Towards next-generation federated learning: A case study on privacy attacks in artificial intelligence systems

Ekta Sharma
School of Mathematics, Physics and Computing
University of Southern Queensland
Springfield, Queensland 4300, Australia
ekta.sharma@unisq.edu.au

Ravinesh C. Deo
School of Mathematics, Physics and Computing
University of Southern Queensland
Springfield, Queensland 4300, Australia
ravinesh.deo@unisq.edu.au

Christopher P. Davey
School of Mathematics, Physics and Computing
University of Southern Queensland
Springfield, Queensland 4300, Australia
chris.davey@unisq.edu.au

Brad D. Carter
Centre for Astrophysics
University of Southern Queensland
Toowoomba, Queensland 4350, Australia
brad.carter@unisq.edu.au

Sancho Salcedo-Sanz
Department of Signal Processing and Communications,
Universidad de Alcalá
Alcalá de Henares, 28805, Spain
sancho.salcedo@uah.es

***Abstract*—Accurate and trust are crucial for ChatGPT and other artificial intelligence (AI) markets. One of the challenges is data leakage, which is frequently overlooked but possesses highly consequential implications. Federated learning (FL) is recognised as a new era of secure AI systems. The market for FL is estimated to reach USD 266.77 million by 2030 according to Polaris Market Research (1). This paper focuses on FL-based approaches for improving AI safety and examines the significance of Deep learning (DL) and its privacy implications. This has been achieved through six models: Federated Convolutional Neural Network (F-CNN), Federated averaging CNN (FA-CNN), Federated Adam (FA), Malicious Generative adversarial network (MGAN), Federated M-GAN (FMGAN) and Conditional GAN (CGAN). The authors analysed MNIST and CIFAR-10 datasets and conducted extensive numerical evaluations to confirm improved user privacy in federated learning for AI models. A case study with fast convergence speed and excellent asymptotic test accuracy was designed to outline White-box attacks on MGAN, FMGAN, and CGAN models. The study also implemented active inference attacks on deep neural networks without sharing raw data through FL. We created 256 synthetic images specifically to test the effectiveness of the original classifier. These counterfeit visuals effectively deceived the classifier, appearing as legitimate representations of true class labels. Trimming shared parameters was ineffective in preventing the attack, revealing limitations in collaborative learning. The generator shows the least loss of 0.0104 encountered of all models in the study. Our Generator is also the fastest after the FMGAN model. FMGAN performs best with maximum accuracy (0.9613) followed by CGAN (0.9208), MGAN (0.9163), FA (0.5148), FCNN (0.4376) and FACNN (0.4285). It also demonstrated high efficiency by successfully attacking in a short timeframe of 0.7459 milliseconds. The Federated approach led by Adam exhibited the longest processing time, at approximately 10.52 minutes. The case study illustrates the risks of surveillance and manipulation by attackers, who pressured participants to disclose confidential information. It also aimed to increase flexibility and robustness. Our work is accessible to diverse audiences, facilitating the adoption and practical applications of deep learning methods for privacy protection by major corporations.**

***Index Terms*—Federated Learning, Artificial Intelligence, Data Security, Attacks, Deep Learning, Machine Learning**

I. Introduction

Artificial Intelligence (AI) is an ever-evolving and progressive technology. The convergence of data and machine learning has undeniably resulted in substantial advancements and exhilarating technological breakthroughs, ultimately reaching unprecedented levels of intelligence. As the internet expands and attains greater complexity, novel devices are interconnected with the network, rendering it susceptible to various security threats. Intrusions are precisely described as endeavours aimed at undermining the security of computers or networks (2)

Deep Learning (DL) a branch of machine learning (ML), uses neural networks inspired by the human brain to solve complex tasks (3). DL surpasses previous ML techniques with multiple hidden layers. DL is highly effective in computer sci-

