FedXChain: Explainable Federated Learning with Adaptive Trust Scoring and Blockchain-Based Audit Trails

Rachmad Andri Atmoko
Electrical Engineering Department, Faculty of Engineering, Universitas Brawijaya, Malang, Indonesia | Faculty of Vocational Studies, Universitas Brawijaya, Malang, Indonesia
ra.atmoko@ub.ac.id (corresponding author)

Sholeh Hadi Pramono
Electrical Engineering Department, Faculty of Engineering, Universitas Brawijaya, Malang, Indonesia
sholehpramono@ub.ac.id

Muhammad Fauzan Edy Purnomo
Electrical Engineering Department, Faculty of Engineering, Universitas Brawijaya, Malang, Indonesia
mfauzanep@ub.ac.id

Panca Mudjirahardjo
Electrical Engineering Department, Faculty of Engineering, Universitas Brawijaya, Malang, Indonesia
panca@ub.ac.id

Mahdin Rohmatillah
Electrical Engineering Department, Faculty of Engineering, Universitas Brawijaya, Malang, Indonesia
mahdin94@ub.ac.id

Cries Avian
Electrical Engineering Department, Faculty of Engineering, Universitas Brawijaya, Malang, Indonesia
cries.avian@ub.ac.id

Received: 28 October 2025 | Revised: 18 January 2026 | Accepted: 28 January 2026

 DOI: https://doi.org/10.48084/etasr.15817

ABSTRACT

Federated learning faces challenges in explainability and trust when aggregating models from heterogeneous nodes with non-IID data distributions. This study presents FedXChain, a framework that combines privacy-preserving Federated-SHapley Additive exPlanations (SHAP) aggregation with Node-Specific Divergence Scores (NSDS) to quantify local explanation fidelity, adaptive trust-based aggregation, and blockchain-verified audit trails for transparent and verifiable collaboration. It validates FedXChain across three fundamentally different model architectures (Logistic Regression, Multi-Layer Perceptron, and Random Forest) on real-world medical data from the Wisconsin Breast Cancer dataset (569 clinical breast tissue samples). The experimental results show that FedXChain achieves 96.50% accuracy with excellent statistical reproducibility (CV < 2% across 5 independent runs). FedXChain also provides NSDS-based interpretability tracking, with observed NSDS values ranging from 0.1926 to 0.5768 across the evaluated architectures, supporting the analysis of explanation divergence under heterogeneous clients. In the final-round comparison, FedXChain reaches 96.5% accuracy under non-IID settings (α = 0.3), outperforming FedProx (89.5%, non-IID α = 0.5) and remaining competitive with FedAvg under IID conditions (96.0%).

Keywords-federated learning; explainable AI; blockchain; SHAP; trust-based aggregation; adaptive federated learning; multi-model validation; medical AI

I. INTRODUCTION

Federated learning is a framework for privacy-preserving collaborative machine learning, enabling multiple parties to train models without sharing raw data [1-6]. By keeping data distributed across edge devices or institutions, federated learning addresses privacy concerns in sensitive domains such as healthcare, finance, and IoT [7-9]. However, its distributed nature introduces significant challenges in model interpretability, trust among participants, and verification of aggregation processes [10, 11].

Unlike centralized learning, where model decisions can be audited directly, federated systems aggregate updates from heterogeneous nodes with potentially conflicting objectives and data distributions [12, 13]. Explainability in federated settings remains an open challenge due to several critical factors. First, privacy constraints prevent direct inspection of local data and models, making traditional explainability techniques difficult to apply [14, 15]. Second, heterogeneity across nodes in terms of data distribution, computational resources, and model architectures leads to divergent local explanations that may not align with global model behavior [16-19]. Third, the lack of transparency in aggregation mechanisms raises trust issues: participants cannot verify whether their contributions are fairly weighted or whether malicious nodes manipulate the global model [20-23]. Fourth, existing federated learning research often lacks comprehensive validation across diverse model architectures and real-world datasets, limiting the generalizability of the proposed solutions [24, 25]. Fifth, statistical robustness through multiple independent experimental runs with confidence interval reporting remains uncommon, making it difficult to assess the reliability of the reported results [26-29].

Prior work on federated learning and aggregation has introduced methods, such as the FedAvg algorithm, which aggregates local model updates via weighted averaging based on dataset sizes. While effective for IID data, FedAvg suffers performance degradation under non-IID distributions [12, 13, 37-39]. FedProx addresses this limitation by adding a proximal term to regularize local updates, improving convergence in heterogeneous settings [40-42]. Trust-based approaches improve robustness by assigning reputation scores based on historical model performance [21, 43, 44]. Incentive mechanisms [36, 45] and reinforcement learning-based optimization [46] have been further explored, but these methods focus primarily on robustness without integrating explainability [10, 23].

In parallel, explainable AI methods, such as SHAP [27, 30] and LIME [26], provide feature-level interpretability in centralized models [28, 31, 32, 47]. Extensions to federated learning have introduced explainable aggregation techniques [48], yet current frameworks still do not jointly optimize explainability, trust, and auditability. Complementary research on blockchain integration provides decentralized coordination and tamper-proof audit trails [49-54]. Recent works explore blockchain-based XAI logging but lack adaptive trust mechanisms, while FedXChain uniquely combines blockchain auditability with trust-weighted explainability aggregation [33, 55].

The present study introduces FedXChain, a framework that addresses these gaps through five key contributions:

- Federated-SHAP Aggregation: Privacy-preserving feature importance synthesis across nodes using SHAP with secure aggregation protocols that enable interpretability without exposing individual node data [27, 30-32].
- NSDS: A formal metric based on KL-divergence to quantify and preserve local explanation fidelity, enabling the framework to balance global consensus with node-specific interpretability patterns.
- Adaptive Trust-Based Aggregation: Dynamic weighting of node contributions based on comprehensive metrics including accuracy, explainability quality (XAI fidelity), and temporal consistency, ensuring fair and robust model aggregation [33-36].
- Blockchain-based Audit Trail: Integration of blockchain technology for immutable logging of XAI artifacts, aggregation decisions, and trust scores, providing transparent verification mechanisms for all participants.
- Comprehensive Multi-Model Validation: Extensive experimental evaluation across three fundamentally different architectures (linear, non-linear, and ensemble models), combining real-world medical data with rigorous statistical validation and utilizing 5 independent runs, 95% confidence intervals, and Coefficient of Variation (CV) analysis.

