WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Defending Against Poisoning Attacks in Federated Learning with Blockchain

[1]Gokul P, [2]Surya S, [3]Thirumalai Vasan L, [4]Vignesh P G, [5]Sundaram M

[1] B.E. Computer Science and Engineering (Final year),

[2] B.E. Computer Science and Engineering (Final year),

[3] B.E. Computer Science and Engineering (Final year),

[4] B.E. Computer Science and Engineering (Final year),

[5] Assistant Professor, B.E, M.E,

Department of Computer Science and Engineering, Pavai College of Technology, Paavai Institutions, Paavai Nagar, NH-44, Pachal -637 018. Namakkal Dist., Tamil Nadu, India

ABSTRACT:

Federated Learning (FL) has revolutionized machine learning by enabling decentralized, privacy-focused data processing. However, FL remains susceptible to data poisoning attacks, especially Label-Flipping Attacks (LFA). In LFA, malicious clients intentionally mislabel local data, leading to inaccuracies in the global model's predictions and compromising its integrity. To address this challenge, this project proposes a blockchain-based federated learning system integrated with a Self-Purified FL (SPFL) model verification mechanism.

KEYWORDS: Data Integrity, Data Manipulation, Model Vulnerability, Security in AI, Training Data poisoning, Poisoning Attack, Adversarial Attacks, Machine Learning Security, Backdoor Attacks

HIGHLIGHTS

- **Federated Learning Security:** A system that enables collaborative machine learning across multiple users without sharing private data.
- **Label-Flipping Attack Protection:** Detects and blocks malicious users who intentionally flip data labels to corrupt the learning process.
- **Blockchain Integration:** Uses blockchain to store and verify model updates, ensuring transparency, immutability, and data integrity.
- **Self-Purified FL (SPFL**): Automatically filters poisoned data before it affects the global model by validating updates with smart algorithms.

INTRODUCTION:

Federated learning is a way to develop and validate AI models from diverse data sources while mitigating the risk of compromising data security or privacy, as the data never leaves individual sites. For AI algorithms, experience comes in the form of large, varied, high-quality datasets. But such datasets have traditionally proved hard to come by, especially in the area of healthcare. Federated learning is a way to develop and validate accurate, generalizable AI models from diverse data sources while mitigating the risk of compromising data security or privacy. It enables AI models to be built with a consortium of data providers without the data ever leaving individual sites. Medical institutions have had to rely on their own data sources, which can be biased by, for example, patient demographics, the instruments used or clinical specializations. Or they've needed to pool data from other institutions to gather all of the information they need, which requires managing regulatory issues. Federated learning makes it possible for AI algorithms to gain experience from a vast range of data located at different sites. The approach enables several organizations to collaborate on the development of models, but without needing to directly share sensitive clinical data with each other. Over the course of several training iterations the shared models get exposed to a significantly wider range of data than what any single organization possesses in-house.

PROBLEM DEFINITION:

Federated Learning (FL) has emerged as a powerful paradigm for decentralized machine learning, allowing multiple clients to collaboratively train a global model without exposing their private data. However, despite its privacy-preserving nature, FL systems are increasingly vulnerable to sophisticated data poisoning techniques, particularly Label-Flipping Attacks (LFA). In these attacks, malicious clients intentionally alter the labels of their local datasets to manipulate the learning process, ultimately causing the global model to produce inaccurate predictions. This undermines the reliability and integrity of the entire FL ecosystem. One of the critical challenges in traditional FL frameworks is the absence of robust data verification mechanisms before model training. As FL relies on trust among distributed participants, it becomes difficult to detect or prevent malicious contributions during the training phase. Without a decentralized and tamper-proof validation method, attackers can easily inject poisoned data and

escape detection, compromising the model's performance and trustworthiness. Moreover, existing FL systems often employ centralized aggregation and verification mechanisms, which introduce a single point of failure and are prone to inefficiencies. These centralized solutions also fail to provide transparency and auditability, which are essential for securing collaborative training environments. The lack of real-time verification and feedback mechanisms further hampers the ability to isolate or eliminate malicious clients before they impact the global model. Another significant limitation is the absence of data integrity enforcement. There is no effective method to ensure that the labels and data used for training remain consistent and unaltered. This leaves the system exposed to not only LFAs but also to other types of data manipulation and integrity breaches. Furthermore, FL systems lack adaptive defense mechanisms capable of identifying malicious patterns in real-time. The inability to automatically purify contributions from clients based on behavior or data anomalies results in compromised training rounds and model deterioration over time. To address these pressing issues, this project introduces a Blockchain-Based Federated Learning architecture integrated with a Self-Purified Federated Learning (SPFL) mechanism. By leveraging the immutable and decentralized nature of blockchain technology, the system securely stores global model labels and training metadata. Whenever a client attempts to train on local data, the blockchain alerts the SPFL module to verify the integrity of labels and data. Only verified and trusted contributions are allowed to participate in the model aggregation process, ensuring a secure, transparent, and tamper-resistant learning environment.

