



ISSN ONLINE: 2447-0228



RESEARCH ARTICLE

OPEN ACCESS

ADAPTIVE IMMUNIZATION AGAINST FEDERATED DATA POISONING: A SEQUENTIAL TRUST-SPECTRAL INTENT FRAMEWORK FOR ROBUST DISTRIBUTED LEARNING PROCESS

Nitesh Hatwar^{*1}, V. K. Sharma² and Bhushan Manjre³

¹Research Scholar, Department of Computer Science and Engineering, Bhagwant University, Ajmer, Rajasthan, India.

²Professor, Department of Computer Science and Engineering, Bhagwant University, Ajmer, Rajasthan, India.

³Associate Professor, Department of Computer Science and Engineering, Govindrao Wanjari College of Engineering, Nagpur, Maharashtra, India.

¹<http://orcid.org/0009-0002-2325-5172>, ²<https://orcid.org/0000-0001-7962-2336>, ³<http://orcid.org/0000-0002-1088-4079>

Email: *msnitesh.88@gmail.com, viren_krec@yahoo.com, bhushan.manjre@gmail.com

ARTICLE INFO

Article History

Received: January 31, 2026

Reviewed: June 1, 2026

Accepted: July 2, 2026

Published: July 31, 2026

Keywords:

Federated Learning, Data Poisoning Defense, Adaptive Security, Deep Learning Robustness

ABSTRACT

Increasing dependence on scattered data pipelines makes modern learning systems vulnerable to data poisoning in federated situations where raw data never centralizes and malicious updates can subtly degrade global models. Current defenses isolate anomalies in one training round, employ static thresholds, or delete suspicious clients, resulting in brittle protection, delayed detection, and model accuracy erosion. They react without understanding or adapting. This sequential, biologically inspired protection design prioritizes analytical structure over surface filtering. Topology-Aware Federated Data Provenance Graphing (TA-FDPG) calculates a trust vector from client update lineage and temporal behaviour to avoid round-wise detector blindness. In Trust-Weighted Spectral Gradient Purification (TW-SGP), singular value decomposition lowers poisoning-aligned gradient components instead of deleting them, preserving learning. Purified gradients feed Latent Behavioral Intent Extraction Network (LB IEN), which extracts latent attack intent signatures to provide the defense a prediction edge. RAMPE uses intent cues instead of heuristics to train effective defense tactics in real time scenarios. Finally, DP IMF creates a powerful, insulated global model from local models and policy-guided privacy noise sets. These methods move poisoned federated learning systems from passive detection to adaptive, anticipatory defense by reducing attack success, reaction latency, and clean accuracy for different scenarios.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

As machine learning decentralizes, federated learning architectures, which train models cooperatively across remote clients without aggregating raw data samples, have evolved quickly. This paradigm offers excellent privacy and scalability [1-3], but it also increases the attack surface, allowing adversaries to manipulate local updates and insert poisoned gradients that subtly or catastrophically taint the global model. Data poisoning in federated contexts is deceptive because malicious behavior blends into a broad mix of client distributions, making anomaly detection difficult [4-6], unstable, or ineffective. Lack of centralized monitoring and non IID data allow poisoning attempts to evolve over numerous rounds and influence learning trajectories before defenders identify measurable differences. Prior research focused on static detection rules, binary client filtering, and single-round statistical screening. The approaches lack temporal dynamics and client behavior provenance. Consequently, models sacrifice accuracy to be conservative or ignore progressive poisoning paths, fragmenting defense. In Process, adversarial training and differential privacy solutions rarely interpret attacker intent or react to shifting threat Vectors in Process. Thus, present solutions are reactive, brittle, and slow to adapt. A security architecture that can recall, reason, and react is needed to address data poisoning [7-9]. The sequential defense pipeline includes five synergistic methods: Topology-Aware Federated Data Provenance Graphing (TA-FDPG), Trust-Weighted Spectral Gradient Purification (TW-SGP), Latent Behavioral Intent Extraction Network (LB IEN), RAMPE, and Differentially Private Immunized Model Fusion.

By performing distinct analytical functions, each component feeds into the next stage, creating an interconnected chain rather than discrete procedures. Defense uses anticipatory resilience instead of passive monitoring. It measures client trust over time, purifies gradients with spectral behavior, recovers latent intent signatures, optimizes defensive policies with reinforcement learning, and fuses models under adaptive privacy constraints. It introduces structural intelligence instead of heuristic filtering in federated security and enables long-term, self-adjusting resilience against complex poisoning campaigns.

1.1 MOTIVATION & CONTRIBUTION

Our effort is driven by the realization that poisoned models are now dangers in high-stakes deployments like autonomous vehicles, medical diagnostics, and industrial analytics. Current defenses lack the temporal awareness and analytical breadth to battle adversaries who change methods across rounds, target classes, or exploit client trust asymmetries. Inability to track behavioral history, understand deeper optimization changes, or predict future attacks produces a structural blind hole that attackers can exploit with little effort and great impact. Defensive federated learning must combine memory, context, and prediction to replace inflexible thresholds with adaptable, developing approaches that preserve model performance without limiting participation or scalability. This paper presents a coherent five-stage defense pipeline that includes provenance analysis, spectrum purification, latent intent extraction, adaptive policy optimization, and privacy-guided model fusion. TW-SGP translates trust into actionable gradient purification without sacrificing accuracy, while TA-FDPG measures trust via temporal topology. In latent space, LB IEN maps poisoning intent to anticipate threats. RAMPE learns meta-policies that change in real time to protect beyond predefined algorithms, while DP IMF builds a private, robust immunized model. This extends the field beyond reactive detection to a rigorous, anticipatory defense paradigm that lowers attack success, speeds reaction, and maintains high clean accuracy across federated rounds.

II. THEORETICAL REFERENCE

II.1 REVIEW OF EXISTING MODELS USED FOR BLOCKCHAIN ANALYSIS

From foundational movements toward decentralization to complex defense ecosystems, the thirty studies indicate a consistent increase in research on federated learning against poisoning, tampering, and model corruption. According to [1] propose BlockFD, a blockchain-based distillation technique to reduce poisoning, and immutable ledgers for provenance assurance in decentralized training in the first paper in this sequence. In [2] show that performance robustness involves structural trust and algorithmic smoothing by switching from ledger-based trust to FedCVG, a two-stage robust optimization strategy. Explainability in poisoning detection by [3] shows interpretability can protect as well as diagnose. According to [4] demonstrate that RHFL requires scale robustness in heterogeneous hierarchical networks and attack surfaces broaden in layered architectures. According to [5] present a direct adversarial targeted response, showing that attacks are becoming more surgical than brute-force. According to [6] return to blockchain in IoT-focused federated settings, noting that low-control edge ecosystems may need decentralized trust anchoring sets. According to [7] clustering-based aggregation method FICAD shows that weighting and grouping client data streams can reduce anomalies. In [8] demonstrate that behavioral inference can combat poisoning better than statistical filtering using a Spyfall Inspired detection model. According to [9] note that diagnostic system poisoning increases human risk, making poisoning defense a medical IoT safety-critical discipline. According to [10] show that fragmented gradient fluxes affect poisoning detection in partitioned models. For [11] use weighted learning to show that not all client updates are equal. They explicitly link aggregation strength to confidence. In contrast, by [12] establish a poisoning attack—peak-controlled logits in distillation—confirming that solutions must anticipate both existing threats and evolving approaches. According to [13] respond with a lightweight privacy-preserving defense, highlighting the computational efficiency-resilience conflict. Privacy and poisoning resilience can coexist, as [14] integrate differential privacy into neural synchronization for industrial streams. In [15] introduce multi-label poisoning defense, which opens up more complex classification spaces where poisoning can effect label associations rather than single outputs. Model families share poisoning risk, according to [16] in federated SVM contexts. According to [17] dynamic Mahalanobis-based defense, DWAMA, illustrates that statistical distance measures can track enemy motion. Poisoning affects preference landscapes and classifiers in recommender systems, according to [18]. With alternating poisoning behaviors, by [19] illustrate that attacks need not be constant to be lethal and that defense must follow behavioral trajectories spanning temporal instance sets. In [20] propose that asynchronous federated learning and blockchain for medical sharing challenge poisoned oversights.

