Team V
Analysis of Secured Federated Learning

Aayush Jannumahanti, Avni Saxena, Shashwat Jain, Aksheetha Muthunooru

Spring 2023
CMSC 691 DATA PRIVACY
Professor Roberto Yus

Motivation

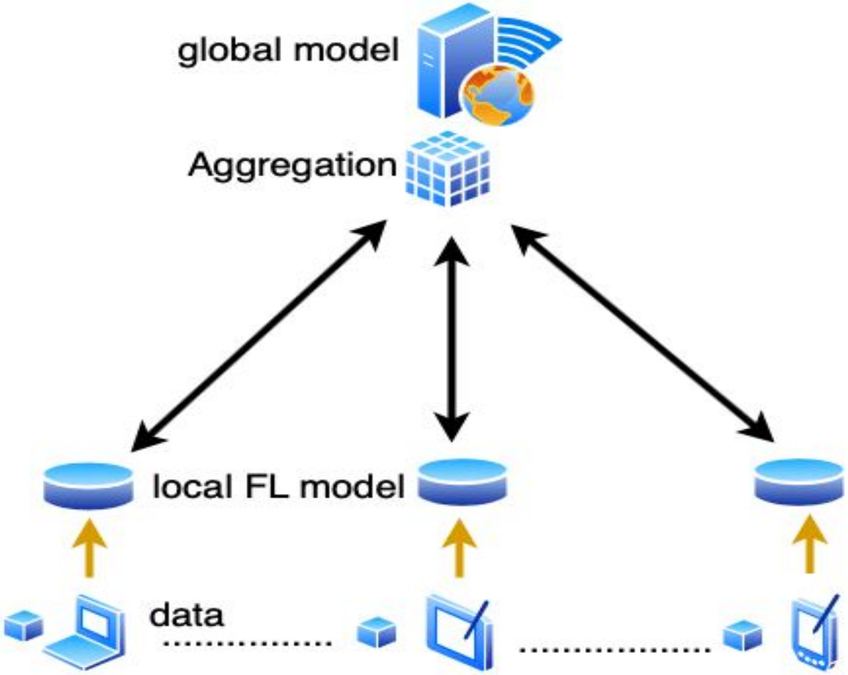
- Federated Learning is a technique used in various industries nowadays:
  - Healthcare: Federated Learning is being used in healthcare to share data while maintaining patient privacy, which can help improve patient outcomes.
  - Financial Services: Federated Learning is being used in financial institutions identify patterns and trends in customer behavior to improve their services and prevent fraud.
- While Federated Learning is a promising approach for training machine learning models on decentralized data sources without compromising privacy, there are still several challenges and limitations to consider.
- To dwell deep in this issue, our project focuses on study to check how protected FL and FL with DP is as a privacy technique.

