A systematic literature review on untargeted model poisoning attacks and defense mechanisms in federated learning

Tabassum Anika*

1Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh.

*Corresponding Author email: bsse0925@iit.du.ac.bd

Submitted: 23 May 2022 Revised: 18 December 2022 Accepted: 27 December 2022

ABSTRACT

In the past few years, Federated Learning has offered an optimistic solution to the privacy concerns of users who use different Machine Learning Models. But there are risks of exploiting the models by inside and outside adversaries. To preserve the data privacy and the model integrity, the Federated Learning model needs to be protected against the attackers. For this, the untargeted model poisoning attack where the model quality is compromised, needs to be detected early. This study focuses on finding various attack, detection and defense mechanisms against untargeted model poisoning attacks. Total 245 studies were found after searching Google Scholar, ScienceDirect and Scopus. After passing the selection criteria, only 15 studies were included in this systematic literature review. We have highlighted the attacks and defense mechanisms found in the related studies. Additionally, further study avenues in the area were recommended.

Keywords: Federated Learning; Untargeted Model Poisoning Attack; Systematic Literature Review

1. INTRODUCTION

In 2016, Konečný and McMahan (2016), introduced Federated Learning (FL) and from then it has gained high attention because of its privacy preserving nature. In Federated Learning, clients do not have to share their private data for training yet they can work together to develop a shared machine learning model. As they train the common model by collaborating on the distributed nodes (Bagdasaryan & Veit, 2020, pp 2938-2948), it solves issues such as security, access rights and data privacy. Clients communicate their local gradients to the parameter server, which collects them and changes the global model for download by local clients.

Though FL has emerged as a viable solution for many situations involving mutually untrusting clients, poisoning attacks are very common in Federated Learning mechanisms. The poisoning attack is a form of attack that makes use of the machine learning model during training. Some poisoning attacks target the model availability, while other attacks such as backdoor attacks target the model integrity. The attack that targets the model integrity is also known as untargeted model poisoning, which attacks to minimize the global accuracy and the global model becomes useless gradually. A backdoor attack, also called targeted attack, involves clients training local models with malicious training data.
So eventually the integrity of the model is compromised. Such an attack can modify a model to give a wrong prediction for a set of data (image, words etc.)

The goal of untargeted poisoning attacks is to generate poisoned local model updates to inject into the system, rendering the parameters that the global model learns virtually unusable. U ntargeted model poisoning attacks are more impactful than other traditional attacks.

Under the right assumptions, robust aggregators can effectively mitigate untargeted attacks (Abadi & Barham, 2016; Blanchard & Mhamdi, 2017; Chen & Su, 2017, pp1-25; Mhamdi & Guerraoui, 2018). For more challenging scenarios regarding malicious clients, Pan et al. (Pan et al., 2020, pp 1641-1658). A design of Justinian’s GAAvernor (GAA), a Gradient Aggregation Agent proposed a defense to restore the robustness of distributed learning.

Mallah et al. (2022), proposed attested FL, a defense mechanism to protect against local model poisoning attack and evaluate the effect of defense mechanism for more challenging scenarios.

No systematic literature review on robust defense mechanisms against untargeted model poisoning attack in Federated Learning has not been proposed yet. This paper was aimed to discuss various attacks on Federated Learning, defense mechanisms working against untargeted model poisoning attacks in Federated Learning. The paper constructs as follows, in Section 1, the study is introduced, in section 2, the methodology of the research, in section 3, the results and discussion. Lastly in section 4, the conclusions and recommendations are discussed.

2. METHODOLOGY

For conducting a systematic literature review, first of all a proper planning is needed. The reason behind the research work needs to be defined properly. Then, deciding the databases to use, looking for the exact keywords and their perfect combination used as search strings. Then after searching the databases with the keywords, some papers are brought. Then these results are screened through the selection criteria. Finally, the included papers are read to fulfill the goal of the research work. The results are analyzed and described in figurative language.

2.1. RESEARCH QUESTIONS

The questions this paper is addressing are as follows:

RQ1: What are the different types of untargeted model poisoning attacks in Federated Learning?

RQ2: What are the existing defense mechanisms to work against untargeted model poisoning attacks in Federated Learning?
2.2. Search Strategy

A search technique was devised to locate relevant articles for this systematic search. This search method was primarily designed for the following databases: ScienceDirect, Scopus, Google Scholar. Various search terms were used for getting the most relevant articles. Such as, "Federated Learning" OR "Federated Learning" AND Attack OR "Federated Learning" AND "Poisoning Attack" OR "Federated Learning" AND "Model Poisoning Attack". The searches covered the articles written in years 2019 to 2022. Total 245 results were found till date 11th May, 2022.

Table 1, 2 and 3 shows the searching keywords and the number of results found in each database.

