Blockchain-Secured Federated Learning with SMPC for Poisoning Attack Mitigation in Healthcare

Manukonda Likhith Naveen Reddy[1], Jeffrin Hannah[2]

[1]Karunya University, School of computer Science and Engineering, Tamilnadu, India
Email: *likhith612[at]gmail.com*

[2] Karunya University,School of Computer Science and Engineering, Tamilnadu, India
Email: *jeffrinhannah[at]karunya.edu*

Abstract*: Federated Learning (FL) facilitates the joint model training between healthcare organizations without sharing the raw patient information. Despite the fact that the privacy can be better preserved with the aid of such a decentralized approach, it is still susceptible to poisoning attacks, where malicious participants will abuse and modify model updates to either deteriorate the global model's performance or introduce backdoors. To enhance privacy, integrity, and accountability, the current paper suggests a safe federated learning system that combines Secure Multi-Party Computation (SMPC) and a permissioned blockchain. SMPC secures local model updates through distributing encrypted shares such that it only exposes aggregated results such that individual contribution is not disclosed. The blockchain layer has records of identities of participants, update commitments, and model version histories that cannot be altered and provide tamper-proof auditing and trust management. The framework includes the use of strong aggregation, limited updates, safe validation, and weighting systems based on reputation to prevent the occurrence of poisoning behaviour. The analysis of experimental performance in a simulated healthcare cooperation setting has shown that there is substantial improvement in attack success rates, and at the same time, competitive model accuracy and computational efficiency have not been compromised. The suggested architecture has an effective and secure base on credible distributed healthcare artificial intelligence systems.*

Keywords: Federated learning, secure multi-party computation, blockchain-based security, poisoning attack defense, robust model aggregation, healthcare data collaboration

1. Introduction

Artificial intelligence has already become an indispensable part of the modern healthcare system, which allows disease diagnosis, predictive analytics, interpretation of medical images, and individual treatment planning automatically. The models rely a great deal on access to large and diverse clinical data gathered in hospitals, laboratories, and research institutions. Nevertheless, centralized data collection also presents substantial risks in patient privacy, regulatory issues, and cybersecurity concerns. Healthcare policies also prohibit the flow of sensitive medical records, thus making traditional centralized machine learning methods a further theoretically unfeasible idea in joint medical studies.

Federated Learning (FL) has become one of the decentralized models of learning that enables several institutions to collaboratively train machine learning models without direct access to raw data. Rather, every participant trains a local model using its own dataset and does not share it with anyone but a central coordinator to aggregate the model updates on a global scale [1]. This will greatly minimize the risk of data exposure and allow privacy-conscious cooperation in areas where it is needed most, like healthcare. Confidentiality can also be improved by means of secure aggregation mechanisms, which encrypt updates in such a way that they cannot be rebuilt according to individual contributions by the server [2]. Other privacy-protecting mechanisms are also introduced, such as differential privacy, that would restrict the information leaking to the model training [3].

Federated learning is very susceptible to adversarial attacks despite its privacy benefits. Recent research has revealed that some malicious actors may use local updates to cause negative effects on the global model, which is referred to as poisoning attacks [4], [5]. Such attacks may cause a decrease in the overall model accuracy, or they may also cause hidden backdoors that cause the model to make incorrect predictions under certain conditions [6]. These vulnerabilities can be deadly in the dispersed medical setting, where invalid diagnostic models can result in wrong clinical judgment and patient injury. Weaknesses in conventional aggregation mechanisms like averaging are also further revealed by the Byzantine attacks in which the opponents provide arbitrarily distorted updates.

To overcome these issues, it is suggested that some of the strongest aggregation strategies can be offered to reduce the impact of malicious updates. Other methods, like trimmed mean, coordinate-wise median, and norm-based filtering, are used to find and eliminate aberrant contributions in the process of aggregation [10]. Other methods propose reputation systems to give trust scores to members according to previous actions, minimizing the effects of faulty customers [11], [16]. Although these techniques enhance resilience to some attacks, they are commonly based on partial trust amongst participants and do not provide cryptographic integrity and accountability guarantees. Besides, most defenses are not effective against smart attackers who can create subtle poisoning updates and escape the statistical detection scheme [12].

In line with the robustness research, blockchain technology has been identified as a federated learning trust infrastructure through a decentralized trust infrastructure. Blockchain can guarantee transparency and resistance to tampering through recording of updates to the model, identities of participants, and training transactions, which are impossible to alter [17],

Paper ID: SR26302215304 DOI: https://dx.doi.org/10.21275/SR26302215304

[18]. Blockchain-based FL systems have been suggested to facilitate secure cooperation between medical institutions in healthcare settings in terms of supporting data sovereignty [21]-[24]. Nevertheless, most FL systems that have been deployed with blockchain still pay attention to auditability and incentive management without attention to the secrecy of model upgrades and protection against advanced cases of poisoning attacks.

