DEFENDING AGAINST POISONING ATTACKS IN FEDERATED LEARNING WITH BLOCKCHAIN

T SUNDARARAJULU [1]
RAJULA BADRINADH REDDY[2], V DIVYA SREE[2], JAINI SAI PAVAN[2], PEDDANABOYIENA TEJASWINI[2], GUDIWADA VISWA TEJA[2]

[1] *Assistant Professor, Department of Computer Science & Information Technology, Siddharth Institute of Engineering & Technology, Andhra Pradesh, India*
[2] *Research Scholar, Department of Computer Science & Information Technology, Siddharth Institute of Engineering & Technology, Andhra Pradesh, India*

ABSTRACT

In the era of deep learning, federated learning (FL) presents a promising approach that allows multiinstitutional data owners, or clients, to collaboratively train machine learning models without compromising data privacy. However, most existing FL approaches rely on a centralized server for global model aggregation, leading to a single point of failure. This makes the system vulnerable to malicious attacks when dealing with dishonest clients. In this work, we address this problem by proposing a secure and reliable FL system based on blockchain and distributed ledger technology. Our system incorporates a peer-to-peer voting mechanism and a reward-and-slash mechanism, which are powered by on-chain smart contracts, to detect and deter malicious behaviours. Both theoretical and empirical analyses are presented to demonstrate the effectiveness of the proposed approach, showing that our framework is robust against malicious client-side behaviours.

Keyword: - Federated learning (FL), decentralized systems, blockchain, malicious clients, model aggregation, serverless architecture, secure machine learning, collaborative learning, robustness, security and trustworthiness.

1. INTRODUCTION

Federated Learning (FL) has emerged as a transformative approach in machine learning, enabling multiple decentralized devices to collaboratively train models without sharing raw data. This decentralized paradigm enhances privacy by keeping data local while still benefiting from collective learning. However, FL is inherently vulnerable to **poisoning attacks**, where malicious participants manipulate model updates to degrade overall system performance or introduce biases. Such adversarial attacks can compromise **data integrity, model accuracy, and trustworthiness**, making it critical to develop **robust defense mechanisms**.

Poisoning attacks in FL can be categorized into **data poisoning and model poisoning**. **Data poisoning** involves injecting corrupted data into local training sets, while **model poisoning** manipulates gradient updates to bias the global model. Given the distributed nature of FL, detecting and mitigating such attacks is challenging, especially in adversarial settings where attackers operate stealthily. Existing defense strategies, including anomaly detection and differential privacy, offer some protection but often suffer from performance trade-offs or lack transparency in model updates.To address these challenges, **blockchain technology** presents a novel and effective solution for securing FL against poisoning attacks. By leveraging blockchain's **decentralized, immutable, and transparent ledger**, federated learning updates can be securely recorded, audited, and verified without relying on a single trusted entity. Smart contracts can be employed to enforce **consensus-based validation mechanisms**, preventing malicious model updates from

being aggregated. Additionally, cryptographic techniques such as **zero-knowledge proofs and homomorphic encryption** can enhance data privacy while maintaining accountability.This research proposes a **blockchain-integrated FL framework** that strengthens model robustness against poisoning attacks by incorporating **secure aggregation, decentralized trust mechanisms, and anomaly detection techniques**. By ensuring **tamper-proof model updates, real-time attack detection, and enhanced privacy**, this hybrid approach enhances the resilience of FL systems against adversarial threats.Through rigorous evaluation, this study aims to demonstrate that blockchain-enhanced FL not only mitigates poisoning attacks but also improves **model accuracy, fairness, and security** in decentralized learning environments. The proposed system offers a **scalable, transparent, and attack-resilient framework** that can be applied to critical domains such as **healthcare, finance, and IoT networks**, where data privacy and integrity are paramount.

