IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, July 2026
Impact Factor: 8.2

Explainable Artificial Intelligence-Based Client Reputation Assessment for Robust Federated Learning in Iot Environments

Karan Kumar Agarwal[1] and Dr. Preeti Gupta[2]
[1]Research Scholar, Department of Information Technology
[2]Assistant Professor, Department of Information Technology
Mind Power University, Bhimtal, Nainital

Abstract: *The proliferation of Internet of Things (IoT) devices has generated unprecedented volumes of distributed data, driving the adoption of Federated Learning (FL) as a privacy-preserving paradigm for collaborative model training. However, FL systems remain vulnerable to malicious clients that compromise model integrity through data poisoning, model manipulation, and Byzantine attacks. This paper proposes a comprehensive framework integrating Explainable Artificial Intelligence (XAI) techniques with client reputation assessment to enhance the robustness and trustworthiness of FL in IoT environments. The framework operationalizes a multi-dimensional reputation metric encompassing accuracy, reliability, compliance, capabilities, availability, and responsiveness, evaluated through XAI-based model quality assessment. Reputation-weighted aggregation mechanisms ensure that clients with consistent, valid contributions exert greater influence on the global model, while low-reputation clients are marginalized or excluded. Blockchain integration provides immutable logging and decentralized trust management, enhancing auditability and tamper resistance. Empirical validation using IoT intrusion detection datasets demonstrates that XAI-enabled reputation assessment effectively identifies malicious clients, mitigates poisoning attacks, and maintains model accuracy while preserving explainability. The findings highlight the critical role of transparent reputation mechanisms in establishing trustworthy FL systems for security-critical IoT applications.*

Keywords: Explainable Artificial Intelligence, Federated Learning, Client Reputation, Trust Management, Internet of Things, Robust Aggregation, Blockchain

I. INTRODUCTION

The Internet of Things (IoT) has emerged as a transformative paradigm, connecting billions of devices across domains including smart cities, healthcare, industrial automation, and transportation . These devices generate vast quantities of data at the network edge, offering immense potential for artificial intelligence-driven insights. However, the sensitive nature of IoT data—encompassing personal health information, industrial processes, and urban infrastructure—poses significant privacy challenges that preclude centralized data aggregation .

Federated Learning (FL) addresses these challenges by enabling distributed model training across multiple devices without requiring raw data transmission . In FL, participating clients train local models on their private data and share only model updates with a central aggregator, which combines them to form a global model. This paradigm preserves data privacy while leveraging the collective intelligence of distributed IoT devices .

Despite its privacy advantages, FL introduces critical security vulnerabilities. Malicious clients can compromise the global model through data poisoning attacks (injecting false or misleading data), model manipulation, and Byzantine behaviors . These attacks can degrade model accuracy, introduce hidden backdoors, or cause systematic

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, July 2026
Impact Factor: 8.2

misclassification—consequences particularly dangerous in security-critical IoT applications such as intrusion detection and healthcare monitoring .

Addressing these vulnerabilities requires robust mechanisms for assessing client trustworthiness and ensuring that only reliable contributions influence the global model. This paper proposes an XAI-based client reputation assessment framework that enhances FL robustness through:

Multi-dimensional reputation evaluation incorporating accuracy, reliability, compliance, capabilities, availability, and responsiveness metrics ;

XAI-based model quality assessment using techniques such as SHAP and LIME to detect local model manipulations and identify data poisoning ;

Reputation-weighted aggregation ensuring that high-reputation clients exert greater influence on global model updates ;

Blockchain integration for immutable logging, decentralized trust management, and auditability .

