DATA POISONING ATTACKS ON FEDERATED MACHINE LEARNING

[1]DR.AR.SIVAKUMARAN, [2]POLNENI ABHINAYA, [3]PENDYALA SWETHA, [4]POKALA MAHITHA

[1]Assistant Professor, Department of Information Technology, **MALLA REDDY ENGINEERING COLLEGE FOR WOMEN**, Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

[2, 3,4] Student, Department of Information Technology, **MALLA REDDY ENGINEERING COLLEGE FOR WOMEN**, Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

ABSTRACT

Federated machine learning which enables resource constrained node devices (e.g., mobile phones and IoT devices) to learn a shared model while keeping the training data local, can provide privacy, security and economic benefits by designing an effective communication protocol. However, the communication protocol amongst different nodes could be exploited by attackers to launch data poisoning attacks, which has been demonstrated as a big threat to most machine learning models. In this paper, we attempt to explore the vulnerability of federated machine learning. More specifically, we focus on attacking a federated multi-task learning framework, which is a federated learning framework via adopting a general multi-task learning framework to handle statistical challenges. We formulate the problem of computing optimal poisoning attacks on federated multi-task learning as a bilevel program that is adaptive to arbitrary choice of target nodes and source attacking nodes. Then we propose a novel systems-aware optimization method, ATTack on Federated Learning (AT2FL), which is efficiency to derive the implicit gradients for poisoned data, and further compute optimal attack strategies in the federated machine learning. Our work is an earlier study that considers issues of data poisoning attack for federated learning. To the end, experimental results on real-world datasets show that federated multi-task learning model is very sensitive to poisoning attacks, when the attackers either directly poison the target nodes or indirectly poison the related nodes by exploiting the communication protocol.

I.INTRODUCTION

WITH the fast development of industry 4.0 and the widespread popularity of industrial Internet of Things (IIOT) applications makes applications such as smart transportation and smart healthcare thrive and also makes the data generated by the industrial devices exponentially grow. Such as autonomous driving technology [1], it needs to train all data generated by sensor and camera devices to build a stable joint model to identify

road conditions. And the distributed IIOT devices can generate a large amount of data in a short time [2]. In order to take into account the efficiency of processing big data and protect the privacy of clients. A novel machine learning paradigm named federated learning (FL) [3] was proposed, which is a new solution based on distributed training to alleviate the performance bottleneck and privacy risk caused by centralized processing. Traditional machine learning methods [4] usually store and run these data centrally, which will generate considerable computational and communication overhead in involving millions of mobile devices or massive data. This makes it unacceptable for sensitive IIOT applications (e.g., autonomous driving, intelligent robots, smart medical) that require real-time data transmission [5]. In addition, relying on centralized storage will cause a huge risk of private leakage [6]. Generally, when FL performs the collaborative training process of multiple distributed participants (e.g., IIOT devices), the sensitive information and private data of each client are kept locally [7]. FL has demonstrated excellent performance in the distributed execution process, while ensuring the privacy of participants by performing independent local training and model updates, so as to implement collaborative calculating in a joint environment that includes malicious participants. This also makes FL attract much attention in many fields including smart healthcare [8] [9], smart feature prediction [10], and Internet of Things in smart homes [11] [12].

The IIOT represents a distributed network composed of intelligent and highly interconnected industrial devices, eac device can act as an FL participant to participate in training and updating [13]. FL improves the performance of the model for IIOT applications through continuous iterative training, and finally obtains a stable global model when the iteration reaches convergence. However, FL greatly exposes its weaknesses to malicious adversaries during the process of performing training [14]. Malicious adversaries can obtain the information of the global model in each round and upload malicious parameters or perform a small part of the beneficial contribution for collaborative training while avoiding anomaly detection as much as possible. For instance, malicious adversaries use contaminated data for training locally [15] [16], or tamper and prune local models for poisoning aggregation [17] [18].