PHASE-1

Federated Learning

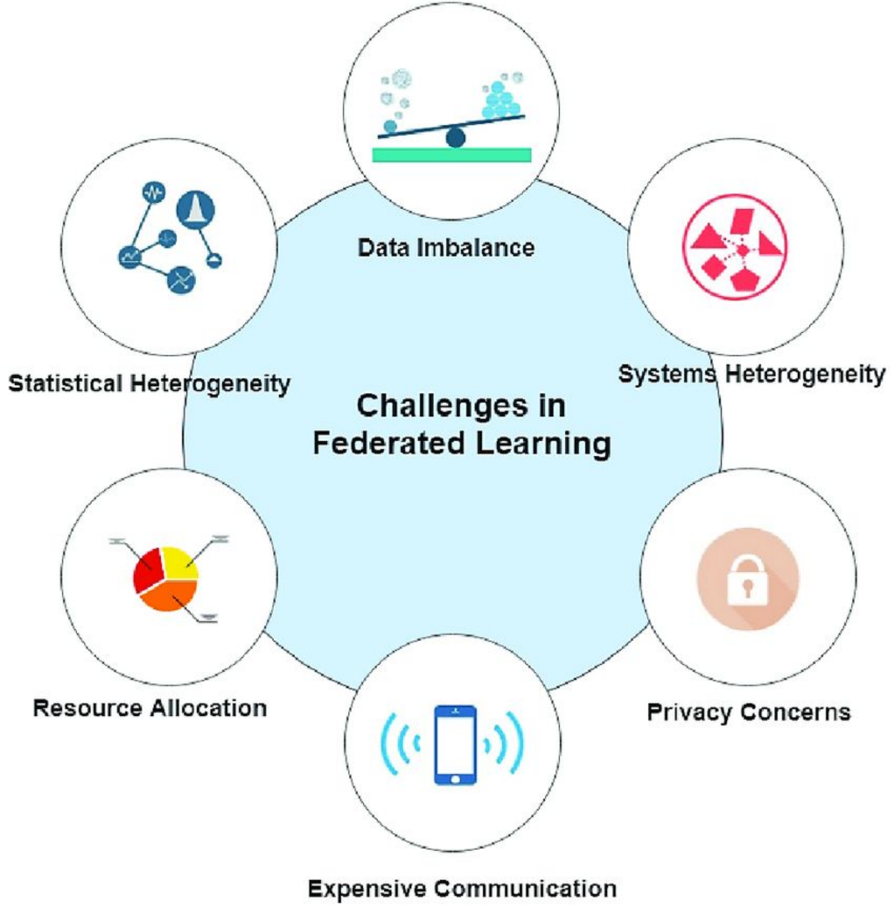
- Federated learning (FL) is a machine learning setting where many clients collaboratively train a model under the orchestration of a central server while keeping the training data decentralized without sharing raw data. It is motivated by the growing need for privacy and security in the era of big data.

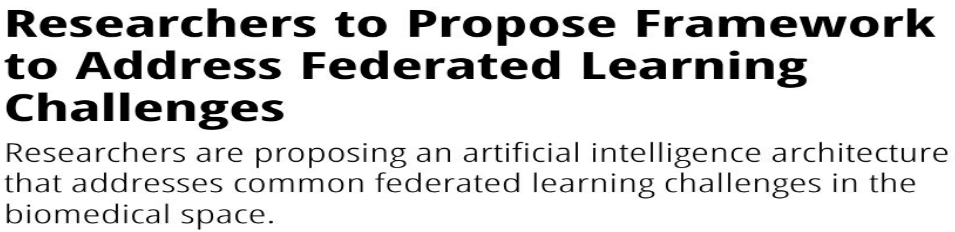
- In FL, multiple parties (client devices) train the model on their own local data. The local models are then sent to a central server, which aggregates them to form a global model. This process is repeated until the global model converges.



Federated learning is a machine learning architecture that enables training of models across multiple devices or servers without the need for data to be centralized. Here are the main components of a federated learning architecture:

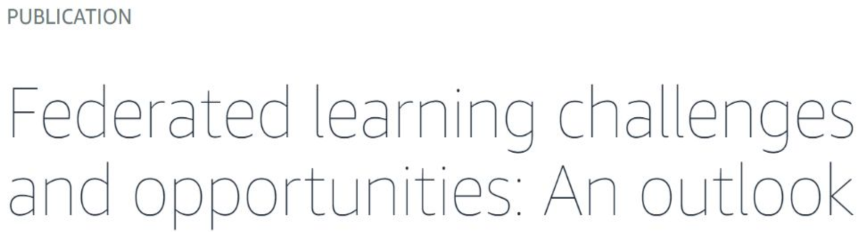
a. **Client devices:** These are the devices (such as smartphones, laptops, or IoT devices) that participate in the federated learning process. Each device has a local dataset that is used to train the model.
b. **Edge servers:** These are the servers that manage the federated learning process. They coordinate the communication between the client devices and aggregate the model updates from the clients.
c. **Central server:** This is the server that stores the global model and distributes it to the edge servers. The central server also manages the overall training process and sends updated models to the edge servers as needed.
d. **Model:** This is the machine learning model that is being trained in the federated learning process. The model is initially trained on the central server and then distributed to the edge servers for further training with local data.
e. **Communication protocol:** This is the protocol that enables communication between the client devices and edge servers. It ensures that the model updates are securely transmitted and that client privacy is protected.