2. LITERATURE SURVEY

[1] Authors: Jakub Konecny. Federated Learning: Strategies for Improving Communication Efficiency. 2022. Federated Learning is a machine learning setting aimed at training a high-quality centralized model while keeping the training data distributed across multiple devices. This paper focuses on improving communication efficiency between clients and servers in federated environments.

[2] Authors: Eider Moore. Communication-Efficient Learning of Deep Networks from Decentralized Data. 2020. This work explores techniques for efficiently training deep learning models using data distributed across mobile devices. It highlights how leveraging local data can enhance model performance and user experience while reducing communication overhead.

[3] Authors: Frank Dabek. Vivaldi: A Decentralized Network Coordinate System. 2021. Vivaldi is presented as a fully decentralized system that estimates network coordinates without relying on fixed infrastructure or designated hosts. The paper demonstrates its applicability in large-scale decentralized systems.

[4] Authors: Arjun Nitin Bhagoji. Analyzing Federated Learning Through an Adversarial Lens. 2020. This paper examines the security and privacy challenges in federated learning, especially under adversarial conditions. It discusses the behavior of agents and the potential risks posed by malicious participants during distributed model training.

3. METHODOLOGY

3.1EXISTING SYSTEM

- Poisoning is the most widespread type of attacks in the history of the learning field.
- In general, poisoning attacks reduce the learning model accuracy by manipulating the learning training process to change the decision boundary of the machine learning system.
- Depending on the goal of poisoning attacks, we classify those attacks into two categories: targeted poisoning attacks and non-targeted poisoning attacks.
- Non-targeted poisoning attacks are designed to reduce the prediction confidence and mislead the output of the ML system into a class different from the original one.
- In targeted poisoning attacks, the ML system is forced to output a particular target class designed by the attacker.

3.1.1DISADVANTAGES OF EXISTING SYSTEM

- Usually, the attackers can produce the poisoned local training updates by injecting new malicious clients into the system or manipulating original clean clients.
- The security problems in the FL systems gain much interest from the ML community. Especially, compared to centralized learning models, implementing an attack on the FL system is much easier because of its loose structure and plenty of spaces between the remote clients and the aggregator.

3.2 PROPOSED METHODOLOGY

- We propose our two-phase defense algorithm Local Malicious Factor (LoMar), which is able to detect the anomalies in FL from a local view, instead of the existing global view.
- The main idea of the proposed LoMar is to evaluate the remote update maliciousness based on the statistical characteristic analysis of the model parameters, which is intuitively motivated by the fact that each remote update in the FL system can be considered as being generated from a specific distribution of the parameters.
- Specifically, once the aggregator receives remote updates from a client, instead of using the whole remote updates set, LoMar performs the feature analysis of this update with its nearest neighbors.

4. SYSTEM DESIGN

It is a process of planning a new business system or replacing an existing system by defining its components or modules to satisfy the specific requirements. Before planning, you need to understand the old system thoroughly and determine how computers can best be used in order to operate efficiently.

