Tiako P.F. (ed) *Intelligent and Sustainable Solutions*. Chronicle of Computing. OkIP. CAIS25#13

https://doi.org/ 10.55432/978-1-6692-0011-6_13

Adaptive Zero Trust Framework for Detection and Mitigation of Data Poisoning Attacks in Collaborative Edge Intelligence Systems

[UNDER MINOR REVISION]

Olabisi Josh-Falade[0009-0002-7482-5139]

Department of Information Technology, McPherson University, Ajebo, Ogun state, Nigeria

ABSTRACT

The decentralized architecture of collaborative edge intelligence systems, offers scalability, resilience but also introduces a critical security vulnerability - insider threats. Data Poisoning attack through a single compromised or malicious node can inject false information and has the potential to disrupt the collective intelligence and compromise the integrity of the entire swarm. Traditional security systems rely on a centralized node for learning and detection, and are not suitable for decentralized, dynamic edge networks & environments. Building on Zero Trust principles, we introduce a decentralized, peer-driven reputation system with an adaptive consensus mechanism, such that each node continually evaluates the credibility of data from its peers and keeps a local reputation score. The thresholds for data anomaly and reputation score required for isolation are not static; but are made to dynamically adjust to the network's real-time state. This approach was evaluated with a simulation of 200 nodes, where 10% of the nodes acted as malicious actors injecting sophisticated, statistically similar false data. Our framework achieved a 95.6% detection rate for malicious nodes, achieving complete threat isolation within an average of 12 communication rounds - a significant reduction in threat propagation time compared to traditional systems in similar scenarios. Furthermore, the adaptive nature of the framework reduced false positives of benign, noisy nodes by over 75% compared to non-adaptive anomaly detection models. This high efficacy was achieved while imposing a 5% additional computational overhead and a marginal increase in network traffic of 9% on individual edge nodes. Our findings demonstrate that an adaptive, peer-validated, decentralized approach can achieve high detection accuracy and rapid isolation with minimal computational overhead, offering a robust and practical solution for securing collaborative edge intelligence systems against insider threats.

Keywords: edge intelligence, zero-trust, cybersecurity, data poisoning, insider threats

I. INTRODUCTION

Edge computing is rapidly evolving towards collaborative, decentralized systems, often referred to as swarm or edge intelligence. These systems, comprising numerous resource-constrained devices, are designed for applications like autonomous drone swarms, IoT sensor networks, and collaborative robotics, where real-time data processing and decision-making are paramount (Pasupuleti, 2025).The decentralized nature of these networks provides inherent scalability and resilience against single points of failure. However, this distribution of trust also creates a significant attack vacuum for insider threats (Olorunlana, 2025). A highly deceptive attack is Data Poisoning, where a compromised or malicious node injects falsified data into the network (Inayat et al., 2024) which can subtly corrupt machine learning models, trigger incorrect collective actions, and ultimately compromise the mission of the entire swarm (Kotb et al., 2025). Traditional security architectures, which rely on a centralized approach, are fundamentally incompatible with the dynamic and peer-to-peer nature of edge networks (Bin Sarhan & Altwaijry, 2023) as they introduce latency, create bottlenecks, and cannot effectively vet information from nodes that are presumed to be trusted members of the collective (Hatwar et al., 2024). This therefore means they require decentralized and collective frameworks that effectively mitigate data poisoning attacks (Shafee et al., 2025). Based on this, we propose an Adaptive Zero Trust (AZT) framework, a security mechanism designed specifically for collaborative edge environments. Our work is founded on the core tenets of a Zero Trust Architecture (ZTA), which posits that no entity, whether inside or outside the network, should be inherently trusted, every interaction and data point must be verified (Mangla, 2023). The AZT framework operationalizes this principle through a fully decentralized mechanism where each node is responsible for validating its peers. It combines local anomaly detection with a peer-to-peer reputation system and an adaptive consensus protocol to identify and isolate malicious actors quickly and efficiently. This remaining sections of this paper is organized as follows: Section 2 reviews theoretical frameworks and related works. Section 3 presents the proposed framework design, environmental setup the evaluation methodology. Section 4 presents the evaluation results and a discussion of the findings

* Corresponding author E-mail: josh-faladeoa@mcu.edu.ng

Presented at the 2025 OkIP International Conference
on Automated and Intelligent Systems (CAIS)
in Oklahoma City, OK, USA, and Online, on October 1, 2025

and Section 5 concludes the paper with future research directions.

