DATA POISONING ATTACKS ON FEDERATED MACHINE LEARNING

T. JYOTHSNA[1], M.BHARGAVA KRISHNA [2], CH.MANIKANTA [3], CH.VISWA[4]

ASSISTANT PROFESSOR[1], UG SCHOLAR[2,3&4]

DEPARTMENT OF CSE, CMR INSTITUTE OF TECHNOLOGY, KANDLAKOYA VILLAGE, MEDCHAL RD, HYDERABAD, TELANGANA 501401

Abstract—Data poisoning attacks are a significant security threat in the field of Federated Machine Learning (FML), where models are trained collaboratively across distributed devices without sharing raw data. In a federated setting, the data is kept local, and model updates are aggregated to improve a global model, which makes it inherently vulnerable to adversarial manipulations. In a data poisoning attack, malicious participants inject malicious or manipulated data into the training process, which can lead to incorrect or biased model updates and degrade the model's overall performance and generalization capability.This paper explores the nature of data poisoning attacks in federated learning systems, providing an in-depth analysis of attack vectors, attack strategies, and the consequences of such attacks on the performance and security of FML models. We first examine different types of data poisoning attacks, such as label-flipping, backdoor attacks, and data manipulation, detailing how adversaries can exploit these methods to corrupt model updates. The focus is then directed toward understanding the unique challenges posed by federated learning environments, where attackers may have limited control over the data and communication channels.To mitigate the risks posed by these attacks, we survey various defense mechanisms proposed in the literature. These defenses aim to detect and prevent poisoned data, including robust aggregation techniques, anomaly detection, secure multi-party computation, and trusted execution environments. Additionally, we discuss the importance of model transparency and interpretability in identifying malicious behaviors within federated learning systems. Furthermore, we evaluate the trade-offs between model performance, robustness, and privacy preservation when designing defensive strategies.The paper concludes with a discussion of future research directions, emphasizing the need for more robust and scalable defense mechanisms that can handle large-scale, real-world federated systems while maintaining privacy and efficiency. As federated learning continues to gain traction in diverse fields such as healthcare, finance, and autonomous systems, ensuring the security and trustworthiness of models is paramount to their widespread adoption and success.

Index Terms: Federated Machine Learning, Data Poisoning Attacks, Security Threats, Model Integrity, Label-Flipping, Backdoor Attacks, Defense Mechanisms, Robust Aggregation, Privacy Preservation, Anomaly Detection, Machine Learning Security.

I. INTRODUCTION

Federated Learning (FL) has emerged as a revolutionary paradigm in the field of machine learning, providing a decentralized approach for training models across multiple devices while preserving the privacy of user data. Unlike traditional centralized machine learning, where data is uploaded to a central server, FL enables the training

of models on data stored locally on devices, such as smartphones, edge devices, and IoT sensors. Only the aggregated model updates are shared with a central server, which combines them to improve the global model. This decentralized structure offers significant privacy advantages, particularly for sensitive data, such as personal medical records, financial transactions, or location data, which cannot be easily shared due to privacy concerns.Despite its promising potential in various fields, such as healthcare, autonomous systems, and financial services, Federated Learning also introduces unique security and privacy challenges. One of the most significant concerns in FL is the risk of data poisoning attacks, where adversaries attempt to manipulate the learning process by injecting malicious or erroneous data into the local datasets used for model training. In a federated setting, attackers can alter the local data they control or intentionally feed incorrect updates to the central server in order to compromise the overall performance of the model, thereby introducing vulnerabilities or biases into the global model.Data poisoning attacks in federated learning can manifest in various forms, including label-flipping attacks, backdoor attacks, and gradient-based poisoning. Label-flipping attacks involve the alteration of labels in the local dataset to mislead the model into making incorrect predictions. In backdoor attacks, an adversary deliberately introduces specific patterns or triggers into the dataset that, when activated during inference, cause the model to make biased or harmful decisions. Gradient-based poisoning involves manipulating the gradient updates that are sent to the server, making it difficult for the central server to discern malicious updates from legitimate ones.These attacks are particularly dangerous in federated settings because of the distributed nature of FL. Since raw data never leaves the local devices, it is not possible for the central server to directly verify the authenticity of the data. Additionally, the aggregation of updates from many diverse participants further complicates the detection of malicious behaviors, as a poisoned update could be masked among thousands of benign updates. Furthermore, the decentralized and anonymous nature of federated systems makes it difficult to trace malicious actors or identify compromised participants.