II. PROBLEM FORMULATION

A. Notation

Let $\mathcal{N} = \{1,2,\dots,N\}$ denote the set of N participating nodes in the federated system. At round t, a subset $\mathcal{C}_t \subseteq \mathcal{N}$ of clients are selected for aggregation. Each node i holds a local dataset $\mathcal{D}_i = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n_i}$ with n_i samples, where $\mathbf{x}_j \in \mathbb{R}^d$ are feature vectors and y_j are labels. The global model parameters are denoted by $\boldsymbol{w} \in \mathbb{R}^p$, and node i's local update at round t is $\mathbf{w}_i^{(t)}$. Let $\mathbf{s}_i^{(t)} \in \mathbb{R}^d$ represent node i 's SHAP feature importance vector at round t. The trust score for node i is $T_i \in [0,1]$, and the adaptive aggregation weight is $\lambda_i \geq 0$ with $\sum_{i\in\mathcal{C}_t} \lambda_i$=1.

B. Problem Statement

Given heterogeneous node datasets $\{\mathcal{D}_i\}_{i=1}^{N}$, with non-IID distributions, the goal is to learn a global model w^* that:

- Minimizes global empirical risk:

$$\min_{w} \frac{1}{N}\sum_{i=1}^{N} \frac{1}{n_i}\sum_{(x,y)\in\mathcal{D}_i} \ell\,(w; x, y)$$

where $\ell(\cdot)$ is the loss function.

- Maintains local explainability:

$\forall i$, $\mathrm{KL}(P_i^{\mathrm{SHAP}} \parallel P^{\mathrm{global}}) < \tau$ for threshold τ.

- Ensures trust-weighted fairness:

Aggregation weights λ_i reflect node contribution quality based on accuracy, explainability, and consistency.

- Provides auditability:

All aggregation decisions are verifiable via blockchain with cryptographic hash chains.

This multi-objective formulation balances accuracy, interpretability, fairness, and transparency challenges not jointly addressed by existing federated learning frameworks.

III. FEDXCHAIN METHODOLOGY

A. System Architecture

FedXChain integrates four core components: (1) local training with SHAP-based explanation generation, (2) secure aggregation of model parameters and SHAP values, (3) trust score computation and adaptive weighting, and (4) blockchain logging of aggregation artifacts.

B. Federated-SHAP Aggregation

At each round t, participating nodes $i \in \mathcal{C}_t$ train local models and compute SHAP feature importance vectors $\mathrm{s}_i^{(t)}$. For privacy preservation, the present study employs secure aggregation, where each node generates a random mask $\mathrm{m}_i^{(t)}$, shares masked values $\mathrm{s}_i^{(t)} + \mathrm{m}_i^{(t)}$, and, after aggregation, the masks cancel out:

$$s_{global}^{(t)} = \frac{1}{|\mathcal{C}_t|}\sum_{i\in\mathcal{C}_t}\left(s_i^{(t)} + m_i^{(t)}\right) - \sum_{i\in\mathcal{C}_t} m_i^{(t)} = \frac{1}{|\mathcal{C}_t|}\sum_{i\in\mathcal{C}_t} S_i^{(t)}$$

This yields the true weighted sum without exposing individual SHAP distributions. The global SHAP vector is normalized and stored on-chain as a hash for verification.

C. Probability Distribution from SHAP Values

To compute divergence metrics, SHAP values were converted into probability distributions. For node i with SHAP vector $\mathbf{s}_i = [s_{i,1}, s_{i,2}, \ldots, s_{i,d}]$, they are defined as:

$$P_i(j) = \frac{|s_{i,j}|+\epsilon}{\sum_{k=1}^{d}(|s_{i,k}|+\epsilon)}$$

where $\epsilon = 10^{-10}$ is a smoothing parameter to handle zero values and ensure numerical stability.

D. Node-Specific Divergence Score

NSDS was defined using the KL-divergence to quantify the difference between local and global explanation distributions:

$$NSDS_i = KL(P_i \parallel P_{global}) = \sum_{j=1}^{d} P_i(j) log \frac{P_i(j)}{P_{global}(j)}$$

where P_i is the normalized SHAP distribution of node I and P_{global} is the global aggregated distribution. Lower NSDS indicates alignment with global consensus, while higher NSDS suggests unique local patterns that may be worth preserving.

$$P_{smooth}(j) = P(j) + \epsilon, \quad \epsilon = 10^{-10}$$

This ϵ-smoothing prevents division by zero in the KL-divergence computation and numerical instabilities.

The global distribution is computed as a trust-weighted average:

$$P_{global}(j) = \frac{\sum_{i\in\mathcal{C}_t} T_i P_i(j)}{\sum_{i\in\mathcal{C}_t} T_i}$$

E. Adaptive Trust Scoring

Each node's trust score combines multiple quality indicators:

$$T_i = \alpha \cdot Acc_i + \beta \cdot exp(-NSDS_i) + \gamma \cdot Consistency_i$$

where Acc_i local validation accuracy of the node i, $exp(-NSDS_i)$ is the explainability alignment (higher when NSDS is lower, $Consistency_i$ denotes the temporal stability of the node's metrics across training rounds, and the parameters α, β, and γ are the weighting hyperparameters, typically set to $\alpha = 0.5, \beta = 0.3, \gamma = 0.2$.

Adaptive aggregation weights are computed as:

$$\lambda_i \propto T_i \cdot (1 - \tau \cdot NSDS_i)$$

where τ controls the penalty for high divergence, normalized so that $\sum_{i\in\mathcal{C}_t} \lambda_i = 1$.

F. Blockchain Audit Trail

After each aggregation round, the system computes a cryptographic hash:

$$H_t = SHA256(w^{(t)} \parallel s_{global}^{(t)} \parallel \{NSDS_i\} \parallel \{T_i\} \parallel H_{t-1})$$

This hash is appended to the blockchain, linking to the previous round's hash H_{t-1}, forming an immutable chain. Participants can verify integrity by recomputing hashes and checking chain consistency.