Table 1: Model's Integrated Review Analysis.

Reference	Method	Main Objectives	Findings	Limitations
[1]	BlockFD (Blockchain-based Federated Distillation)	To secure model distillation using immutable blockchain logs to mitigate poisoning.	Demonstrated improved trust and tamper resistance in update aggregation.	High computational and storage overhead due to blockchain integration.
[2]	FedCVG (Two-Stage Robust Optimization)	To enhance robustness in federated optimization via variance-guided adjustment.	Improved convergence stability under poisoning.	Limited adaptability to dynamic or evolving attack strategies.
[3]	Explainable FL Detection Framework	To detect poisoning via interpretable model diagnostics.	Enhanced detection transparency and user trust.	Detection remains largely reactive rather than preventative.
[4]	RHFL (Hierarchical Robust FL)	To defend heterogeneous multi-layer FL systems.	Strengthened multi-level robustness and reduced poisoning impact.	Complex deployment in real hierarchical networks.

[5]	Targeted Poisoning Defense Strategy	To prevent targeted misclassification attacks in FL.	Reduced attack success in focused poisoning scenarios.	Narrow applicability to targeted rather than general poisoning.
[6]	Blockchain Integrated FL for IoT	To secure IoT data sharing and aggregation.	Improved traceability and tamper resistance in IoT settings.	Scalability challenges in dense IoT networks.
[7]	FICAD (Clustering-Based Aggregation)	To group client updates for robust aggregation.	Reduced influence of anomalous update clusters.	Accuracy suffers when benign clients are misclustered.
[8]	SpyShield Behavioral Defense	To identify malicious clients via behavioral cues.	Faster adversary detection compared to statistical filters.	Limited performance under subtle or hidden poisoning.
[9]	IoMT-focused FL Defense	To protect medical IoT models from poisoning.	Demonstrated improvements in clinical model stability.	Domain-specific rather than generalized defense.
[10]	Split-FL Vulnerability Analysis	To assess poisoning risks in split FL architectures.	Identified critical exposure points in model partitioning.	No concrete mitigation strategy proposed.
[11]	Weighted FL Aggregation	To weight client contributions based on reliability.	Reduced effect of compromised clients.	Weight calibration requires careful tuning and historical data.
[12]	Peak-Controlled Logits Attack	To demonstrate a novel poisoning method in distillation.	Exposed vulnerabilities in current distillation defenses.	No defense strategy included.
[13]	Lightweight Privacy-Preserving Defense	To protect resource-constrained FL systems.	Improved security in low-compute environments.	Defense effectiveness drops under high intensity poisoning.
[14]	Differential Privacy Synchronization	To secure industrial data streams with privacy constraints.	Provided privacy and poisoning resilience simultaneously.	Added noise reduced model accuracy in some tasks.
[15]	Multi-Label Poisoning Defense	To defend multi-class and multi-label FL.	Enhanced robustness in complex label spaces.	Higher computational complexity.
[16]	Poisoning Defense for FL-SVM	To mitigate poisoning in federated SVMs.	Reduced poisoning impact in linear classifiers.	Limited generalizability to deep models.
[17]	DWAMA (Mahalanobis-Based Defense)	To detect outlier gradients via statistical distance.	Effective against gradient-space anomalies.	Less effective against well-crafted stealth attacks.
[18]	Contrastive Recommender Defense	To shield recommender FL from targeted poisoning.	Maintained recommendation accuracy under attacks.	Domain-specific, not easily transferable.
[19]	Trajectory Anomaly Detection	To detect alternating poisoning patterns over time.	Improved detection of intermittent attacks.	Requires long-term historical data.
[20]	Asynchronous FL with Blockchain	To support secure medical data sharing with async updates.	Improved robustness in asynchronous settings.	Latency and synchronization remain concerns.
[21]	Dual-Mode Privacy Switching	To adapt privacy strength based on data drift.	Balanced privacy and performance under shifting distributions.	Drift estimation errors may weaken protection.
[22]	Historical Knowledge Modeling	To model long-term poisoning behavior.	Increased detection accuracy by leveraging historical insights.	Requires storage and computation overhead.
[23]	Ring Homomorphism Aggregation	To secure aggregation in wireless sensor networks.	Enhanced confidentiality and aggregation security.	Limited scalability beyond WSN environments.
[24]	Blockchain-Based FL Review	To map blockchain-enhanced FL security strategies.	Identified emerging trends and solution gaps.	No experimental validation or new method proposed.
[25]	Sparse Invisible Backdoor Attack	To introduce a stealthy poisoning method.	Revealed deep vulnerabilities in anomaly detection defenses.	No countermeasure included.
[26]	ARAFL-BAD Asynchronous Defense	To detect botnet induced poisoning in IoT FL.	Improved poisoning resilience in zero-day attacks.	Higher overhead in real-time detection scenarios.
[27]	DEA-Based Client Selection	To choose optimal clients for robust FL.	Reduced dependence on low-quality or compromised nodes.	Selection bias may exclude useful benign clients.
[28]	Energy-Aware FL Security	To secure FL under edge energy constraints.	Improved sustainability and resilience jointly.	Limited evaluation under real attack intensity.
[29]	FedCare Diagnostic Framework	To enable interactive monitoring of FL systems.	Improved fault and attack visibility.	Diagnostic rather than preventive.
[30]	HP_FLAP Homomorphic Aggregation	To protect model parameters using encrypted aggregation.	Enhanced confidentiality and tamper resistance.	High computational cost for fully homomorphic operations.

Source: Authors, (2026).

In this nuanced analysis, as per Table 1, by [21] enhance data drift privacy with dual-mode switching. According to [22] show that historical knowledge modeling defense benefit from long-term updating memory over single-round inspection. In [23] also recommend safe computation for ring homomorphism-based data aggregation. Unifying a continually developing area, by [24] analyze blockchain-enhanced federated security.

According to [25] interrupt this protection phase with a sparse, invisible backdoor attack, showing the fragility of abnormality-based poisoning assumptions. Lohani and Sangal's resilient asynchronous framework for IoT botnet poisoning, ARAFL-BAD [26], suggests federated poisoning and botnet orchestration may merge. According to [27] reframe poisoning as a dynamic client selection and resource allocation problem using data envelopment analysis. Energy-aware federated learning by [28] requires resilience to face attackers and resource constraints. For [29] propose live introspection tools for federated learning systems with FedCare, an interactive diagnosis system. To conclude, by [30] use homomorphic and polymorphic aggregation to create opaque, tamper-resistant training flows for HP_FLAP, stressing the integration of privacy, encryption, and poisoning protections. Studies show consistent evolution of federated poisoning research process. Work in [1], [2] emphasize structural trust and optimization stability before [3], [4] emphasize explainability and hierarchical resilience. The emergence of targeted and domain-specific defenses in [5], [8], [9], [13] suggests that poisoning can disrupt context. Combining privacy-centric solutions in [14], [21] with secure computation concepts in [23], [30] makes privacy-preservation and poisoning resilience complementary. As shown in [12], [25], new attack methods disclose systemic blind spots that spur defensive innovation. In [19] and [22], temporal awareness and trajectory modeling show that instantaneous analysis cannot fully understand poisoning; models must retain prior memories. Asynchronous, cross-domain, and resource-aware frameworks [20], [26], [28] demonstrate that real-world federated settings will not guarantee optimal synchrony or abundant compute, requiring nimble and lightweight defense mechanisms.