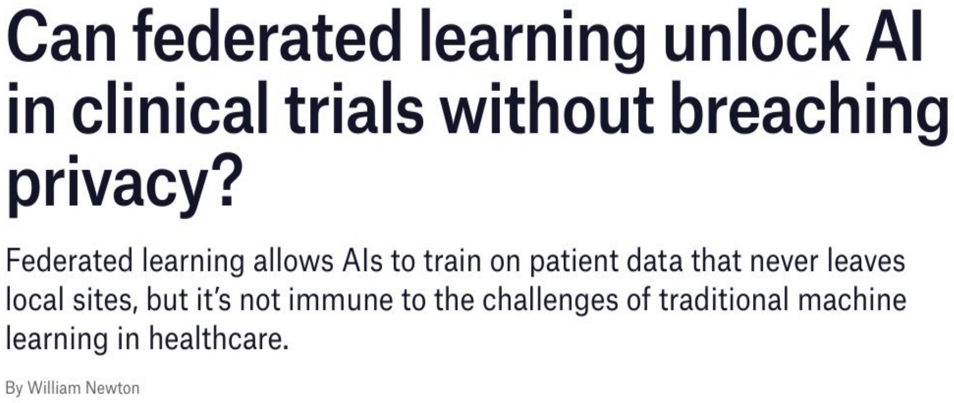
Researchers to Propose Framework to Address Federated Learning Challenges

Researchers are proposing an artificial intelligence architecture that addresses common federated learning challenges in the biomedical space.



PUBLICATION

Federated learning challenges and opportunities: An outlook



Can federated learning unlock AI in clinical trials without breaching privacy?

Federated learning allows AIs to train on patient data that never leaves local sites, but it's not immune to the challenges of traditional machine learning in healthcare.

By William Newton

PHASE-2

Privacy - Introducing FL with Differential privacy

- Federated learning with differential **privacy adds random noise** to gradients during local training on each device. The edge server aggregates the noise gradients and updates the global model.
- This preserves client data privacy while still allowing effective model training.
- This method is useful in applications like healthcare and finance where data privacy is critical.
- However, it can increase computational requirements and training time. Therefore, careful consideration of the specific application and data privacy requirements is necessary.

Implementation of Differential Privacy in Federated Learning

1. **Local Training:** Each device or client in the Federated Learning network trains a model based on its local data. This is the same as traditional Federated Learning.
2. **Noise Addition:** After training the local model, instead of sending the exact model updates to the server, each device adds some random noise to its model updates. This noise is generated according to a specific statistical distribution.
3. **Aggregation:** The server collects the noisy updates from all the devices. The server does not have access to the exact updates from any device, only the noisy versions.
4. **Global Model Update:** The server averages the noisy updates to create the global model update, then applies this update to the global model.
5. **Iteration:** Steps 1-4 are repeated for several rounds until the global model's performance is satisfactory.