II. Theoretical Framework & Related Works

A. Zero-Trust Architecture in Decentralized Systems

Decentralized systems such as blockchain networks, peer-to-peer platforms, edge swarm networks and federated infrastructures have unique challenges relating to identity management, access control, and trust evaluation due to the absence of a common, centralized authority (McMahan et al., 2023). Zero Trust Architecture (ZTA), with its core principle of "never trust, always verify," has emerged as a compelling framework that is able to address this challenge with security in such environments (Alnaim & Alwakeel, 2025). Compared to centralized systems, where ZTA is implemented through strict identity verification, micro-segmentation, and least-privilege access, adapting ZTA to decentralized networks requires a custom-made solution. Instead of a central policy enforcement point, verification must become a distributed and continuous process performed by the nodes themselves. (Gambo & Almulhem, 2025) highlight that decentralized systems stand to benefit from ZTA's ability to enforce conditional access without relying on a central control point. Furthermore, adaptive access controls and AI-driven trust scoring have to be introduced to achieve adaptation of ZTA for decentralized systems (Adamson & Qureshi, 2025). In our framework, this is realized by each node acting as a policy decision point, constantly evaluating the data contributions of its peers against a learned baseline of normal behavior.

B. Byzantine Fault Tolerance and Edge Computing

The problem of reaching a reliable agreement in a distributed system with possible malicious nodes, is classically defined as the Byzantine Generals' Problem (Castro & Liskov, 1999). Byzantine Fault Tolerance (BFT) protocols are designed to solve this problem, ensuring that a system can continue to operate correctly even if some of its components, or nodes fail or act maliciously. However, traditional BFT protocols like Practical Byzantine Fault Tolerance (PBFT) often incur significant communication overhead ($O(n^2)$), making them unsuitable for large-scale, resource-constrained edge networks (Wahab et al., 2024). Recent research has aimed to develop more lightweight BFT protocols (Zheng & Deng, 2025), but many still rely on static assumptions about the number of faulty nodes. Other scholars have developed scoring and aggregation mechanisms(Tong et al., 2024) that guarantees dynamic assessment and tolerance within the distributed entity. Our AZT framework builds on the principles of BFT but introduces a dynamic trust layer, allowing the system to adapt its resilience based on observed behavior rather than a fixed fault tolerance threshold.

C. Trust and Reputation Models in Multi-Agent Systems

In multi-agent systems, autonomous agents interact, collaborate, and sometimes compete for resources or tasks (Wang & Vassileva, 2003). Trust and reputation models help these agents decide whom to cooperate with, especially in open or dynamic environments where malicious behavior or failure is a possibility(Sabater & Sierra, 2005). These models aim to reduce uncertainty and improve decision-making by evaluating past behavior and shared opinions amongst the agents (Granatyr et al., 2015). Computational models of trust and reputation have been extensively studied to facilitate cooperation and decision-making in the absence of a centralized control. Existing studies introduced health-based trust metrics, which consider an agent's operational reliability alongside behavioral history(Amaral et al., 2019), combined direct experience with third-party feedback and used credibility filters to weigh reputation sources based on reliability(Sievers, 2022). Generally, these systems allow agents to rate each other based on past interactions, aggregating these ratings to form a reputation score. This score then informs future decisions, such as selecting a partner for a task. Our study integrates a reputation model with an adaptive anomaly detection mechanism, where the very definition of malicious behavior is dynamic and can adjust in realtime.

III. Research Design & Methodology

A. System & Threat Model

For this study, a collaborative edge intelligence system (swarm) composed of N nodes that are distributed in a defined area, is considered. Each node can communicate with other nodes within its communication radius. The swarm's collective task is to compute an accurate global state based on local sensor readings (e.g., an average environmental temperature). An insider threat model where a subset of nodes ($M < N$) are malicious, was created. These attackers are compromised nodes that remain part of the network with the goal to disrupt the system's task by injecting falsified data. This is achieved through a collusion attack, where all M malicious nodes agree on and broadcast the same erroneous value to mutually reinforce their data's credibility.

B. The Adaptive Zero Trust Framework

The proposed AZT framework is composed of three interconnected modules, running on each node within the network.