The existing works [12] [19] have shown that controlling more malicious IIOT devices or using more direct poisoning attacks during the execution of FL is more destructive to the global model. Due to network, communication, power, and other issues in a heterogeneous federated environment, many IIOT devices are at risk of offline. Malicious participant will virtualize multiple malicious nodes in this unstable communication network.

With more significant damage to the construction of the shared global model, this byzantine fault tolerance [20] problem usually uses the technology of fusion sybil-based attacks. In addition, in the process of malicious participants performing poisoning attacks, they usually use mislabeled samples for training or upload the poisoned models to the central server for aggregation. Compared with the independent attacks by a single malicious participant, the collusion attacks by multiple malicious participants have a higher attack success rate and can better obscure their attack behavior. Meanwhile, due to the characteristics of data privacy protection, the central server cannot verify the local data of all participants, and the parameter transmission process of all participants is anonymous, which provides more possibilities for malicious participants to launch malicious attacks.

Therefore, in order to better focus on the implementation of poisoning attacks in the IIOT-FL system, in this work, we introduce an efficient sybil-based collusion attacks (SCA) scheme. We represent the malicious IIOT device as a malicious participant in our system. Precisely, first, in the FL computing environment that we set, all the participants can only control local data and cannot access the data of other participants. This enables them to better manipulate local data for poisoning training without being detected. The most commonly used data poisoning methods are backdoor poisoning attacks [15] and label flipping attacks [16]. This paper uses label flipping attacks to conduct poisoning training on the massive data generated by IIOT devices, aiming to make the global model misclassify the selected attack class samples. However, the attack effect achieved by such an attack is insufficient. Second, we use the cloning properties of sybil that all sybil nodes virtualized by malicious participants will perform the same malicious operations during the training process and have equal attack influence. We consider making the malicious model has a higher probability to be aggregated during FL aggregation. Finally, we collude with all malicious participants to launch the collusion attacks, aiming to replace the global model using the poisoning model. Meanwhile, such collusion attacks can better obscure their attack behavior. We utilize Fashion MNIST and CIFAR-10 datasets to represent the data generated by IIOT devices and conduct experiments. In summary, our contributions to this work are mainly in four-fold as below.

• We explore sybil-based collusion attacks of IIOT data poisoning for the IIOT-FL application, and implement poisoning training and model collusion attacks in this IIOT-FL system.

• We make minimal malicious assumptions for malicious adversaries and integrate the label flipping poisoning attacks to make the global model misclassify the selected

attack class samples while maintaining the main task accuracy of other non-attack classes.

• We further propose an efficient sybil-based collusion attacks (SCA) method, which aims to make the poisoning collusion models to be aggregated with greater probability during aggregation, and successfully obscure their attack behavior.

• We utilize F-MNIST and CIFAR-10 datasets to represent the data generated by IIOT devices. Exhaustive experimental analysis demonstrates that our SCA has superior performance than the state-of-the-art.

The remainder of this paper is organized below. Section II primarily describes the background knowledge and corresponding comments of the relevant literature. Section III mainly focuses on introducing the FL problem formulation and threat model. Section IV discusses the proposed SCA based on the IIoT-FL computing environment. Section V analysis the performance results of our proposed SCA in multiple attack scenarios. Section VI summarizes the full paper.