Table 1. Keyword search in ScienceDirect

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Filters</th>
<th>Number of results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federated Learning</td>
<td>None</td>
<td>4,435</td>
</tr>
<tr>
<td>&quot;Federated Learning&quot; AND Attack</td>
<td>None</td>
<td>461</td>
</tr>
<tr>
<td>&quot;Federated Learning&quot; AND&quot; Poisoning Attack&quot;</td>
<td>None</td>
<td>68</td>
</tr>
<tr>
<td>&quot;Federated Learning&quot; AND&quot; Model Poisoning Attack&quot;</td>
<td>None</td>
<td>17</td>
</tr>
<tr>
<td>&quot;Federated Learning&quot; AND&quot; Model Poisoning Attack&quot;</td>
<td>from 2019 – to 2022</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 2. Keyword search in Google Scholar

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Filters</th>
<th>Number of results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federated Learning</td>
<td>None</td>
<td>141,000</td>
</tr>
<tr>
<td>&quot;Federated Learning&quot; AND Attack</td>
<td>None</td>
<td>8,300</td>
</tr>
<tr>
<td>&quot;Federated Learning&quot; AND&quot; Poisoning Attack&quot;</td>
<td>None</td>
<td>786</td>
</tr>
<tr>
<td>&quot;Federated Learning&quot; AND&quot; Model Poisoning Attack&quot;</td>
<td>from 2019 – to 2022</td>
<td>204</td>
</tr>
</tbody>
</table>

Table 3. Keyword search in Scopus

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Filters</th>
<th>Number of results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federated Learning</td>
<td>None</td>
<td>4,040</td>
</tr>
<tr>
<td>&quot;Federated Learning&quot; AND Attack</td>
<td>None</td>
<td>494</td>
</tr>
<tr>
<td>&quot;Federated Learning&quot; AND&quot; Poisoning Attack&quot;</td>
<td>None</td>
<td>104</td>
</tr>
<tr>
<td>&quot;Federated Learning&quot; AND&quot; Model Poisoning Attack&quot;</td>
<td>from 2019 – to 2022</td>
<td>26</td>
</tr>
</tbody>
</table>

2.3. Selection Criteria

The irrelevant research papers were deduced in this process. Based upon relevancy and some other parameters some inclusion criteria and exclusion criteria were set. And each criterion was considered to select the papers.

2.3.1. Inclusion Criteria:

- Published research papers or review papers that worked with the specific study area.
- Papers published from 2019 to 2022.
- Papers written in English language only.

2.3.2. Exclusion Criteria:

- Research papers outside of the study scope.
- Papers that are duplicate
2.4. STUDIES SELECTION

Initially three databases were searched with the search strings repeatedly to narrow down the search results. Total 245 results were found and were recorded in an Excel sheet. After searching for duplicates, 40 papers were removed by excel. Then with the remaining 205 papers, 3 were eliminated as they were book chapters, 1 was eliminated for not being written in English language. Then the remaining 201 papers were selected for screening in the next level. The papers were checked for relevancy to the study by the title and abstract. Finally, 31 papers were extracted. After reading the full-text of the papers, 11 papers remained. After doing the manual search using forward referencing, 4 papers were added to the selection list. These 15 papers were selected for the final stage for the study. Table 4 shows the results retrieved from the search space. Figure 1 shows the flow of selection using PRISMA diagram.

Table 4. Results of Search Space

<table>
<thead>
<tr>
<th>Database</th>
<th>Date Search</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>ScienceDirect</td>
<td>11/05/2022</td>
<td>15</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>11/05/2022</td>
<td>204</td>
</tr>
<tr>
<td>Scopus</td>
<td>11/05/2022</td>
<td>26</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>245</strong></td>
</tr>
</tbody>
</table>

**Fig. 1. Selection Process using PRISMA diagram**

2.5. DATA EXTRACTION AND REPORTING

After the selection process, the full-text of the 15 retrieved papers were read. The research questions mentioned above were studied. The publication year, Author(s) name, title and
abstract of the papers, the poisoning attacks in Federated Learning and the defense mechanisms described in the papers were extracted.

2.6. Results and Discussion

The findings of the study are examined in this section of the report. During the search, 245 publications were found in databases, and 11 of them were chosen for analysis based on the study’s criteria. Moreover, 4 external papers were chosen by manual search. Table 5 shows the names of the authors, the years of publication, the title of the research study, and the model poisoning attacks and defense mechanisms in Federated Learning discussed in the research publications. This table answers the Research Questions. The analysis includes document type, publication year, and source of research data. The publication year, databases are also analyzed here.