4.1 SYSTEM ARCHITECTURE

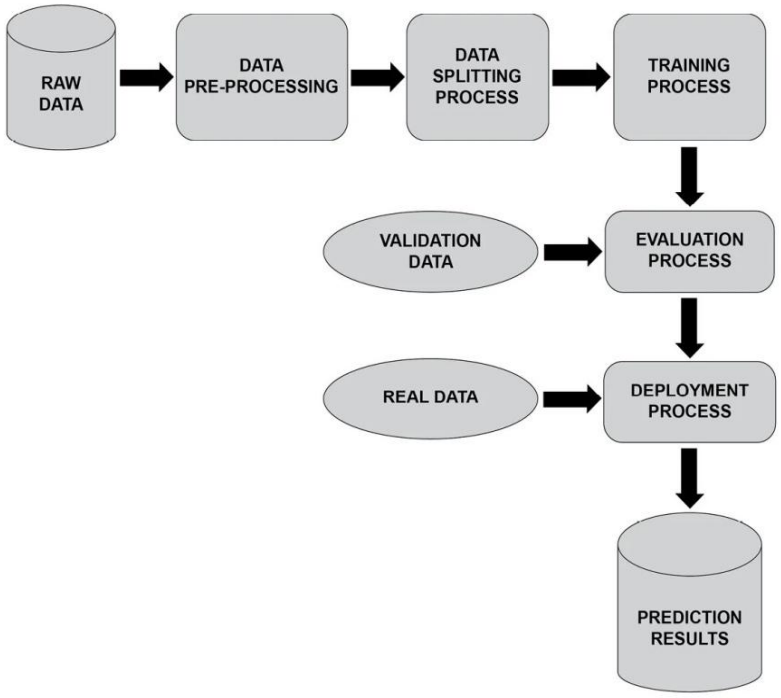

Fig. System Architecture

4.2 MODULES

In this Project , There are Two Modules. They are:

- Service Provider

- User
- Dataset Loading Module
- Training and Evaluation Module
- User Interface Module

4.2.1 MODULES DESCRIPTION

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as

1. Login
2. Browse Datasets & Train & Test Datasets
3. View Trained & Tested Accuracy in Bar chart
4. View Trained & tested Accuracy results
5. View predicted poisoning Attack status type
6. View Predicted poisoning Attack status type ratio
7. Download predicted datasets
8. View Predicted poisoning attack status type ratio results
9. View all remote users
10. Logout

User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like

1. Register
2. Login
3. Predict poisoning Attack status type
4. View your Profile
5. Logout

Dataset Loading Module:

Facilities the loading of datasets for training and evaluation process

Provides graphical representation

Training and Evaluation Module

Training the Model

Evaluate the model with Precision and Recall

User Interface Module

Provides a user friendly interface for users to interact with system

Display functionalities like signup, login, prediction and training and accuracy results

5. RESULTS AND DISCUSSION

EXECUTION PROCEDURE

The Execution procedure is as follows :

1. In this research work with data with attributes are observable and then all of them are floating data. And there's a decision class/class variable. This data was collected from Kaggle machine learning repository.
2. In this research 70% data use for train model and 30% data use for testing purpose.
3. Logistic Regressionis used as Classifier .
4. In the classification report we were able to find out the desired result
5. In this analysis the result depends on some part of this research. However, which algorithm gives the best true positive, false positive, true negative, and false negative are the best algorithms in this analysis.

