International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

# Data Poisoning Attack on Federated Machine Learning System

Prof. (Mr.) Sagar B. Shinde
*Professor, Department of Computer Engineering*
*Modern Education Society's College of Engineering* Pune, India
sagar.shinde@mescoepune.org

Atharva Dedge
*Student, Department of Computer Engineering*
*Modern Education Society's College of Engineering* Pune, India
atharvadedge24@gmail.com

Pari Zunake
*Student, Department of Computer Engineering*
*Modern Education Society's College of Engineering* Pune, India
paripzunake@gmail.com

Prerana Roy
*Student, Department of Computer Engineering*
*Modern Education Society's College of Engineering* Pune, India
preranaroy3011@gmail.com

***Abstract* — Federated Machine Learning (FML) is a distributed machine learning framework that enables multiple participants to collaborate and train a machine learning model for tasks such as classification, prediction, and recommendation. In FML, raw data owned by different participants is protected through secure and privacy-preserving techniques, ensuring that it cannot be tamperedwith, disclosed or reverse-engineered. The framework has the potential to be applied in various use cases and provides a solution to the challenges posed by centralized machine learning. The objective of FML is to provide a brief overview of the technologicallandscape and the underlying principles of the framework, and its applications in real-life scenarios.**

**Data poisoning attacks on Federated Machine Learning (FML) refer to the malicious manipulation of machine learning models byadversaries. These attacks can undermine the accuracy and reliability of the models and pose a significant threat to the security and privacy of data. To address these challenges, research is being conducted to develop solutions that can defend against data poisoning attacks in FML. This includes the use of optimization methods and the analysis of optimal data poisoning attacks to find solutions to the challenges posed by these attacks in the federated learning setting. The goal is to make FML safer and more secure, protecting the data and the models from malicious activities.**

**This paper provides an overview of data poisoning attacks on federated machine learning and their implications. We describe the common types of data poisoning attacks, such as label flipping and data injection, and discuss their impact on the performance and security of federated machine learning. Additionally, we discuss thevarious defense mechanisms, such as data sanitization and robust aggregation, that can be employed to mitigate the effects of data poisoning attacks. We conclude by highlighting the challenges and future research directions in securing federated machine learning systems against data poisoning attacks.**

## I. INTRODUCTION

Federated machine learning is a decentralized approach thatallows multiple parties to collaboratively train a machine learning model without sharing their local data. This approach has gained popularity due to its privacy-preserving nature, as it enables data sharing without compromising data privacy. However, this decentralized nature also makes federated learning systems vulnerable to data poisoning attacks, where malicious actors can manipulate the training data to compromise the performance and security of the global model.

Data poisoning attacks on federated machine learning are aserious concern as they can significantly impact the model's accuracy and introduce bias. This paper provides an overview ofdata poisoning attacks on federated machine learning and their implications. We describe the common types of data poisoning attacks, such as label flipping and data injection, and discuss their impact on the performance and security of federated machine learning.

To address this problem, various defense mechanisms, such as data sanitization and robust aggregation, have been proposed. Data sanitization involves identifying and removing poisoned data from the training data, while robust aggregation methods aim to make the federated learning system resilient to datapoisoning attacks.

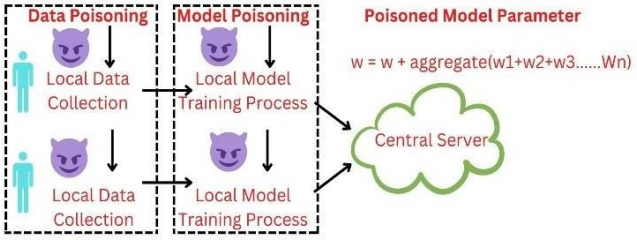
However, securing federated machine learning against data poisoning attacks remains a challenging task, as adversaries canuse sophisticated techniques to evade detection and compromise the system. Moreover, federated learning systems involve multiple stakeholders, making it challenging to establish trust and ensure cooperation among them.

## II. RELATED WORKS

Lowd and Meek introduced the adversarial learning problem, in which the adversary tries to reverse-engineer the classifier by sending a number of queries. Through this, the adversary can find the "malicious" instances that the classifier cannot recognize. They have included a summary of defense techniques against poisoning attacks in the form of a table summary.