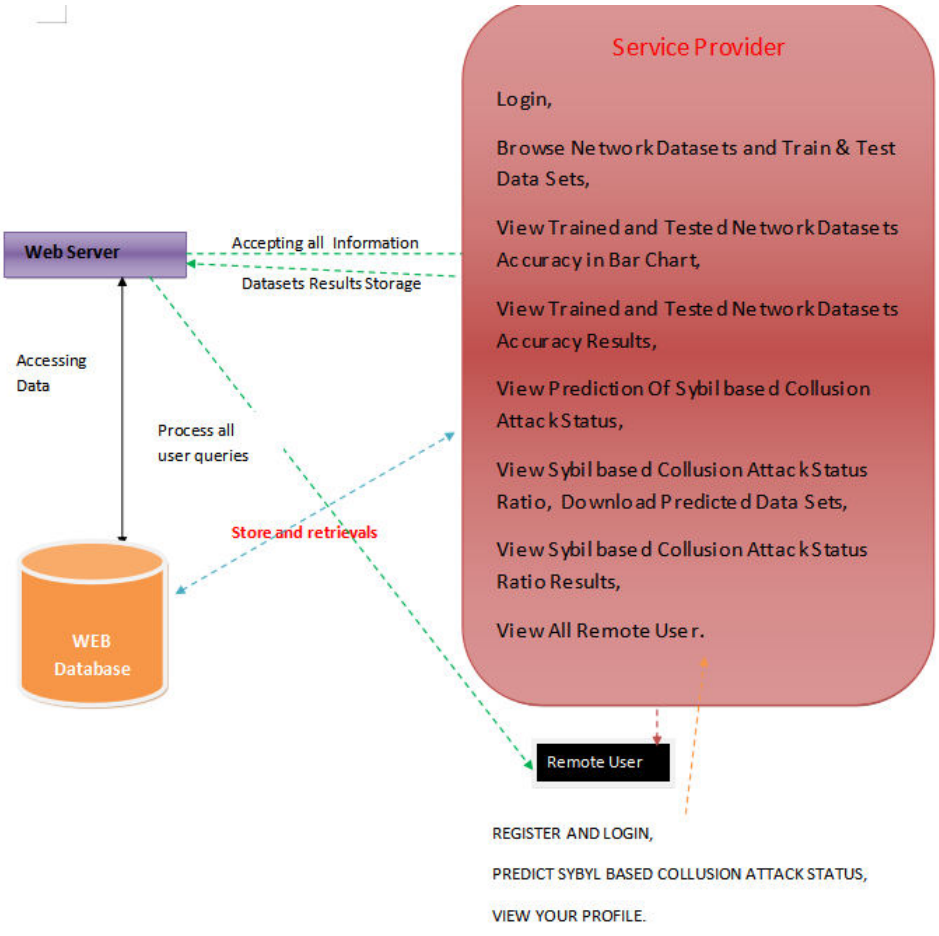

Fig1 :System architecture

II.EXISTING SYSTEM

For the data poisoning attacks, it has become an urgent research field in the adversarial machine learning, in which the target is against machine learning algorithms [4], [10]. The earlier attempt that investigates the poisoning attacks on support vector machines (SVM) [5], where the adopted attack uses a gradient ascent strategy in which the gradient is obtained based on properties of the SVM's optimal solution.

Furthermore, poisoning attack is investigated on many machine learning models, including autoregressive models [1], matrix factorization based collaborative filtering [13] and neural networks for graph data [22]. In addition to single task learning models, perhaps [21] is the most relevant work to ours in the context of data poisoning attacks, which provides the first study on one much challenging problem, i.e., the vulnerability of multi-task learning. However, the motivations for [21] and our work are significantly different as follows:

_ The data sample in [21] are put together, which is different from the scenario in federated machine learning, i.e., machine learning models are built based on datasets that are distributed across multiple nodes/devices while preventing data leakage.

_ The proposed algorithm in [21] is based on optimization method of current multi-task learning methods, which is not suited to handle the systems challenges in federated learning, including high communication cost, etc. Handling these challenges in the setting of data poisoning attacks is a key component of this work. For the federated machine learning, its main purpose is to update classifier fast for modern massive datasets, and the training data it can handle are with the following properties [12]: 1) Non-IID: data on each node/device may be drawn from a different distribution; 2) Unbalanced: the number of training samples for different nodes/devices may vary by orders of magnitude. Based on the distribution characteristics of the data, federated learning [18] can be categorized into: 1) horizontal (sample-based) federated learning, i.e., datasets share the same feature space but different in samples.

The representative work is a multi-task style federated learning system [15], which is proposed to allow multiple nodes to complete separate tasks while preserving security and sharing knowledge; 2) vertical (feature-based) federated learning, i.e., two datasets share the same sample ID space but differ in feature space. Several privacy-preserving machine learning methods have been presented for vertically partitioned data, e.g., secure linear regression [8], gradient descent methods [9]; 3) federated transfer learning, i.e., two datasets differ not only in samples but also in feature space.