This research was conducted by the first author from the University of Southern Queensland for National Intelligence Postdoctoral Grant (project number NIPG-2022-009) and funded by the Australian Government. The first and second authors acknowledge the support and contribution of the Office of National Intelligence for the project. The authors thank the editor and all reviewers for their suggestions.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00259

ence and ideal for cyber security, as it can learn, process, and extract insights from vast data sets. However, emerging attacks challenge the domain's information privacy. Nevertheless, DL engenders privacy concerns as the trained model encapsulates crucial details regarding the training set, consequently facilitating the extraction of sensitive data. Processing extensive data sets involves time, resources, and computational power investment. An exemplification of recent and groundbreaking advancements in Machine Learning is the emergence of a novel and revolutionary concept known as Federated Learning (FL). FL can be described as the decentralised variant of Machine Learning and serves as a crucial methodology for mitigating the challenges posed by machine learning. It allows for creating hyper-personalised models tailored to individual users and intricately distributes the voluminous data across several devices and servers (4). Figure 1 illustrates the General architecture of FL where we have shown 'N' clients. Additionally, applying FL ensures minimal delays and upholds strict privacy preservation measures.

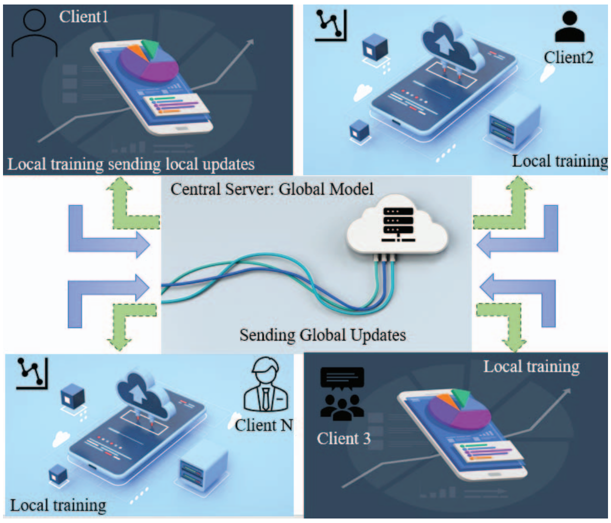

Fig. 1: General Federated Learning Architecture

Consequently, in a collaborative manner, it processes the data without any explicit sharing, meticulously updating the data residing on each device and aggregating it on the main server to yield and construct a significantly enhanced model (5). In this paper, our objective is to address a significant question, namely: what is the potential threat to privacy posed by DL algorithms when utilising data for training deep neural networks? In simpler terms, we aim to quantify the extent to which DL algorithms inadvertently disclose information about the specific data samples used for training. Also, how FL affects AI systems, the challenges, open issues, and future directions. The presence of adversarial examples (AEs) (6) poses a serious challenge to the widespread adoption of security-sensitive applications based on Deep Neural Networks (DNNs). AE attacks exploit the peculiar behaviour of DNNs, wherein a seemingly harmless alteration in input data can deceive a well-trained DNN (7). The susceptibility of DNNs to AE attacks and their counterintuitive responses undermine user confidence in the decision-making capabilities of these systems (8). As a result, it becomes imperative and highly crucial to gain a comprehensive comprehension of AE attacks and take tangible steps to enhance the reliability of DNNs in real-world scenarios. In consideration of the capabilities of opponents, two distinct categories of attacks can be identified: white-box attacks and black-box attacks. Under white-box settings, the opponent possesses complete system access regarding both the target model and data information. Conversely, black-box settings involve opponents who utilise AE transfer-ability to apply the created Adversarial Examples (AE) to an unfamiliar deployed model. From an empirical perspective, white-box attacks exhibit superior potency when confronted with the task of compromising a resilient model compared to black-box attacks (9). Future direction and challenges for intrusion detection systems were covered by FL in (10). The authors in (11) demonstrate the capability of participants with conflicting interests in the FL scenario to efficiently execute active membership inference attacks against other participants, despite the global model attaining significant accuracy in predictions. Data poisoning attacks against FL systems were studied by (12). Our Contributions for this paper are:

- Analyse the emerging branch of FL towards optimising the AI training process with the open issues and challenges.
- Elucidate the importance and privacy of DL through several models empowered with FL.
- Implement active inference attacks on DL networks without sharing raw data through FL. We discuss a method for conducting malicious attacks on Generative Adversarial Networks (GANs), which are renowned for their ability to perform implicit density estimation in distributed DL scenarios and check their efficiency and feasibility.
- Analyse the attack strategy and benchmark with other models to see their resilience against model inversion attacks.
- Present a tactic of incorporating deception within collaborative learning, whereby the adversary skilfully misleads the victim into divulging highly precise sensitive data.