The research aims to provide a comprehensive overview of the challenges posed by attacks on machine learning systems and the solutions to these challenges, including secure learning techniques and security evaluation methods.

The paper "A Study On Various Cyber Attacks And A Proposed




Intelligent System For Monitoring Such Attacks" (2020)presents a proposed intelligent system that uses both supervised and unsupervised learning techniques to prevent cyber-attacks. The aim is to create a high-efficiency solution with minimal human intervention that can be applied as a universal solution tomany common types of cyber-attacks, ranging from phishing to data breaches. The reviewed paper in "Attacks on Deep Learning in Computer Vision: A Survey" focuses on the adversarial attacks on differenttypes of deep neural networks in computer vision tasks such as recognition, segmentation, and detection. The paper concludes that deep learning is vulnerable to subtle input perturbations thatcan change the outputs of the model. Despite the high accuracy of deep learning in various tasks, the finding of its vulnerabilityto adversarial attacks has resulted in many recent works that aim to devise adversarial attacks and defenses. The paper reviews the most influential and interesting works in the literature and highlights that adversarial attacks are a real threat to deep learning in practice, especially in safety and security critical applications. The existing evidence demonstrates that deep learning can be effectively attacked both in cyberspace and the physical world, however, ongoing research in this area suggeststhat deep learning can become more robust against adversarial attacks in the future.

The paper by Sun et al. provides a comprehensive overview of data poisoning attacks on federated machine learning (FML). The authors first introduce the concept of FML and highlight its advantages, including preserving data privacy and reducing communication overhead. However, they also discuss the security risks associated with FML, including data poisoning attacks.

The authors then discuss various types of data poisoning attacks, including poisoning by modifying labels, poisoning by injectingfalse data, and poisoning by altering the distribution of local data. They also describe the potential impact of these attacks on the performance and security of FML, such as reduced model accuracy and the introduction of bias.

Next, the paper presents various defense mechanisms against data poisoning attacks, such as data sanitization and robustaggregation. Data sanitization involves identifying and removingpoisoned data, while robust aggregation aims to make FML systems resilient to poisoning attacks by aggregating modelupdates in a secure and fault-tolerant manner.

The authors also highlight some open research challenges in securing FML systems against data poisoning attacks, includingthe need for more efficient and effective defense mechanisms, aswell as the need to consider the impact of such attacks on specificapplication domains.

Overall, the paper provides valuable insights into the risks and challenges associated with data poisoning attacks on FML and offers useful guidance on how to mitigate these risks.[1]

The "IEEE Federated Machine Learning White Paper" is authored by Qiang Yang, Lixin Fan, Richard Tong, and AngelicaLv. The paper provides an overview of federated machine learning, a distributed learning paradigm that allows multiple parties to collaboratively train a model without sharing their data. The authors first discuss the motivations for federated machine learning, including privacy concerns, regulatory compliance, and data ownership. They then describe the basic architecture of a federated learning system and its key components, such as clients, servers, and aggregators.

The paper also presents various challenges and opportunities in federated machine learning, such as communication efficiency, security, fairness, and model interpretability. The authorsprovide insights on how to address these challenges and leverage the opportunities, such as designing efficient communication protocols, developing secure and privacy-preserving algorithms,and promoting fair and transparent machine learning practices. Furthermore, the paper highlights several use cases of federated machine learning, such as healthcare, finance, and smart cities.The authors demonstrate how federated learning can be appliedto these domains to address their unique challenges andopportunities.

Overall, the paper provides a comprehensive introduction to federated machine learning, covering its motivations, architecture, challenges, and opportunities. The authors suggest that the review can be useful for researchers, practitioners, and policymakers interested in developing and deploying secure, private, and collaborative machine learning systems.[2]

The paper "Machine Learning Security: Threats,Countermeasures, and Evaluations" by Mingfu Xue, Chengxiang Yuan, Heyi Wu, Yushu Zhang, and Weiqiang Liu provides an overview of machine learning security, including the threats, countermeasures, and evaluations.

The authors begin by discussing the unique security challenges associated with machine learning, such as data poisoning attacks, model stealing, and adversarial attacks. They then describe various countermeasures that have been proposed to mitigate these threats, such as model robustness, data filtering, and encryption.