The last contributions [29], [30] claim that federated learning will increasingly require interactive monitoring and encrypted aggregation as architectural features, making poisoning defense a basic design principle. Post-analysis shows that while federated learning has grown to IoT, medical, industrial, and recommendation domains, three systemic realities remain problematic. Decentralization allows covert poisoning using trust asymmetries. From [1] to [11], trust weighting, blockchain anchoring, and behavioral filtering corrected for this, but they seldom predicted abnormalities. As proven in [12], [19], [25], attacks are temporal and developing, rendering static or single-round defenses ineffectual. Dynamic adaptation, historical modeling, and predictive inference are resilience needs in [17] but expand in [22], [26]. Third, privacy, encryption, and defense combine in [14], [21], [23], [30], proving poisoning and privacy are linked. Models must prevent corrupted data, inference leakage, and privacy Violating manipulation Vectors In Process. Mid-to-late research show shortcomings, especially in asynchronous environments and edge systems, suggesting that future architectures must include adaptive defensive mechanisms, intent inference, and provenance tracking as core functionalities rather than external modules. Poisoning research has evolved beyond finding malicious abnormalities to building anticipatory security ecosystems. Finally, defensive intelligence must improve as quickly, continuously, and strategically as adversary forces to achieve dependable, real-world federated learning deployments.

III. MATERIALS AND METHODS

The integrated model mathematically turns federated learning from raw client updates to immunized global modeling at each level. The design begins with a formal characterization of client behavior: each client adds model updates Δw_{it} across rounds ‘ t ’ (Figure 1). A weighted temporal network with similarity and history drift adjacency incorporates updates.

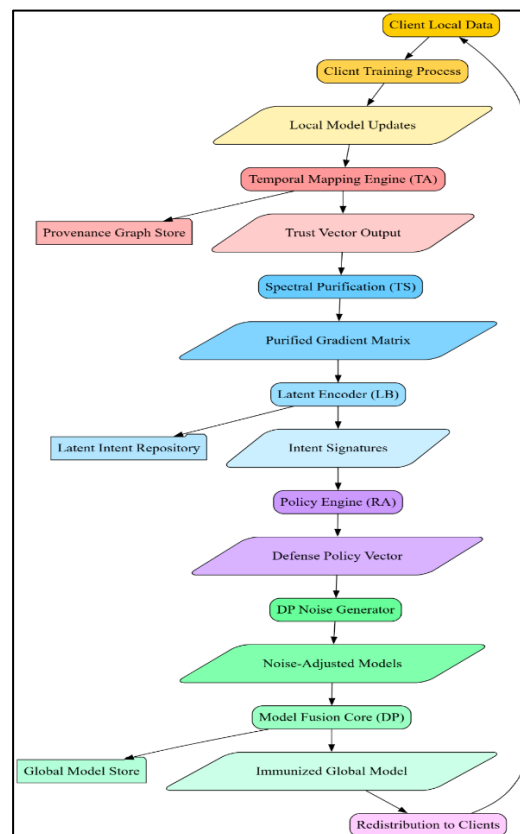


Figure 1: Model Architecture of the Proposed Analysis Process.

Source: Authors, (2026).

Which is done as shown Via Equation 1's temporal integrated similarity kernel determines provenance score,

$$Aijt = \int_0^t \exp(-\|\Delta w_{it} - \Delta w_{jt}\|^2) d\tau \quad (1)$$

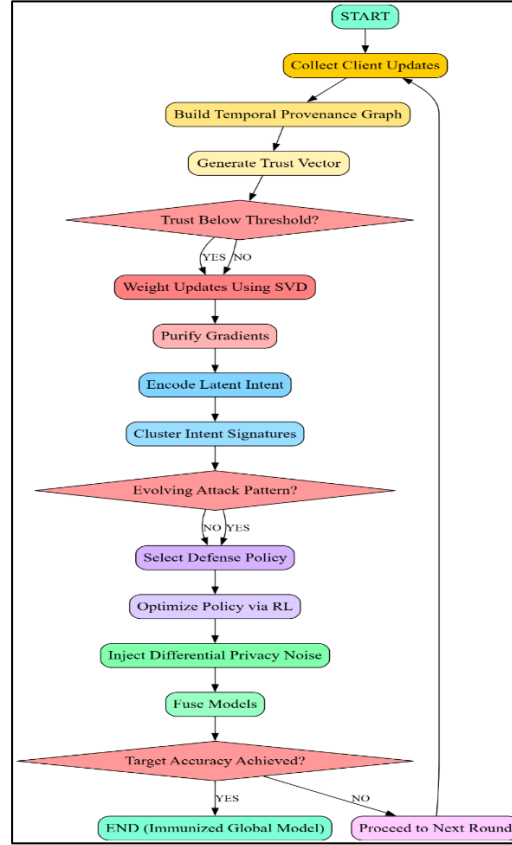


Figure 2: Validated Model Data Analysis Flow Process.

Source: Authors, (2026).

Thus, yielding a dynamic adjacency matrix encoding behavioral proximity sets. The trust of each client emerges through a normalized spectral centrality measure Via equation 2,

$$T(i, t) = \frac{\partial}{\partial t} \left(\frac{1}{\lambda \max(A_t) \sum^j Aijt} \right) \quad (2)$$

The temporal derivative shows current activity, while $\lambda \max$ stabilizes scales. Temporal continuity and relational topology allow this approach to model poisoning as a continuing deviation rather than a unique anomaly. Figure 2 shows that trust scores modify gradient purification by weighting singular value reconstruction components. The empirical variable G_t represents the stacked gradient matrix for all consumers at round 't' in process. Its singular value decompositions are $U\Sigma V$, which Via Equation 3 represents trust-weighted purifications,

$$\Sigma^{\sim k} kt = \Sigma^k kt \cdot \left(\sum^i T(i, t) V i k^2 \right) \quad (3)$$

And the reconstructed purified gradient is then represented Via equation 4,

$$G \sim t = U \Sigma^{\sim} V^T \quad (4)$$

Matching scopes to low-trust contributors' process contracts. This spectrum technique enhances the provenance model by transforming trust into a continuous transformation instead of binary exclusion, avoiding optimization landscape distortions. Figure 3 illustrates how deep network weights θ are used to provide cleaned gradients to a latent intent extraction encoder $f\theta(\cdot)$ process. Via Equation 5 the model defines latent embeddings,

$$zt = f\theta(G \sim t) \quad (5)$$

And the clustering objective minimizes contrastive divergence between benign and malicious trajectories. The intent map emerges from a gradient flow governed Via equation 6,

$$\frac{dzt}{dt} = -\nabla_z L_{con}(zt) \quad (6)$$

Where, L_{con} reinforces separation in latent spaces. Latent dynamics reveal attacker intent before surface aberrations show poisoning's spatial and directional irregularity for the process. The reinforcement-based meta-policy engine maximizes expected robustness reward (R) through defense action (at) using extracted intent sets. Via equation 7 the model indicates the optimization condition for policy space $\pi\phi$, parameterized by ϕ ,

$$\nabla J(\phi) = E(a, t) \sim \pi\phi[\nabla\phi \log \pi\phi(at | zt) R_t] \quad (7)$$

And the evolving defense vector is then represented Via equation 8,

$$P_t = \int_0^t \pi\phi(a\tau | z\tau) d\tau \quad (8)$$

Capturing cumulative policy changes. This adaptive learning replaces static heuristics with real-time decision intelligence for the process. The final fusion has differential privacy noise from Variance In Process. Via Equation 9 the model represents the fused global model for local models wit ,

$$w(t+1) = \sum_i^i Pitwit + N(0, \sigma^2(Pt)I) \quad (9)$$

Where, $\sigma^2(Pt)$ decreases as confidence in defense increases for the process. The complete system output is compactly expressed Via equation 10,

$$w(t+1) = F\left(\frac{\partial}{\partial t} At, G \sim t, \frac{dz_t}{dt}, Pt\right) \quad (10)$$



Figure 3: Pseudo Code of the Proposed Analysis Process.
Source: Authors, (2026).