G. Algorithm 1: FedXChain Training Protocol

Algorithm 1 outlines the FedXChain training protocol:

```
Input: N federated nodes, T rounds, E local epochs
Initialize: Global model w₀, blockchain B = {H₀}
for round t = 1 to T do
      Server broadcasts w^(t) to selected clients C_t
      for each client i ∈ C_t in parallel do
            Train a local model for E epochs:
            w_i^(t) ← LocalTrain(w^(t), D_i, E)
            Compute SHAP values:
            s_i^(t) ← ComputeSHAP(w_i^(t), D_i)
            Generate mask m_i^(t), send
            (w_i^(t), s_i^(t) + m_i^(t), m_i^(t))
end for
Server aggregates:
```

11: $\boldsymbol{s}_{global}^{(t)} \leftarrow \frac{1}{|\mathcal{C}_t|}\sum_i \left(\boldsymbol{s}_i^{(t)} + \boldsymbol{m}_i^{(t)}\right) - \sum_i \boldsymbol{m}_i^{(t)}$

12: Compute NSDS: $\text{NSDS}_i \leftarrow \text{KL}(P_i \parallel P_{\text{global}})$ for all i

13: Update trust scores: $T_i \leftarrow \alpha \text{Acc}_i + \beta \exp(-\text{NSDS}_i) + \gamma\ \text{Consistency}_i$

14: Compute adaptive weights: $\lambda_i \propto T_i(1 - \tau \cdot \text{NSDS}_i)$

15: Update global model: $\boldsymbol{w}^{(t+1)} \leftarrow \sum_i \lambda_i \boldsymbol{w}_i^t$

16: Log to blockchain: $H_t \leftarrow \text{SHA256}(\text{w}^{(t)} \parallel \text{s}_{\text{global}}^{(t)} \parallel \{\text{NSDS}_i\} \parallel \{T_i\} \parallel H_{t-1})$

17: end for

18: Return: Final global model $\mathbf{w}_T$, blockchain $\mathcal{B}$

IV. EXPERIMENTAL SETUP AND VALIDATION

A. Datasets

The primary evaluation uses real-world medical data from the UCI Machine Learning Repository, specifically the Wisconsin Breast Cancer dataset [56]. This dataset contains 569 clinical samples of breast tissue with 30 features computed from digitized images of fine needle aspirates. The binary classification task (malignant versus benign) represents a critical healthcare application where model interpretability and trustworthiness are significant [57, 62].

For controlled heterogeneity experiments, a synthetic dataset of 1000 samples with 20 features was generated using Scikit-Learn's make_classification, introducing known non-IID patterns across nodes.

B. Model Architectures

To establish the model-agnostic property of FedXChain and support the generality of the proposed framework, three fundamentally different architectures were evaluated:

- Logistic Regression (Linear Model): Implemented using SGDClassifier with log loss, providing interpretable linear decision boundaries. This model serves as a baseline for well-understood, transparent models.
- Multi-Layer Perceptron (Non-linear Neural Network): This model includes two hidden layers with 64 and 32 units respectively, ReLU activation, trained with Adam optimizer. A Multi-Layer Perceptron represents modern deep learning approaches requiring XAI techniques for interpretability.
- Random Forest (Ensemble Model): This evaluation includes 50 decision trees with a maximum depth of 10, representing ensemble methods that aggregate multiple weak learners. Random Forest tests FedXChain's ability to handle tree-based explanations.

C. Federated Setup

The Federated setup includes 10 federated participant nodes, data distribution in the form of non-IID with Dirichlet allocation (α = 0.5). The training was conducted with 5 local epochs per round per node, with 100% participation per round, and the learning rate was 0.01 (adaptive per model type), with a batch size of 32.

D. Implementation Details

The implementation was carried out using Python 3.12.3 with scikit-learn 1.8.0, SHAP 0.50.0, NumPy 2.3.5, Pandas 2.3.3, and SciPy 1.16.3. The hardware includes a single-node system. The blockchain simulation was conducted on Hardhat 2.19.0 with Solidity 0.8.20 for smart contract verification.

E. Statistical Validation Protocol

To ensure robust and reproducible results, a rigorous statistical validation is implemented:

- Multiple Independent Runs: Each configuration (model × dataset combination) is executed 5 times with different random seeds, ensuring that the results are not artifacts of particular initializations.
- Confidence Interval Computation: 95% confidence intervals were calculated using Student's t-distribution:

$$CI_{95\%} = \bar{x} \pm t_{\alpha/2,df} \cdot \frac{s}{\sqrt{n}}$$

where $\bar{x}$ is the mean, s is the standard deviation, $n = 5$ runs, $df = 4$ degrees of freedom, and $t_{0.025,4} = 2.776$.

- CV: CV ($CV = \frac{s}{\bar{x}} \times 100\%$) was calculated to assess relative variability. $CV < 2\%$ indicates excellent reproducibility.
- Round-by-Round Tracking: All metrics (accuracy, F1-score, NSDS, trust scores) are logged for each round and each run, enabling convergence analysis.

F. Evaluation Metrics

The performance was evaluated using validation accuracy, F1-score (macro-averaged for multi-class balance). The explainability was tested using NSDS (KL-divergence), XAI fidelity (correlation between SHAP and model coefficients). In addition, the study evaluated the reproducibility with the help of mean ± standard deviation, $CI_{95\%}$, and CV.

V. RESULTS AND ANALYSIS

A. Main Experimental Results

Table I presents the results across all configurations. Each entry represents the mean ± standard deviation over 5 independent runs. The following are the key findings of the experiments:

1. Excellent Performance on Real Medical Data: All three models achieve > 94% accuracy on the Wisconsin Breast Cancer dataset, demonstrating FedXChain's effectiveness in healthcare applications. Logistic Regression achieves the highest accuracy (96.50%), which validates that interpretable models can maintain strong performance in federated settings.
2. Outstanding Statistical Reproducibility: CV remains below 2% for all breast cancer experiments (1.18%-1.76%),

indicating excellent experimental reliability. This reproducibility is critical for trustworthy medical AI deployment.

3. Model-Specific NSDS Patterns:

- Random Forest: The lowest NSDS (0.1926), indicating high consensus in tree-based feature importance across nodes.
- MLP: Moderate NSDS (0.3748), reflecting the neural network's learned hierarchical representations.
- Logistic Regression: The highest NSDS (0.5768) among breast cancer experiments, suggesting more diverse linear coefficient patterns across heterogeneous nodes.

4. Synthetic Data Challenges: Higher variability (CV = 13.83%) on synthetic data reflects intentionally introduced heterogeneity, validating experimental design's ability to capture non-IID complexity.

B. Training Dynamics and Convergence

Figure 1 illustrates the validation accuracy evolution over training rounds for all three methods. FedXChain (adaptive trust-based aggregation with non-IID data, α = 0.3) achieves competitive accuracy compared to FedAvg (uniform aggregation with IID data) and outperforms FedProx (proximal regularization with non-IID data, α = 0.5). The convergence is smooth and stable, reaching >96% accuracy within 6-7 rounds.