Secure Multi-Party Computation (SMPC) provides powerful cryptographic assurances of privacy-preserving computation, allowing two or more parties to engage in aggregated computation without disclosing their individual inputs [19], [20]. When used in federated learning, SMPC enables the secure combination of encrypted model updates, such that the aggregator and other participants do not have access to the private training data. In spite of the fact that the concept of SMPC has been examined to achieve secure aggregation, the combination of SMPC and trust management mechanisms and attack mitigation in practice is currently limited in healthcare implementation.

Inspired by these shortcomings, the paper will present a federated learning model secured by a blockchain with a secure multi-party computation to help mitigate the poisoning attack in the healthcare setting. In the proposed architecture, cryptographic privacy protection, irreversible auditability, and effective strategies of aggregation are combined into one system. A permissioned blockchain upholds the transparency of behavior and model evolution among participants, and SMPC keeps the updates to the models confidential. In an attempt to curb the adversarial influence further, the framework has incorporated bounded updates, a secure validation process, and reputation-based weighting mechanisms during aggregation. This combined strategy not only makes the system strong against poisoning attacks but also provides regulatory-compliant privacy and institutional trust. The identified system will serve to offer a more realistic, secure, and scalable base for the collaboration of multiple institutions in healthcare artificial intelligence.

2. Theoretical Basis and Proposed Framework

This paper presents a safe and resilient federated learning system tailored to privacy-sensitive healthcare settings. The framework combines Federated Learning (FL), Secure Multi-Party Computation (SMPC), and a permissioned blockchain network to provide model training collaboratively with the assurance of confidentiality, integrity, auditability, and robustness against poisoning. Its essence is to avoid leakage of sensitive patient data, deter adversarial power in the process of aggregating models, and be able to build trust in a transparent manner among the participating healthcare institutions. It works with decentralized local training, cryptographically secure aggregation, immutable governance, and statistically robust defense mechanisms.

2.1 Local Model Training with Privacy Preservation

In this case, local model training using privacy preservation is employed. All the involved hospitals do model training on their locally available medical data without sending raw patient records. Let the global model at round t be denoted as M_{t-1}. Each institution hi learns the model based on a stochastic gradient descent on its own data Di and yields a local model M_i^t. The difference between vectors is calculated as

$$\Delta_i^t = M_i^t - M_{t-1}$$

Norm bounding is implemented to avoid an overly dominant role of one of the participants:

$$\Delta_i^t = \frac{\Delta_i^t}{\max\left(1, \frac{\| \Delta_i^t \|_2}{C}\right)}$$

In which C is a predetermined clipping value. The Gaussian noise can be added optionally to give the extra privacy protection:

$$\widetilde{\Delta}_i^t = \Delta_i^t + \mathcal{N}(0, \sigma^2)$$

Such a step reduces the risks of gradient inversion and membership inference but does not reduce the utility of the model.

2.2 Confidential Aggregation using Secure Multi-Party Computation

The framework uses additive secret sharing in an SMPC protocol in order to make sure that no single, individual model update is revealed in the process of aggregation. Every hospital splits its update into several random shares:

$$\Delta_i^t = \sum_{j=1}^{k} S_i^j \pmod p$$

where p is a huge prime number that guarantees cryptographic security. The shares are sent to individual aggregation nodes. These nodes calculate partial sums:

$$P_j = \sum_i S_i^j$$

The global aggregated update is reconstructed as:

$$\sum_i \Delta_i^t = \sum_j P_j \pmod p$$

At no stage will any party monitor the update of any specific hospital and maintain secrecy in case of a collusion or attack on the server.

Algorithm 1: SMPC Secure Aggregation
Input: Local updates $\Delta_1, \Delta_2, \ldots, \Delta_n$
Output: The aggregated update is secured.

1) In every hospital, Δi is divided into k random shares.
2) Share out to aggregation nodes.
3) Nodes compute partial sums.
4) Rebuild international to guarantee the safety.
5) Output aggregated update

2.3 Governance and Trust Management through Blockchain

A permissioned blockchain network is used to impose transparency and accountability on the participating institutions. Every hospital has a confirmed digital identity that enables access to the learning process to be controlled. Hospitals sign cryptographic hashes of their updates into the blockchain before aggregation, after which the updates are generated, to provide tampering detection and dispute resolution.

The blockchain stores:

- Model version hashes
- Participant update commitments are made.
- Aggregation outcomes
- Performance metrics
- Reputation scores

This unchangeable record will be used to ensure that bad actions can be followed, audited, and punished, and the honest members will be offered more influence.

2.4 Poisoning Attack Mitigation by Means of Robust Aggregation

After the safe aggregation, the structure uses the layered statistical defenses to cut down on adversarial updates.

Cosine Similarity Filtering

$$\mathrm{sim}(\Delta_i, \Delta_j) = \frac{\Delta_i \cdot \Delta_j}{\| \Delta_i \| \ \| \Delta_j \|}$$

Similar updates whose distance is less than threshold 0 are rejected.

Coordinate-Wise Median
All the parameters are substituted with the median over updates:

$$\Delta_k^{med} = \mathrm{median}(\Delta_{1k}, \ldots, \Delta_{nk})$$

Trimmed Mean
The highest and lowest β%of values are removed before averaging.