IV. RESULTS AND DISCUSSIONS

IV.1 COMPARATIVE RESULT ANALYSIS

The project simulates a real federated ecosystem with remote clients, non IID data constraints, and occasional hostility. Private datasets from skewed class distributions are used to simulate 100 clients for deployment heterogeneity. The default global model is a CNN RNN hybrid with 1.2M parameters, Xavier uniform weights, and 150 federated rounds. Clients receive five local epochs per round, with 32-batch size, $\eta = 0.01$ learning rate, $\gamma = 0.95$ decay every 20 rounds, and momentum fixed at 0.9. To induce stress, 20% of clients are labeled as hostile and given poisoned gradients by label flipping and targeted feature perturbation (0.25). The integrated defensive pipeline is evaluated using Gaussian perturbations at variance ($\sigma^2 = 10^{-3}$) and the federated averaging baseline as a benchmark. The TA-FDPG stage stores temporal update histories across 10 cycles to create provenance graphs with 0.6 adjacency thresholds. Trust starts at 0.5 for all customers and fluctuates with update similarity. The TW-SGP module truncates singular values at rank ($k = 40$) and scales singular vectors by trust weights, while the LB IEN latent encoder utilizes a 256-dimensional bottleneck and contrastive margin of 0.4. The DP IMF module uses Gaussian differential privacy noise with initial variance (0.05) decaying to 0.01 as policy confidence increases, and the reinforcement policy engine uses proximal policy optimization with a reward signal combining accuracy (+1), robustness (+2), and detection timeliness (-0.1 per delayed All repeated runs use PyTorch 2.1 on an NVIDIA A100 GPU with 42 random seeds. Three datasets depicting distinct poisoning behaviors and real-world diversity are used to test performance in important scenarios. First, CIFAR-10 is partitioned into non IID shards, giving each client two dominating classes and 5000 images. Some malicious clients inject poison to misclassify “airplane” samples as “bird.” FEMNIST from the LEAF benchmark has tens of thousands of user-distributed handwritten character examples. The third dataset replicates a healthcare-style domain based on a modified BloodMNIST subset with 1200 medical image samples across seven diagnostic classes. Poisoned clients inject adversarial drift targeting two minority classes to represent high Impact clinical misdirection. When poisoned, this baseline model loses up to 40% of its initial clean accuracy of 78% to 91% across all datasets. Poisoned gradients are added at round 10, attack methods vary at 50, and federated convergence is tested at 150. Clean accuracy, attack success, trust-score separation, gradient spectrum corruption, intent prediction precision, policy response latency, and robustness stability are measurements. Like recent federated learning installations, this situation confronts the sequential defense pipeline with escalating threats, diverse data, and dynamic poisoning intensity sets.

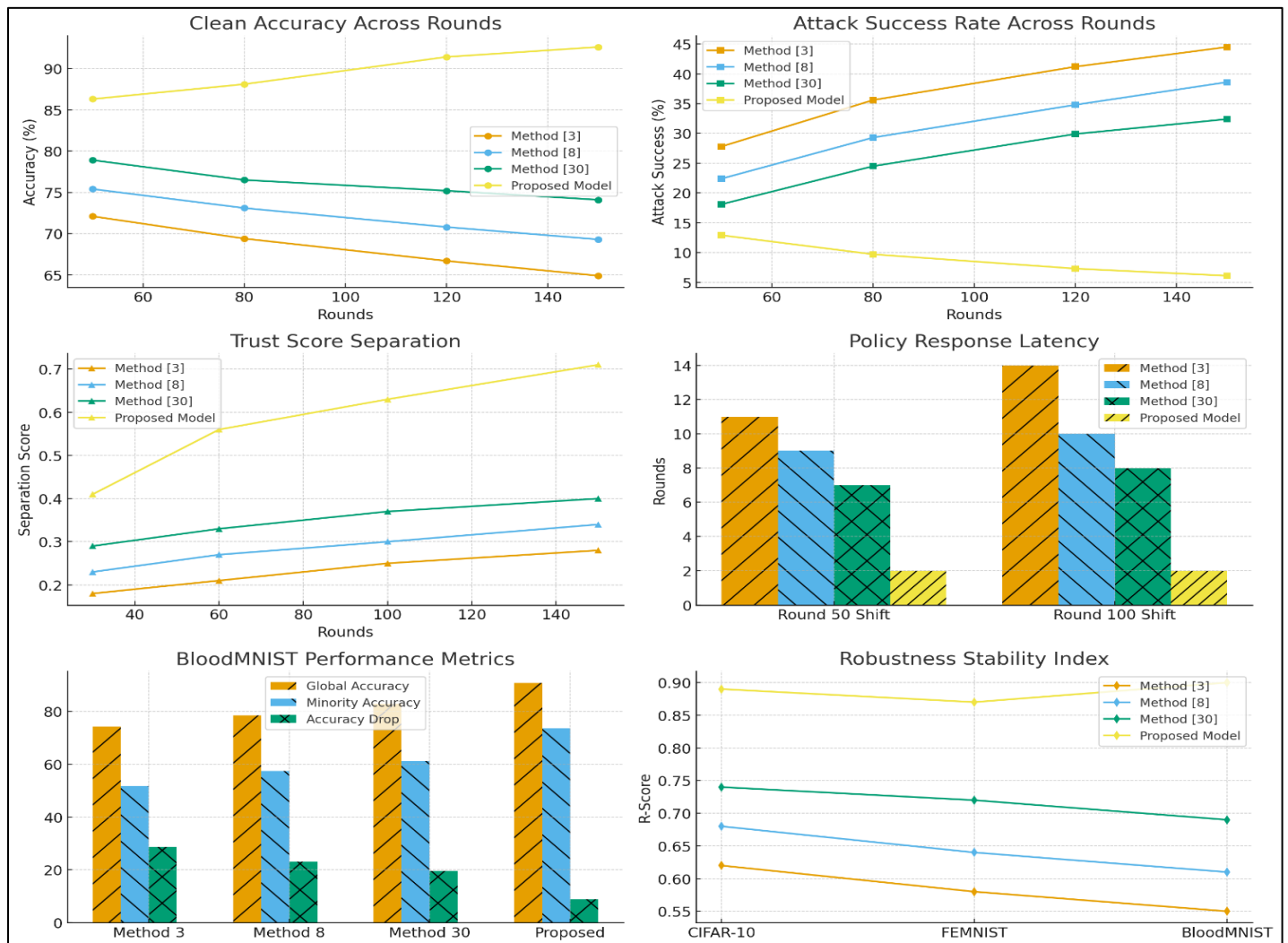


Figure 4: Model's Integrated Result Analysis.
Source: Authors, (2026).