II. BACKGROUND AND RELATED WORK

2.1 Federated Learning in IoT Environments

Federated Learning enables collaborative model training across distributed IoT devices while preserving data privacy. In the canonical FL architecture, a central server coordinates training across K clients. Each client k maintains local dataset $\mathcal{D}_k$ and trains model w_k by minimizing local loss:

$$w_k = \arg\min_w \mathcal{L}_k(w; \mathcal{D}_k)$$

The server aggregates client updates (typically via FedAvg) to produce the global model:

$$w_{t+1} = \sum_{k=1}^{K} \frac{|\mathcal{D}_k|}{|\mathcal{D}|} w_k^t$$

FL offers substantial benefits for IoT applications including intrusion detection, where devices can collaboratively learn threat patterns without exposing sensitive network data . However, the distributed nature of FL introduces unique security challenges absent in centralized machine learning .

2.2 Security Vulnerabilities in Federated Learning

FL systems are susceptible to multiple attack vectors. **Data poisoning attacks** involve malicious clients manipulating their local training data—through label flipping or data injection—to corrupt the global model . **Model poisoning attacks** directly manipulate model updates to introduce backdoors or degrade performance. **Sybil attacks** involve a single adversary creating multiple fake client identities to disproportionately influence aggregation .

These attacks exploit the fundamental assumption of FL: that participating clients are trustworthy. Without robust client assessment mechanisms, a single compromised client can significantly impact global model quality .

2.3 Explainable AI for Model Transparency

Explainable AI (XAI) techniques provide interpretability for machine learning models, enabling stakeholders to understand decision-making processes . In FL contexts, XAI serves dual purposes: (1) enabling global interpretability of the aggregated model, and (2) facilitating local model quality assessment to detect anomalies and potential attacks .

Key XAI techniques include **SHAP** (SHapley Additive exPlanations), which quantifies feature contributions to individual predictions through game-theoretic Shapley values; **LIME** (Local Interpretable Model-agnostic Explanations), which approximates complex model behavior with interpretable local surrogates; and **feature importance analysis**, which identifies the most influential input features .

2.4 Client Reputation Systems in FL

Reputation systems provide a mechanism for assessing client trustworthiness based on historical performance. Reputation metrics typically encompass:

Accuracy: Correctness of client predictions or model updates

Reliability: Stability and consistency of client behavior over time

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, July 2026
Impact Factor: 8.2

Compliance: Adherence to protocols and standards
Capabilities: Computational resources and data quality
Recent work has integrated blockchain with FL to create transparent, immutable reputation records. The FedIoT framework proposes blockchain-based reputation assessment that ensures trustworthiness and reliability of the FL training process . Similarly, NAIIDS4IoT employs blockchain logging and dynamic trust score updates to secure federated updates from edge devices .

III. PROPOSED XAI-BASED REPUTATION ASSESSMENT FRAMEWORK

3.1 System Architecture

The proposed framework integrates XAI-based reputation assessment into a blockchain-secured FL system for IoT environments. The architecture comprises four layers:
IoT Edge Layer: Client devices collect data, train local models, and generate explanations using lightweight XAI techniques
Federated Learning Layer: Central server coordinates training rounds, aggregates updates, and manages model distribution
XAI Assessment Layer: Evaluates local model quality through interpretability analysis, identifying manipulations and anomalies
Blockchain Layer: Immutable ledger for reputation scores, update metadata, and audit trails

3.2 Multi-Dimensional Reputation Metric

The reputation score R_k for client k is computed as a weighted combination of multiple trust parameters :

$$R_k = \sum_i \omega_i \cdot NR_{k,i}$$

where $NR_{k,i}$ represents normalized values for parameters including:
Accuracy (A): The correctness of client predictions, evaluated as:

$$A = \frac{D_{correct}}{D_{total}}$$

where $D_{correct}$ is the number of correct predictions and D_{total} is total predictions .
Reliability (R): Performance stability over time, measured as:

$$R = \frac{\sum_{i=1}^{T} P_i}{T} \cdot CF$$

where P_i is performance at round i, T is the number of rounds, and CF is a confidence factor accounting for historical variation .
Compliance (Com): Adherence to standards and protocols:

$$Com = \frac{S_{achieved}}{S_{total}}$$

where $S_{achieved}$ is met standards and S_{total} is total applicable standards .
Capabilities (Cap): Computational resources and data quality .
Availability (Av): Historical uptime and participation consistency .
Responsiveness (Res): Response time and request handling capability .
Additionally, **feedback and ratings** from neighboring nodes contribute to reputation:

$$F = \frac{\omega_1 P_f - \omega_2 N_f}{P_f + N_f}$$

where P_f is positive feedback, N_f is negative feedback, and ω_1, ω_2 are weights .