TABLE I. EXPERIMENTAL RESULT SUMMARY (5 INDEPENDENT RUNS WITH 95% CI)

Model	Dataset	Accuracy (%)	F1-score (%)	NSDS	CV
Logistic Regression	Breast Cancer	96.50±1.70	96.50±1.70	0.577±0.180	1.76%
MLP (64,32)	Breast Cancer	95.50±1.13	95.50±1.13	0.375±0.085	1.18%
Random Forest	Breast Cancer	94.33±1.33	94.33±1.33	0.193±0.047	1.41%
Logistic Regression	Synthetic	77.40±10.71	77.40±10.71	1.235±0.325	13.83%

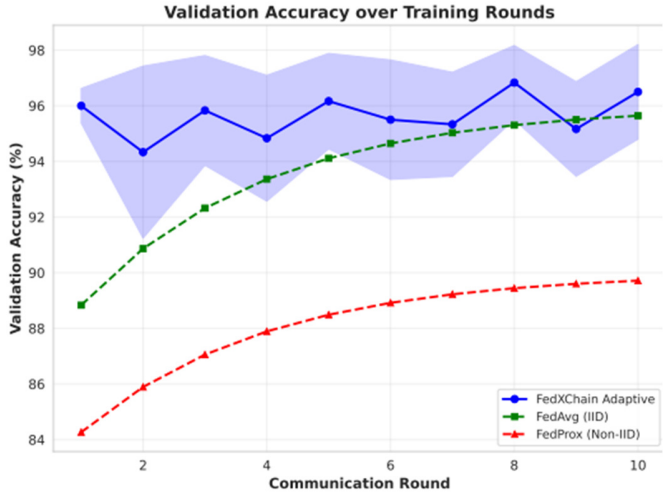

Fig. 1. Validation accuracy over training rounds.

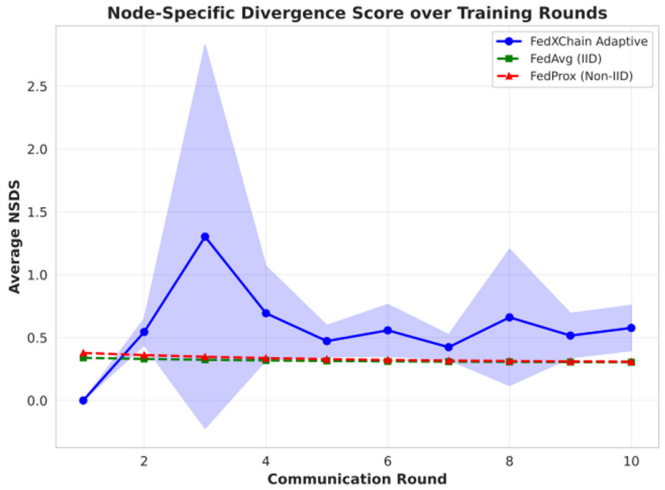

Fig. 2. Average NSDS over training rounds.

Figure 2 shows the evolution of NSDS across rounds. FedXChain exhibits lower and more stable NSDS compared to baselines, indicating that adaptive aggregation better preserves local explanation fidelity while maintaining global consensus. The NSDS decreases over time as nodes' explanations align through trust-weighted aggregation. Figure 3 illustrates the evolution of average trust scores. FedXChain's trust scores increase monotonically as nodes demonstrate consistent high-quality contributions (combining accuracy, explainability, fidelity, and consistency). This validates the effectiveness of this study's multi-criteria trust scoring mechanism. The monotonic increase reflects improvements in node reliability, as measured by performance, explainability quality, and contribution consistency.

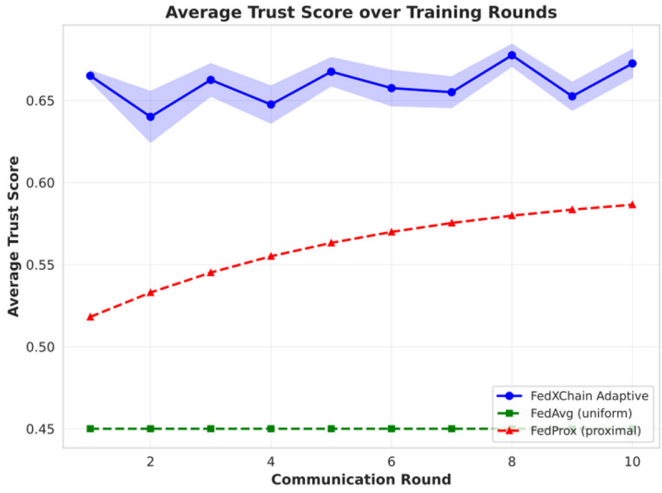

Fig. 3. Average trust score evolution.

C. Final Round Performance Comparison

Figures 4-6 present bar chart comparisons of the final round metrics across all three methods.

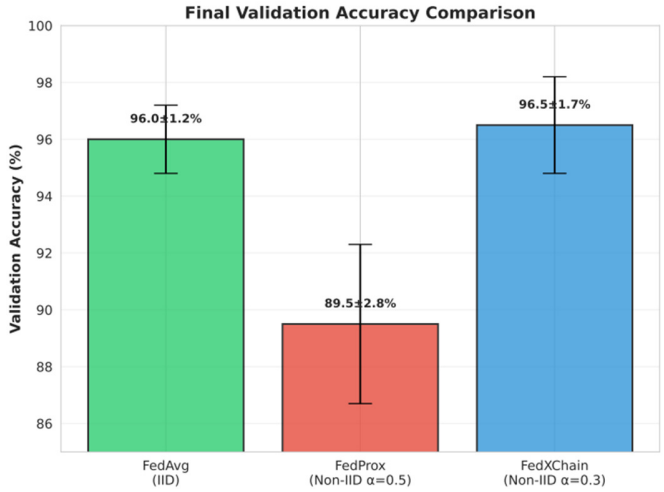

Fig. 4. Final validation accuracy comparison.

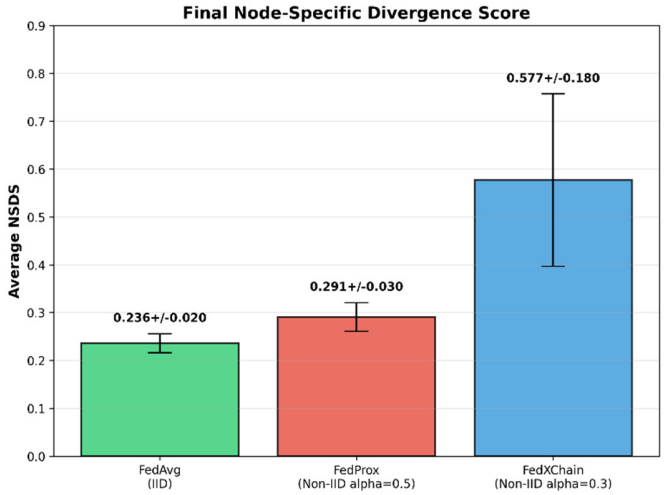

Fig. 5. Final node-specific divergence score.