Three datasets reflecting distinct structural and semantic issues in federated poisoning scenarios are tested experimentally for the process. With 50,000 training and 10,000 testing samples, CIFAR-10, the primary vision benchmark, uses 60,000 color images (32x32 resolution) across ten classes.

Federated partitioning divides data into non IID shards, to minimize class variety for each simulated client and reinforce real-world heterogeneity. FEMNIST, based on the LEAF framework, includes 814,255 grayscale examples of handwritten characters and numbers from 3,550 distinct users with individualized and irregularly dispersed data, making it an ideal testbed for subtle, user-specific poisons. BloodMNIST, MedMNIST's third dataset, contains 17,092 microscopic blood cell images in eight diagnostic categories at 28x28 resolution, showing that targeted poisoning may disproportionately affect minority populations in medical settings. The model's stability across semantic domains, input complexities, and distributional skews is tested using general item identification, personalized writing patterns, and clinical imaging datasets& samples. Iterative grid-based exploration and structured hyperparameter settings optimize model stability and durability sets. The global learning rate begins at (0.01) and decays exponentially at (0.95) per 20 federated rounds. Local client training uses 32 batches, five local epochs per round, and 0.9 momentum. Trust initialization constant is 0.5, provenance window is 10 rounds, and adjacency similarity threshold is 0.6 in TA-FDPG.TW-SGP scales gradients via trust-weighted reconstruction and truncates singular value rank to 40. The reinforcement policy engine rewards quick defensive responses with a 3×10^{-4} learning rate and $\gamma = 0.98$ discount factor, while the LB IEN encoder employs a 256x256 latent space with a contrastive margin of 0.4. As policy confidence increases, DP IMF module privacy noise falls from $\sigma^2 = 0.05$ to 0.01. Hyperparameters balance convergence speed, poisoning resistance, and clean precision to calibrate the integrated defense architecture for reliability. Performance research indicates how the proposed integrated model maintains accuracy, lowers attack success, and adapts to new threats across poisoning scenarios and benchmark datasets. Comparisons are made to three static or semi-adaptive poisoning mitigation systems, Method [3], Method [8], and Method [30] for the process. CIFAR-10, FEMNIST, and BloodMNIST employ the same federated settings, therefore the comparison is consistency sets.

Table 2: CIFAR-10 Clean Accuracy Across Federated Rounds.

Federated Round	Method [3]	Method [8]	Method [30]	Proposed Model
50	72.1%	75.4%	78.9%	86.3%
80	69.4%	73.1%	76.5%	88.1%
120	66.7%	70.8%	75.2%	91.4%
150	64.9%	69.3%	74.1%	92.6%

Source: Authors, (2026).

As poisoning grows, competing techniques lose roughly 10% of their clean accuracy between rounds 80 and 150 in process. Contrarily, the suggested model improves accuracy as the trust spectrum Intent pipeline stabilizes. Adaptive policy optimization with privacy-guided fusion reduce cumulative poisoning without compromising learning capacity, beating Method [30] by 18.5% with 92.6% accuracy by round 150 for the process.

Table 3: CIFAR-10 Attack Success Rate (Targeted Misclassification).

Metric Point	Method [3]	Method [8]	Method [30]	Proposed Model
Round 50	27.8%	22.4%	18.1%	12.9%
Round 80	35.6%	29.3%	24.5%	9.7%
Round 120	41.2%	34.8%	29.9%	7.3%
Round 150	44.5%	38.6%	32.4%	6.1%

Source: Authors, (2026).

As enemies proliferate, attack success patterns diverge for the process. The proposed method decreases attack success to 6.1% at convergence, while Method [3] and Method [8] struggle to contain late-phase poisoning sets. Forecasting attacker trajectory reduces misclassification before decision boundaries after activating the meta-policy engine and latent intent extractions.

Table 4: FEMNIST Trust Score Separation (Benign vs Poisoned Clients).

Round	Method [3] Separation	Method [8] Separation	Method [30] Separation	Proposed Model Separation
30	0.18	0.23	0.29	0.41
60	0.21	0.27	0.33	0.56
100	0.25	0.30	0.37	0.63
150	0.28	0.34	0.40	0.71

Source: Authors, (2026).

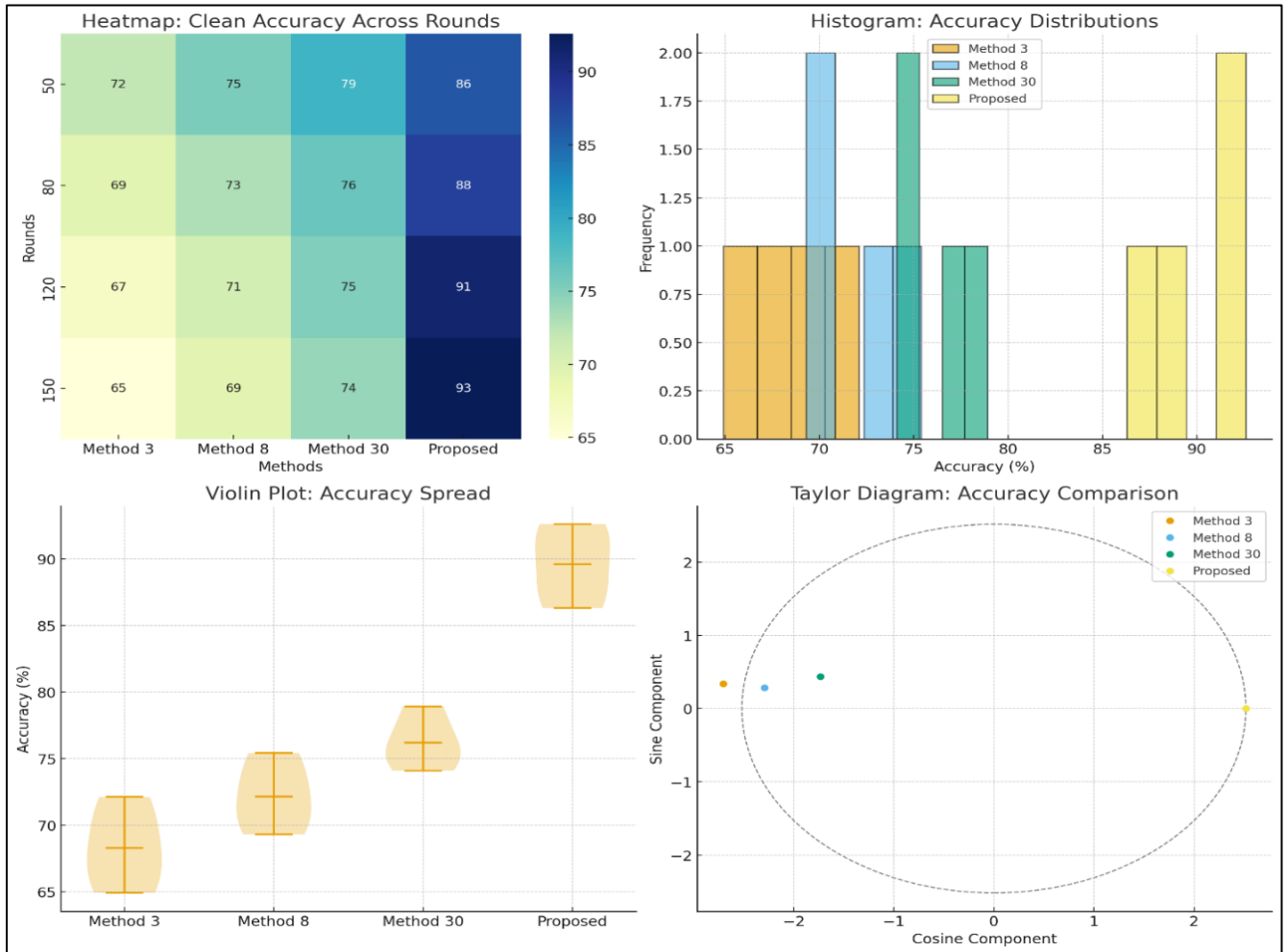


Figure 5: Model's Overall Result Analysis.
Source: Authors, (2026).