Disadvantages

1) .The system doesn't have data poisoning attack model on federated
machine learning.

2). There is no technique called Data Integrity Check on data poisoning attacks.

III. PROPOSED SYSTEM

1. The system proposes a bilevel optimization framework to compute optimal poisoning attacks on federated machine learning. To our best knowledge, this is an earlier attempt to explore the vulnerability of federated machine learning from the perspective of data poisoning.

2. The proposed system derives an effective optimization method, i.e., Attack on Federated Learning (AT2FL), to solve the optimal attack problem, which can address systems challenges associated with federated machine learning.

3. The proposed system demonstrates the empirical performance of our optimal attack strategy, and our proposed AT2FL algorithm with several real-world datasets. The experiment results indicate that the communication protocol among multiple nodes opens a door for attacker to attack federated machine learning.

Advantages

- The system is more effective due to federated machine learning.
- The proposed system handles both direct attack and indirect attack.

IV.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 Network Datasets and Train & Test Data Sets, View Trained and Tested Network Datasets Accuracy in Bar Chart, View Trained and Tested Network Datasets Accuracy Results, View Prediction Of Sybil based Collusion Attack Status, View Sybil based Collusion Attack Status Ratio,Download Predicted Data Sets , View Sybil based Collusion Attack Status Ratio Results, 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 sybyl based collusion attack status, view your profile.

V.CONCLUSION

This paper analyzed the security vulnerabilities of joint training in the IIOT-FL system, then proposed a sybil-based collusion attacks (SCA) approach for the vulnerabilities. Meanwhile, we also gave further details on the execution of related algorithms, model architecture, and analysis of the effectiveness of the experiment. In this work, malicious participants in our federated system can virtualize multiple Sybil nodes and perform malicious collusion attacks. The purpose is to make the local poisoning model be aggregated with a greater possibility. They aim to make the samples of the selected attack class be misclassified, while other non-attack classes maintain similar accuracy as before. Compared with the state of- the-art, our SCA can achieve a more substantial attack effect under the condition of fewer malicious participants performing collusion, and can successfully obscure their attack behavior. Extensive experimental results show that our SCA has a more robust attack performance on several evaluation metrics.

VI.REFERENCES

[1] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, D. Niyato and H. V. Poor, “Federated learning for industrial internet of things in future industries,” IEEE Wireless communications magazine, 2021.

[2] P. Zhang, C. Wang, C. Jiang, and Z. Han. ”Deep reinforcement learning assisted federated learning algorithm for data management of IIoT,” IEEE Transactions on Industrial Informatics (TII), 2021.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273-1282.

[4] Y. Roh, G. Heo, and S. E. Whang, “A survey on data collection for machine learning: A big data - AI integration perspective,” IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 33, no. 4, pp. 1328-1347, 2021.

[5] B. Jia, X. Zhang, J. Liu, Y. Zhang, K. Huang, and Y. Liang, “Blockchainenabled federated learning data protection aggregation scheme with differential privacy and homomorphic encryption in IIoT,” IEEE Transactions on Industrial Informatics (TII), 2021.

[6] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, “A survey on security and privacy of federated learning,” ELSEVIER Future Generation Computer Systems (FGCS), vol. 115, pp. 619-640, 2021.

[7] T. Li, A. K. Sahu, A. Talwalker, and V. Smith, “Federated learning: Challenges, methods, and future directions,” IEEE Signal

Processing Magazine, vol. 37, no. 3, pp. 50-60, 2020.

[8] W. S. Zhang, T. Zhou, Q. H. Lu, X. Wang, C. S. Zhu, H. Y. Sun, Z. P. Wang, S. K. Lo, and F. Y. Wang, "Dynamic fusion-based federated learning for COVID-19 detection," IEEE Internet of Things Journal (IoTJ), 2021.