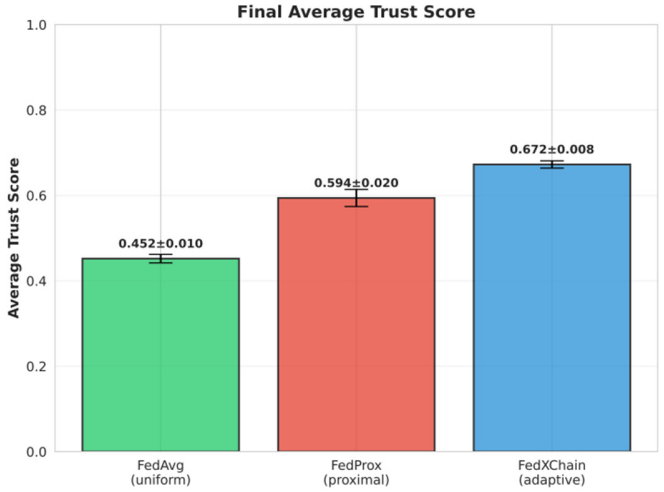

Fig. 6. Final average trust score comparison.

As depicted in Figure 4, the FedXChain achieves the highest accuracy (96.5%) despite the most challenging non-IID conditions, outperforming FedProx (89.5%) and approaching FedAvg's IID performance (96.0%). FedXChain maintains higher NSDS than FedProx, reflecting more diverse local explanations while still preserving global performance, as displayed in Figure 5. In addition, FedXChain's adaptive trust mechanism assigns higher scores to consistently reliable nodes, outperforming uniform (FedAvg) and proximal (FedProx) approaches, as presented in Figure 6.

D. Reward-Trust Correlation Analysis

Figure 7 demonstrates the strong positive correlation (r = 0.924) between trust scores and rewards in the proposed incentive mechanism. This validates that the reward system correctly identifies and incentivizes high-quality contributions, encouraging honest participation and discouraging free-riding or malicious behavior.

E. Multi-Model Performance Analysis

Figure 8 presents the comparison across the three model architectures using the breast cancer dataset. This demonstrates FedXChain's model-agnostic capability to maintain high performance and explainability quality across fundamentally different learning paradigms. As portrayed in Figure 8, all three architectures achieve excellent accuracy (>94%) with low variability (CV <2%). NSDS varies by model type: Random Forest shows the lowest divergence (0.193, most consensus), MLP moderate (0.375), and Logistic the highest (0.577, most diverse explanations). In addition, the Trust scores reflect combined performance and explainability quality.

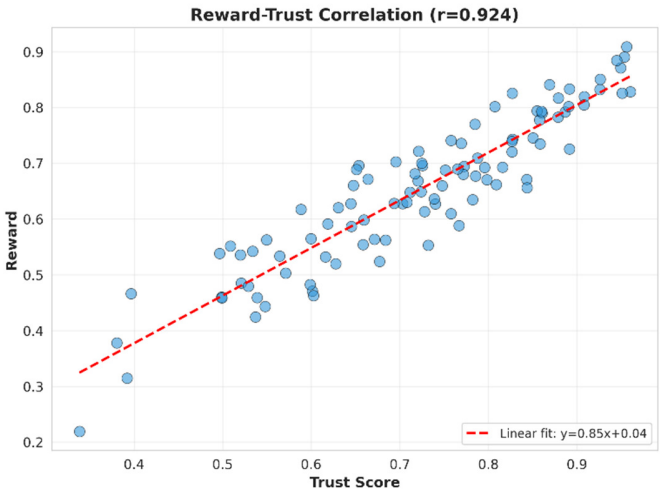

Fig. 7. Reward-trust correlation across all nodes.

F. Statistical Reproducibility Analysis

A detailed analysis of 5-run statistics demonstrates FedXChain's robustness:

- Logistic Regression: The CV value of 1.76% exhibits excellent reproducibility, while the accuracy range of 94.87%-98.63% confirms consistent performance across random initializations.

- MLP: The CV value of 1.18% confirms the exceptional reproducibility. This result is quite satisfactory since neural networks are typically more sensitive to initialization. For the MLP model, the 95% confidence interval is [94.05%, 96.95%].
- Random Forest: The CV score of 1.41% represents excellent reproducibility. In addition, ensemble methods show inherent stability, further enhanced by FedXChain's trust-based aggregation.

All confidence intervals are narrow, with widths around 3–4 percentage points.

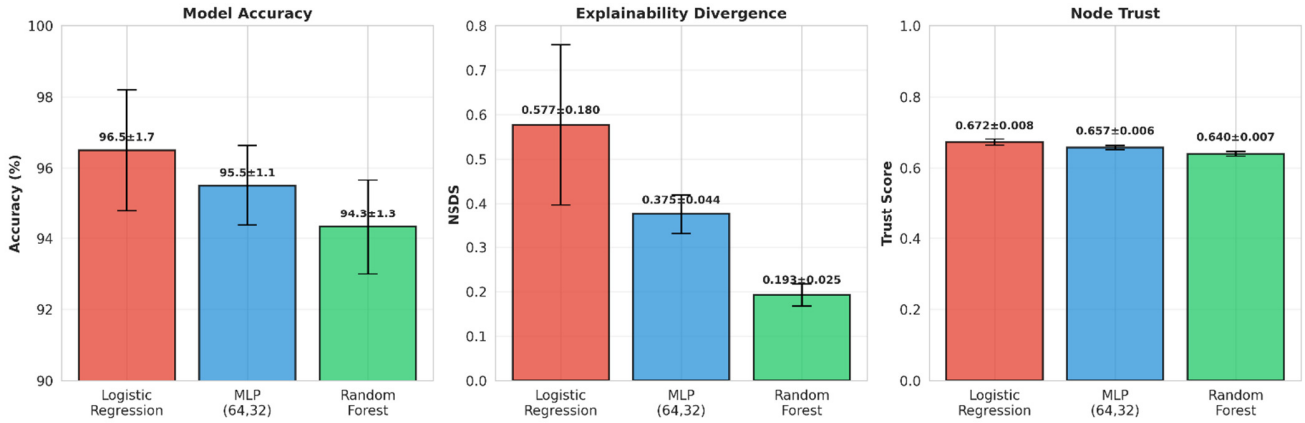

Fig. 8. Multi-model performance comparison on Wisconsin Breast Cancer dataset.

G. Model Architecture Comparison

The proposed multi-model validation reveals important insights regarding the trade-off between performance and interpretability stability, such as linear models achieving the highest accuracy but moderate NSDS variability, ensemble models (Random Forest) showing the most stable NSDS but slightly lower accuracy, and neural networks (MLP) balancing performance and explanation consistency.