The structure of this paper is as follows: Section II provides an overview of the Federated Models and Methodology, including the case study models. Following that, the research data and model architectures are examined. Section III discusses the details of experiments with the time taken for training, results, as well as the computational intensity of the decoding models. Finally, Sections IV and V present the concluding remarks and suggestions for future research.

II. Federated Models and Methodology

It is imperative to highlight that the approach to FL possesses the capability to be customised and adapted in various ways. For this paper we have developed the following models:

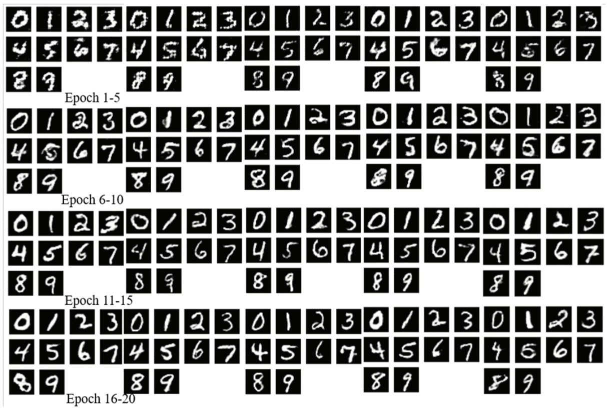

Fig. 2: The generator successfully employs its capability to mislead the discriminator by convincingly presenting the fabricated images as genuine numerical representations. The classifier correctly classifies fake images, hence the model succeeds in deceiving. Rows of images depict 20 epochs in a set of 5.

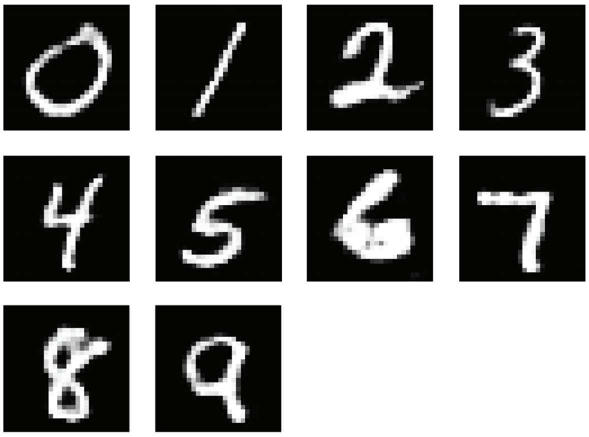

Fig. 3: The generator successfully generates images of numbers 0 to 9.

A. Federated Convolutional Neural Network (FCNN)

We developed and implemented a streamlined simulation pipeline using a CNN trained with FL. This model trained the CIFAR-10 dataset federated over 10 clients, updating model parameters. Each pairing had 5000 samples with 80% testing, 10% validation, and 10% testing samples. In the Centralised training phase, we trained on a singular location with one train loader and one validation loader to simulate the current reality of machine learning projects. By training the simple CNN on CIFAR-10 split for 5 epochs, we obtained a substandard test set accuracy of about 41%. This model serves as a reference point to establish a basic and centralised training pipeline for future FL models and case studies. The CIFAR-10 dataset was split into multiple partitions to simulate scattered data from different organisations or in a cross-silo FL setting. This resulted in ten training sets and ten validation sets representing ten distinct organisations. In the FL system, each organisation acts as a client, resulting in ten clients connected to the FL server. In FL, the server shares global model parameters with the client. The client modifies the parameters by training the local model with its data and then sends these updated parameters back to the server. Alternatively, the client can send only the gradients instead of the entire set of parameters. We use two helper functions: one to update the local model with received parameters from the server, and another to obtain the updated parameters from the local model. Finally, the current local model parameters are returned, which were received from the server, used for local training, and then sent back as the updated parameters. In another aspect, the model parameters are received from the server and evaluated using local data. The evaluation result is then sent back to the server.