Table 5. List of selected papers

<table>
<thead>
<tr>
<th>Title</th>
<th>Author and Year</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Towards multi-party targeted model poisoning attacks against federated learning systems (Chen &amp; Tian, 2021)</td>
<td>Chen et al., (2021)</td>
<td>Various stealthy metrics are included to get around various detection methods on the central parameter server to detect untargeted model poisoning attack.</td>
</tr>
<tr>
<td>SecFedNIDS: Robust defense for poisoning attack against federated learning-based network intrusion detection system (Zhang &amp; Zhang, 2022, 154-169)</td>
<td>Zhang et al., (2022)</td>
<td>To detect attack it offers effective low-dimensional representations of the uploaded local model parameters using a gradient-based important model parameter selection technique. The proposed defense mechanism uses identifying poisoned models and preventing them from joining the global intrusion detection model, an online unsupervised poisoned model detection approach was developed.</td>
</tr>
<tr>
<td>Dynamic defense against byzantine poisoning attacks in federated learning (Rodríguez-Barroso &amp; Martínez-Cámara, 2022, pp 1-9)</td>
<td>Rodríguez et al., (2022)</td>
<td>DDaBA (Dynamic Defense Against Byzantine Attacks) is a dynamic aggregation operator that dynamically picks the clients to be aggregated and discards those that are considered hostile to defend against poisoning attacks.</td>
</tr>
<tr>
<td>Untargeted Poisoning Attack Detection in Federated Learning via Behavior Attestation (Mallah, Lopez, Marfo &amp; Farooq, 2022)</td>
<td>Mallah et al., (2022)</td>
<td>Detects untargeted attacks in FL by observing worker behavior over time using state persistent analysis. Proposes a three line defense mechanism stated below: attestedFL-1, which tracks the local model's time convergence with the global model. Then during the training of a node, attestedFL-2 was used to track the angular distance between successive local model updates. Finally attestedFL-3 is used to remove local model updates from workers whose performance on a quasi-validation dataset doesn't really improve as compared to their own previous performance.</td>
</tr>
<tr>
<td>Desmp: Differential privacy exploited stealthy model poisoning attacks in federated learning (Hossain, Islam, Badsha &amp; Shi, 2021, pp 167-174)</td>
<td>Hossain et al., (2021)</td>
<td>The attack occurs by manipulating FL models by interfering with the communication channel, sieving the model parameters, and then inserting fake noises into the parameters. The defense strategy is based on Reinforcement Learning which intelligently picks the differential privacy level for the clients' model update.</td>
</tr>
<tr>
<td>Title</td>
<td>Author and Year</td>
<td>Findings</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>---------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>TESSERACT: GRADIENT FLIP SCORE TO SECURE FEDERATED LEARNING</td>
<td>Sharma et al., (2021)</td>
<td>Certain patterns of flips of the sign of gradients across various parameters and over numerous clients should be common to detect bad clients that exploit federated learning. The defense mechanism decreases the weight of malicious clients when time approaches zero.</td>
</tr>
<tr>
<td>AGAINST MODEL POISONING ATTACKS (Sharma, Chen, Zhao, Qiu, Chaterji, &amp; Bagch, 2021)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPAF: Model Poisoning Attacks to Federated Learning based on Fake Clients (Cao &amp; Gong, 2022)</td>
<td>Cao et al., (2022)</td>
<td>Based on fake clients and requires no extra knowledge about the FL system beyond received global models during training. Depending on fake clients, the attack mechanism requires no extra information about the Federated Learning System</td>
</tr>
<tr>
<td>Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning (Shejwalkar &amp; Houmansadr)</td>
<td>Shejwalkar et al., (2021)</td>
<td>A complete collection of threat models for model poisoning attacks along two dimensions of adversary information: knowledge of the updates shared by benign clients and knowledge of the server’s AVR algorithm. The defense mechanism is robust AVR, which does dimensionality reduction by divide and conquer (DnC). It effectively detects malicious model updates also spectral analysis of input updates. SparseFed mitigates attackers by using gradient sparsification. It proposes a defense mechanism which works at training time.</td>
</tr>
<tr>
<td>Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Production Federated Learning (Shejwalkar &amp; Houmansadr, 2021)</td>
<td>Shejwalkar et al., (2021)</td>
<td>The attacking technique is a gradient ascent algorithm which fine tunes the global model. Then to bypass the robustness criterion of the target AVR, they adjust the L2-norm of the corresponding poisoned update.</td>
</tr>
<tr>
<td>PDGAN: A Novel Poisoning Defense Method in Federated Learning Using Generative Adversarial Network (Zhao, Chen, Zhang, Wu, Teng &amp; Yu, 2020, pp 595-602)</td>
<td>Zhao et al., (2020)</td>
<td>Poisoning defense generative adversarial network (PDGAN), is a detection mechanism is based on accuracy auditing, which can defend poisoning attacks in FL.</td>
</tr>
<tr>
<td>Certifiably robust federated learning against backdoor attacks (Xie, Huang, Chen, Y &amp; Li, 2021)</td>
<td>Xie et al., (2021)</td>
<td>Dynamic Defense Against Byzantine Attacks (DDaBA) works with a federated aggregation operator based on a Induced Ordered Weighted Averaging (IOWA). It can select which users need to be aggregated dynamically, and find out the adversarial users. Certifiably Robust Federated Learning (CRFL) proves robustness using Markov Kernel to analyze the training dynamics of the aggregated model.</td>
</tr>
</tbody>
</table>