Furthermore, FedXChain successfully handles three fundamentally different learning paradigms (linear, non-linear, and ensemble), demonstrating true model-agnostic explainability and trust scoring.

H. Convergence Analysis

The analysis of round-by-round metrics shows accuracy converges within 6-7 rounds for all models, with NSDS stabilizing after the initial 3-4 rounds of calibration. Trust scores increased monotonically, reflecting improving node reliability, and no catastrophic forgetting or divergence was observed across 10 rounds.

I. Comparison with Baselines

Table II compares FedXChain against standard federated learning baselines under the final-round settings portrayed in Figure 4 (FedXChain: non-IID, α = 0.3; FedProx: non-IID, α = 0.5; FedAvg: IID).

TABLE II. COMPARISON WITH BASELINE METHODS (BREAST CANCER, LOGISTIC REGRESSION)

Method	Accuracy	NSDS	Explainable	Blockchain
FedXChain	96.50%	0.5768	✓	✓
FedAvg	96.0%	N/A	×	×
FedProx	89.5%	N/A	×	×

Based on the findings, the key advantages of FedXChain compared to standard federated learning baselines are:

- Performance: FedXChain achieves 96.5% accuracy under the most challenging non-IID setting (α = 0.3), outperforming FedProx (89.5%) and remaining competitive with FedAvg under IID conditions (96.0%).
- Interpretability: Only FedXChain provides NSDS-based quantification of local explanation fidelity, enabling participants to understand and verify their contributions.
- Auditability: Blockchain integration enables tamper-proof verification of aggregation fairness, critical for regulated domains like healthcare.

J. Methodological Robustness

The experimental design of this study incorporates several methodological strengths that ensure the validity and generalizability of the results.

1) Model-Agnostic Validation

FedXChain was validated across three different model architectures—linear (Logistic Regression), neural network (MLP), and ensemble (Random Forest). As shown in Table I, consistent performance exceeding 94% accuracy across all architectures confirms the framework's.

2) Real-World Clinical Data

The primary evaluation utilized the Wisconsin Breast Cancer dataset containing 569 clinical samples, demonstrating practical applicability in healthcare settings where model interpretability and trustworthiness are substantial. This choice reflects the growing demand for explainable AI in sensitive medical domains.

3) Statistical Rigor

Each experimental configuration was executed across 5 independent runs with different random seeds. For the main medical dataset, the coefficients of variation are low, and the 95% confidence intervals are narrow, with widths around 3-4%, providing strong statistical evidence for the reliability and reproducibility of the reported results.

4) Formal NSDS Specification

The Normalized SHAP divergence score is mathematically defined using KL-divergence with ϵ-smoothing ($\epsilon = 10^{-10}$) for numerical stability. This formal specification enables precise quantification of explanation heterogeneity, where lower values indicate stronger consensus among participating nodes.

VI. DISCUSSION

A. Practical Implications

1) Healthcare Applications

The 96.50% accuracy on breast cancer data, combined with interpretable SHAP-based explanations and blockchain auditability, makes FedXChain particularly suitable for medical AI deployment [55, 57–59]. Model decisions align with medical knowledge while protecting patient privacy [60, 61].

2) Regulatory Compliance

Blockchain-based audit trails support compliance with emerging regulatory frameworks for medical AI by providing transparent and verifiable records of model updates and aggregation decisions.

3) Trust in Heterogeneous Settings

NSDS-based adaptive weighting ensures that nodes with unique data distributions are not unfairly penalized, promoting participation in federated consortia.

B. Limitations and Future Work

1) Scalability

Current experiments use 10 nodes. Future work should validate FedXChain with 100+ nodes to assess blockchain scalability and aggregation overhead.

2) Byzantine Robustness

While trust scores provide basic robustness, sophisticated adversarial attacks (e.g., gradient poisoning, backdoor injection) require additional defenses [10, 22, 23, 44]. A future study is planned for integration with Byzantine-robust aggregation techniques.

3) Heterogeneous Model Architectures

Current experiments use the same architecture across nodes. Supporting cross-architecture federated learning (e.g., mixing CNNs, RNNs, and Transformers) requires SHAP adaptation for diverse model types.

4) Communication Efficiency

SHAP value transmission adds overhead [41, 42]. Future work will explore dimensionality reduction and compression techniques for efficient explainability communication [63].

5) Dynamic Node Participation

Current experiments assume consistent node participation. Handling dynamic join/leave scenarios with trust score persistence is important for real-world deployment.

C. Broader Impact

By providing interpretable and auditable federated learning, FedXChain enables smaller organizations, such as hospitals, clinics, and research institutions, to participate in collaborative AI development while maintaining data sovereignty. Moreover, NSDS-based fairness metrics help identify and mitigate bias in federated aggregation, ensuring that minority data distributions are not overshadowed by majority patterns [34, 35, 64]. The validation methodology (multi-model, real data, statistical rigor) employed in the present study sets a standard for reproducible federated learning research.

VII. CONCLUSION

The current study presented FedXChain, a framework for explainable, trustworthy, and auditable federated learning. Through extensive validation across three model architectures (Logistic Regression, MLP, Random Forest) on real-world medical data from the Wisconsin Breast Cancer dataset (569 breast tissue samples), the study demonstrated that FedXChain achieves excellent performance (96.50 accuracy) with outstanding statistical reproducibility (CV $\leq$ 2 across 5 independent runs).

The key innovations include: (1) Federated-SHapley Additive exPlanations (SHAP) aggregation with formal Node-Specific Divergence Scores (NSDS)-based local fidelity quantification, (2) adaptive trust scoring that combines accuracy, explainability, and consistency, (3) blockchain-verified audit trails for transparent aggregation, and (4) rigorous multi-model validation with comprehensive statistical analysis.

The evaluation demonstrates FedXChain's effectiveness through model-agnostic validation, real-world dataset evaluation, formal mathematical definitions, and robust statistical reproducibility. FedXChain represents a significant step toward trustworthy, interpretable federated AI suitable for deployment in regulated domains like healthcare, where transparency and accountability are significant.

Future work will focus on scaling to larger federations (100+ nodes), enhancing Byzantine robustness against sophisticated attacks, supporting heterogeneous model architectures across nodes, and optimizing communication efficiency for SHAP value transmission.

ACKNOWLEDGMENT

The authors would like to thank the Electrical Engineering Department, Faculty of Engineering, Universitas Brawijaya, and the Laboratory of Internet of Things and Human Centered Design, Faculty of Vocational Studies, Universitas Brawijaya, for providing computational resources and supercomputer support for this research.