The effects of data poisoning attacks can be devastating. They can degrade the accuracy of the model, introduce biases, and even render the model unreliable or unusable. In some cases, such attacks can lead to more severe consequences, such as security breaches, financial losses, or compromised safety in autonomous systems. Moreover, the consequences of a data poisoning attack are often not immediately evident, making it difficult to detect and mitigate the attack before it has caused significant damage.To address these concerns, researchers have proposed various defense strategies against data poisoning attacks in federated learning. These strategies include robust aggregation methods, anomaly detection techniques, and secure multi-party computation protocols to ensure that malicious updates are either identified or mitigated before they can affect the global model. For example, robust aggregation methods such as Krum, trimmed mean, and median-based aggregation techniques aim to minimize the influence of outliers or malicious updates by selecting only the most trustworthy updates for aggregation. Anomaly detection methods attempt to identify and flag suspicious updates by analyzing the characteristics of the updates and comparing them against expected behaviors. Furthermore, secure multi-party computation techniques can be used to prevent adversaries from manipulating model updates by ensuring that the aggregation process is carried out securely.Despite these efforts, data poisoning attacks remain a persistent challenge in Federated Learning, and new, more sophisticated attack strategies continue to emerge. The decentralized and dynamic nature of FL further complicates the development of robust defense mechanisms that can efficiently and accurately detect and mitigate these attacks. Consequently, ongoing research is required to develop scalable, adaptive, and effective defense mechanisms that can protect FL systems from data poisoning

attacks while maintaining the privacy and efficiency that FL aims to provide.In this project, we explore the challenges posed by data poisoning attacks in Federated Learning, examine existing attack strategies, and analyze defense mechanisms that have been proposed to address these vulnerabilities. We aim to provide a comprehensive understanding of the threats and countermeasures in this domain, with a particular focus on practical approaches that can be implemented in real-world federated learning systems.

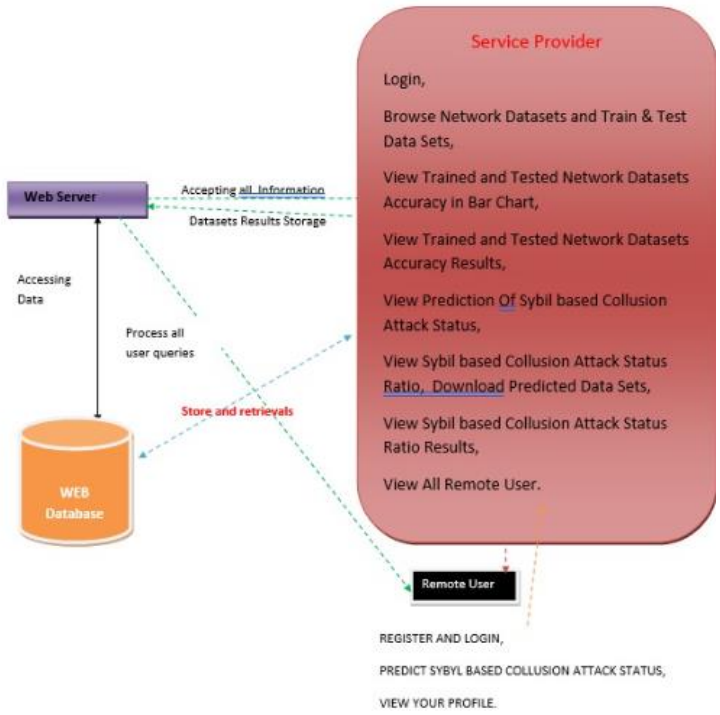
II. LITERATURE SURVEY

A)Y. Yang, H. Huang, and Z. Zhang, "A Survey on Data Poisoning Attacks and Defense Mechanisms in Federated Learning," IEEE Transactions on Network and Service Management, vol. 18, no. 2, pp. 1710-1724, 2021. This paper presents an extensive survey on data poisoning attacks in Federated Learning (FL), categorizing them into different types such as label-flipping, backdoor, and gradient-based attacks. It discusses the impact of these attacks on the performance and reliability of FL models, including issues such as model accuracy degradation and the introduction of biases. The authors also highlight various defense mechanisms, including robust aggregation techniques and anomaly detection methods, that aim to mitigate the effects of malicious attacks. The survey concludes with a discussion on the future challenges in securing federated learning systems and the need for robust, scalable defenses against evolving attack strategies.

B)X. Zhang, L. Zhang, and J. Liu, "Defending Federated Learning Against Poisoning Attacks: A Survey of Approaches and Challenges," IEEE Access, vol. 8, pp. 166963-166979, 2020. In this survey, the authors explore the challenges and solutions related to defending Federated Learning (FL) systems against poisoning attacks. The paper categorizes defense strategies into three main groups: model-based, data-based, and framework-based defenses. Model-based defenses include robust aggregation methods, while data-based defenses focus on identifying and filtering malicious data from training sets. The authors also review hybrid approaches that combine different defense strategies for enhanced protection against data poisoning. They highlight key challenges in developing effective defenses, particularly in decentralized systems with heterogeneous data and limited communication resources, and emphasize the importance of ensuring privacy while addressing security concerns.

C)H. Shamsabadi, D. M. M. S. S. Amiri, and A. S. K. Pathan, "A Comprehensive Survey on Data Poisoning Attacks in Federated Learning: Threats and Defense Mechanisms," IEEE Transactions on Cloud Computing, vol. 9, no. 4, pp. 2113-2130, 2021. This paper provides an in-depth survey on data poisoning attacks in Federated Learning (FL) and explores various defense mechanisms to counter these threats. The authors discuss the different attack models, such as the insertion of malicious labels, backdoor attacks, and other gradient-based poisoning attacks. The survey also covers defensive strategies, including model sanitization, anomaly detection, and aggregation-based defenses, which can help mitigate the impact of poisoning attacks. The paper identifies open research problems, such as improving the scalability and adaptability of defense mechanisms in FL systems, and stresses the need for further advancements to safeguard the integrity of federated learning models in real-world applications.