UMBC

PHASE-3

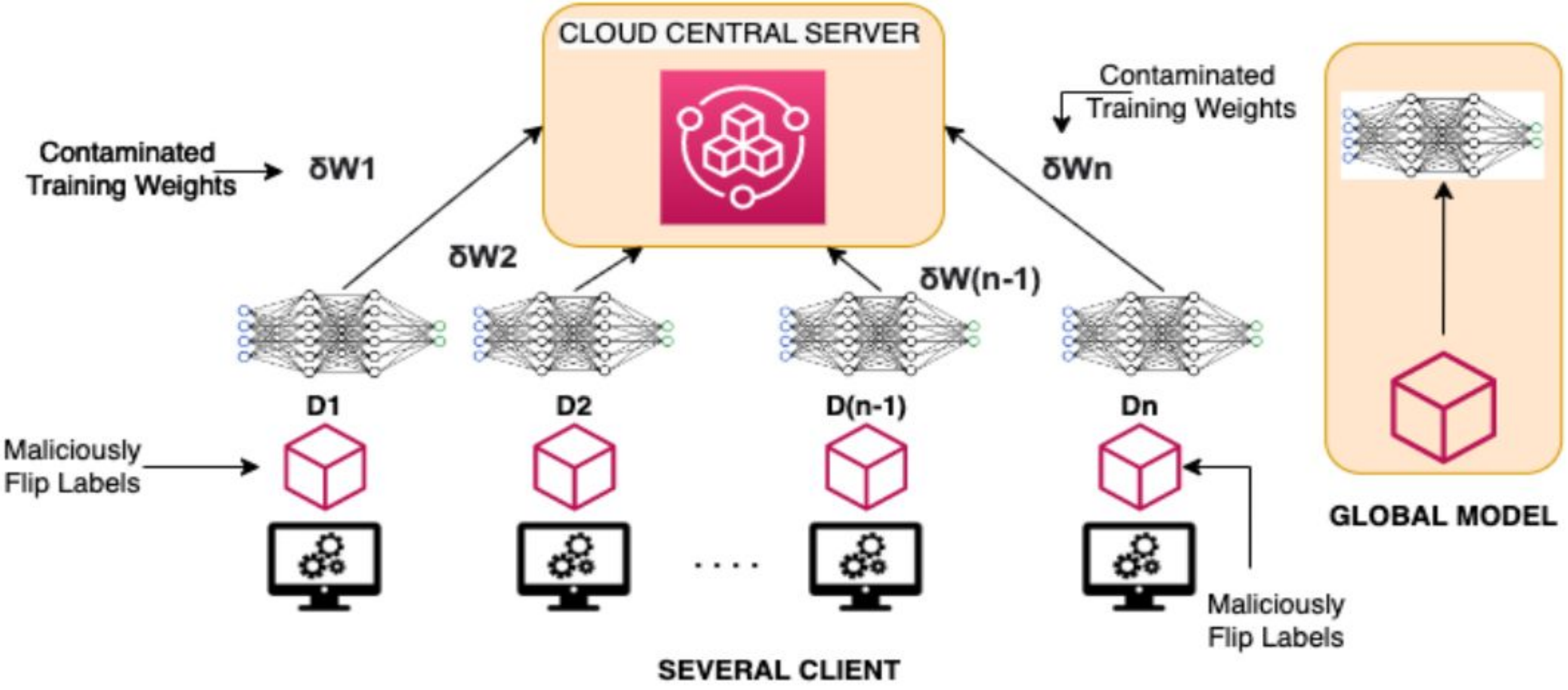
Data Privacy attacks

- In Federated Learning (FL), privacy is a critical concern since multiple parties collaborate to train a machine learning model without sharing raw data. Privacy attacks can compromise user data and undermine the integrity of the model.
- To mitigate privacy risks, it is essential to understand the different types of privacy attacks that can occur in FL. These attacks can target various aspects of the FL process, including the model, the data, and the communication channels.
- In this study we have explored the different types of privacy attacks on FL and FL with Differential privacy .
    - Data Poisoning (Label Flipping)
    - Back Door Attack

Label Flipping Attack

- A label flipping attack is a type of poisoning attack where an adversary intentionally mislabels their training data, hoping to degrade the performance of the global model. In the context of federated learning, this can be accomplished by a **malicious client flipping the labels of their local training data**.
- To implement a label flipping attack, you would need to modify the client data during preprocessing. Here's how you could do this:
- Identify which client(s) will be the adversary. You can choose this randomly, or based on some other criteria.
- During preprocessing, when the labels are being assigned, flip the labels for the adversary client(s). For a binary classification task, you can just invert the label (i.e., change 0s to 1s and vice versa). For a multi-class classification task like image classification, you can change each label to some other class.