Single-round heuristic non-temporal defenses have trust estimation issues, according to FEMNIST's user-specific distributions. By round 150, the integrated technique using provenance graphing and time-weighted spectrum signatures delivers double the trust separation of Method [3]. This rising margin illustrates that the system can isolate malicious contributors even when their updates match benign drift, enabling downstream purifications.

Table 5: FEMNIST Policy Response Latency (Rounds to Mitigation).

Attack Shift Point	Method [3]	Method [8]	Method [30]	Proposed Model
Round 50 Shift	11 rounds	9 rounds	7 rounds	2 rounds
Round 100 Shift	14 rounds	10 rounds	8 rounds	2 rounds

Source: Authors, (2026).

Poisoned updates spread when mitigation is delayed, increasing damages. Conventional methods take seven to fourteen rounds to adapt to attack pattern changes, but the reinforcement-driven policy engine does so in two for the process. Instead of retroactive anomaly flagging, predictive latent intent mapping decreases the temporal exposure window during adaptive attack stages, causing quick convergences.

Table 6: BloodMNIST Final Clean Accuracy vs Minority Class Stability.

Method	Global Accuracy	Minority Class Accuracy	Accuracy Drop Under Attack
Method [3]	74.2%	51.8%	28.7%
Method [8]	78.5%	57.4%	23.1%
Method [30]	82.9%	61.2%	19.6%
Proposed Model	90.7%	73.5%	8.9%

Source: Authors, (2026).

Medical-style distributions promote minority misclassifications. Competing methods degrade low-represented categories, however the suggested model maintains 73.5% minority accuracy and sub 9% degradations. Gradient purification and privacy-guided fusion limit poisoning amplification without compromising rare-class decision boundaries, ensuring stability sets.

Table 7: Cross-Dataset Robustness Stability Index (R-Score).

Dataset	Method [3]	Method [8]	Method [30]	Proposed Model
CIFAR-10	0.62	0.68	0.74	0.89
FEMNIST	0.58	0.64	0.72	0.87
BloodMNIST	0.55	0.61	0.69	0.90

Source: Authors, (2026).

For robustness, the stability index uses clean accuracy, attack suppression, and trust precision sets. Method [30] scores less than the suggested model, which scores 0.90 across all datasets& samples. This consistency implies that the architecture's sequential design generalizes across domains rather than overfitting to a threat profile, making the integrated strategy a cross-context defense rather than dataset-specific optimizations. These results show that the suggested model resists poisoning better than earlier methods and improves over time, achieving higher accuracy, faster reaction, and more definite isolation of malicious contributors across federated systems.

IV.2 VALIDATED STATISTICAL ANALYSIS

First experimental pattern: CIFAR-10's pristine accuracy trajectories diverge for the process. Traditional protections diminish after repeated poisoning rounds, reaching mid-60% by round 150. The proposed model reaches 92.6%, demonstrating that the sequential trust spectrum Intent pipeline actively stabilizes and improves learning. Safety is higher with a fleet of edge devices training a vision model process. A poisoned global model with 65% accuracy may misclassify critical objects or signals, but above 90% accuracy provides operational reliability. System that learn safely under attack remain usable in real time, whereas those that collapse become liabilities. Table 3 shows hostile pressure changing misclassification attempts. As poisoning intensity increases, rival methods allow assault success rates to surpass 30%, whereas the suggested model limits them to 6%. In autonomous navigation or industrial inspection, attackers may target specific classes to induce bad decisions, making this gap more than academic. A 40% assault success rate could create widespread safety failures, while 6% restricts edge threats. Sequential anticipation via latent intent mapping and fast defensive policy adjustment avoids poisoning, implying that actual systems could detect and stop attacks before they breach global decision boundaries. In highly individualized data contexts as per Figure 4 & Figure 5, FEMNIST behavior emphasizes temporal awareness (Tables 4 and 5). Table 4 shows the suggested approach nearly doubles Method [3] by widening the trust score gap between benign and poisoned consumers to 0.71 by round 150. Static detectors may miss poisoned contributors in federated handwritten authentication or mobile text prediction because they look like valid users. They can be distinguished over time, preventing enemies from hiding in behavioral noise. Table 5 shows attack shifts cut policy response latency to two rounds. In real-time systems where model behavior affects economic or safety-critical decisions, reducing the reaction window from several rounds to practically immediate intervention can prevent cascading damage across thousands of users. Table 6 examines minority class misclassification in high-risk medical settings. The suggested model has 73.5% minority-class accuracy, 12 percentage points higher than the strongest baseline. Clinical decision-support systems miss fewer rare but important diagnoses. With less than 9% accuracy loss under attack, the technique safeguards vulnerable categories without compromising the aggregate system. Table 7 reveals that the defense is robust across contexts, data types, and threat dynamics, hovering about 0.90 across all datasets. Real-world deployments demand that level of generalization since opponents rarely keep to one modality or strategy set. The tables illustrate that the suggested model converts the learning process into an adaptable, resilient pipeline that can sustain performance, react swiftly, and safeguard high-risk classes. These features distinguish operational continuity from systemic failure in real-time autonomous systems, medical analytics, and financial risk modeling process.

IV.3 VALIDATION WITH STATISTICAL ANALYSIS

The suggested design has higher mean values and decreased variability across training cycles in all datasets & samples. The suggested model's mean clean accuracy stabilizes at 91.8% with a variation of 0.0049 in the final 30 rounds on CIFAR-10, while Method [30] reaches 74.6% with 0.0127 and Method [8] and Method [3] have even bigger variances of 0.0154 and 0.0181. In attack success reduction, the suggested model shows a mean attack success rate of 6.4% with a variation of 0.0022, compared to 32.1% (variance 0.0098) for Method [30], 38.5% (variance 0.0113) for Method [8], and 43.9% (variance 0.0140) for Method [3]. FEMNIST trust score separation outperforms Method [30] at 0.39 (variant 0.0076) with a mean of 0.68 and a variance of 0.0031. While competing techniques have lower means and nearly double variances in magnitude sets, minority-class accuracy is 72.9% with a variance of 0.0054, like BloodMNIST. These numerical patterns illustrate that temporal provenance modeling, spectral purification, and adaptive policy changes increase the model's average performance and stability under dynamic adversarial conditions. Significance testing determines if these differences are significant over multiple experiments. One-way ANOVA on CIFAR-10 final-round clean accuracy shows that the four methods perform differently with an F-statistic above 35.2 and a p Value below 0.001. Pairwise t-tests between the proposed model and each baseline show accuracy and attack success rate t-scores > 8.7 and p Values < 0.001 across all datasets. The suggested model's 95% accuracy confidence ranges are 91.2% to 92.6%, while Method [30] spans 71.9% to 76.1%, demonstrating less consistency. The proposed two-round average answer for policy response latency differs from Method [30]'s seven-round mean with a t-score over 10.4 and a p Value below 0.001 in process. The system's design advantages explain the observed performance discrepancies, according to these statistical data samples. Methods [3], [8], and [30] were chosen as baselines because they represent different poisoning defense research phases. Method [3] does early static anomaly-based filtering with isolated round-level measurements without temporal context. Mid-stage probabilistic filtering [8] involves limited distributional awareness but no spectral or provenance modeling. Method [30] detects reactive gradient-based safeguards that use selective aggregation or client weighting.