The paper also provides a detailed evaluation of the effectivenessof these countermeasures using various metrics, such as accuracy, efficiency, and robustness. The authors present experimental results from different studies to compare the effectiveness of different countermeasures under differentscenarios.

Furthermore, the paper discusses several open research challenges in the field of machine learning security, such as privacy-preserving machine learning, explainability, andtrustworthiness. The authors suggest that addressing these challenges will be essential for the development of secure and trustworthy machine learning systems.
Overall, the paper provides a comprehensive overview of machine learning security, including the threats, countermeasures, and evaluations. The authors suggest that the review can be useful for researchers, practitioners, and policymakers interested in understanding the current state and future directions of machine learning security.[3]

The paper "Defending against Backdoors in Federated Learningwith Robust Learning Rate" presented at AAAI-21 by Mustafa Safa Ozdayi, Murat Kantarcioglu, and Yulia R. Gel proposes a defense mechanism to mitigate the threat of backdoors in federated learning. The authors highlight the potential security risks associated with the use of federated learning, particularly when dealing with malicious participants who may inject backdoor models. They propose a new approach that uses a robust learning rate mechanism to identify and excludeparticipants who may be contributing to a backdoor attack. The proposed method is evaluated on various datasets, and the results demonstrate its effectiveness in defending against backdoor attacks in federated learning. The authors suggest that theirproposed method can be used in various settings, includinghealthcare, finance, and Internet of Things (IoT) applications, where federated learning is commonly used. [4]

The paper "A Study on Various Cyber Attacks and A Proposed Intelligent System For Monitoring Such Attacks" presented at the 2018 3rd International Conference on Inventive Computation Technologies (ICICT) by Atul S Choudhary, Pankaj P Choudhary, and Shrikant Salve discusses various types of cyber attacks and proposes an intelligent system to monitor and detect them. The authors provide an overview of different cyber-attacks, such as Denial of Service (DoS), Distributed Denial of Service (DDoS), Man-in-the-Middle (MitM), and Phishing attacks, and their potential impacts on businesses and individuals. They also review different security mechanisms used to prevent these attacks, including firewalls, intrusion detection and prevention systems, and antivirus software.
The proposed intelligent system is based on machine learning algorithms, including decision trees and random forests, to detectand classify cyber-attacks in real-time. The system is designed toanalyze network traffic data, identify patterns and anomalies, and generate alerts when an attack is detected. The authors validate the effectiveness of their proposed system by conducting experiments on a simulated network environment, and the resultsshow that the system is able to accurately detect and classify various types of cyber-attacks. The authors suggest that their proposed system can be useful for organizations to improve theircybersecurity posture and mitigate the risk of cyber-attacks.[5]

The paper "Preventing Data Poisoning Attacks by Using Generative Models" presented at the 2019 1st International Informatics and Software Engineering Conference (UBMYK) by Merve Aladag, Ferhat Ozgur Catak, and Ensar Gul proposes a method to prevent data poisoning attacks using generative models. The authors discuss data poisoning attacks in machine learning, where an attacker introduces malicious data into the training dataset to manipulate the model's output. The authors propose using generative models to detect and remove these malicious data points from the training dataset. The proposed method involves training a generative model on the original dataset, identifying the malicious data points using clustering techniques, and generating new data points to replace the identified malicious data points. The authors evaluate the proposed method on a simulated dataset and demonstrate that it can effectively prevent data poisoning attacks. They suggest thattheir proposed method can be used in real-world machine learning applications to improve the security and robustness of the models against malicious attacks.[6]