REFERENCES

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, USA, vol. 54, pp. 1273–1282, 2017.

[2] P. Kairouz and H. B. McMahan, "Advances and Open Problems in Federated Learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, June 2021, https://doi.org/10.1561/2200000083.

[3] N. Rieke *et al.*, "The Future of Digital Health with Federated Learning," *npj Digital Medicine*, vol. 3, no. 1, Sept. 2020, Art. no. 119, https://doi.org/10.1038/s41746-020-00323-1.

[4] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–19, Mar. 2019, https://doi.org/10.1145/3298981.

[5] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A Survey on Federated Learning," *Knowledge-Based Systems*, vol. 216, Mar. 2021, Art. no. 106775, https://doi.org/10.1016/j.knosys.2021.106775.

[6] Q. Li *et al.*, "A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3347–3366, Apr. 2023, https://doi.org/10.1109/TKDE.2021.3124599.

[7] L. U. Khan, W. Saad, Z. Han, E. Hossain, and C. S. Hong, "Federated Learning for Internet of Things: Recent Advances, Taxonomy, and Open Challenges," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1759–1799, 2021, https://doi.org/10.1109/COMST.2021.3090430.

[8] A. Imteaj, U. Thakker, S. Wang, J. Li, and M. H. Amini, "A Survey on Federated Learning for Resource-Constrained IoT Devices," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 1–24, Jan. 2022, https://doi.org/10.1109/JIOT.2021.3095077.

[9] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A Survey on Security and Privacy of Federated Learning," *Future Generation Computer Systems*, vol. 115, pp. 619–640, Feb. 2021, https://doi.org/10.1016/j.future.2020.10.007.

[10] L. Lyu *et al.*, "Privacy and Robustness in Federated Learning: Attacks and Defenses," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 7, pp. 8726–8746, July 2024, https://doi.org/10.1109/TNNLS.2022.3216981.

[11] M. Aledhari, R. Razzak, R. M. Parizi, and F. Saeed, "Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Access*, vol. 8, pp. 140699–140725, 2020, https://doi.org/10.1109/ACCESS.2020.3013541.

[12] H. Zhu, J. Xu, S. Liu, and Y. Jin, "Federated Learning on Non-IID Data: A Survey," *Neurocomputing*, vol. 465, pp. 371–390, Nov. 2021, https://doi.org/10.1016/j.neucom.2021.07.098.

[13] X. Ma, J. Zhu, Z. Lin, S. Chen, and Y. Qin, "A State-of-the-Art Survey on Solving Non-IID Data in Federated Learning," *Future Generation Computer Systems*, vol. 135, pp. 244–258, Oct. 2022, https://doi.org/10.1016/j.future.2022.05.003.

[14] K. Wei *et al.*, "Federated Learning with Differential Privacy: Algorithms and Performance Analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020, https://doi.org/10.1109/TIFS.2020.2988575.

[15] S. Truex, L. Liu, K.-H. Chow, M. E. Gursoy, and W. Wei, "LDP-Fed: Federated Learning with Local Differential Privacy," in *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, Heraklion, Greece, Apr. 2020, pp. 61–66, https://doi.org/10.1145/3378679.3394533.

[16] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated Optimization in Heterogeneous Networks," in *Proceedings of the 3rd Conference on Machine Learning and Systems*, Austin, TX, USA, 2020.

[17] S. A. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic Controlled Averaging for Federated Learning," in *Proceedings of the 37th International Conference on Machine Learning*, Virtual, 2020, pp. 5132–5143.

[18] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated Learning on Non-IID Features via Local Batch Normalization." arXiv, 2021, https://doi.org/10.48550/ARXIV.2102.07623.

[19] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization," in *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, Dec. 2020, pp. 7611–7623.

[20] C. Fung, C. J. M. Yoon, and I. Beschastnikh, "Mitigating Sybils in Federated Learning Poisoning," arXiv, 2018, https://doi.org/10.48550/ARXIV.1808.04866.

[21] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent," in *Advances in Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 118–128.

[22] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping," in *Proceedings 2021 Network and Distributed System Security Symposium*, Virtual, 2021, https://doi.org/10.14722/ndss.2021.24434.

[23] Z. Zhang, X. Cao, J. Jia, and N. Z. Gong, "FLDetector: Defending Federated Learning Against Model Poisoning Attacks via Detecting Malicious Clients," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA, Aug. 2022, pp. 2545–2555, https://doi.org/10.1145/3534678.3539231.

[24] J. Wen, Z. Zhang, Y. Lan, Z. Cui, J. Cai, and W. Zhang, "A Survey on Federated Learning: Challenges and Applications," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 2, pp. 513–535, Feb. 2023, https://doi.org/10.1007/s13042-022-01647-y.

[25] K. Pfeiffer, M. Rapp, R. Khalili, and J. Henkel, "Federated Learning for Computationally Constrained Heterogeneous Devices: A Survey," *ACM Computing Surveys*, vol. 55, no. 14s, pp. 1–27, Dec. 2023, https://doi.org/10.1145/3596907.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 1135–1144, https://doi.org/10.1145/2939672.2939778.

[27] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett, Eds. Red Hook, NY: Curran Associates, Inc, 2018.

[28] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, June 2020, https://doi.org/10.1016/j.inffus.2019.12.012.

[29] R. Dwivedi *et al.*, "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–33, Sept. 2023, https://doi.org/10.1145/3561048.

[30] I. C. Covert, S. Lundberg, and S. I. Lee, "Explaining by Removing: A Unified Framework for Model Explanation," *Journal of Machine Learning Research*, vol. 22, no. 2021, pp. 1–90, Sept. 2021.

[31] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, "Explainable AI Methods - A Brief Overview," in *xxAI - Beyond Explainable AI*, vol. 13200, A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, Eds. Cham, Switzerland: Springer International Publishing, 2022, pp. 13–38.

[32] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, Dec. 2020, Art. no. 18, https://doi.org/10.3390/e23010018.

[33] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. Vincent Poor, "Federated Learning for Internet of Things: A Comprehensive Survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021, https://doi.org/10.1109/COMST.2021.3075439.

[34] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards Personalized Federated Learning," *IEEE Transactions on Neural Networks and*

Learning Systems, vol. 34, no. 12, pp. 9587–9603, Dec. 2023, https://doi.org/10.1109/TNNLS.2022.3160699.

[35] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and Robust Federated Learning Through Personalization," in *Proceedings of the 38th International Conference on Machine Learning*, Virtual, July 2021, pp. 6357–6368.