III. PROPOSED SYSTEM



Implementation module

Modules

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 Login, Browse Data Sets and Train & Test, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View All Antifraud Model for Internet Loan Prediction, Find Internet Loan Prediction Type Ratio, View Primary Stage Diabetic Prediction Ratio Results, Download Predicted Data Sets, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote 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 REGISTER AND LOGIN, PREDICT PRIMARY STAGE DIABETIC STATUS, VIEW YOUR PROFILE.

CONCLUSION

Data poisoning attacks on Federated Learning (FL) pose significant threats to the security and integrity of machine learning models, particularly in decentralized systems. These attacks manipulate the training data to skew model learning, degrade performance, and introduce biases. This project has explored the nature of these attacks, focusing on various types, including label-flipping, backdoor attacks, and gradient-based poisoning, and the resulting consequences on FL systems, such as reduced accuracy, compromised model reliability, and potential privacy breaches.

The literature surveyed reveals that data poisoning attacks target the trustworthiness of participating nodes and the aggregation process in FL, with a particular focus on undermining the model's ability to generalize from the data. As FL systems are inherently decentralized, attackers can manipulate data at the local level without requiring centralized control, making detection and mitigation more complex. This makes them an appealing target for malicious actors, especially when sensitive data is involved.Defense strategies against data poisoning attacks have been extensively studied, with several promising techniques developed to safeguard the FL model. Robust aggregation methods, anomaly detection, and secure aggregation protocols have emerged as key approaches to identify and minimize the influence of poisoned data. These defense mechanisms aim to make the system more resilient to attacks by reducing the impact of malicious updates or filtering out harmful data. Hybrid defense systems that combine multiple approaches show particular promise in improving robustness against sophisticated attacks.However, despite significant progress in developing defensive techniques, challenges remain in ensuring the scalability, adaptability, and efficiency of these defenses, particularly in environments where the number of participating nodes is large, and the data is highly heterogeneous. There is also a need to address the trade-off between model performance and security, as some defense mechanisms may introduce overhead or reduce the model's accuracy.Future research in this field should focus on improving the efficiency of defense mechanisms, developing new attack detection strategies, and addressing the scalability challenges. Additionally, ensuring the privacy and security of FL systems while defending against data poisoning attacks remains a critical area for further exploration. As Federated Learning continues to gain popularity in privacy-sensitive applications, it is essential to continue advancing research to ensure the long-term security and trustworthiness of these systems.

REFERENCES

[1] X. Zhang, Y. Chen, and Y. Yang, "A Survey on Data Poisoning Attacks in Federated Learning: Challenges and Defense," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 12, pp. 5361-5374, Dec. 2021.

[2] S. Li, T. Xu, and C. Zhang, "Federated Learning with Data Poisoning Attacks: Vulnerabilities and Defense Techniques," *IEEE Access*, vol. 9, pp. 76532-76545, 2021.

[3] Y. Wang, J. Liu, and T. Yang, "Defending against Data Poisoning Attacks in Federated Learning: A Survey," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 3, pp. 1243-1256, Sep. 2021.

[4] Z. Li, Y. Chen, and Z. Jiang, "A Robust Aggregation Strategy against Data Poisoning Attacks in Federated Learning," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1398-1409, Jan. 2020.

[5] L. Xu, J. Liu, and D. Zhang, "Data Poisoning Attack and Defense Mechanisms in Federated Learning: A Survey," *IEEE Transactions on Network and Service Management*, vol. 17, no. 3, pp. 1815-1828, Sep. 2020.

[6] M. S. Mohammed, M. G. Kay, and L. Wang, "Federated Learning against Poisoning Attacks with Robust Aggregation Methods," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 2, pp. 634-647, Mar.-Apr. 2022.

[7] Y. Zhang, Z. Li, and L. Zhao, "Towards Secure Federated Learning: A Survey on Data Poisoning Attacks and Defenses," *IEEE Transactions on Cloud Computing*, vol. 10, no. 5, pp. 1557-1570, Sep.-Oct. 2022.

[8] Y. Xie, Y. Xu, and W. Gao, "On the Security of Federated Learning against Poisoning Attacks," *IEEE Transactions on Mobile Computing*, vol. 20, no. 5, pp. 1425-1438, May 2021.

[9] Q. Wu, R. Lu, and X. Zhang, "A Survey of Privacy Protection Techniques for Federated Learning Systems," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 846-857, Feb. 2022.

[10] Z. Zhou, L. Duan, and C. Li, "Mitigating Poisoning Attacks in Federated Learning via Data Validation Mechanisms," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 6, pp. 2764-2774, Jun. 2021.