International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 5, 2025
https://theaspd.com/index.php

A Method To Study The Robustness Of ML Models Against Adversarial Attacks

Archana Yashwant Panpatil[1], Dr. Vishesh Pratap Gaikwad[2]
[1,2]Department of CSE, Sardar Vallabhbhai National Institute of Technology (SVNIT), Surat, Gujarat
[1]archanapanpatil6@gmail.com

Abstract
This invention presents a novel method for evaluating and enhancing the robustness of machine learning (ML) models against adversarial attacks. Adversarial attacks, which involve small perturbations to input data that mislead models into making incorrect predictions, pose significant risks, particularly in safety-critical applications such as autonomous vehicles, healthcare systems, and security frameworks. Traditional methods for assessing model robustness, such as accuracy and precision, fail to account for adversarial vulnerabilities, leaving ML systems susceptible to exploitation. The proposed method introduces a comprehensive evaluation framework that rigorously tests models under various adversarial attack scenarios, providing a more accurate and realistic assessment of their resilience.
This approach ensures that models are subjected to a diverse range of adversarial threats, helping identify weaknesses that may not be apparent under standard conditions. In addition, the invention incorporates refined adversarial training techniques that expose models to a broad spectrum of adversarial examples, enabling them to learn robust patterns while maintaining performance on non-adversarial inputs. Complementing adversarial training, advanced optimization techniques are employed to enhance the model's inherent resistance to adversarial perturbations, thereby improving overall security. A significant contribution of this invention is the real-time adversarial attack detection system, which allows models to identify and mitigate adversarial manipulations during deployment, adding an extra layer of protection. Moreover, the invention supports custom defense mechanisms tailored to specific machine learning architectures, ensuring that defense strategies are optimized for different model types. This method offers a scalable, adaptable, and practical solution for enhancing the security of machine learning models, thereby making them more reliable and resilient against adversarial threats in real-world applications.
Keywords: - *Adversarial Robustness, Model Vulnerability, Defence Mechanisms, Perturbation Analysis.*

I. INTRODUCTION

The basic premise of machine learning is to teach computers to analyze data, identify patterns, and then apply that knowledge to solve problems or make informed decisions. Methods based on machine learning, as opposed to symbolic reasoning, seek patterns in data and utilize them to make predictions. When scientists realized that statistical approaches could help them deal with complicated and ambiguous data, a paradigm shift occurred. When it comes to learning, machine learning places a premium on data. Algorithms may learn representations, patterns, and correlations from massive datasets by training on them. Machine learning encompasses a wide range of techniques, including supervised, unsupervised, and reinforcement learning. Although conventional machine learning methods have found applications in specific contexts, recent advancements in deep learning—specifically, multi-layer neural networks—have enabled previously unimaginable improvements in domains such as image and speech recognition, NLP, and beyond. Domain experts manually identify meaningful features from the data via feature engineering, a fundamental stage in conventional machine learning. Algorithms used for training and prediction take these features as input. Linear regression, decision trees, k-nearest neighbors, support vector machines, and other similar algorithms are examples of classical machine learning. While these algorithms perform admirably on some tasks, they may struggle to grasp hierarchical representations that are particularly complex. Automatic learning of hierarchical and abstract representations from raw data is hindered by traditional machine learning models' reliance on handcrafted features. The capacity of deep learning, a branch of machine learning, to handle complicated tasks and make sense of massive volumes of data has led to its meteoric rise in popularity in the past few years.

Adversarial Attacks on Machine Learning Models

Machine learning (ML) models can be vulnerable to adversarial attacks when their inputs are intentionally designed to deceive or mislead them into producing unexpected or inaccurate results. Attacks like these

take advantage of holes in ML algorithms, intense learning models, by subtly changing the input data in ways that people can't see, but which drastically change the model's predictions. For example, an adversary could manipulate a few pixel-level alterations to cause a picture-recognition classifier to identify a stop sign as a speed restriction sign incorrectly.

In general, these kinds of attacks fall into two categories: white-box and black-box. When an attacker uses white-box attacks, they have complete access to the model's parameters, training data, and architecture, which allows them to manipulate it more precisely. On the other hand, black-box assaults depend on studying outputs to deduce effective perturbations, rather than having access to the model's internal workings. Some of the most well-known methods for creating adversarial instances include the Carlini and Wagner (C&W) attacks, Projected Gradient Descent (PGD), and the Fast Gradient Sign Method (FGSM).

Threats from adversaries are a significant concern in many ML-using industries, including those dealing with autonomous cars, financial forecasting, image recognition, and natural language processing. Misinformation, incorrect diagnoses, or compromised systems are all possible outcomes of such assaults, making the consequences all the more severe in high-stakes settings like healthcare diagnostics and cybersecurity.

Adversarial training, defensive distillation, input preprocessing, and model verification procedures are just a few of the defense mechanisms that researchers have devised to combat these dangers. The subject of antagonistic research, on the other hand, is dynamic and ever-changing due to the cat-and-mouse character of the game. Given the growing integration of AI into sensitive and mission-critical applications, ML systems must be durable, interpretable, and safe.