OBJECTIVE:

The objective of the project is to develop a Blockchain-Based Federated Learning system integrated with a Self-Purified Federated Learning (SPFL) mechanism to enhance the security, integrity, and accuracy of collaborative machine learning. The system aims to prevent Label-Flipping Attacks by verifying and validating local training data before it is aggregated into the global model. By leveraging blockchain technology, the project ensures tamper-proof storage of labels and model updates, enabling transparent and decentralized verification without relying on a central authority. This approach helps in aggregating only trustworthy contributions, thereby improving the reliability and performance of the global model while maintaining data privacy across distributed clients.

SUMMARY OF ISSUES:

- Latency in Real-Time Training
- High Computational and Time Overhead
- Complex Implementation and Maintenance

EXISTING SYSTEM:

- **Existing Federated Learning (FL)** – Decentralized training without centralized data collection but vulnerable to attacks.
- **Anomaly Detection-Based Approaches** – Detects abnormal model updates but lacks efficiency against sophisticated Label-Flipping Attacks (LFA).
- **Reputation-Based Client Scoring** – Assigns trust scores to clients, but attackers can still manipulate scores over time.
- **Differential Privacy (DP) Methods** – Adds noise to data for privacy but does not directly prevent malicious data poisoning.

DISADVANTAGES

- Fails to detect and prevent Label-Flipping Attacks.
- Lacks proper validation of local training data.
- Relies on a central server, creating a single point of failure.
- Secure aggregation and privacy methods increase computational costs.
- Reputation-based systems can be manipulated, reducing trust.

PROPOSED SYSTEM:

- The proposed system introduces a **Blockchain-Based Federated Learning framework integrated with a Self-Purified Federated Learning (SPFL) mechanism** to address the security vulnerabilities of traditional federated learning, particularly against Label-Flipping Attacks (LFA). In this system, blockchain technology is used to store and verify model updates and data labels, ensuring immutability, transparency, and tamper-proof data integrity.
- The SPFL component acts as a verification layer that detects and filters out poisoned or mislabeled data from malicious clients before they are included in the global model training. By analyzing update patterns and comparing them with expected behaviors, the system identifies suspicious contributions and prevents them from influencing the final model.
- This decentralized architecture eliminates reliance on a central authority, enhances trust among participants, and ensures that only verified, high-integrity data contributes to the global model. As a result, the proposed system significantly improves model security, reliability, and performance in collaborative machine learning environments.

ADVANTAGES

- Blockchain ensures tamper-proof storage of global model labels.
- SPFL detects and eliminates malicious clients.
- Only verified local data is used for training.
- Eliminates the need for a central authority, increasing trust.
- Aggregates only reliable updates, improving overall performance.

SYSTEM REQUIREMENT SPECIFICATION:

This chapter outlines the essential requirements for the successful implementation and operation of the **Blockchain-Based Federated Learning System with Self-Purified FL (SPFL)**. It details both hardware and software prerequisites necessary for the application to function correctly. The Software Requirement Specification (SRS) is explained comprehensively, including an overview of the system and a breakdown of both functional and non-functional requirements.

The SRS document provides a full description of the data, functions, and behaviors expected from the software system under development. It serves as a foundational document in the software development life cycle, acting as a reference for all stages including design, implementation, testing, and maintenance. This document ensures that the system being developed aligns with the objectives and expectations set forth during the requirement analysis phase.

Requirement analysis identifies and analyzes the conditions the system must satisfy to be considered successful. These requirements must be **clearly documented, measurable, testable**, and directly linked to the project's core goals. They should be defined to a level of detail that supports accurate system design and implementation.

The SRS functions as a **blueprint** for the entire development process. The primary purposes of preparing this document are:

- To facilitate clear and consistent communication between stakeholders such as the customer, analyst, system developer, and maintenance teams.
- To provide a solid foundation for the design and development phases.
- To support effective system testing and validation procedures.
- To guide and control the evolution and enhancement of the system over time.