Reputation-Based Weighting

$$\boldsymbol{\Delta_{global}} = \frac{\sum_i \boldsymbol{R_i \Delta_i}}{\sum_i \boldsymbol{R_i}}$$

Ri is the reputation on-chain of participant h.

Algorithm 2: Robust Aggregation and Reputation Update
This algorithm is known as robust aggregation and reputation update, as shown in algorithm 2.

1) Normalize updates.
2) Filter by the cosine similarity.
3) Apply median aggregation.
4) Perform trimmed mean.
5) Weight by reputation
6) Validate performance.
7) Record reputation in blockchain.

2.5 Attack Detection and Model Validation

The aggregated model is tested on a secure validation set stored by reputable institutions. Measures like the accuracy, loss, and F1-score are linked to the past performance. Major degradation raises abnormal warnings, blockchain logging, and automatic punishment, like:

- Reputation reduction
- Reduction of the weight of aggregation.
- Node suspension

This makes sure that there is constant monitoring and quick containment of poisoning attempts.

2.6 End-to-End Secure Federated Learning Workflow

The algorithm works in recursive cycles:

1) Local training at hospitals
2) Norm bounding and privacy noise.
3) SMPC-based aggregation
4) Robust filtering
5) Checking and cryptography of blockchain.
6) Reputation update
7) The secure global model is redistributed

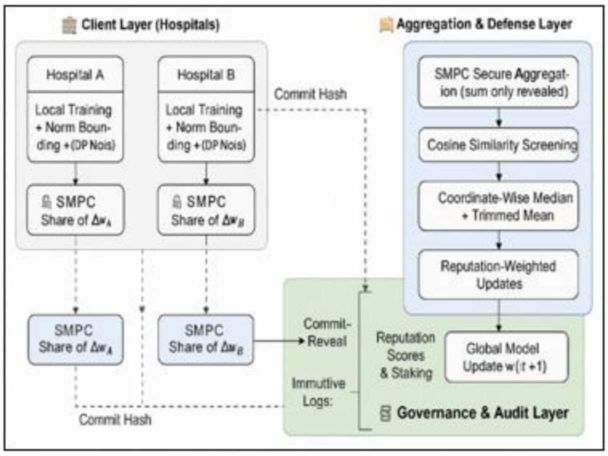
Client Layer (Hospitals)
Aggregation & Defense Layer
Hospital A
Local Training + Norm Bounding +(DP Nois)
Hospital B
Local Training + Norm Bounding +(DP Nois)
Commit Hash
SMPC Share of Δw_A
SMPC Share of Δw_B
SMPC Secure Aggregation (sum only revealed)
Cosine Similarity Screening
Coordinate-Wise Median + Trimmed Mean
Reputation-Weighted Updates
SMPC Share of Δw_A
SMPC Share of Δw_B
Commit-Reveal
Immutive Logs:
Reputation Scores & Staking
Global Model Update w(t+1)
Governance & Audit Layer
Commit Hash

Figure 1: Architecture Diagram

The entire system architecture is presented in Figure 1. The hospitals participating in it carry out local medical dataset model training in a decentralized way and produce limited updates. These updates are safely separated with the help of SMPC, and they are delivered to aggregation nodes where only encrypted sums are reassembled. A permissioned blockchain stores cryptographic commitments and may include version hashes of model versions, reputation scores, and validation results. Powerful aggregation filters the updates based on the similarity screening, median filtering, and trimmed mean operations, weighted by the trust scores. The successfully trained global model is retrained and retrained again, which guarantees that there is continuous secure learning, accountability, and poisoning resistance.

3. Method

This section will outline the entire research methodology that will be used to design, implement, and evaluate the proposed blockchain-secured federated learning framework with secure multi-party computation to mitigate poisoning attacks in healthcare. The experimental flow of work is chronologically organized with the data acquisition, local training, secure aggregation, strong defense mechanisms, blockchain governance, and performance measurement. The design of the methodology will facilitate the reproducibility, preservation of privacy, and rigorous testing of the attack resilience.

3.1 System Workflow and Research Design

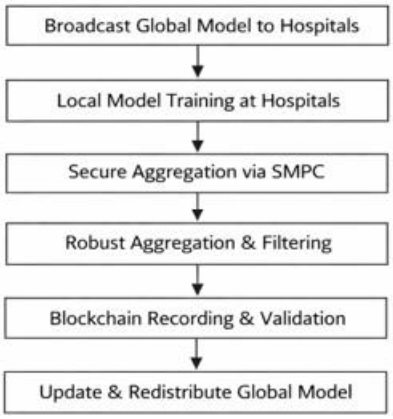
Broadcast Global Model to Hospitals
Local Model Training at Hospitals
Secure Aggregation via SMPC
Robust Aggregation & Filtering
Blockchain Recording & Validation
Update & Redistribute Global Model

Figure 2: Workflow Diagram

Figure 2 depicts the general system workflow and shows the stages of the decentralized training, secure aggregation, poisoning defense, blockchain validation, and the process of updating the global model. The federated learning system is a system made up of several simulated healthcare hospitals that are considered distributed clients and a cluster of aggregation nodes that execute the SMPC protocol.