Purpose of the study

Computer vision, autonomous systems, healthcare, and natural language processing are just a few of the many areas where machine learning (ML) has achieved significant success recently. Despite these improvements, adversarial attacks are still a major threat to ML models, especially DNNs. These kinds of assaults involve intentionally tampering with input data, causing models to make erroneous predictions, even when the changes are imperceptible to humans. Particularly when used in situations where safety is paramount, this flaw makes one very wary of the dependability and security of ML systems.

A crucial area of research now is understanding and assessing how well ML models can withstand these hostile assaults. However, while numerous studies have examined attack-generating methods, far fewer have systematically evaluated the robustness of models under various attack scenarios. This effort is further complicated by the diversity of ML architectures and the complexity of adversarial behaviors, highlighting the necessity for a consistent and thorough evaluation approach.

In this research, we offer a framework for evaluating ML models that takes into account various attack routes, model kinds, and defense mechanisms to assess their resilience against adversarial attacks. By simulating white-box and black-box attacks, the proposed method can quantitatively and qualitatively evaluate a model's resilience to these types of threats. The study helps build safer and more resilient machine learning systems by shedding light on how models behave in hostile environments.

II. REVIEW OF RELATED STUDIES

Ajayi, Joaja. (2025). Increasingly, machine learning (ML) is powering autonomous systems in the real world, encompassing applications such as self-driving cars, drones, and robotic agents. While these systems are efficient and autonomous, they are vulnerable to adversarial attacks, which are minor, deliberately designed perturbations that might mislead the system because they rely on ML models. Examining how well ML models in autonomous systems withstand adversarial attacks is the focus of this research. We examine white-box and black-box attacks in training and inference settings, simulate attacks in a virtual autonomous vehicle, and assess the efficacy of current protection strategies. Hybrid defense systems offer a promising path toward practical implementation, as the results demonstrate trade-offs among model accuracy, resilience, and computational cost. Finally, we propose some next steps for research into making autonomous systems more resilient, with an emphasis on integrated system design and real-world testing.

Avilov, Fedor et al., (2024). The modeling of disordered crystals with the aid of artificial intelligence is a proven method for creating new materials, but there has been limited discussion or resolution of the issues with its stability, reliability, and robustness. In this study, we emphasize it by training several machine learning models on nested intermetallic approximants of quasicrystal datasets. Our quantitative and qualitative analysis of the prediction discrepancy demonstrates that a variety of plausible adjustments to the training set can produce an entirely new collection of anticipated novel materials. We also demonstrated the value of pre-training and suggested sequential training as a straightforward method for improving stability.

Bayani, Samir et al. (2024). Today, modern financial management and investment decision-making are incomplete without financial forecasting. The complexity and unpredictability of financial markets render conventional methods of financial forecasting frequently inadequate. Research Design/Results: Applying machine learning techniques is a great way to boost the efficiency and accuracy of economic forecasting, by looking at its strengths, weaknesses, and potential for the future. Within their research, they account for both linear and nonlinear approaches. Here, we focus on penalized regressions and ensembles of models as examples of linear approaches. Boosted trees, random forests, and other tree-based methods, as well as feed-forward and recurrent deep and shallow neural networks, are all considered in the study. Additionally, they consider hybrid and ensemble models, which merge characteristics of multiple alternatives. Policy, Practice, and Theory Consequences: A concise synopsis of the evaluation tools for exceptional predictive capacity is offered. The study concludes by discussing potential applications of machine learning in economics and finance, and we present an example that utilizes high-frequency financial data (Benti, Chaka, and Semie, 2023).

Freiesleben, Timo & Grote, Thomas. (2023). In contemporary Machine Learning (ML), the term "robustness" is frequently used. But context and community determine its meaning. Either researchers define robustness narrowly in technical terms like adversarial robustness, natural distribution changes, or performativity, or they don't define robustness at all. To help bring together various branches of robustness research, this paper provides a conceptual understanding of the term, aiming to establish a common vocabulary. Our robustness metric is the degree to which a robustness target is relatively unaffected by targeted modifications to a modifier. Robustness to distribution shifts, prediction robustness, and algorithmic explanations are only a few of the subtypes of robustness that our account encompasses. Lastly, we separate robustness as a distinct epistemic term and set it apart from related central ideas in ML like uncertainty, generalization, and extrapolation.

Henry, John et al. (2022). Using physiological data collected from wearable sensors, numerous recent studies have focused on the detection of negative emotional states, including anxiety and stress. Various publications have documented excellent accuracy when using features extracted from sensor signals, such as skin temperature, heart rate, and skin conductance, in conjunction with machine learning classifiers. The question of whether these models are field-deployable, however, is little discussed. In this research, we assess the transferability of models trained on cardiac signals for anxiety and stress detection using publicly available data from two big experimental investigations. We select the cardiac signal because widely used properties of heart-rate variability can be extracted from several sensor modalities, enabling us to assess model robustness both within and across experimental environments. We demonstrate that models can frequently train on proxies within the noise of lower-quality data, and that reliable classification beyond the original experimental setting depends on high-quality training data with minimal artifacts. Additionally, our findings highlight the importance of training on data from a variety of emotional states to reduce the likelihood of incorrect classifications from hidden areas of the feature space.