SYSTEM ARCHITECTURE:

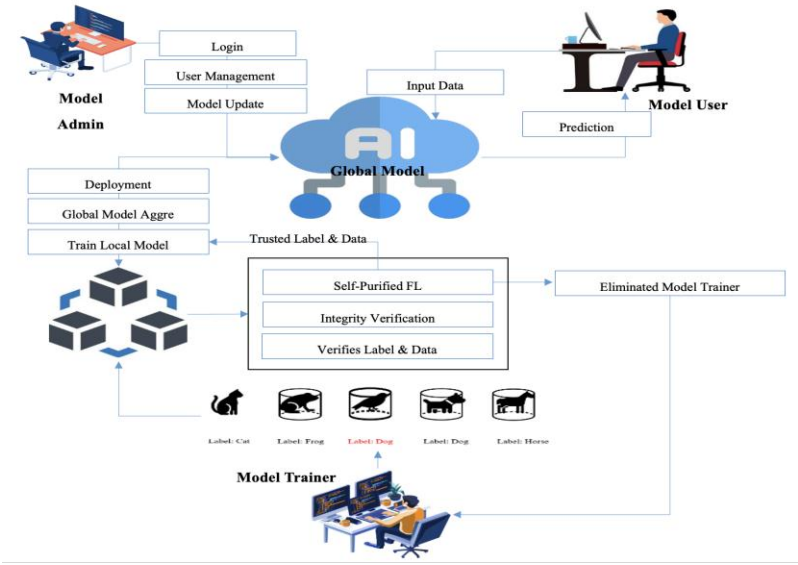
HARDWARE REQUIREMENT

- **Processor:** Intel Core i5 or higher, AMD Ryzen 5 or Higher
- **RAM:** 8 GB or more
- **Storage:** 250 GB SSD or higher
- **Operating System:** Windows 10 or 11

SOFTWARE REQUIREMENT

- **Programming Language:** Python
- **Frameworks:** Flask, TensorFlow, Scikit-learn
- **Blockchain Tool:** Web3.py
- **Database:** MySQL
- **Web Technologies:** HTML, CSS, JavaScript, Bootstrap
- **Other Tools:** WampServer, pip for dependency management

SYSTEM ARCHITECTURE:



PROCEDURE

- Admin initiates the global training process and uploads the initial global model
- Registered model trainers log in and download the global model for local training
- Each trainer trains the model locally using private datasets without sharing raw data
- Trainers submit their locally trained model updates through the system interface
- SPFL (Self-Purified Federated Learning) verifies the updates for malicious behavior (e.g., label-flipping attacks)
- Verified updates are hashed and logged on the blockchain to ensure transparency and tamper-proof storage
- Only verified and clean updates are aggregated into the global model using secure aggregation algorithms
- The updated global model is evaluated for performance and accuracy
- Admin monitors training progress, SPFL decisions, and blockchain verification via a real-time dashboard
- Once training rounds are complete, the final global model is deployed and made available for real-world use

CONCLUSIONS:

In conclusion, this project demonstrates a robust and secure framework that integrates Blockchain with Federated Learning to mitigate Label-Flipping Attacks using a Self-Purified FL (SPFL) mechanism. It ensures privacy-preserving local model training while validating model updates for integrity through blockchain verification. By filtering out malicious clients and maintaining an immutable log of legitimate contributions, the system significantly enhances the reliability and trustworthiness of the global model. The modular design supports real-time dashboards, user management, and secure deployment, making it a powerful solution for decentralized and secure AI model training in sensitive domains such as healthcare, finance, and IoT.

REFERENCE:

1. Y. Zhao, J. Zhang and Y. Cao, "Manipulating vulnerability: Poisoning attacks and countermeasures in federated cloud–edge–client learning for image classification", Knowl.-Based Syst., vol. 259, Jan. 2023.
2. G. Xia, J. Chen, C. Yu and J. Ma, "Poisoning attacks in federated learning: A survey", IEEE Access, vol. 11, pp. 10708-10722, 2023.
3. S. Bansal, M. Singh, M. Bhadauria and R. Adalakha, "Federated learning approach towards sentiment analysis", Proc. 2nd Int. Conf. Technol. Adv. Comput. Sci. (ICTACS), pp. 717-724, 2022.
4. X. Liu, H. Li, G. Xu, Z. Chen, X. Huang and R. Lu, "Privacy-enhanced federated learning against poisoning adversaries", IEEE Trans. Inf. Forensics Security, vol. 16, pp. 4574-4588, 2021.
5. D. Li, W. E. Wong, W. Wang, Y. Yao and M. Chau, "Detection and mitigation of label-flipping attacks in federated learning systems with KPCA and K-means", Proc. 8th Int. Conf. Dependable Syst. Their Appl. (DSA), pp. 551-559, 2021