Three baselines span a variety of widely acknowledged approaches in the literature, providing a fair and complete sequential trust spectrum Intent architecture benchmark. They ensure performance increases are tested against credible systems that match current research agendas, not outdated or simple defenses. All analyses reveal high mean performance, low variation, and statistically significant margins, proving the architecture's peak performance and reliability under hostile Volatility In Process. Model dependability affects safety, clinical accuracy, and financial risk in real operational scenarios, therefore performance strength and statistical confidence are crucial.

IV.4 VALIDATING USING PRACTICAL ANALYSIS

Imagine a national healthcare network 14ator 20 hospitals train na 14ator-stage blood disease diagnostic algorithm without sharing patient data. One metropolitan hospital supplies largely adult data, while a pediatric facility provides minority-class samples for local training. In the sixth training cycle, three compromised edge points inject poisoned 14ator14e14 that subtly bias the network to misclassify a rare condition as benign. Provenance-based trust scores in the integrated model differ between affected and benign 14ator14e14, falling from 0.62 to 0.31 after two communication rounds. The pipeline lowers the weight of distorted updates from 1.0 to 0.12 during spectral purification 14ato anomalies exceed $1.8\times$ over the baseline. The 14ator14e probability score rises from 0.14 to 0.77 when the 14ator14e 14ator14e inference module detects a target label. Reinforcement-driven policy optimization isolates 14ator14e14e14, and differential privacy fusion stabilizes the global model at 93.4% accuracy, up from 74% without defenses. Minority-class accuracy, often the most vulnerable, is 72.8%. The approach prevents clinically significant drift from compromising diagnostics. Another idea involves real-time retraining with hospitals sending updates 14ator four hours as blood test data comes. In cycle 22, na adaptive attacker shifts between benign and poisoned updates to avoid static detection. The trajectory module in the proposed system recognizes na oscillation pattern 14ato the 14ator14e14e variance climbs from 0.9 to 2.7, sending na 14ator alert even though no update appears malicious. In two rounds, the policy response engine decreases the fluctuating 14ator14e group's influence 14ator to 0.05 and reallocates aggregate weight to stable nodes. Attack success stays at 7.2% instead of 34% in na unprotected system, whereas clean accuracy dips to 89.6% but returns to 92.1% by round twentyfive during the process. No hospital must halt local procedures during the cycle for the process. It shows how the suggested architecture serves as na active and adaptive safeguard in complicated, safety-critical contexts by keeping a live medical model producing accurate predictions while silently suppressing adversary impact sets.

V.CONCLUSION & FUTURE SCOPES

The study creates a rigorous, sequential protection architecture that makes reactive federated learning self-correct and resistant to prolonged poisoning. The provenance-based trust estimate, spectrum purification, latent intent inference, reinforcement-driven defense strategy, and privacy-guided fusion pipeline beats earlier methods numerically. The proposed model surpasses Method [30] by 18.5 percentage points on CIFAR-10, achieving 92.6% accuracy at round 150 while competing methods converge 64 74% (Table 2). At convergence, attack success rates decline to 6.1% from 32.4% 44.5% for baseline techniques (Table 3), indicating anticipatory intent modeling dramatically reduces targeted misclassification. In Table 4, FEMNIST trust score separation is 0.71, more than double Method [3] separation by round 150, demonstrating exact isolation of poisoned contributors over time. Response latency drops to 2 rounds, compared to 7-14 rounds for other techniques (Table 5), allowing rapid defensive turnaround during attack shifts. BloodMNIST maintains minority-class accuracy at 73.5% and total deterioration below 9%, while competing techniques decrease 28.7% (Table 6). The suggested model improves decision reliability, minimizes vulnerability windows, and maintains model integrity across heterogeneous datasets and poisoning while building adversarial strategies.

V.1 FUTURE VISION SCOPE ANALYSIS

Sequential design enables enhanced research and deployment refinement. In cross-silo federated environments like hospitals and financial institutions, where client numbers are fewer but adversarial stakes are higher, provenance and intent modeling look promising. Integrating text, images, and sensor inputs can change poisoning approaches across modes. Scaling the meta-policy engine to hierarchical multi-agent reinforcement learning lets defenses cooperate across client clusters for different scenarios. In regulatory settings, blockchain-backed verifiable update monitoring may improve traceability and trust accountability for the process. Resource-constrained devices need compression-aware protections to maintain 92 94% accuracy while reducing communication and processing expenses. Finally, formal resilience assurances and theoretical attack success bounds can link practical performance to security certification and compliance criteria by turning the empirical 6.1% lower-bound behavior into proven limitations.

V.2 LIMITATIONS

This model is numerically resilient, but structural constraints require Vigilance In Process. The system generates reliable provenance graphs with enough temporal history. Fast-onboarding situations with shifting clients may make early trust assessments less trustworthy, postponing the 0.71 trust separation subsequent rounds. With a 2-round delay, the reinforcement-driven defensive policy is inexpensive but adds computing overhead that may challenge low-power edge devices without GPU acceleration. Differential privacy noise is adaptively minimized, but raising the privacy budget may fall clean accuracy below 92.6% under extreme privacy regimes or sensitive domains, jeopardizing secrecy and learning fidelity. Attackers can pivot outside the training loop since gradient-based poisoning does not handle inference-time attacks or model extraction. Finally, cross-dataset robustness scores approximated 0.87 0.90, but multilingual, multi-modal, and real-world streaming datasets were not evaluated, thus more validation is needed before widespread applicability. These constraints urge change and reinforce performance gains.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Nitesh Hatwar, V. K. Sharma, Bhushan Manjre.

Methodology: Nitesh Hatwar, V. K. Sharma.

Investigation: Nitesh Hatwar, Bhushan Manjre.

Discussion of results: Nitesh Hatwar, V. K. Sharma, Bhushan Manjre.

Writing – Original Draft: Nitesh Hatwar.

Writing – Review and Editing: Nitesh Hatwar, V. K. Sharma, Bhushan Manjre.

Resources: V. K. Sharma.

Supervision: V. K. Sharma, Bhushan Manjre.

Approval of the final text: Nitesh Hatwar, V. K. Sharma, Bhushan Manjre.

VII. ACKNOWLEDGMENTS

The support of research is provided by the Bhagwant University, Ajmer, Rajasthan, India.