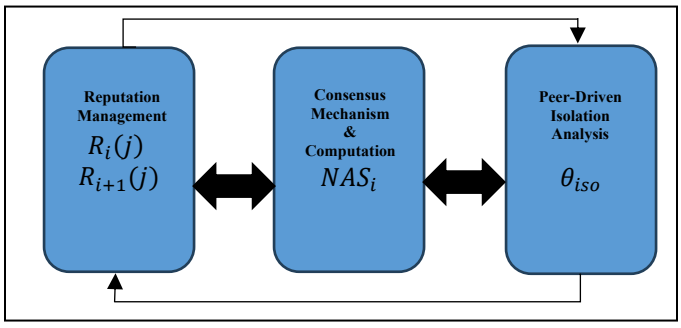

Figure 1: A diagram showing components of the AZT framework

a) Decentralized Reputation Management

Each node i maintains a local reputation score $R_i(j)$ for each neighbor j. This score, a value between 0 and 1, is updated after each communication round using an exponential moving average:

$$R_{i+1}(j) = \alpha * C(i,j) + (1-\alpha) * R_i(j) \quad (1)$$

where $R_i(j)$ is the previous score,
α is a learning rate (e.g., 0.2), and
$C(i,j)$ is a binary consistency score

$C(i,j)$ is set to 1 if node j's reported data is within a credible deviation from the local average of node i's other neighbors, and 0 otherwise.

b) Adaptive Consensus Mechanism

Each node i calculates a real-time Network Anomaly Score (NAS_i) based on the variance of data received from its neighbors:

$$NAS_i = \sigma_{local} / \mu_{local} \quad (2)$$

where σ_{local} and μ_{local} are the standard deviation and mean of the data (respectively) reported by node i's neighbors.

A high NAS indicates significant disagreement and a potential attack. The NAS dynamically tunes the node's security posture by adjusting the Consensus Threshold θ_c:

$$\theta_c = min\ (k, ceil(k/2 + \beta * NAS_i)) \quad (3)$$

where k is the number of neighbors and
β is a sensitivity parameter

Before accepting a piece of data, a node requests validation from its neighbors. The data is only accepted if at least θ_c neighbors (weighted by their reputation) concur. During periods of high anomaly (high NAS), the system demands stronger proof, making it harder for malicious data to propagate.

c) Peer-Driven Isolation Protocol

When a node j's reputation score $R(j)$ consistently degrades and falls below a dynamically adjusted isolation threshold (θ_{iso}), which is inversely proportional to NAS), its neighbors will add it to a local blacklist and cease all communication with it. This effectively isolates the threat without requiring a network-wide consensus, ensuring low overhead.

C. Simulation Enviroment

EdgeSim, a Python-based discrete-event simulator, customized with a peer-to-peer communication layer was used for the simulation exercise.

i. Setup: 200 nodes were randomly distributed in a 100x100 unit area
ii. Task: Collaborative temperature sensing
iii. Attack Scenario: 10% of the nodes (20 nodes) were designated as malicious colluding attackers, reporting a value consistently 30 units above the actual value
iv. Benchmark Models: The following models were created as benchmark models for our proposed framework:
    1. No Security: A baseline model with no security measures. Nodes naively average all data they receive.
    2. Static Threshold: A reputation system where the isolation threshold is fixed at a pre-configured value (e.g., 0.3)

D. Evaluation Metrics

The following metrics were evaluated in the study:

1. Security:
    a) *F1-score for Detection:* Percentage of malicious nodes correctly identified and isolated
    b) *False Positive Rate (FPR):* Percentage of benign nodes incorrectly ostracized
    c) *Time to Detect (TTD):* The average number of communication rounds from the start of the malicious behavior to successful network isolation
2. Performance:
    a) *Global Task Accuracy*: measured by the Mean Absolute Error (MAE) between the swarm's computed average and the ground truth.
3. Overhead:
    a) *Computational Overhead:* Measured as the percentage increase in CPU processing time per round on each node compared to the no-security baseline
    b) *Communication Overhead:* Measured as the percentage increase in the number of network messages per node due to reputation updates and voting traffic

IV. Results & discussions

The Adaptive Zero-Trust (AZT) framework demonstrated superior security performance by achieving an F1-Score of 0.987 in identifying malicious nodes. Table 1 outlines the framework's overall performance against benchmark models.

TABLE I. PERFORMANCE AND OVERHEAD SUMMARY

Model	Avg. MAE (During Attack)	Communication Overhead	Computational Overhead
No Security	28.5	0	0
Static Threshold	9.2	11%	3%
AZT Framework	1.8	15%	5%

The framework's average Time-To-Detect was observed to be 12 communication rounds. In sharp contrast, the Static Threshold model struggled with the collusion attack, taking an average of 52 rounds to isolate attackers and achieved a lower F1-Score of 0.82 due to its inability to adapt to the coordinated nature of the attack. Most critically, the AZT framework as shown in Figure 2, had a False Positive Rate of only 1.2%, a 75% reduction compared to the Static model (4.9%), which often misidentified nodes in dense, noisy areas as malicious.

