Hathout, P. Shepherd, T. Dagiuklas, E. al., J. Rodriguez, Adaptive trust management for data poisoning attacks in MEC-based FL infrastructures, IEEE Open Journal of the Communications Society 6 (2025) 3140–3160. URL: http://dx.doi.org/10.1109/OJCOMS.2024.3523368. doi:10.1109/

OJCOMS.2024.3523368.

[17] G. Rjoub, O. A. Abdel-Wahab, J. Bentahar, A. Bataineh, Trust-driven reinforcement selection strategy for federated learning on IoT devices, Computing 104 (2024) 1689–1705. URL: http://dx.doi.org/10.1007/s00607-022-01078-1. doi:10.1007/s00607-022-01078-1.

[18] T. P. Chander, B. K. Sarkar, Federated learning with dynamic trust scoring for unreliable or adversarial nodes, Journal of Computational Analysis and Applications (JoCAAA) 33 (2024) 2935–2952. URL: https://eudoxuspress.com/index.php/pub/article/view/4573.

[19] R. Taiello, S. Cansiz, M. Vesin, F. Cremonesi, L. Innocenti, M. Önen, M. Lorenzi, Enhancing privacy in federated learning: Secure aggregation for real-world healthcare applications, arXiv [cs.CR] (2024). URL: http://dx.doi.org/10.48550/arXiv.2409.00974. doi:10.48550/arXiv.2409.00974. arXiv:2409.00974.

[20] J. Liu, C. Chen, Y. Li, L. Sun, Y. Song, J. Zhou, B. Jing, D. Dou, Enhancing trust and privacy in distributed networks: a comprehensive survey on blockchain-based federated learning, Knowl. Inf. Syst. 66 (2024) 4377–4403. URL: http://dx.doi.org/10.1007/s10115-024-02117-3. doi:10.1007/s10115-024-02117-3.

[21] Govarthan, Bakiyalakshmi, Banu, D. Dharshini, A scalable blockchain-enabled federated learning framework for privacy, trust, and efficiency in edge environments, in: 2025 International Conference on Digital Innovations for Sustainable Solutions (ICDISS), IEEE, 2025, pp. 1–6. URL: http://dx.doi.org/10.1109/icdiss68238.2025.11320699. doi:10.1109/icdiss68238.2025.11320699.

[22] X. Li, W. Wu, A blockchain-empowered multiaggregator federated learning architecture in edge computing with deep reinforcement learning optimization, IEEE Trans. Comput. Soc. Syst. 12 (2025) 645–657. URL: https://ieeexplore.ieee.org/document/10739775/. doi:10.1109/tcss.2024.3481882.

[23] O. Manjang, Y. Zhai, J. Shen, A. Sarwar, X. He, H. Wang, L. Zhu, Efficient hierarchical federated learning with pareto-optimal bi-level reinforcement learning, IEEE Internet Things J. 12 (2025) 30710–30724. URL: http://dx.doi.org/10.1109/jiot.2025.3572900. doi:10.1109/jiot.2025.3572900.