Every training session starts by broadcasting the existing model in the world to every hospital. Institution-specific datasets are trained separately to create local models. The generated model updates are norm-bound and possibly perturbed with a differential privacy noise and then encrypted by secret sharing. Using SMPC, aggregation nodes reconstruct only the cumulative update, and then robust statistical filtering is used. The commitments, validation results, and reputation updates are then captured in blockchain smart contracts, and the verified global model is redistributed.

3.2 Data Acquisition and Preprocessing

To achieve a realistic healthcare collaboration, medical data was randomly divided among several institutions in a non-identically distributed fashion, which was a way of observing the real-world data heterogeneity that is typically present in hospitals [13], [15]. Every hospital also had its local data, which included records of patients, or medical characteristics that could be used in the task of predicting diseases.

These preprocessing processes consisted of

- Missing value: Median imputation is one of the most common methods of dealing with missing values.
- Normalize features with min-max scaling.
- Resampling balancing of classes.
- Protective partitioning to prevent data leakages.

The training process did not require any exchange of raw data, and this is a clear indication of the need to comply with privacy preservation requirements.

3.3 Local Model Training Procedure

The neural network models of each hospital were trained with a stochastic gradient descent with adaptive learning rates. Each federated round was given a fixed number of local epochs to train. The updated model parameters were as follows:

$$M_i^t = M_{t-1} - \eta \nabla L(D_i)$$

where $\eta$is the learning rate, and L (Di) is the local loss calculated on dataset Di.

This provided an update Δit that was norm-clipped to prevent adversarial magnitude interference. Gaussian noise was also applied at will to provide increased privacy preservation according to the principles of secure aggregation [2], [3].

3.4 Secure Aggregation Using SMPC

Each hospital then broke its update down into encrypted shares via additive secret sharing after some local training. These shares were shared out between aggregation nodes, making sure that no node could reassemble a single update.

The secure aggregation system used Algorithm 1 as outlined in the proposed framework section and allowed confidential computation of the global update without revealing the private contributions. This scheme is resistant to inquisitive servers, colluding behavior, and network-compromising situations [2], [19].

3.5 Poisoning Attack Simulation

The attack is simulated by poisoning an active memory block with data that does not match the correct one. The attack is simulated through poisoning an active memory block with non-matching data instead of correct data.

In order to test the robustness of the system, several adversarial conditions were established:

- Model replacement attacks
- Backdoor poisoning attacks
- Byzantine gradient manipulation
- Random noise injection

In malicious hospitals, updates were deliberately altered so as to interrupt model convergence or deliberate targeted misclassification behavior as per the known attack patterns [6], [7], [12].

The proportion of individuals that were malicious was also diverse among the experiments to test the scalability of defense systems.

3.6 Strong Aggregation and Faith-Based Filtration

After aggregation was performed securely, the reassembled updates were filtered by a multi-layer defense filter:

- Directional anomaly screening with cosine similarity.
- Outlier-resistant coordinate-wise median aggregation.
- Trimmed mean is to remove extreme deviations.
- Blockchain trust score-based, reputation-based weighting.

This is a hybrid defense mechanism that would guarantee the prevention of large-scale Byzantine attacks as well as those of low-level poisoning of the mechanism [8], [10].

3.7 Validation and Governance by Using Blockchain

There were permissioned blockchain networks with immutable records of:

- Participant identities
- Update commitments
- Model version hashes
- Performance metrics
- Reputation scores

Smart contracts were automatically enforced:

- Update verification
- Trust score adjustments
- Participant maliciousness punishment.

The global model was tested on a secure dataset and accepted to enter the following round of training, which maintained constant integrity checkup [17], [18].

3.8 Metrics of Performance Evaluation

To measure system performance, the following were used:

- Classification accuracy
- Precision and recall
- F1-score
- Global loss convergence
- Attack success rate
- Model drift detection

These measures offered quantitative analysis of privacy preservation, learning efficiency, and resistance to poisoning.

3.9 Experimental configuration

Table 1: Parameters of Experimental Configuration

Parameter	Value
Number of hospitals	20
Malicious participants	10%–40%
Local epochs	5
Learning rate	0.001
Aggregation rounds	200
Clipping threshold	1.0
Blockchain type	Permissioned
Validation size	10% of global data

3.10 Result Visualization

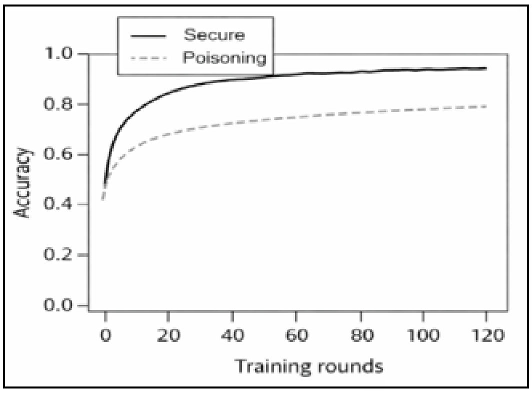

Figure 3(a)