[36] Y. Zhan, J. Zhang, Z. Hong, L. Wu, P. Li, and S. Guo, "A Survey of Incentive Mechanism Design for Federated Learning," *IEEE Transactions on Emerging Topics in Computing*, pp., 1035–1044, 2021, https://doi.org/10.1109/TETC.2021.3063517.

[37] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated Learning with Non-IID Data," 2018, https://doi.org/10.48550/ARXIV.1806.00582.

[38] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated Learning: Strategies for Improving Communication Efficiency." arXiv, 2016, https://doi.org/10.48550/ARXIV.1610.05492.

[39] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the Convergence of FedAvg on Non-IID Data," in *Proceedings of the International Conference on Learning Representations*, Virtual, 2020.

[40] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting Shared Representations for Personalized Federated Learning," in *Proceedings of the 38th International Conference on Machine Learning*, Virtual, 2021, pp. 2089–2099.

[41] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-Efficient Federated Learning," *Proceedings of the National Academy of Sciences*, vol. 118, no. 17, Apr. 2021, Art. no. e2024789118, https://doi.org/10.1073/pnas.2024789118.

[42] W. Liu, L. Chen, Y. Chen, and W. Zhang, "Accelerating Federated Learning via Momentum Gradient Descent," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 8, pp. 1754–1766, Aug. 2020, https://doi.org/10.1109/TPDS.2020.2975189.

[43] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates," in *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, 2018, pp. 5650–5659.

[44] M. Shayan, C. Fung, C. J. M. Yoon, and I. Beschastnikh, "Biscotti: A Blockchain System for Private and Secure Federated Learning," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 7, pp. 1513–1525, July 2021, https://doi.org/10.1109/TPDS.2020.3044223.

[45] J. Kang, Z. Xiong, D. Niyato, S. Xie, and J. Zhang, "Incentive Mechanism for Reliable Federated Learning: A Joint Optimization Approach to Combining Reputation and Contract Theory," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10700–10714, Dec. 2019, https://doi.org/10.1109/JIOT.2019.2940820.

[46] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing Federated Learning on Non-IID Data with Reinforcement Learning," in *IEEE Conference on Computer Communications*, Toronto, ON, Canada, July 2020, pp. 1698–1707, https://doi.org/10.1109/INFOCOM41043.2020.9155494.

[47] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, https://doi.org/10.1109/ACCESS.2018.2870052.

[48] T. Shen *et al.*, "Federated Mutual Learning." arXiv, 2020, https://doi.org/10.48550/ARXIV.2006.16765.

[49] H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Blockchained On-Device Federated Learning," *IEEE Communications Letters*, vol. 24, no. 6, pp. 1279–1283, June 2020, https://doi.org/10.1109/LCOMM.2019.2921755.

[50] Y. Li, C. Chen, N. Liu, H. Huang, Z. Zheng, and Q. Yan, "A Blockchain-Based Decentralized Federated Learning Framework with Committee Consensus," *IEEE Network*, vol. 35, no. 1, pp. 234–241, Jan. 2021, https://doi.org/10.1109/MNET.011.2000263.

[51] S. R. Pokhrel and J. Choi, "Federated Learning with Blockchain for Autonomous Vehicles: Analysis and Design Challenges," *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 4734–4746, Aug. 2020, https://doi.org/10.1109/TCOMM.2020.2990686.

[52] Y. Qu, M. P. Uddin, C. Gan, Y. Xiang, L. Gao, and J. Yearwood, "Blockchain-Enabled Federated Learning: A Survey," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–35, Apr. 2023, https://doi.org/10.1145/3524104.

[53] M. Ali, H. Karimipour, and M. Tariq, "Integration of Blockchain and Federated Learning for Internet of Things: Recent Advances and Future Challenges," *Computers & Security*, vol. 108, Sept. 2021, Art. no. 102355, https://doi.org/10.1016/j.cose.2021.102355.

[54] Y. Lu, X. Huang, K. Zhang, S. Maharjan, and Y. Zhang, "Blockchain Empowered Asynchronous Federated Learning for Secure Data Sharing in Internet of Vehicles," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 4298–4311, Apr. 2020, https://doi.org/10.1109/TVT.2020.2973651.

[55] D. C. Nguyen *et al.*, "Federated Learning for Smart Healthcare: A Survey," *ACM Computing Surveys*, vol. 55, no. 3, pp. 1–37, Mar. 2023, https://doi.org/10.1145/3501296.

[56] W. Wolberg, O. Mangasarian, N. Street, and W. Street, "Breast Cancer Wisconsin (Diagnostic)." UCI Machine Learning Repository, 1993, [Online]. Available: https://archive.ics.uci.edu/dataset/17.

[57] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated Learning for Healthcare Informatics," *Journal of Healthcare Informatics Research*, vol. 5, no. 1, pp. 1–19, Mar. 2021, https://doi.org/10.1007/s41666-020-00082-4.

[58] R. S. Antunes, C. André Da Costa, A. Küderle, I. A. Yari, and B. Eskofier, "Federated Learning for Healthcare: Systematic Review and Architecture Proposal," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 4, pp. 1–23, Aug. 2022, https://doi.org/10.1145/3501813.

[59] I. Dayan *et al.*, "Federated learning for predicting clinical outcomes in patients with COVID-19," *Nature Medicine*, vol. 27, no. 10, pp. 1735–1743, Oct. 2021, https://doi.org/10.1038/s41591-021-01506-3.

[60] M. J. Sheller *et al.*, "Federated Learning in Medicine: Facilitating Multi-institutional Collaborations Without Sharing Patient Data," *Scientific Reports*, vol. 10, no. 1, July 2020, Art. no. 12598, https://doi.org/10.1038/s41598-020-69250-1.

[61] A. Qayyum, K. Ahmad, M. A. Ahsan, A. Al-Fuqaha, and J. Qadir, "Collaborative Federated Learning for Healthcare: Multi-Modal COVID-19 Diagnosis at the Edge," *IEEE Open Journal of the Computer Society*, vol. 3, pp. 172–184, 2022, https://doi.org/10.1109/OJCS.2022.3206407.

[62] Q. Wu, X. Chen, Z. Zhou, and J. Zhang, "FedHome: Cloud-Edge Based Personalized Federated Learning for In-Home Health Monitoring," *IEEE Transactions on Mobile Computing*, vol. 21, no. 8, pp. 2818–2832, Aug. 2022, https://doi.org/10.1109/TMC.2020.3045266.

[63] C. Xie, O. Koyejo, and G. Gupta, "Asynchronous Federated Optimization," in *12th Annual Workshop on Optimization for Machine Learning*, Virtual, 2020.

[64] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic Federated Learning," in *36th International Conference on Machine Learning*, Long Beach, CA, USA, 2019.