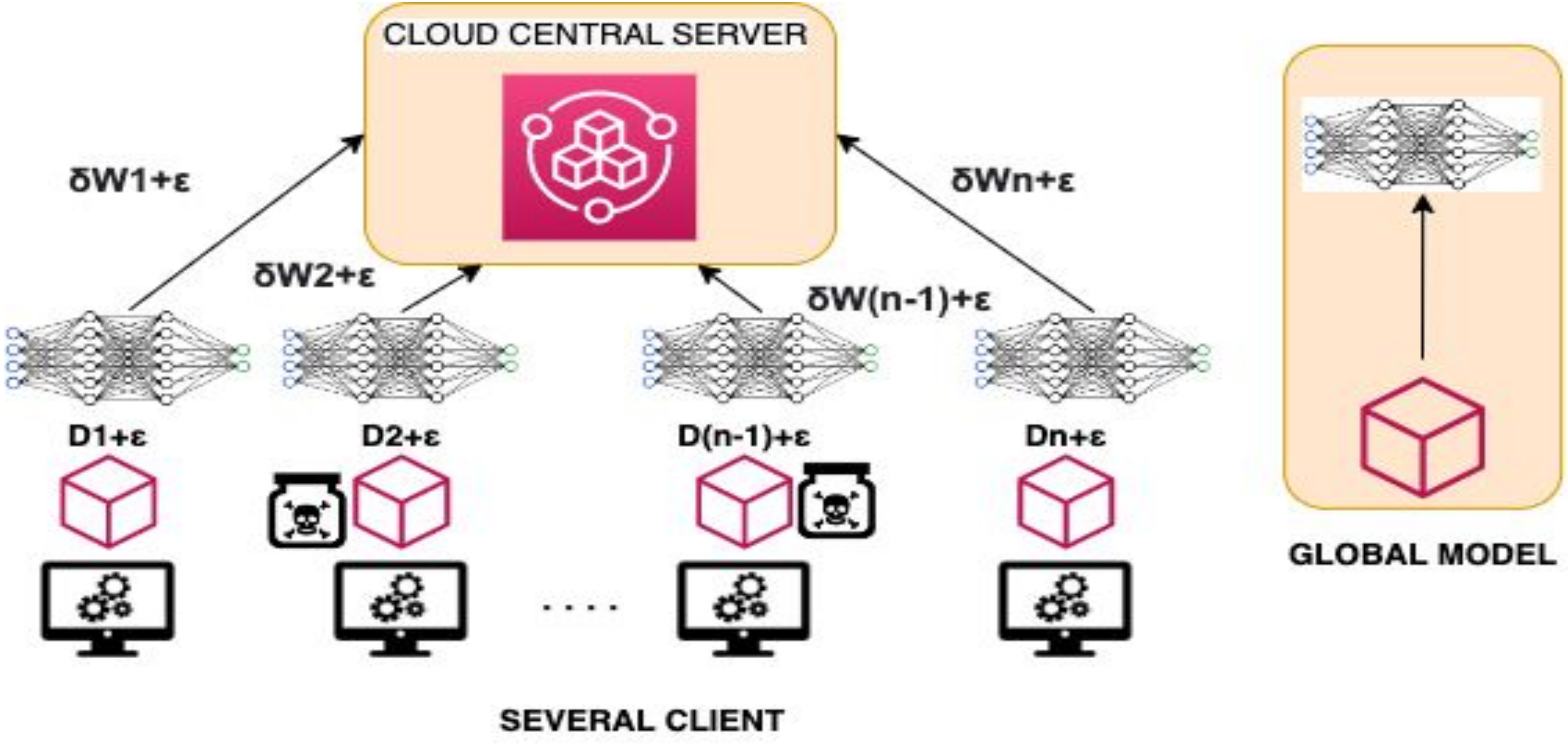
Federated Learning - Label Flipping Attack Architecture