B. Federated averaging CNN (FA-CNN)

This model uses the FL strategy of utmost centrality with steps including client sampling, model distribution, aggregation, evaluation, and other consequential tasks. The strategy embodies the FL algorithm - FedAvg, designed for efficient distributed training on a large scale involving numerous clients (13). Clients ensure data privacy and security by storing it locally. A central parameter server mediates coordination and communication among the clients. Clients are created locally for training, and their metrics are consolidated efficiently. A global model is formed by averaging the models of participating clients. The performance of the global model is evaluated using a small, centrally located set of tests. If no centrally located dataset is available, the evaluation of the global model must be decentralised.

C. Federated Adam (FA)

This proposed framework enhances the FL approach with a tailored FedAdam strategy. We customised the FL system using CIFAR-10 training and test sets divided into ten smaller subsets. Our system includes novel functionalities such as parameter initialisation, custom strategy selection, model evaluation options, and value exchange with clients.

D. Case Study with Malicious Generative adversarial network (MGAN), Federated MGAN (FMGAN), and Conditional GAN (CGAN)

A generative adversarial network (GAN)(14) is an advanced ML model that captures participant data. It uses two neural networks (discriminative and generative) to enhance prediction accuracy. The discriminative network distinguishes real from generated images, while the generative network learns from random noise to mimic the training set. Assuming 10 clients with different classes of data, we used weight averaging aggregation instead of uploading or downloading specific parameters. However, this aggregation is difficult to converge due to the non-independent and identically distributed nature of the data. To address this, we follow the strategy of a warm-up training with 5% of the data to improve accuracy(15). We also ensure fairness by using the same number of epochs as other models. The process ends when the discriminative network can no longer distinguish between samples from the original database and the generator's samples. The classifier and discriminator models are distinct, with one for the original classifier and one for training the generator. The models produce outputs across 11 classes, with the final class 10 used for categorising counterfeit images. During training, there are two instances of gradient backpropagation. The first instance includes genuine data, authentic labels, and falsely generated images labelled as 10. This helps the discriminator distinguish between real and fake images. In the second instance, the generator creates images for each authentic label to imitate the genuine data and deceive the discriminator. Both generator and discriminator losses are returned for plotting. We then use the generator to produce images for a range of target labels and assess whether the original classifier can correctly classify the generated images. If the classifier accurately recognises the generated images, our attempt to deceive it is successful. This is demonstrated in Figure2. If the classifier fails to recognise the images as numerals, the generator fails. The losses endured during discriminator training should closely resemble the losses during generator training for effective deception. The generated models were evaluated based on loss and accuracy. We generated 256 artificial images to attack the original classifier. It is important to note that these counterfeit images effectively deceive the classifier.

E. Data and Model Architecture

We conducted a numerical assessment of the performance of the suggested algorithms by measuring the test accuracy on two extensively acknowledged datasets: MNIST (Modified National Institute of Standards and Technology database)(16), CIFAR10 (Canadian Institute for Advanced Research)(17). MNIST is a large database of grayscale handwritten images of digits ranging from 0 to 9 that is commonly used for training various image processing systems. The dataset is comprised of a total of 60,000 training data records and 10,000 test records. MNIST database and the CIFAR-10 dataset are widely used for training and testing in machine learning and computer vision algorithms.

A collection of six functional models was created in this study. They are: Federated CNN (F-CNN), Federated averaging CNN (FA-CNN), Federated Adam (FA), Malicious Generative adversarial network (MGAN), Federated M-GAN (FMGAN), Conditional GAN (CGAN).

Intel i9 Generation with Windows platform was used to design all the models with a memory of 16 GB and a processor of 3.5 gigahertz. Tensor Flow, Scikit-learn and Keras were utilised as open-source libraries in the modelling process with the Python programming language. In FL, the effective establishment of distinct data partitions for each client is accomplished through a set of partitions in the ratio of 80% training, 10% validation, and 10% testing subsets. It is to be noted that there is no standard rule for data partitioning. We generated small training and test sets for each edge device and encapsulated each set into a PyTorch data loader or Jupyter Notebook. Some models were also run in Google Colaboratory giving access to powerful computing resources, such as GPUs. For our simulations, we avoided performing any preliminary processing on the data. The exclusive manipulation conducted on the data entailed adjusting the scale of each image to fall within the prescribed range of -1 and $+1$, following the methodology described in the published research (18). This conformed to the sophisticated generator model which incorporates a hyperbolic tangent tanh activation function in its concluding layer (18). Consequently, this results in outputs that strictly adhere to the normalisation range between -1 and $+1$:

$$I_{NORM} = (I - I_{MIN})/(I_{MAX} - I_{MIN}) \tag{1}$$

In equation 1 $I_{MIN}, =$ The minimum value of input, $I_{MAX}, =$ The maximum values of input received, and $I_{NORM} =$ The normalised input.