---

122
2.7. **YEAR-BASED PUBLICATIONS**

Figure 2 reports the number of publications in this topic from 2019 to 2022. The number of publications published on the model poisoning attack in Federated Learning is increasing, according to the graph. In the last year, 2021, model poisoning attack and defense mechanisms in Federated Learning have been studied the most.

![Number of Publications](image)

**Fig. 2.** Selected paper over the year

2.8. **RESEARCH DATA SOURCE**

ScienceDirect, Google Scholar and Scopus were used as databases to collect the papers. From the figure 3 we can see that about 33% of the papers were retrieved from Google Scholar and ScienceDirect each. But Scopus did not give that much information. Only one paper, after the searching and selection criteria, was used from Scopus. 4 papers were searched and retrieved manually from forward referencing.

![Selected Paper from Different Databases](image)

**Fig. 3.** Selected paper from different databases
3. SOLUTIONS TO RESEARCH QUESTIONS

3.1. WHAT ARE THE DIFFERENT TYPES OF UNTARGETED MODEL POISONING ATTACKS IN FEDERATED LEARNING?

There are various mechanisms proposed for untargeted model poisoning attack in Federated Learning. Some use detection methods on local model parameters. Some use it on global parameters, certain patterns of flips of the sign of gradients across various parameters and over numerous clients. Attack can also occur by manipulating models through interference in the communication channel and then inserting fake noises into the parameters, gradient ascent algorithm, implanting each adversarial party with a local trigger design, and after assembling designs an attack.

3.2. WHAT ARE THE EXISTING DEFENSE MECHANISMS TO WORK AGAINST UNTARGETED MODEL POISONING ATTACKS IN FEDERATED LEARNING?

There are many detection policies against untargeted model poisoning attacks. Such as, online unsupervised poisoned model detection approach, dynamic aggregation operator that dynamically picks the aggregated client and discards hostile ones, by observing worker behavior over time, by doing spectral analysis of input updates.

The significant defense mechanisms are a three line defense mechanism which removes local model updates of workers whose performance do not improve as compared to their previous performance, a reinforcement learning which picks differential privacy level for the clients’ model update, by decreasing the weight of malicious clients when time approaches zero, by dimensionality reduction by divide and conquer. Training time defense mechanisms by mitigating attackers using gradient sparsification, reputation-based iterative filtering technique, dynamically selecting which users need to be aggregated, and increasing robustness using Markov Kernel also give good defenses.

4. CONCLUSION AND LIMITATIONS

This study was conducted to write a systematic literature review on different untargeted model poisoning attacks and defense mechanisms in Federated Learning. After the searching and selection, finally 15 papers were selected which were related to the attack and defenses of untargeted model poisoning attacks in Federated Learning. As Federated Learning is a new topic in modern research, the research area is still growing. But some challenges were faced because of this inadequate data.

There are some limitations. As this paper only focused on English research, some researches of other languages may have missed out. Three databases were used, but Scopus, being a very large database, gave very few related search results. Also, the databases update frequently and keeping track of the latest literature is not easy. As we used complex keywords to search, we got very limited results from the databases.
5. FUTURE WORK

The future work for this kind of systematic literature review can be broadening the search
throughout more databases to extract more related studies. In the present work, only
full-text articles were retrieved, by relaxing this, more relevant studies can be achieved. Also,
future works can remove the constraint of using the research papers only. Books and
review papers can also be of great use. Moreover, more intelligent techniques can be used
to do analysis on this type of literature review.

Author Contributions:

Being a single author, I, Anika Tabassum, did all the work of this research.

Funding:

This research received no external funding.

Institutional Review Board Statement:

The study did not involve any humans or animals.

Informed Consent Statement:

Not applicable

Acknowledgments:

Not Applicable

Conflicts of Interest:

The authors declare no conflict of interest.

Reference:

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M.,
Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan,

learning. In International Conference on Artificial Intelligence and Statistics (pp. 2938-2948). PMLR.


attack in federated learning. In 2019 IEEE 25th International Conference on Parallel and
Distributed Systems (ICPADS) (pp. 233-239). IEEE.

Cao, X., & Gong, N. Z. (2022). Mpaf: Model poisoning attacks to federated learning based on fake clients.
In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(pp. 3396-3404).