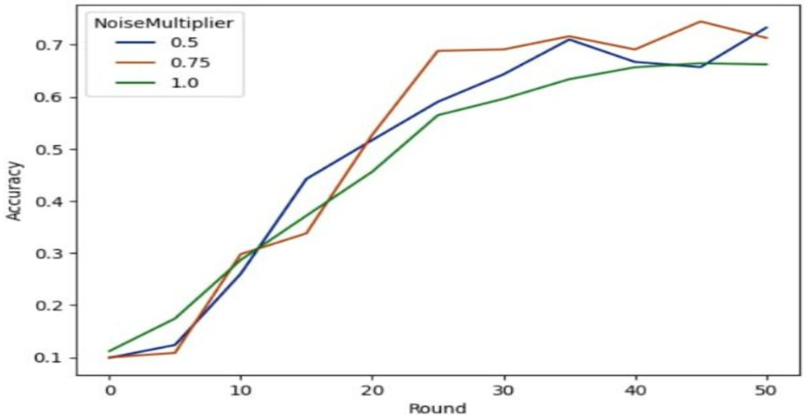
Back-Door Attack

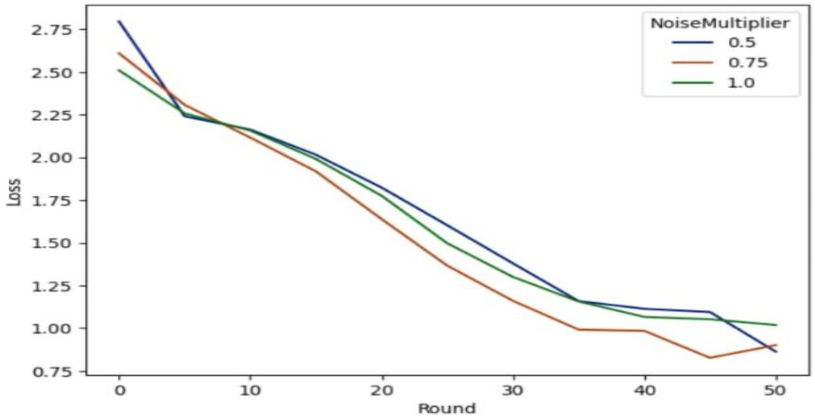
- A backdoor attack in Federated Learning (FL) involves a malicious client device installs secret access points or 'backdoors' in a federated learning model, which can manipulate the global model's performance.
- In a federated learning scenario, a backdoor attack could be implemented by having one or more malicious clients modifying their local data in a specific way.
- Suppose we have a model trained to classify images of different animals. In a backdoor attack, the attacker might introduce a unique marker (like a small red dot) in some of the training images. The attacker also modifies the labels for these images to some other class (like "cat"), regardless of what animal is actually in the image. After the model is trained and deployed, the attacker can control its output for specific inputs by including the red dot in those images.
- To implement this attack in the TensorFlow Federated tutorial for image classification, you can modify the preprocessing function in a similar way to the label-flipping attack.

Federated Learning - Backdoor Attack Architecture



Federated Learning model with Differential Privacy





- The loss graph of any federated learning model typically shows a decreasing trend, as the model becomes more accurate with additional rounds of training.
- So, federated learning with differential privacy also aims to achieve the same goal while ensuring that the client's private data remains protected. **Differential privacy involves adding noise to the model updates.. This added noise can make the model updates less accurate, resulting in a higher loss. Thus, the graphs are jagged. This is the trade off we need to think about while implementing DP in FL.**
- We have analyzed this technique for different values of noise, to see how much utility is decreasing with different levels of noise. Therefore, we need to settle on a certain value of noise where we can get **optimal value of privacy vs utility.**

Label Flipping Attack on Federated Learning Model FL

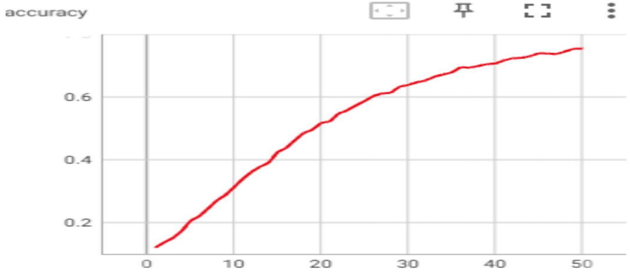


Figure 1(a)

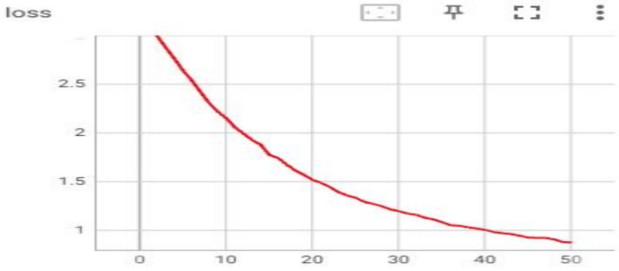


Figure 1(b)