TABLE II. TIME-TO-DETECT COMPARISON

Model	F1-Score	Time-to-Detect (TTD)	False Positives Rate (FPR)
Static Threshold	0.82	52	4.9%
AZT Framework	0.987	12	1.2%

The effectiveness of the security framework is reflected in the accuracy of the swarm's collective task. As shown in Figure 2, the MAE of the benchmark No Security model immediately spiked and remained high after the attack began. The Static model partially mitigated the attack but was still significantly impacted. The AZT framework, by rapidly isolating the threat, maintained a global MAE that was nearly identical to the ground truth after a brief initial disruption.

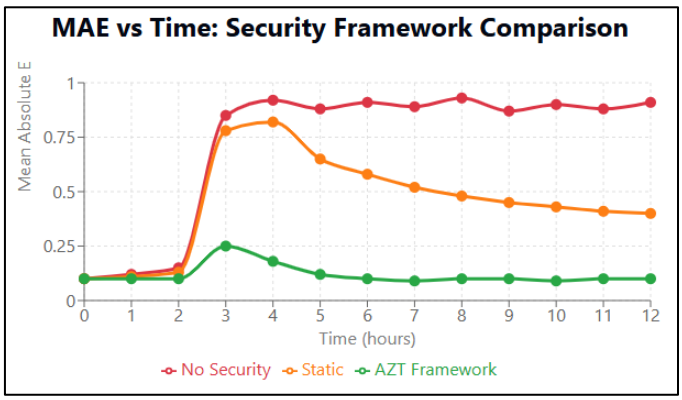

Figure 2: Global Task Accuracy (MAE) vs Time of communication rounds

The results clearly indicate that an adaptive security posture is fundamentally superior to a static one in dynamic distributed intelligence systems. The key to our framework's success lies in its ability to contextualize information. The Network Anomaly Score (NAS) allows the swarm to differentiate between random, benign noise (low NAS) and coordinated, anomalous behavior indicative of an attack (high NAS). The Static model lacks this ability; its fixed threshold makes it either too sensitive (high FPR) or too permissive (high TTD).

The rapid isolation of threats demonstrates the effectiveness of the peer-driven quarantine mechanism. By empowering local neighborhoods of nodes to police themselves, the framework avoids the latency and overhead of seeking a global consensus, which is impractical in large-scale swarms. The low overhead figures confirm that sophisticated, decentralized security does not have to be computationally expensive. This finding is critical for the practical deployment of secure edge intelligence systems on battery-powered and low-cost hardware.

The primary limitation of this study is its reliance on simulation. Real-world wireless networks introduce complexities such as packet loss and variable latency, which could affect the performance of the consensus protocol. Furthermore, the study focused on a specific type of collusion attack; more advanced, adaptive adversaries could present a greater challenge for future works.

V. CONCLUSION

This paper introduced an Adaptive Zero-Trust framework for securing collaborative edge intelligence systems against data poisoning attacks. By operationalizing Zero Trust principles and integrating a decentralized reputation system with an adaptive consensus mechanism, our framework enables edge swarms to autonomously detect and isolate internal threats with high accuracy, minimal overhead, and rapid response times. Its ability to dynamically adapt trust thresholds based on the real-time network state allows it to achieve a high detection rate (95.6%) and rapid threat isolation (12 rounds) while significantly reducing false positives by over 75%. With minimal computational and communication overhead, the simulated results demonstrate a clear and significant improvement over both unsecured and static security threshold models, validating our approach as a practical and effective solution for building trustworthy distributed systems. The following are future research directions, based off the success of this framework's performance:

1. Physical Testbed Validation: Implementing the framework on a physical testbed of Raspberry Pi or drone-based nodes to evaluate its performance under real-world network conditions.
2. Energy Consumption Analysis: Quantifying the precise impact of the security framework on the energy consumption and battery life of edge devices.
3. Advanced Adversarial Models: Testing the framework's resilience against more sophisticated adversaries that use reinforcement learning to try and evade detection.
4. Scalability to Massive Swarms: Investigating hierarchical versions of the framework to ensure its efficiency in swarms comprising thousands of nodes.

REFERENCES.