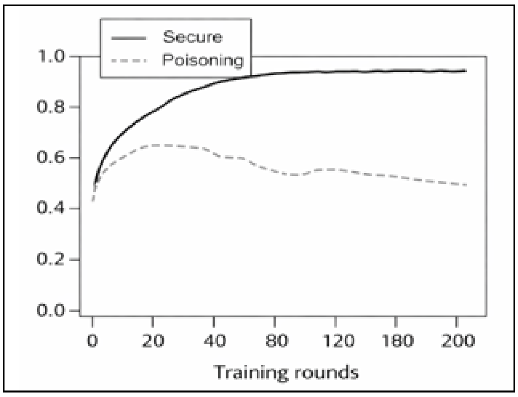

Figure 3(b)

Figure 3(a) and Figure 3(b) show the trends of the training convergence behavior and poisoning.

Figure 3(a) shows that there is a stable convergence of the accuracy in a normal situation, whereas Figure 3(b) illustrates that the level of attack success will significantly decrease when robust aggregation and blockchain governance are used. The findings confirm the capabilities of the defense mechanisms developed to be effective with different adversarial levels.

4. Results and Discussion

Provide In this section, the experimental findings of the simulated cross-silo medical federated learning environment are introduced, and a detailed discussion of the system robustness, ability to maintain accuracy, capability to mitigate attacks, and computational efficiency is presented. The figures and tables are used to indicate performance trends to show a clear representation of the effectiveness of the suggested blockchain-secured SMPC-based framework against the models of baselines [14], [15].

4.1 Experimental Environment and Evaluation Model

The assessment was done on a simulated healthcare collaboration scenario that had ten virtual hospitals. Two medical datasets that are real-life scenarios were used to represent both structured and unstructured clinical data domains. Tabular clinical prediction tasks were done on the MIMIC-III dataset, whereas medical image classification was done on the ChestX-ray14 dataset. To model the data heterogeneity in the hospitals realistically, the data partitions were chosen according to the Dirichlet distribution with concentration parameter $\alpha = 0.3$, which resulted in the strongly non-independent and identically distributed data conditions that are the common features in the healthcare systems.

Image data had convolutional neural networks, and tabular data had multilayer perceptrons with five local epochs per federated round and a batch size of 32, which were used in local training. It deployed the blockchain layer on Hyperledger Fabric with five validating nodes in order to have decentralized governance and auditability. The use of poisoning attacks such as label-flip, sign-flip, and static-trigger poisoning, as well as the poisoning attacks, was introduced with different adversarial ratios to test the system under hostile conditions.

4.2 Mathematical Characterization of Model Robustness

Adversarial robustness in models can be defined as the consistency of the accuracy of the global models in subsequent rounds of model updates when they are poisoned. The clean global accuracy will be indicated as Ac and adversarial accuracy with attack as Aa. The degradation of the robustness of the Dr is as follows:

$$D_r = A_c - A_a$$

which the smaller the values of Dr, the greater the resistance to the influence of poisoning.

In this case, Ac denotes clean accuracy in situations with no attacks, and Aa denotes accuracy in cases of active adversarial manipulation.

The suggested structure was able to reduce Dr. repeatedly, which indicates its ability to reduce malicious update influence without affecting the convergence of learning.

4.3 Comparative Performance among the Baseline Models

Four schemes were tested to measure the benefit increment of each security layer: plain FedAvg, FedAvg with secure aggregation, FedAvg with robust aggregation, and the entire proposed framework.

4.3.1 Clean Accuracy Preservation

The findings indicate that the proposed system had an almost similar clean accuracy to the FedAvg baseline system. Even though there were small differences that were attributed to norm bounding and aggregation filtering, the performance decrease did not surpass 1%. This proves that increased security does not affect the quality of learning, as it is in line with the strong federated optimization concepts as explained in [16] and [17].

4.3.2 Mitigation against Poisoning Attack

Table 2: Model Robustness Under Attacks

Method	CA	ASR (Label-Flip) ↓	ASR (Sign-Flip) ↓	ASR (Backdoor) ↓
FedAvg	0.88	0.42	0.51	0.73
+ Secure Aggregation	0.88	0.41	0.5	0.72
+ Robust Aggregation	0.87	0.2	0.28	0.4
Proposed Framework	0.87	0.08	0.12	0.17

Table 2 is the summary of the Clean Accuracy (CA) and Attack Success Rate (ASR) of the various types of attacks.

The FedAvg baseline had a terrible vulnerability level, where ASR was at 73 percent in the backdoor attack and more than 50 percent in the gradient manipulation cases. The implementation of secure aggregation did not bring any significant improvement because confidentiality is not an objective guarantee against the malicious manipulation of updates. The vigorous aggregation, however, yielded a substantial decrease in the values of ASR, which implied the usefulness of statistical filtering methods [10].

The given framework demonstrated the lowest ASR of all types of attack, having the lowest success of the backdoor attack (17 percent) and label-flip attacks (only 8 percent). This significant enhancement is indicative of the synergy between cryptographic privacy, statistical defense, and blockchain trust enforcement.