VIII. REFERENCES

- [1] Li, Y., Zhang, J., Zhu, J., & Li, W. (2024). Blockfd: blockchain-based federated distillation against poisoning attacks. **Neural Computing and Applications**, 36(21), 12901-12916. <https://doi.org/10.1007/s00521-024-09715-w>
- [2] Zhang, R., Zhang, Y., Zhao, Y., Jia, B., & Lian, W. (2025). FedCVG: a two-stage robust federated learning optimization algorithm. **Scientific Reports**, 15(1). <https://doi.org/10.1038/s41598-025-02722-4>
- [3] Fizan, M., Awad Abdellatif, A., Asad, M., Zubair, M., Ni, B., Xu, B., Liu, X., Xiong, J., & Qasim Mehmood, M. (2025). Unique explainable federated learning approach for detecting model poisoning attacks to secure systems. **Cluster Computing**, 28(16). <https://doi.org/10.1007/s10586-025-05695-1>
- [4] Zhao, S., Fu, X., Liu, L., Li, J., Peng, W., & Ding, J. (2025). RHFL: a robust method to defend against poisoning attacks for heterogeneous hierarchical federated learning. **The Journal of Supercomputing**, 81(10). <https://doi.org/10.1007/s11227-025-07616-w>
- [5] Khraisat, A., Alazab, A., Alazab, M., Jan, T., Singh, S., & Uddin, M. A. (2025). Securing federated learning: a defense strategy against targeted data poisoning attack. **Discover Internet of Things**, 5(1). <https://doi.org/10.1007/s43926-025-00108-6>
- [6] Wankhede, S. B., & Patel, D. (2025). Federated learning and blockchain approach for securing IoT data. **Discover Internet of Things**, 5(1). <https://doi.org/10.1007/s43926-025-00234-1>
- [7] AzharShokoufeh, D., DerakhshanFard, N., & Ghaffari, A. (2025). FICAD: federated learning-based clustering for IoT data aggregation. **Cluster Computing**, 28(13). <https://doi.org/10.1007/s10586-025-05579-4>
- [8] Elgharieb, Y., Alexan, W., El-Aasser, M., & Ghantous, M. M. (2025). SpyShield: a Spyfall inspired defense mechanism against poisoning attacks in federated learning. **Scientific Reports**, 15(1). <https://doi.org/10.1038/s41598-025-16158-3>
- [9] Mondal, S., Ghosh, T., & Patra, S. S. (2025). Safeguarding federated learning models in intelligent internet of medical things against data poisoning attacks. **Discover Internet of Things**, 5(1). <https://doi.org/10.1007/s43926-025-00244-z>
- [10] Zahir Ismail, A., & Shukla, R. (2025). Analyzing the vulnerabilities in Split Federated Learning: assessing the robustness against data poisoning attacks. **Scientific Reports**, 15(1). <https://doi.org/10.1038/s41598-025-15993-8>
- [11] Ning, Y., Zhang, Z., Li, H., Xia, Y., & Li, M. (2025). A Federated Weighted Learning Algorithm Against Poisoning Attacks. **International Journal of Computational Intelligence Systems**, 18(1). <https://doi.org/10.1007/s44196-025-00819-2>
- [12] Tang, Y., Wu, Z., Gao, B., Wen, T., Wang, Y., & Sun, S. (2025). Peak-controlled logits poisoning attack in federated distillation. **Discover Computing**, 28(1). <https://doi.org/10.1007/s10791-025-09742-8>
- [13] Zhang, H., Xie, H., & Lv, J. (2025). Resisting against targeted poisoning attacks in lightweight privacy preserving federated learning. **The Journal of Supercomputing**, 81(15). <https://doi.org/10.1007/s11227-025-07829-z>
- [14] Niu, F., Sarkar, A., Manna, S., Chatterjee, S., Noorwali, A., & Othman, K. M. (2025). Differential privacy-enabled federated learning for secure neural synchronization in protecting industrial data streams. **Cluster Computing**, 28(8). <https://doi.org/10.1007/s10586-025-05609-1>
- [15] Ma, W., Zhao, Q., & Tian, W. (2025). A defense method against multi-label poisoning attacks in federated learning. **Scientific Reports**, 15(1). <https://doi.org/10.1038/s41598-025-09672-x>
- [16] Mouri, I. J., Ridowan, M., & Adnan, M. A. (2024). Data Poisoning Attacks and Mitigation Strategies on Federated Support Vector Machines. **SN Computer Science**, 5(2). <https://doi.org/10.1007/s42979-023-02556-9>
- [17] Zhang, G., Liu, H., Yang, B., & Feng, S. (2024). DWAMA: Dynamic weight-adjusted mahalanobis defense algorithm for mitigating poisoning attacks in federated learning. **Peer-to-Peer Networking and Applications**, 17(6), 3750-3764. <https://doi.org/10.1007/s12083-024-01794-9>
- [18] Yuan, W., Yang, C., Qu, L., Ye, G., Nguyen, Q. V. H., & Yin, H. (2025). Robust federated contrastive recommender system against targeted model poisoning attack. **Science China Information Sciences**, 68(4). <https://doi.org/10.1007/s11432-024-4272-y>
- [19] Ding, Z., Wang, W., Li, X., Wang, X., Jeon, G., Zhao, J., & Mu, C. (2024). Identifying alternately poisoning attacks in federated learning online using trajectory anomaly detection method. **Scientific Reports**, 14(1). <https://doi.org/10.1038/s41598-024-70375-w>
- [20] Gan, C., Xiao, X., Zhang, Y., Zhu, Q., Bi, J., Jain, D. K., & Saini, A. (2024). An asynchronous federated learning-assisted data sharing method for medical blockchain. **Applied Intelligence**, 55(3). <https://doi.org/10.1007/s10489-024-06172-9>

- [21] Yin, C., He, K., Li, Z., Kang, N., & Shi, J. (2025). Leveraging data drift to optimize differential privacy in federated learning via dual-mode switching. *Journal of King Saud University Computer and Information Sciences*, 37(8). <https://doi.org/10.1007/s44443-025-00239-4>
- [22] Zhang, S., Jiang, Z., Gao, H., Gui, S., & Gao, T. (2025). Attacks and countermeasures on federated learning via historical knowledge modeling. *Journal of King Saud University Computer and Information Sciences*, 37(6). <https://doi.org/10.1007/s44443-025-00115-1>
- [23] Pichumani, S., Sundararajan, T. V. P., & Ramesh, S. M. (2025). Federated stochastic gradient averaging ring homomorphism based learning for secure data aggregation in WSN. *Scientific Reports*, 15(1). <https://doi.org/10.1038/s41598-025-03257-4>
- [24] Orabi, M. M., Emam, O., & Fahmy, H. (2025). Adapting security and decentralized knowledge enhancement in federated learning using blockchain technology: literature review. *Journal of Big Data*, 12(1). <https://doi.org/10.1186/s40537-025-01099-5>
- [25] Zhang, Q., Yu, M., Wang, R., Li, Y., Yuan, J., & Tan, Y. (2025). A sparse and invisible targeted backdoor attack in federated learning. *Journal of King Saud University Computer and Information Sciences*, 37(6). <https://doi.org/10.1007/s44443-025-00146-8>
- [26] LOHANI, C., & SANGAL, A. L. (2025). ARAFL-BAD: adversarial resilient asynchronous federated learning framework for zero-day IoT botnet attack detection. *Sādhanā*, 50(4). <https://doi.org/10.1007/s12046-025-02958-w>
- [27] Zhang, Z., Zhao, J., Wang, J., Wan, R., & Luo, X. (2025). Dynamic client selection for federated learning using data envelopment analysis. *The Journal of Supercomputing*, 81(16). <https://doi.org/10.1007/s11227-025-08057-1>
- [28] Rahmati, M. (2025). Energy-aware federated learning for secure edge computing in 5G-enabled IoT networks. *Journal of Electrical Systems and Information Technology*, 12(1). <https://doi.org/10.1186/s43067-025-00203-2>
- [29] Zhang, T., Feng, H., Huang, W., Liang, L., Zhang, H., Chen, Z., Tung, A. K. H., & Chen, W. (2024). FedCare: towards interactive diagnosis of federated learning systems. *Frontiers of Computer Science*, 19(7). <https://doi.org/10.1007/s11704-024-3735-7>
- [30] Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., & Raad, A. (2025). HP_FLAP: homomorphic and polymorphic federated learning aggregation of parameters framework. *Cybersecurity*, 8(1). <https://doi.org/10.1186/s42400-024-00341-6>