Label Flipping Attack on Federated Learning Model with Differential Privacy

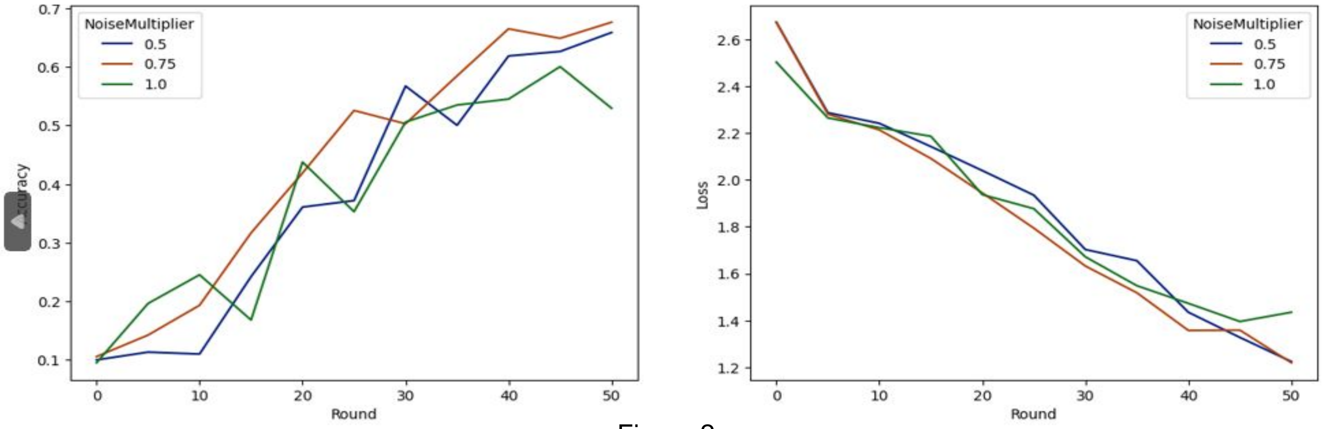


Figure 2

Label Flipping Attack - Comparative Analysis -

- To understand the Label Flipping attack, we compare the label flipping attack on a normal federated learning model with a federated learning model that has differential privacy implemented in it.
- As can be seen from graph 1(above) - attacked FL model with no differential privacy, the loss value is equivalent to 0.9 at the convergence of the model. Now comparing the same with graph below, we can see that the loss value at convergence is 1.1 which is almost equivalent. So from here, we can see that in case of graph 2, even after adding noise to the data (because differential privacy is used here) to increase the user level privacy, we are achieving comparable level of performance in both the scenarios.
- After further increasing the level of attack on the model by increasing the value of label flipping factor to 0.7, the loss value does not converge as it does in the previous case. Hence, if the level of attack to the model is increased, the model performance decreases. This can be seen from the graph where the loss value is equivalent to 2.1 at convergence of the model.