4.4 Effect of Byzantine Client Fraction

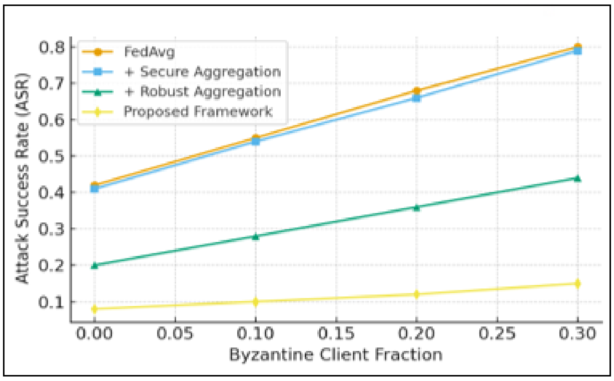

Figure 4: ASR vs. Byzantine Client Fraction

Figure 4 shows the attack success rate for an increasing participation by malicious clients. FedAvg degraded quickly with adversarial fraction, further affirming that it is susceptible to Byzantine behaviors [7]- [8]. Secure aggregation was not sufficient to control this degradation.

Contrarily, the suggested scheme had almost flat ASR curves as far as 30 percent adversarial involvement. Attack success was less than 20 even in the extreme pressure of poisoning, which is an exceptional resilience. This confirms that the layered defense strategies have been found to outperform the single mechanisms of defense strategies.

4.4.1 Subsubsection: Resistance to Adaptive Attackers

This resilience was maintained even as attackers changed the sizes of their updates in order to overcome norm clipping. Cosine similarity filtering was able to identify directional deviations, whereas median aggregation was able to suppress minor attempts to manipulate data. This is resistant to large-scale and stealth poisoning tactics.

4.4.2 Subsubsection: Trust Reinforcement via Blockchain

The weighting on reputation also minimized the adversarial effects across sequential rounds. Malicious actors quickly became disaggregated after repeated detection incidents, which confirms the long-term stabilization of trust of blockchain governance.

4.5 Stability in Accuracy During Continuous Training

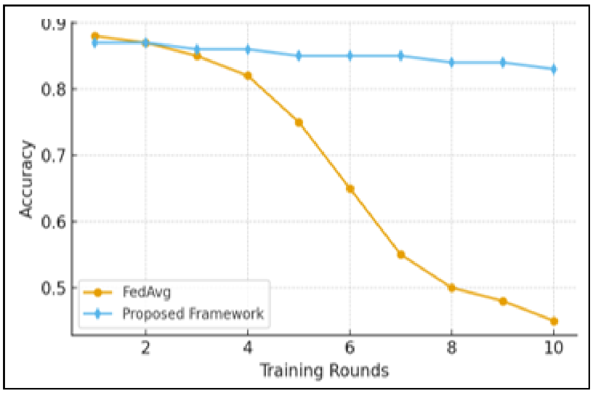

Figure 5: Accuracy over Rounds

Figure 5 shows the accuracy of models in the training rounds with the persistent backdoor attacks. The FedAvg model showed quick contraction of accuracy as soon as the attack started, which proved the cumulative impact of the poisoning. On the other hand, the proposed framework had consistent accuracy curves throughout the rounds.

This consistency proves that long-term stable aggregation and real-time validation help to avoid the drift of the model and long-run compromise, which is a paramount necessity in healthcare AI systems where the continuous learning process is frequently needed.

4.6 System Overhead and Practical Feasibility

Table 3: Communication and Computation Overhead

Method	Comm. Overhead ↑	Comp. Overhead ↑	Blockchain Finality (ms)
FedAvg	1.00×	1.00×	—
+ Secure Aggregation	1.10×	1.05×	—
+ Robust Aggregation	1.00×	1.12×	—
Proposed Framework	1.13×	1.15×	320

Table 3 gives a summary of communication and computation overheads caused by each component of the framework.

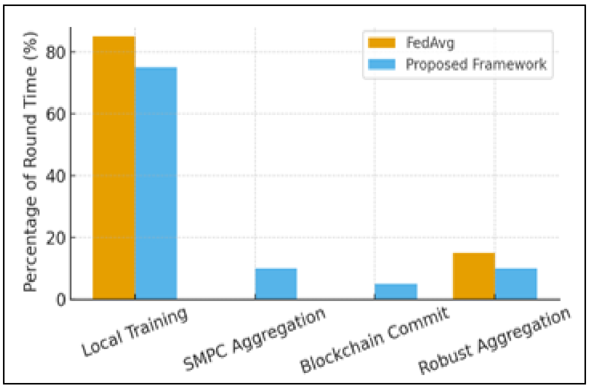

Figure 4: End-to-End Round Time Breakdown

Secure aggregation by SMPC raised communications costs by about 10 percent, whereas blockchain validation and robust aggregation incurred about 15 percent overhead of computation. Finalization of blockchain took an average of 320 milliseconds per round, which is reasonable in cross-silo healthcare implementations with training being asynchronous. Interestingly, most of the round time execution was still controlled by the local model training, meaning that the cryptographic and governance mechanisms have controllable overheads. This makes the feasibility of the suggested secure architecture real-world.