[9] M. Parimala, M. S. Swarna, P. V. Quoc, D. Kapal, M. Praveen, T. Gadekallu, and T. H. Thien, "Fusion of federated learning and industrial internet of things: A survey," arXiv preprint arXiv:2101.00798, 2021.

[10] M. X. Duan, K. L. Li, A. J. Ouyang, K. N. Win, K. Q. Li and Q. Tian, "EGroupNet: A feature-enhanced network for age estimation with novel age group schemes," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), vol. 16, no. 2, 2020.

[11] U. M. Aivodji, S. Gambs, and A. Martin, "IOTFLA : A secured and privacy-preserving smart home architecture implementing federated learning," in Proceedings of the 2019 IEEE Security and Privacy Workshops (SPW), IEEE, 2019, pp. 175-180.

[12] Y. Song, T. Liu, T. Wei, X. Wang, Z. Tao, and M. Chen, "FDA3 : Federated defense against adversarial attacks for cloud-based IIoT applications," IEEE Transactions on Industrial Informatics (TII), 2020.

[13] M. Rehman,and A. Dirir. "TrustFed: A framework for fair and trustworthy cross-device federated learning in IIoT," IEEE Transactions on Industrial Informatics (TII), 2021.

[14] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 10, no. 2, pp. 1-19, 2019.

[15] E. Bagdasaryan, A. Veit, Y. Q. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR, 2020, pp. 2938-2948.

[16] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data poisoning attacks against federated learning systems," in Proceedings of the European Symposium on Research in Computer Security (ESORICS), Springer, 2020, pp. 480-501.

[17] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in 2019 IEEE Symposium on Security and Privacy (SP), IEEE, 2019.

[18] Y. Jiang, S. Q. Wang, V. Valls, B. J. Ko, W. H. Lee, K. K. Leung, and L. Tassiulas, "Model pruning enables efficient federated learning on edge devices," arXiv preprint arXiv:1909.12326, 2019.

[19] Y. Qu, S. Pokhrel, S. Gary, L. Gao, and Y. Xiang. "A blockchained federated learning framework for cognitive computing in industry 4.0 networks," IEEE Transactions on Industrial

Informatics (TII), vol. 17, no. 4, pp. 2964-2973, 2020.

[20] M. H. Fang, X. Y. Cao, J. Y. Jia, and N. Gong, "Local model poisoning attacks to byzantine-robust federated learning," in Proceedings of the 29th USENIX Security Symposium (USENIX Security), USENIX, 2020.

[21] Y. Liu, R. H. Zhao, J. W. Kang, A. Yassine, D. Niyato, and J. L. Peng, "Towards communication-efficient and attack-resistant federated edge learning for industrial internet of things," arXiv preprint arXiv:2012.04436, 2020.

[22] C. L. Xie, K. L. Huang, P. Y. Chen, and B. Li, "DBA: Distributed backdoor attacks against federated learning," in Proceedings of the International Conference on Learning Representations (ICLR), 2020.

[23] Z. T. Sun, P. Kairouz, and H. B. McMahan, "Can you really backdoor federated learning?" arXiv preprint arXiv:1911.07963, 2019.

[24] C. Fung, C. M. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," arXiv preprint arXiv:1808.04866, 2018.

[25] Y. P. Jiang, Y. Li, Y. P. Zhou, and X. Zheng, "Mitigating sybil attacks on differential privacy based federated learning," arXiv preprint arXiv:2010.10572, 2020.

Dr.AR.SIVAKUMARAN, has been working as a Associate Professor in Department of Information Technology, Malla Reddy Engineering College for Women, Secunderabad, Telangana, India, since 2019. He received his Doctorate Degree from Anna University, Chennai, Tamil Nadu. He received M.Tech(CSE) Degree from Motilal Nehru National Institute of Technology (NIT), Allahabad, Uttar Pradesh. He has a Good Academic and Research Experience of more than 23 years. His current area of research includes Web Mining, AI, NLP, Deep Learning and Machine Learning. He has published many papers in Scopus, UGC Care List and reputed International Journals. He has five patent publications