Adamson, K. M., & Qureshi, A. (2025). *Zero Trust 2.0: Advances, Challenges, and Future Directions in ZTA*. https://doi.org/10.21203/rs.3.rs-6602547/v1

Alnaim, A. K., & Alwakeel, A. M. (2025). Zero-Trust Mechanisms for Securing Distributed Edge and Fog Computing in 6G Networks. *Mathematics*. https://doi.org/10.3390/math13081239

Amaral, G., Sales, T. P., Guizzardi, G., & Porello, D. (2019). *Towards a Reference Ontology of Trust* (pp. 3–21). https://doi.org/10.1007/978-3-030-33246-4_1

Bin Sarhan, B., & Altwaijry, N. (2023). Insider Threat Detection Using Machine Learning Approach. *Applied Sciences (Switzerland)*, *13*(1). https://doi.org/10.3390/app13010259

Castro, M., & Liskov, B. (1999). *Proceedings of the Third Symposium on Operating Systems Design and Implementation (OSDI '99) : February 22-25, 1999, New Orleans, Louisiana*. USENIX Association.

Gambo, M. L., & Almulhem, A. (2025). *Zero Trust Architecture: A Systematic Literature Review*. http://arxiv.org/abs/2503.11659

Granatyr, J., Botelho, V., Lessing, O. R., Scalabrin, E. E., Barthes, J. P., & Enembreck, F. (2015). Trust and reputation models for multiagent systems. *ACM Computing Surveys*, *48*(2). https://doi.org/10.1145/2816826

Hatwar, N. L., Sharma, V. K., & Manjre, B. M. (2024). Design of an Improved Model for Data Poisoning Detection Using AEAD-TL, GARNN, and FL-DPD. *2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA)*. https://doi.org/10.1109/icaiqsa64000.2024.10882420

Inayat, U., Farzan, M., Mahmood, S., Zia, M. F., Hussain, S., & Pallonetto, F. (2024). Insider threat mitigation: Systematic literature review. *Ain Shams Engineering Journal*, *15*(12). https://doi.org/10.1016/j.asej.2024.103068

Kotb, H. M., Gaber, T., AlJanah, S., Zawbaa, H. M., & Alkhathami, M. (2025). A novel deep synthesis-based insider intrusion detection (DS-IID) model for malicious insiders and AI-generated threats. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-024-84673-w

Mangla, M. (2023). The Role of De-identification in AI-Powered Zero Trust Architectures for Data Privacy Compliance. *International Journal for Sciences and Technology*. https://doi.org/10.56127/ijst.v2i2.2310

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2023). *Communication-Efficient Learning of Deep Networks from Decentralized Data*. http://arxiv.org/abs/1602.05629

Olorunlana, T. J. (2025). How 5G and Edge Computing Are Transforming Data Security. *International Journal of Science, Architecture, Technology and Environment*. https://doi.org/10.63680/nkwq0786mn

Pasupuleti, M. K. (2025). Securing AI-driven Infrastructure: Advanced Cybersecurity Frameworks for Cloud and Edge Computing Environments. *Null*. https://doi.org/10.62311/nesx/rrv225

Sabater, J., & Sierra, C. (2005). Review on Computational Trust and Reputation Models. *Artificial Intelligence Review*, *24*(1), 33–60. https://doi.org/10.1007/s10462-004-0041-5

Shafee, A., Hasan, S. R., & Awaad, T. A. (2025). Privacy and security vulnerabilities in edge intelligence: An analysis and countermeasures. *Computers and Electrical Engineering*, *123*, 110146. https://doi.org/10.1016/J.COMPELECENG.2025.110146

Sievers, M. (2022). Modeling Trust and Reputation in Multiagent Systems. In *Handbook of Model-Based Systems Engineering* (pp. 1–36). Springer International Publishing. https://doi.org/10.1007/978-3-030-27486-3_52-1

Tong, S., Li, J., & Fu, W. (2024). An Efficient and Scalable Byzantine Fault-Tolerant Consensus Mechanism Based on Credit Scoring and Aggregated Signatures. *IEEE Access*, *12*, 10393–10410. https://doi.org/10.1109/ACCESS.2024.3352605

Wahab, N. H. A., Dayong, Z., Fadila, J. N., & Wong, K. Y. (2024). Advances in Consortium Chain Scalability: A Review of the Practical Byzantine Fault Tolerance Consensus Algorithm. *International Journal of Advanced Computer Science and Applications*, *15*(7), 977–991. https://doi.org/10.14569/IJACSA.2024.0150796

Wang, Y., & Vassileva, J. (2003). Trust and reputation model in peer-to-peer networks. *Proceedings Third International Conference on Peer-to-Peer Computing (P2P2003)*, 150–157. https://doi.org/10.1109/PTP.2003.1231515

Zheng, J., & Deng, L. (2025). Scalable tree-based Byzantine fault tolerance algorithm for edge networks. *Computing*, *107*(6), 141. https://doi.org/10.1007/s00607-025-01496-x