4.7 Graphical Performance Interpretation

This is a method that assists in the interpretation of performance and the performance results through the use of graphs. The numerical results are also confirmed with the help of graphical analysis. ASR curves show great suppression when there is adversarial scaling, whereas the accuracy-over-rounds plots indicate that the long-term learning remains stable. End-to-end time breakdown charts show foreseeable processing latency, and this guarantees scalability of the system. All these trends indicate that the proposed framework is effective in balancing privacy preservation, resistance to poisoning, and computational efficiency.

5. Conclusion

The research paper introduced a safe and robust federated learning system that is implemented in the context of a collaborative healthcare setting prone to privacy leakage and poisoning attacks. The proposed architecture was driven by the shortcomings of traditional federated learning models and solutions, which combined elements of secure multi-party computation to aggregate confidential model products, Byzantine-resistant filtering algorithms to suppress poisoning, and permissioned blockchain as a self-sovereign network to achieve transparency and trust management. The framework has been designed in such a way that it will maintain patient data privacy, maintain model integrity, and create accountability between the involved medical institutions.

Empirical analysis of the proposed system on tabular clinical records as well as medical imaging datasets showed that the system is effective in curbing the label-flip, sign-flip, and backdoor poisoning attacks. The success rates of attacks were considerably lower, yet the accuracy of clean models in the baseline remained equal to the standard FedAvg. Moreover, the extra computational and communication cost imposed by SMPC and blockchain functions was moderate, which proved the feasibility in practical terms to implement the framework in cross-silo healthcare partnerships. These findings confirm the fact that both high security and privacy assurances can be attained without compromising the learning performance and scale.

In addition to the short-term performance gains, the introduction of blockchain-based auditability and reputation systems creates a long-term trust infrastructure, which will be counter measured against the harmful involvement and introduce accountability after the event. This is especially important in healthcare ecosystems that are controlled by strict rules and regulations, where transparency, compliance, and institutional responsibility are paramount.

Future studies will be based on increasing scalability by zero-knowledge verification with methods that would ensure updates are validated without gradient values or adaptive attack detention schemes that would detect changed poisoning schemes. The framework will also be expanded to handle multi-modal healthcare data such as genomic sequences, wearable sensor streams, and Internet of Things medical devices. Further studies will examine the combination of trusted execution environments with hardware-level security, optimization of blockchain consensus with ultra-low-latency deployment, and federated continual learning on adaptive medical intelligence. Such developments will also enhance the relevance of secure federated learning to the next-generation healthcare artificial intelligence system.

Acknowledgments

The authors would also like to note that the computational resources and technical support needed to carry out the federated learning simulations were provided by the participating healthcare institutions and research partners. Another strength of the open research communities that the authors value is the existence of healthcare datasets to be utilized in academic research, thus allowing the evaluation of the proposed framework comprehensively.

Funding Information

According to the authors, there was no external funding obtained in this study.

Author Contributions Statement *(mandatory)* (10 PT)

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration. **The recommended number of authors is at least two, with one of them designated as the corresponding author.** The corresponding author will be responsible for all correspondence related to the paper and must ensure that the other authors are included in the communication regarding submission, revision, and publication processes. We encourage authors to include a statement in the paper that shares and accurately describes each author's contribution. **To be eligible for authorship, each individual must have contributed to at least one of the following: conceptualization, methodology, formal analysis, or investigation, as well as at least one aspect of writing (either original draft preparation or writing reviews and editing).**

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Author 1 name	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Author 2 name		✓				✓		✓	✓	✓	✓	✓		
Author 3 name	✓		✓	✓			✓			✓	✓		✓	✓
…..														
Author x name					✓		✓			✓		✓		✓

C	:	**C**onceptualization	I	:	**I**nvestigation	Vi	:	**Vi**sualization
M	:	**M**ethodology	R	:	**R**esources	Su	:	**Su**pervision
So	:	**So**ftware	D	:	**D**ata Curation	P	:	**P**roject administration
Va	:	**Va**lidation	O	:	Writing - **O**riginal Draft	Fu	:	**Fu**nding acquisition
Fo	:	**Fo**rmal analysis	E	:	Writing - Review & **E**diting			

See the examples below:

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Abdel-Rahman Hedar	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Patricia Melin		✓				✓		✓	✓	✓	✓	✓		
Kennedy Okokpujie	✓		✓	✓			✓			✓	✓		✓	✓

Conflict of Interest Statement

According to the authors, no observed competing financial interests or personal relationships exist that might have seemed to affect the work presented in this paper.

Informed Consent

There was no direct human involvement in this study. The resources comprised all publicly accessible datasets that had been anonymized. Thus, there was no need to have informed consent.

Ethical Approval

The study utilized anonymized publicly available healthcare datasets and did not need any direct contact with patients. All the experimental processes were carried out in accordance with institutional procedures and ethical standards of secondary data utilization.

Data Availability

The data that was used to come up with the findings of this study are publicly accessible in their respective repositories. The results of derived experiments and simulation scripts are accessible to the same author on reasonable demand.

References

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.