Fig. Home

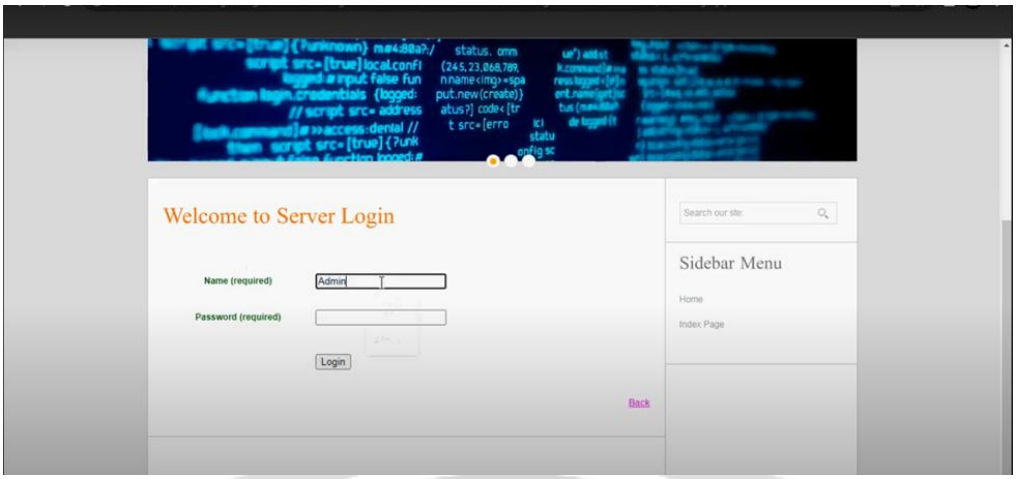

Fig . Server Login

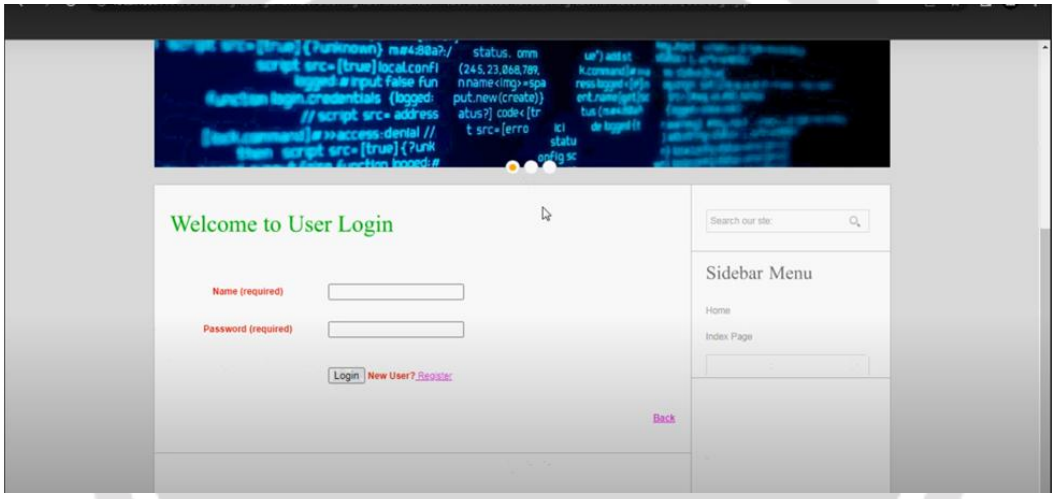

Fig. User Login

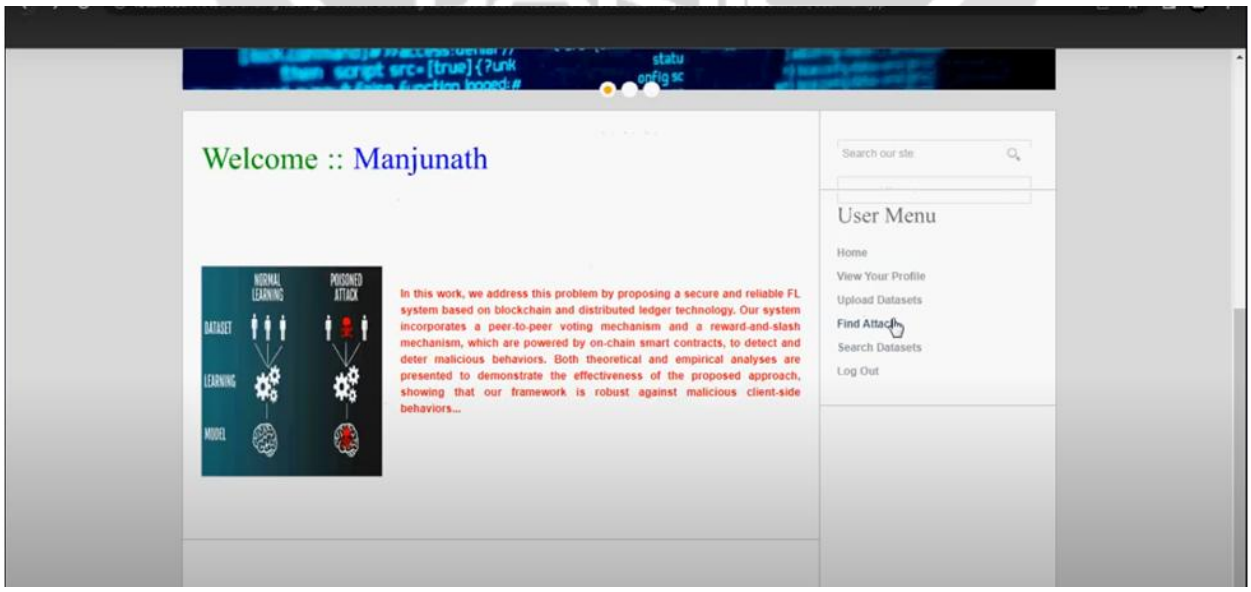

Fig. User Home page

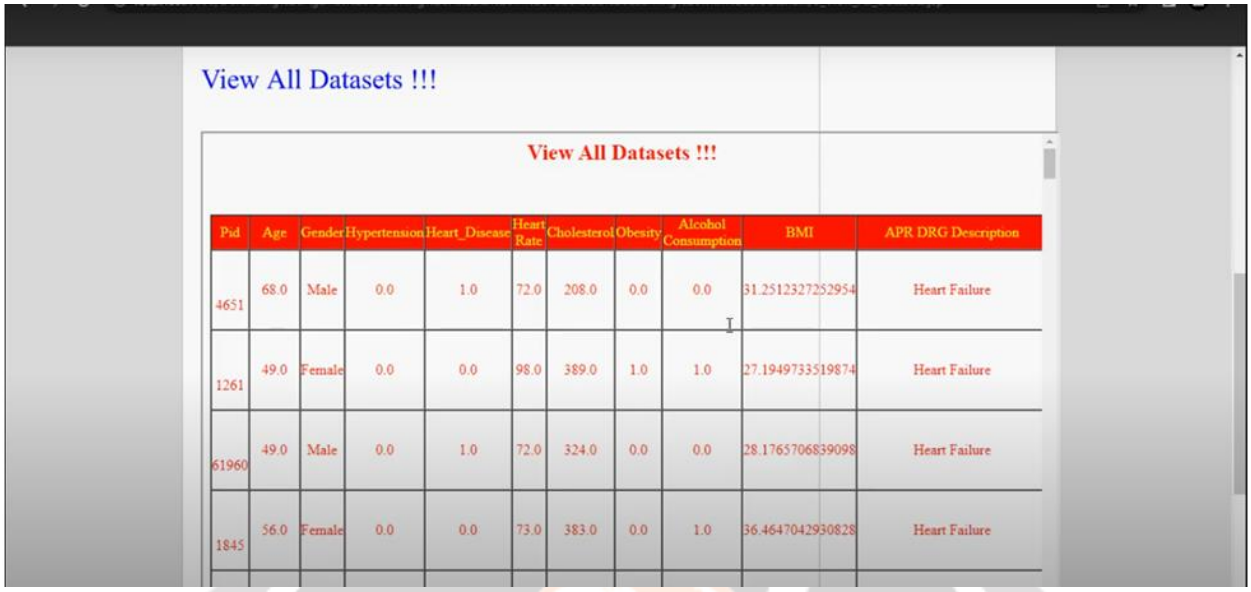

Fig. View all Dataset

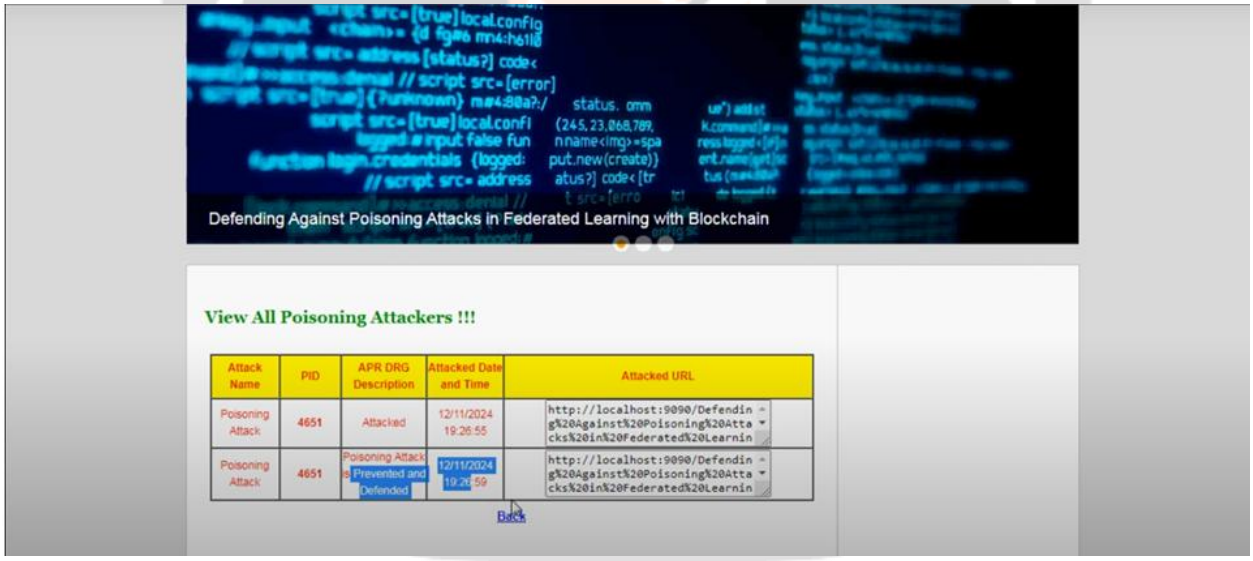

Fig.output

6. CONCLUSION

Federated Learning (FL) is a powerful paradigm for decentralized machine learning, offering privacy benefits by keeping data local. However, its open and distributed nature makes it vulnerable to **poisoning attacks**, where adversarial participants manipulate model updates to degrade performance or introduce biases. Traditional defensive mechanisms, such as anomaly detection and robust aggregation techniques, offer partial solutions but often lack **scalability, transparency, and tamper resistance**.In this study, we proposed a **blockchain-integrated federated learning framework** to mitigate poisoning attacks effectively. Blockchain