Label Flipping Attack on Federated Learning Model with Differential Privacy

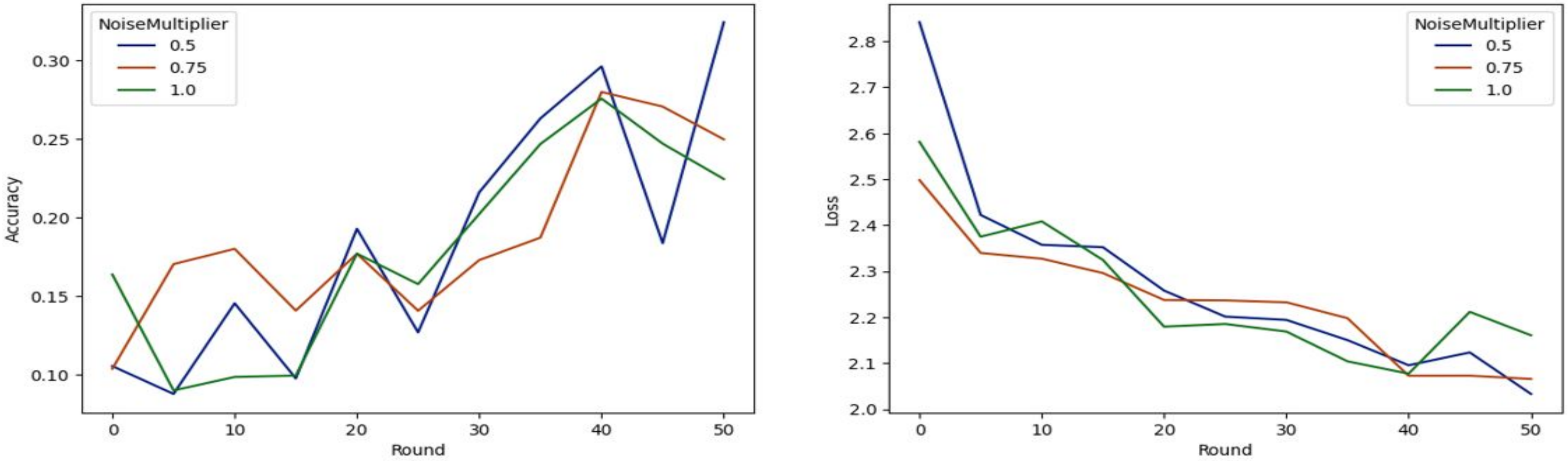


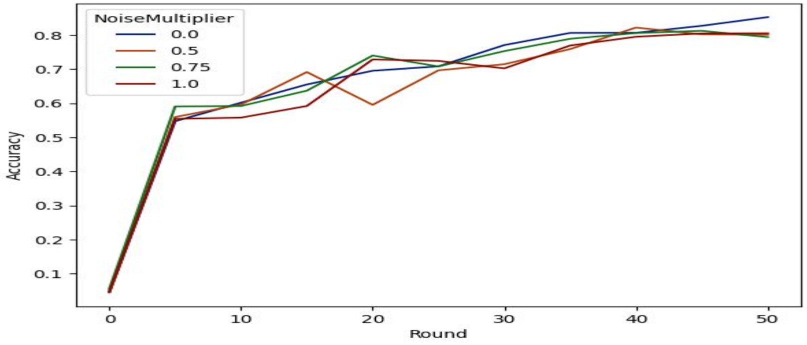
Figure 3

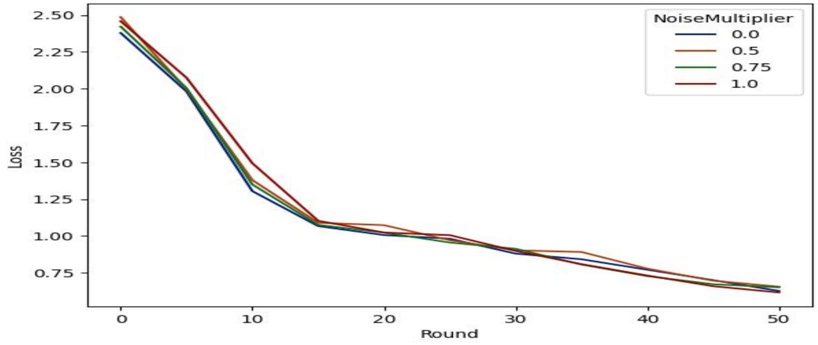
- We have also taken in account different probability factor while performing label flipping attack to study how privacy is getting affected if probability of an attacker maliciously flipping the labels of their local data is increasing.
- A small probability factor of 0.3 in the Figure 2 (earlier slide) may lead to a slow and gradual degradation of the model's performance, while a higher probability 0.7 (above) can cause a more immediate and severe impact on the model's performance.

Future Scope

We have implemented backdoor attack on federated learning with differential privacy model. But in order to completely understand the backdoor attack, a comparative case study needs to be done to show the variation of the utility and privacy, just as in the case of label flipping attack on FL.

Below are the result of backdoor attack on FL+DP -





Note - In case of FL with DP, not all types of attacks can be applied on the model, like membership inference attack, model inversion attack etc.

References

- The TensorFlow Federated documentation provides an overview of Federated Learning and Federated Learning with Differential Privacy and their implementation in TFF:https://www.tensorflow.org/federated/federated_learning
- https://www.tensorflow.org/federated/tutorials/privacy/federated_learning_with_differential_privacy
- In a paper titled "Deep Learning with Differential Privacy", Abadi et al. (2016) discussed the concept of differential privacy and its impact on the accuracy of machine learning models. The paper provides a theoretical analysis of the effect of differential privacy on the loss function of a model:Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308-318).
- A recent paper titled "Federated Learning with Differential Privacy: Algorithms and Performance" by Yu et al. (2021) provides a detailed analysis of the impact of differential privacy on the convergence rate and accuracy of federated learning models. The paper includes experimental results showing the effect of different levels of privacy budget on the loss versus rounds graph:Yu, W., Li, X., Cheng, X., & Liu, Y. (2021). Federated Learning with Differential Privacy: Algorithms and Performance. arXiv preprint arXiv:2102.09696.

THANK YOU