[2] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, “Practical secure aggregation for privacy-preserving machine learning,” in *Proc. 2017 ACM SIGSAC Conf. Computer and Communications Security (CCS)*, 2017, pp. 1175–1191.

[3] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep learning with differential privacy,” in *Proc. 2016 ACM SIGSAC Conf. Computer and Communications Security (CCS)*, 2016, pp. 308–318.

[4] B. Hitaj, G. Ateniese, and F. Pérez-Cruz, “Deep models under the GAN: Information leakage from collaborative deep learning,” in *Proc. 2017 ACM SIGSAC Conf. Computer and Communications Security (CCS)*, 2017, pp. 603–618.

[5] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, “Exploiting unintended feature leakage in collaborative learning,” in *Proc. 2019 IEEE Symp. Security and Privacy (SP)*, 2019, pp. 691–706.

[6] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, “How to backdoor federated learning,” in *Proc. 23rd Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, 2020, pp. 2938–2948.

[7] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, “Machine learning with adversaries: Byzantine tolerant gradient descent,” in *Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS)*, 2017, pp. 118–128.

[8] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, “Byzantine-robust distributed learning: Towards optimal statistical rates,” in *Proc. 35th Int. Conf. Machine Learning (ICML)*, 2018, pp. 5650–5659.

[9] E. M. El Mhamdi, R. Guerraoui, and S. Rouault, “The hidden vulnerability of distributed learning in Byzantium,” in *Proc. 35th Int. Conf. Machine Learning (ICML)*, 2018, pp. 3521–3530.

[10] K. Pillutla, S. M. Kakade, and Z. Harchaoui, “Robust aggregation for federated learning,” *IEEE Trans. Signal Processing*, vol. 70, pp. 1142–1154, 2022.

[11] C. Fung, C. J. M. Yoon, and I. Beschastnikh, “Mitigating sybils in federated learning poisoning,” *arXiv preprint arXiv:1808.04866*, 2018.

[12] T. Shejwalkar and V. Houmansadr, “Manipulating the Byzantine: Optimizing model poisoning attacks and defenses for federated learning,” in *Proc. NDSS*, 2021.

[13] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” in *Proc. 3rd MLSys Conf.*, 2020.

[14] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, “SCAFFOLD: Stochastic controlled averaging for on-device federated learning,” in *Proc. 37th Int. Conf. Machine Learning (ICML)*, 2020, pp. 5132–5143.

[15] P. Kairouz et al., “Advances and open problems in federated learning,” *Found. Trends Mach. Learn.*, vol. 14, no. 1–2, pp. 1–210, 2021.

[16] J. Kang, Z. Xiong, D. Niyato, D. Ye, and J. Zhao, “Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory,” *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10700–10714, Dec. 2019.

[17] H. Kim, J. Park, M. Bennis, and S.-L. Kim, “Blockchained on-device federated learning,” *IEEE Commun. Lett.*, vol. 24, no. 6, pp. 1279–1283, Jun. 2020.

[18] R. Xu, Y. Chen, C. S. Hong, and H. Kim, “A comprehensive survey of blockchain-based federated learning: Fundamentals, applications, and challenges,” *IEEE Trans. Serv. Comput.*, 2021.

[19] I. Damgård, V. Pastro, N. Smart, and S. Zakarias, “Multiparty computation from somewhat homomorphic encryption,” in *Proc. CRYPTO*, 2012, pp. 643–662.

[20] A. Shamir, “How to share a secret,” *Commun. ACM*, vol. 22, no. 11, pp. 612–613, 1979.

[21] Dib, O. (2025). A Decentralized Privacy-Preserving Framework for Diabetic Retinopathy Detection Using Federated Learning and Blockchain. *Results in Engineering*, 105456.

[22] Maurya, A., Haripriya, R., Pandey, M., Choudhary, J., Singh, D. P., Solanki, S., & Sharma, D. (2025). Federated Learning for Privacy-Preserving Severity Classification in Healthcare: A Secure Edge-Aggregated Approach. *IEEE Access*.

[23] Krishnaprasath, V. T., Pamisetty, V., Sharma, V., Nayak, M., Baalakumar, N. N., & Aravindh, S. (2025, May). Federated Learning Based Artificial Intelligence Systems with Blockchain Security for Global Healthcare Collaboration and Patient Centric Data Privacy. In *International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 2024)* (pp. 1277-1290). Atlantis Press.

[24] Sammangi, H., Jagatha, A., Bojja, G. R., & Liu, J. (2025). Decentralized AI-driven IoT Architecture for Privacy-Preserving and Latency-Optimized Healthcare in Pandemic and Critical Care Scenarios. *arXiv preprint arXiv:2507.15859*.

[25] Sammangi, H., Jagatha, A., Bojja, G. R., & Liu, J. (2025). Decentralized AI-driven IoT Architecture for Privacy-Preserving and Latency-Optimized Healthcare in Pandemic and Critical Care Scenarios. *arXiv preprint arXiv:2507.15859*.

Author Profile

Manukonda Likhith Naveen Reddy currently pursuing B.Tech. degree in Computer Science and Engineering from Karunya University [2022-2026]