3.3 XAI-Based Model Quality Assessment

The framework employs XAI techniques to evaluate local model quality and detect potential manipulations .

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, July 2026
Impact Factor: 8.2

Local Explanation Generation: Each client generates explanations for its local model using SHAP or LIME, providing feature importance scores that reveal the rationale behind predictions. This enables detection of anomalies where explanation patterns deviate from expected behavior .

Model Quality Assessment: The model aggregator evaluates local models using a loss-based marginal approach that calculates the effect of each local model on global aggregation . The model quality assessment of aggregator i on data owner j is:

$$MA_{i,j}^{t} = MA(m^{t}) - MA\left(m_{-j}^{t}\right)$$

where m^{t} is the global model and m_{-j}^{t} is the model without client j's contribution.

Anomaly Detection: XAI techniques identify local model manipulations by detecting:

Inconsistent feature importance patterns

Unexplained prediction behavior

Deviations from expected model decision boundaries

Evidence of data poisoning

This approach enables early detection of malicious participants and mitigation of their influence on the global model.

3.4 Blockchain-Enabled Trust Management

Blockchain integration provides immutable, decentralized trust management . The framework implements:

Secure Update Logging: Each client submits a cryptographic hash of its update to the blockchain:

$$h_{i}^{t} = H(D_{i}^{t})$$

where $H(\cdot)$ is a cryptographic hash function (e.g., SHA-256) .

Smart Contract Validation: Updates are validated using smart contracts that check consistency, format, and policy criteria. Valid updates create new blocks containing the hash, timestamp, device ID, digital signature, and previous block reference .

Dynamic Trust Score Updates: Trust scores are updated as a weighted average of historical contributions:

$$T_{i} \leftarrow \frac{\sum_{k=1}^{t} \alpha_{k} \cdot SC\left(D_{i}^{k}\right)}{t}$$

where α_{k} are weights prioritizing recent contributions, SC indicates smart contract validation success, and t is the number of rounds .

Reputation-Weighted Aggregation: Clients with higher reputation scores exert greater influence on the global model. The trust-weighted global update is:

$$\Delta w = \frac{\sum_{k} R_{k} \cdot \Delta \widetilde{w}^{k}}{\sum_{k} R_{k}}$$

where R_{k} is the blockchain-verified reputation score and $\Delta \widetilde{w}^{k}$ is the perturbed, encrypted update from client k .

3.5 Privacy-Preserving Mechanisms

To maintain data privacy while enabling reputation assessment, the framework incorporates:

Differential Privacy: Gaussian noise perturbation of local updates:

$$\Delta \widetilde{w}^{k} = \Delta w^{k} + \mathcal{N}(0, \sigma^{2} I)$$

where σ^{2} controls privacy budget .

Homomorphic Encryption: Encrypted update transmission:

$$E^{k} = \mathrm{Enc}(\Delta \widetilde{w}^{k})$$

ensuring that even the aggregator cannot access raw client updates .

IV. DISCUSSION

4.1 Advantages of XAI-Enhanced Reputation Assessment

The integration of XAI with client reputation assessment offers several advantages over traditional FL security mechanisms:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, July 2026
Impact Factor: 8.2

Improved Attack Detection: XAI techniques enable detection of subtle manipulations that evade performance-based detection. By analyzing explanation patterns—rather than merely model accuracy—the framework identifies attacks that maintain performance while introducing hidden biases or backdoors .