The paper "Towards Security Threats of Deep Learning Systems:A Survey" published in the IEEE Transactions on Software Engineering by Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, and Jinwen He provides a comprehensive survey of securitythreats related to deep learning systems. The authors provide anoverview of the basics of deep learning and its growing importance in various domains. They then discuss various security threats to deep learning systems, including adversarial attacks, data poisoning attacks, model stealing attacks, and backdoor attacks. The authors analyze the potential impact of these attacks on deep learning systems and highlight their challenges and limitations.
The paper also covers various countermeasures that have been proposed to mitigate these security threats, including adversarial training, input sanitization, and model watermarking. The authors provide an assessment of the effectiveness of thesecountermeasures and identify areas for further research.
Overall, the paper provides a comprehensive overview of the security threats to deep learning systems and highlights the needfor further research and development of robust and secure deep learning algorithms and architectures. The authors suggest that the survey can be useful for researchers, practitioners, and policymakers in understanding the security challenges and potential solutions related to deep learning systems. [7]
The paper "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey" by Naveed Akhtar and Ajmal Mian provides a survey of the threat of adversarial attacks on deep learning in computer vision. The authors provide an overview of deep learning in computer vision, including its applications and limitations. They then discuss adversarial attacks, which are designed to exploit the vulnerabilities of deep learning systemsby adding small perturbations to input images. The authorsexplain how adversarial attacks can lead to misclassification and compromise the security and privacy of deep learning systems. The paper covers various types of adversarial attacks, including gradient-based attacks, optimization-based attacks, and transferability attacks. The authors also describe various defense mechanisms that have been proposed to mitigate these attacks, such as adversarial training, defensive distillation, and gradient masking.
The authors provide a critical analysis of the effectiveness of these defense mechanisms and highlight their limitations and challenges. They suggest that the development of robust and secure deep learning algorithms and architectures is crucial to address the threat of adversarial attacks in computer vision.
Overall, the paper provides a comprehensive survey of the threatof adversarial attacks on deep learning in computer vision and highlights the need for further research and development of effective defense mechanisms. The authors suggest that the

survey can be useful for researchers, practitioners, and policymakersin understanding the challenges and potential solutions related to the security of deep learning systems in computer vision.[8]

The paper "Data poisoning attacks on multi-task relationship learning" by M. Zhao, B. An, Y. Yu, S. Liu, and S. J. Pan, presented at the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), between differententities across multiple tasks, which is becoming increasingly important in various domains.

The paper describes how data poisoning attacks can be carried out in this setting by an attacker introducing malicious data points into the training dataset to manipulate the learned relationships. The authors propose a poisoning attack model and present an optimization algorithm to generate the malicious datapoints. They also propose a defense mechanism based on outlierdetection and removal to mitigate the impact of data poisoning attacks.

The authors evaluate the effectiveness of their proposed attack and defense mechanisms on real-world datasets and demonstratethat the attack can significantly degrade the performance of multi-task relationship learning models. They also show that their defense mechanism can effectively detect and remove the malicious data points and improve the robustness of the models against data poisoning attacks.

Overall, the paper highlights the vulnerability of multi-task relationship learning to data poisoning attacks and proposes a novel attack and defense mechanism to address this issue. The authors suggest that their findings can be useful for developing more secure and robust multi-task learning algorithms in the future.[9]

The paper "Machine learning and its applications: A review" by Sheena Angra and Sachin Ahuja provides a comprehensive review of machine learning and its applications. The authors begin by introducing the concept of machine learning and its various types, such as supervised, unsupervised, and reinforcement learning. They also discuss the main challenges associated with machine learning, such as overfitting, underfitting, and bias.

The paper then presents a detailed review of the applications of machine learning in various fields, including healthcare, finance, transportation, and security. The authors describe how machine learning algorithms are being used to solve complex problems and improve decision-making processes in these domains. Theyprovide specific examples of applications, such as predicting disease outbreaks, detecting fraud, optimizing traffic flow, and identifying security threats.

The authors also discuss the future potential of machine learning, including the development of more advanced algorithms and the integration of advancement of machine learning will lead to significant improvements in various fields and have a profound impact on society.

Overall, the paper provides a comprehensive review of machine learning and its applications, highlighting its potential to solve complex problems and improve decision-making processes in various domains. The authors suggest that the review can be useful for researchers, practitioners, and policymakers interested in understanding the current state and future potential of machine learning.[10]

## III. CONCLUSION

Machine learning security is a critical and emerging research area that aims to develop algorithms, models, and systems that can defend against various attacks and threats to machine learning systems.

Adversarial attacks, such as data poisoning, model evasion, andmembership inference, pose significant challenges to machine learning security, and various countermeasures have been proposed to address these attacks, such as robust learning rate, generative models, and adversarial training.