All models were limited in depth beyond linear layers, pooling, and dropout, allowing for fair comparisons between models. The CNN used a three-layer configuration for increased capacity and was tested for 20 epochs to ensure accuracy per theoretical guidelines. Subsequently, the performance of each model was benchmarked. Additionally, the authors took the crucial aspect of overfitting in the model formulation. It is important to acknowledge that underfitting may occur when a machine learning framework does not converge within the expected timeframe, indicating that the framework may not have the required flexibility to capture the most significant data patterns. The study employed a grid search methodology to facilitate successful model learning. By evaluating model accuracy and loss, the optimal architecture was determined. The authors acknowledged the importance of activation functions (AFs) in constructing a strong model. As such, activation functions such as ReLU, sigmoidal, SoftMax, and hyperbolic tangent were utilised to assess the performance of the models and determine the most suitable option (19).

In this study, it should be emphasised that the loss function utilized is cross-entropy, with target values within the set of 0 and 1. Additionally, it is worth noting that a comprehensive search for optimal hyper-parameters may be labour-intensive, given that each model demands around 15 hours (20). However, through careful parameter selection, the authors were able to significantly reduce the overall computation time for the model to below 15 minutes.

III. Results and Discussion

Table I displays the models created, highlighting the optimal attack model in red. Our comparative approaches are evaluated against the model that demonstrated the highest performance across all models. It is observed that the evaluation methods on the client side and server side exhibit notable differences. The centralised or server-side evaluation is relatively straightforward, comparable to the evaluation tasks performed in centralised machine learning. This evaluation allows us to assess the model after each training round using the server-side dataset. Moreover, the advantage lies in the constant accessibility of the entire evaluation dataset.

TABLE I: Models Developed in the study. Most feasible model is shown in RED

S.No.	**Model**	**Accuracy**	**Loss**	**Time (in seconds)**
1	Case Study Model: Federated M-GAN (FMGAN)	0.9613	0.1898	0.7459 milli-sec
2	Case Study Model: Conditional GAN (CGAN)	0.9208	0.0104	9.7301 milli-sec
3	Case Study Model:Malicious Generative adversarial network (MGAN)	0.9163	0.2013	18.89
4	Federated Adam (FA)	0.5148	0.0426	631.69
5	Federated CNN (FCNN)	0.4376	0.0616	196.93
6	Federated averaging CNN (FACNN)	0.4285	0.0489	369.25

On the other hand, Federated Evaluation or client-side evaluation is an intricate yet powerful approach. In the domain of FL, clients follow similar fundamental principles to those found in a traditional centralised setup. However, there is a notable distinction in that they operate on a smaller dataset that has not been encountered before.

It removes the requirement of a centralised dataset and empowers the evaluation to be conducted over a more extensive range of data, yielding more realistic outcomes. This method proves essential in numerous models to attain truly representative evaluation results. Nevertheless, it does come with a drawback. The dynamic nature of the evaluation dataset, which is subject to changes as clients become unavailable and the dataset held by each client evolves over successive rounds, can lead to varying evaluation results, even if the model remains unaltered.

In FCNN, we deployed 10 clients on a single machine. This setup can potentially result in resource depletion, even when only a fraction of clients are actively involved. This configuration enabled seamless resource sharing between the server and the 10 clients, encompassing CPU, GPU, and memory. However, it should be emphasised that employing such a setup on a single machine may rapidly deplete the available memory resources, even if only a subset of clients actively participated in each round of FL. Additionally, it is important to recognise that our model operates on various machines, allowing the server and clients to function independently. The client actively trains the model on the device and later sends the resulting model to the server. Subsequently, the server evaluates the overall performance of the global model on the client's validation set while assessing the degree of personalisation accomplished. After this analysis, the server provides the client with suitable parameters for initialising their local model. The parameters of the client's model are then converted into NumPy arrays and promptly sent to the server.