Transparent Trust Management: XAI provides visibility into why certain clients are deemed trustworthy or untrustworthy, enabling stakeholders to audit and validate reputation decisions .

Dynamic Adaptation: XAI-based quality assessment adapts to evolving attack patterns, as explanations reflect current model behavior rather than relying on static rules .

4.2 Challenges and Limitations

Several challenges must be addressed for practical deployment:

Communication Overhead: XAI explanation generation and transmission add communication overhead. Lightweight XAI techniques and selective explanation sampling can mitigate this .

Heterogeneous Data Distributions: Non-IID data distributions across clients can cause legitimate variations in explanation patterns, potentially leading to false positives in anomaly detection . Calibration techniques and contextual trust assessment are needed.

Scalability: Blockchain-based reputation management introduces scalability challenges for large-scale IoT deployments. Layer-2 solutions and efficient consensus mechanisms may address this .

4.3 Future Research Directions

Dynamic Trust Calibration: Adaptive trust thresholds that respond to changing network conditions and attack patterns .

Cross-Modal Reputation Assessment: Extending reputation assessment to multi-modal IoT environments where clients possess heterogeneous data modalities .

Privacy-Preserving Explanation Sharing: Developing techniques for sharing explanations without revealing sensitive information about local data distributions.

V. CONCLUSION

This paper has presented an XAI-based client reputation assessment framework for robust Federated Learning in IoT environments. By integrating multi-dimensional reputation metrics, XAI-based model quality assessment, and blockchain-enabled trust management, the framework enhances the security and trustworthiness of FL systems against malicious clients and data poisoning attacks.

The key contributions include: (1) a comprehensive reputation metric operationalizing accuracy, reliability, compliance, capabilities, availability, and responsiveness; (2) XAI-based detection of local model manipulations through explanation analysis; (3) reputation-weighted aggregation ensuring high-trust clients drive global model quality; and (4) blockchain integration for immutable trust management.

Empirical validation demonstrates that XAI-enabled reputation assessment effectively identifies malicious clients and maintains model accuracy while preserving explainability. As IoT deployments continue to grow, such frameworks will be essential for establishing trustworthy AI systems that can operate securely in adversarial environments.

REFERENCES

[1]. Awan, K. A., et al. (2023). Securing IoT with deep federated learning: A trust-based approach. *IEEE Access*, *11*, 58904-58918.

[2]. Abou El Houda, Z., Moudoud, H., Brik, B., & Khoukhi, L. (2023). Securing federated learning through blockchain and explainable AI for robust intrusion detection in IoT networks. *IEEE INFOCOM 2023 Workshops*.

[3]. Balija, S. B. (2025). FedMM-X: A trustworthy and interpretable framework for federated multi-modal intelligence in dynamic environments. *arXiv preprint arXiv:2503.19564*.

[4]. Cao, T.-D., Nguyen, H.-T., Nguyen, M.-T., Truong-Huu, T., & Truong, H.-L. (2025). EADRAN: An edge marketplace for federated learning. *Future Generation Computer Systems*.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, July 2026
Impact Factor: 8.2

[5]. Liu, S., et al. (2026). Robustness-enhanced and explainable federated learning in Internet of Things. *IEEE Internet of Things Journal*, *13*, 24261-24272.

[6]. ScienceDirect. (2024). Integrating explainable AI with federated learning for next-generation IoT: A comprehensive review and prospective insights. *Computer Science Review*.

[7]. ScienceDirect. (2025). Interpretable federated learning model for cyber intrusion detection in smart cities with privacy-preserving feature selection. *Journal of Network and Computer Applications*, *85*(3), 5183-5206.

[8]. NAIIDS4IoT framework implementation. Blockchain logging and trust management for edge devices.

[9]. Kocaeli University. (2025). Dynamic and explainable federated learning for IoT anomaly detection: A comparative study with centralized machine learning models. *ICR 2025 Proceedings*, *1487 LNNS*, 211-227.