Federated machine learning is a promising approach for collaborative and privacy-preserving machine learning, and it can help address various challenges in traditional centralized machine learning, such as data privacy, data ownership, and regulatory compliance.

The challenges and opportunities of machine learning security require a multi-disciplinary approach that integrates techniques from machine learning, cryptography, system security, and human-computer interaction.

Overall, the papers demonstrate the importance and urgency of machine learning security and provide valuable insights and solutions for developing secure and trustworthy machine learning systems.

These attacks can have serious consequences for the integrity and accuracy of the learning model, and they can compromise the privacy and security of the data owners who contribute their datato the federated learning system.

To mitigate data poisoning attacks on federated learning systems, various techniques have been proposed. One approach is to use outlier detection algorithms to identify and remove malicious data before it is used in the training process. Another approach isto use robust aggregation techniques that can detect and mitigatethe effects of poisoned data on the learning model.

Moreover, the use of differential privacy techniques can also help protect the privacy of the data owners and make it more difficult for attackers to extract sensitive information from the training data.

Overall, data poisoning attacks on federated learning systems represent a significant challenge for machine learning security, and more research is needed to develop effective countermeasures that can mitigate these attacks and ensure the privacy, security, and integrity of federated machine learning systems.

## IV. ACKNOWLEDGEMENT

Thanks to *Prof. (Mr.) Sagar Shinde* for his valuable contributionin developing this article.

We would like to extend our sincere gratitude for his unwaveringleadership, ongoing monitoring, enthusiastic support, and wise counsel, advice and efficient monitoring throughout the whole project phase.

We owe a great deal of gratitude to *Dr. N. F. Shaikh*, the Head of Computer Engineering Department, for her timely encouragement and on-going direction throughout the project.

## V. REFERENCES

1) [1] Gan Sun, Member, IEEE, Yang Cong, Senior Member, IEEE, Jiahua Dong, Qiang Wang, Lingjuan Lyu, and Ji Liu, "Data Poisoning Attacks on Federated Machine Learning," in IEEE Internet of Things Journal, vol. 9, no. 5, pp. 4305-4316, May 2022, doi: 10.1109/JIOT.2022.3080566.

2) [2] Qiang Yang, Lixin Fan, Richard Tong, Angelica Lv,White Paper - IEEE Federated Machine Learning, 2021, IEEE FEDERATED MACHINE LEARNING WHITEPAPER Authored by Qiang Yang Lixin Fan Richard Tong Angelica Lv

3) [3] Mingfu Xue, Chengxiang Yuan, Heyi Wu, Yushu Zhang, Weiqiang Liu, ”Machine Learning Security:Threats, Countermeasures, and Evaluations”, 2020, IEEE Access ( Volume: 8)

4) [4] Mustafa Safa Ozdayi, Murat Kantarcioglu, Yulia R.Gel, ”Defending against Backdoors in FederatedLearning with Robust Learning Rate”, 2021, The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

5) [5] Atul S Choudhary, Pankaj P Choudhary, Shrikant Salve, ”A Study On Various Cyber Attacks And A Proposed Intelligent System For Monitoring Such Attacks”, 2020, 2018 3rd International Conference on Inventive Computation Technologies (ICICT)

6) [6] Merve Aladag, Ferhat Ozgur Catak, Ensar Gul,”Preventing Data Poisoning Attacks By Using Generative Models”, 2020, IEEE 2019 1st InternationalInformatics and Software Engineering Conference (UBMYK)

7) [7] Yingzhe He, Guozhu Meng, Kai Chen1, Xingbo Hu,Jinwen He, ”Towards Security Threats of DeepLearning Systems: A Survey”, 2020, IEEE 33 Transactions on Software Engineering ( Volume: 48, Issue: 5, 01 May 2022)

8) [8] Naveed Akhtar, Ajmal Mian, ”Threat of AdversarialAttacks on Deep Learning in Computer Vision: A Survey”, 2018, Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey

9) [9] M. Zhao, B. An, Y. Yu, S. Liu, and S. J. Pan, ”Datapoisoning attacks on multi-task relationship learning”, 2018, The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)

10) [10] Sheena Angra and Sachin Ahuja, ”Machine learning and its applications: A review”, 2017, 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)