In FACNN, FedAvg proves to be a straightforward and highly potent model for experimenting. In this simulation, the FedAvg mechanism has proficiently and impartially chosen ten clients from a pool of ten eligible candidates. These selected clients trained the model, and their individual parameter updates are then seamlessly merged by FedAvg. This leads to a fresh global model, which is utilised in the next round of FL. The simulation does not consolidate metrics in the generic metrics, such as the accuracy key. Due to the potential heterogeneity of metrics and the inclusion of non-metric value pairs, the framework lacks the automatic handling ability for such cases. To address and consolidate these custom metrics, metric aggregation functions are needed. These functions are invoked by the strategy during fit or evaluate operations. Our model adopts the weighted average function to aggregate custom evaluation metrics and derive a solitary accuracy metric spanning all clients on the server side. In FA, The improvement provided our system with adaptability and the execution of personalised client-side operations by transmitting arbitrary values.

The case study investigates MGAN, FMGAN, and CGAN. It presents and implements active inference attacks on deep neural networks in a cooperative environment and shows possible improvement without raw data sharing through FL. The objective is to enhance flexibility and strength to improve the CGAN. This involves the periodic transmission of model parameters to a central server and the aggregation of local models into global models. This iterative process continues until the desired accuracy of the learning model is achieved. With each utilisation of the generator, it successfully produces novel random noise to ensure an extensive range of random noise observed by the generator. The definition of an image by the generator is regulated by the concatenation of the one-hot encoded image index and the random noise. Consequently, the generator is effectively conditioned on the one-hot encoded image while employing random noise to introduce diversification in the generated images. It should be duly highlighted that these forged visuals effectively mislead the classifier, as it perceives them as authentic representatives of the accurate class labels.

The exceptional level of precision by the generator signifies its ability to generate images that closely resemble the original training images. Moreover, it has successfully deceived the independently trained classifier using the original training data. This situation can still be viewed as a white box attack since we possess knowledge of the network structure of the original classifier, which we utilise for training our discriminator. Adopting a different architecture for the classifier can be seen as a black box attack by training the generator to mimic the original training data. This has been clearly shown in Figure2 and Figure3. In the latter, the generator successfully generates images of numbers 0 to 9 making the model successful.

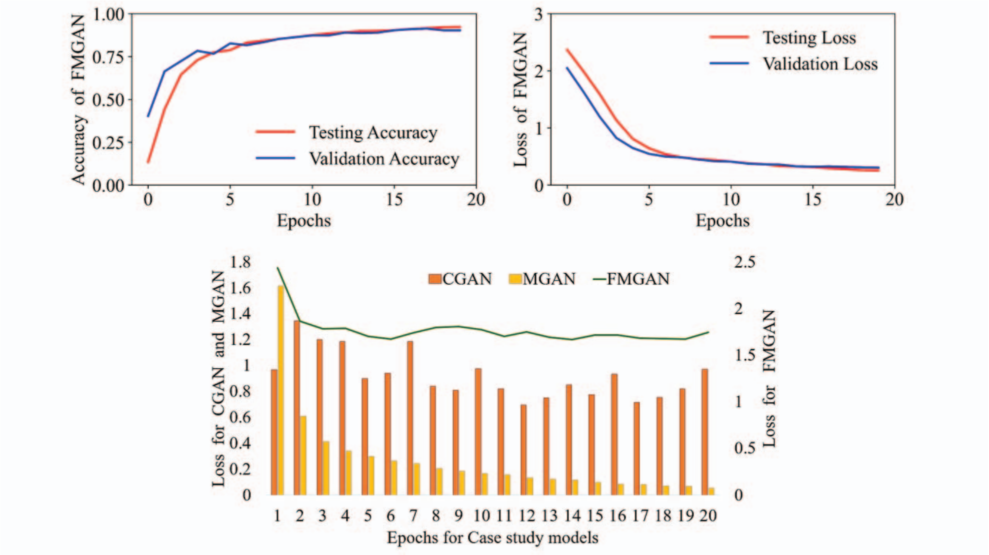

Fig. 4: Testing and Validation Accuracy of FMGAN model. The loss of all models used in the Case study is also shown. FMGAN shows maximum accuracy and CGAN shows minimum overall loss.