technology enhances **data integrity, accountability, and decentralized trust** by maintaining an immutable ledger of model updates. Smart contracts enable **automated verification of contributions**, while consensus mechanisms help prevent malicious actors from compromising the model. Furthermore, **cryptographic techniques** such as homomorphic encryption and zero-knowledge proofs ensure data privacy while maintaining security.

7. REFERENCE

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[2] A. Vaswani et al., "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.

[3] V. Mnih et al., "Human-level control through deep reinforcement learn ing," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[4] European Commission, "General data protection regulation," 2016. Accessed: Apr. 25, 2023. [Online]. Available: https://ec.europa.eu/info/ law/law-topic/data-protection/data-protection-eu_en

[5] U.S. Department of Health and Human Services, "Health insur ance portability and accountability act," 2017. Accessed: Apr. 25, 2023. [Online]. Available: https://www.cdc.gov/phlp/publications/topic/ hipaa.html

[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, PMLR, 2017, pp. 1273–1282.

[7] M. Li, D. G. Andersen, A. J. Smola, and K. Yu, "Communication efficient distributed machine learning with the parameter server," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2014, pp. 19–27.

[8] Y. Qu, M. P. Uddin, C. Gan, Y. Xiang, L. Gao, and J. Yearwood, "Blockchain-enabled federated learning: A survey," *ACM Comput. Surv.*, vol. 55, no. 4, pp. 1–35, 2022.

[9] G. Wood, "Ethereum: A secure decentralised generalised transaction ledger," *Ethereum Project Yellow Paper*, vol. 151, pp. 1–32, 2014.

[10] X. Chen, J. Ji, C. Luo, W. Liao, and P. Li, "When machine learning meets blockchain: A decentralized, privacy-preserving and secure design," in *Proc. IEEE Int. Conf. Big Data*, Piscataway, NJ, USA: IEEE, 2018, pp. 1178–1187.