Models of the case study perform better as compared to other Federated approaches. Figure4 illustrates the Loss of all case study models as well as the accuracy of the FMGAN model. Figure5 shows the accuracy of case study models. FMGAN performs best with maximum accuracy (0.9613) followed by CGAN showing 0.9208 and MGAN with 0.9163 accuracy. FMGAN is also the fastest model with an attack completed in 0.7459 milliseconds. The generator effectively tricks the discriminator into believing that the generated images represent authentic numbers. The generator was also trained using concatenated labels and input noise and it shows the least loss of 0.0104 encountered of all models in the study. Our Generator is also the fastest after the FMGAN model. Figure6 illustrates the accuracy of the remaining models. FA shows maximum accuracy with 0.5148 followed by FCNN showing 0.4376 and FACNN with 0.4285. This case study shows the possibility of unwarranted surveillance and manipulation by potentially harmful attackers on the model's progression, thereby pressuring participants to disclose confidential information related to their datasets. The trimming the shared parameters fails to counteract the attack since it continues to be formidable given the adequate accuracy of localised models. In the context of collaborative learning, it is crucial to acknowledge that any user infringes upon the privacy of fellow users within the system, even without the direct involvement of the service provider. Conversely, this case study underscores the notion that collaborative learning may be considered less preferable when compared with the centralised learning approach it seeks to replace.

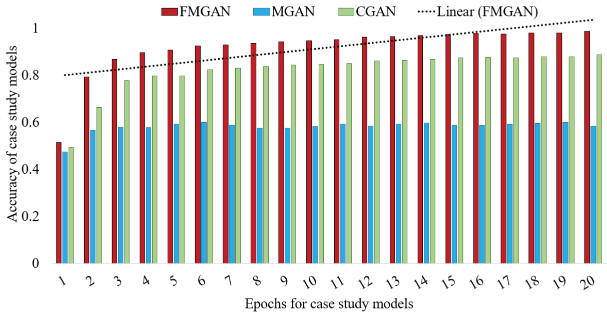

Fig. 5: Accuracy of case study models. The trend line shows the maximum accuracy of the FMGAN model.

IV. Conclusions

Federated Learning has shown significant potential in preserving user privacy and generating robust models by aggregating results and identifying common patterns from a large user base. This approach facilitates self-training and secure user data storage, resulting in an increasingly intelligent system capable of continuously testing itself. Consequently, our training and testing processes become more sophisticated with this approach. The study moves to ensure the security and privacy of AI systems. However, as this field is still in its early stages and faces various challenges regarding design and deployment, the effective approach to address these challenges is to create a clear definition of the FL problem and develop a refined data pipeline suitable for productive implementation.

A range of unique technical challenges have been successfully identified in this study and addressed through the

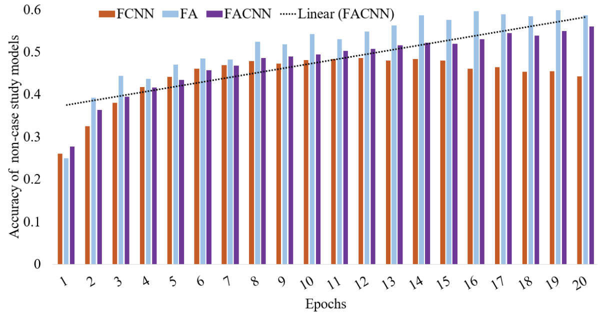

Fig. 6: Accuracy of remaining models. FA shows maximum accuracy followed by FCNN and FACNN models.

adaptation of methods such as Federated Average Fed Adam and observed in malicious attacks, thereby implementing on various models including several Federated versions, CNN, GAN, and C-GAN. Our work effectively showcases the utilisation of FL while maintaining good training performance and significantly reducing training time, surpassing conventional methods for convergence time and test accuracy. Moreover, the proposed algorithms exhibit fewer hyper-parameters, offering the potential for considerably expedited training. The models and case studies highlighted in this document serve as an initial step to elucidate and assess the successful execution and functioning of FL.

These attacks allow us to measure the amount of private information that can be leaked from the parameters and parameter updates of trained models during the training process. We have created inference algorithms for both centralised and FL, considering passive and active attackers with varying levels of prior knowledge.

V. Future Research

In this study, some areas were left unresolved for further investigation. The potential scenarios can also be broadened by utilising a variety of complex datasets such as CIFAR100, Llama, CLIP, and ImageNet. Future researchers may also consider examining privacy risks both before and following the implementation of proposed Federated Learning models. This analysis could further strengthen the adaptability and efficiency of the models. The goal is to develop effective strategies to mitigate attacks on Low Earth Orbit (LEO) Satellites. Potential solutions may include leveraging advancements in secure multiparty computation or encryption techniques. However, it is important to note that privacy-preserving collaborative learning can circumvent the need for these resource-intensive cryptographic methods. Secondly, the proposed solutions derived from these alternatives would still be susceptible to specific attacks as discussed. So, another potential strategy to consider entails embracing differential privacy at different levels of granularity. Implementing differential privacy at the user or device level would provide robust protection against the attacks formulated in this case study.

In conclusion, we anticipate that this work will have a meaningful impact in real-world scenarios and prove advantageous for leading organizations as they evaluate federated or decentralised deep learning approaches to safeguard user privacy.

References

[1] Polaris, “Federated learning market share, size, trends, industry analysis report, by application (industrial internet of things, drug discovery, risk management, augmented and virtual reality, data privacy management, others); by industry vertical; by region; segment forecast, 2022 - 2030,” report, Polaris Market Research, 2022.

[2] P. Maniriho, L. J. Mahoro, E. Niyigaba, Z. Bizimana, and T. Ahmad, “Detecting intrusions in computer network traffic with machine learning approaches,” *International Journal of Intelligent Engineering Systems*, vol. 13, no. 3, 2020.

[3] E. Sharma, R. C. Deo, J. Soar, R. Prasad, A. V. Parisi, and N. Raj, “Novel hybrid deep learning model for satellite based pm10 forecasting in the most polluted australian hotspots,” *Atmospheric Environment*, vol. 279, p. 119111, 2022.

[4] P. M. Mammen, “Federated learning: Opportunities and challenges,” *arXiv preprint arXiv:2101.05428*, 2021.

[5] N. Khajehali, J. Yan, Y.-W. Chow, and M. Fahmideh, “A comprehensive overview of iot-based federated learning: Focusing on client selection methods,” *Sensors*, vol. 23, no. 16, p. 7235, 2023.

[6] J. Shen and N. Robertson, “Bbas: Towards large scale effective ensemble adversarial attacks against deep neural network learning,” *Information Sciences*, vol. 569, pp. 469–478, 2021.

[7] Y. Wang, J. Liu, X. Chang, R. J. Rodríguez, and J. Wang, “Di-aa: An interpretable white-box attack for fooling deep neural networks,” *Information Sciences*, vol. 610, pp. 14–32, 2022.

[8] F. Croce and M. Hein, “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks,” in *International conference on machine learning*, pp. 2206–2216, PMLR.

[9] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519.

[10] S. Agrawal, S. Sarkar, O. Aouedi, G. Yenduri, K. Piamrat, M. Alazab, S. Bhattacharya, P. K. R. Maddikunta, and T. R. Gadekallu, “Federated learning for intrusion detection system: Concepts, challenges and future directions,” *Computer Communications*, 2022.

[11] 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)*, pp. 739–753, IEEE.

[12] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data poisoning attacks against federated learning systems," in *Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part I 25*, pp. 480–501, Springer.

[13] T. Sun, D. Li, and B. Wang, "Decentralized federated averaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4289–4301, 2022.

[14] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[15] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[16] Y. LeCun, "The mnist database of handwritten digits," *http://yann.lecun.com/exdb/mnist/.*, 1998.

[17] A. Krizhevsky and G. Hinton, "Convolutional deep belief networks on cifar-10," *Unpublished manuscript*, vol. 40, no. 7, pp. 1–9, 2010.

[18] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[19] J. Pomerat, A. Segev, and R. Datta, "On neural network activation functions and optimizers in relation to polynomial regression," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 6183–6185, IEEE, 2019.

[20] D. Justus, J. Brennan, S. Bonner, and A. S. McGough, "Predicting the computational cost of deep learning models," in *2018 IEEE international conference on big data (Big Data)*, pp. 3873–3882, IEEE, 2018.