Publication, Research. (2020). A significant concern in recent years has been the vulnerability of AI models to malicious attacks. Particularly in safety-critical applications such as autonomous vehicles, healthcare, and finance, these attacks can pose significant security hazards by manipulating input data to deceive AI models into making incorrect predictions or classifications. This study examines the theoretical foundations and practical implications of adversarial attacks on AI models, exploring their characteristics in the process. We review the various adversarial attack types, including poisoning and evasion attempts, and examine the methods used to enhance the resilience of AI models. Additionally, we provide in-depth

coverage of defensive mechanisms, including adversarial training, robust optimization, and approaches to detection. Finally, the article explores potential avenues for further study, as well as the challenges of developing AI models that are resistant to hostile manipulation.

III. PROPOSED MODEL

Figure 1 illustrates our proposed adversarial training method, which iteratively repeats. The diagram here serves to illustrate the two stages of our proposal strategy: 1) adversarial example testing to assess the global model's robustness and 2) federated adversarial training.

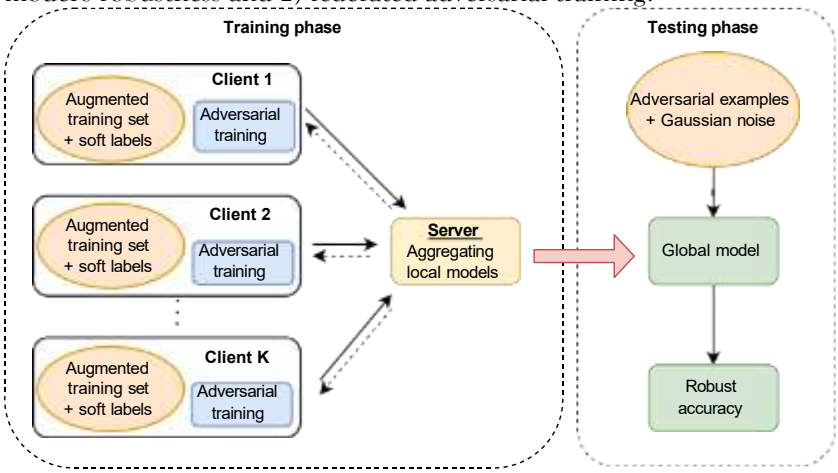

Figure 1: Federated Adversarial Training Framework: Model Aggregation and Robust Evaluation Pipeline

We begin the training phase by augmenting each client's local data, as illustrated in Figure 2. By using adversarial examples made from PGD, we enlarge the training set. We don't create malicious pictures with a lot of disturbances, as people can see them.

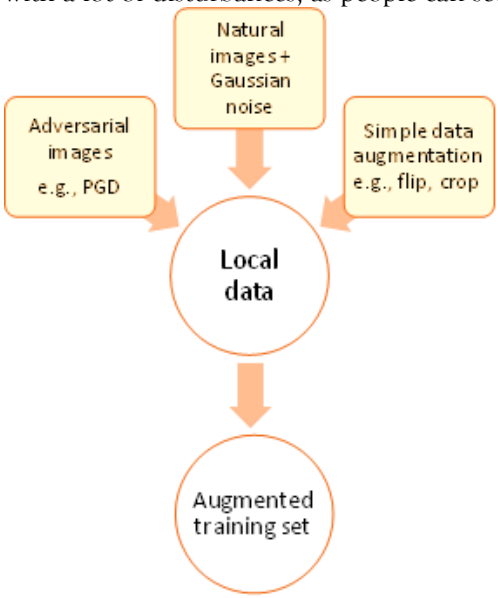

Figure 2: Local Data Augmentation Process Using Natural, Adversarial, and Augmented Images

Our goal is to find a way to protect ourselves from malicious images that are invisible to the naked eye. To strengthen our model against random noise, we create more examples by introducing Gaussian noise to normal or natural photos. To further lessen the likelihood of overfitting, the training set is subjected to basic data processing techniques such as horizontal flipping and random cropping with padding. Then, following the method outlined by Muller et al., we employ soft labeling for the target values rather than hard labeling. We therefore assign the accurate label a high likelihood, closer to 1, and the other labels a very low probability, while keeping the total probability at 1. This approach avoids assigning a probability of '1' to the actual label and '0' to all the other labels.

To promote generality and prevent model overfitting, soft labelling is commonly employed. For every input, Equation 10 specifies soft labels. Think about a piece of data that falls into category c, where c is an integer from 1 to N, the total number of categories. The i-th element of y can represent its ground truth label.

$$y_i = \begin{cases} 1 & \text{if } i = c \\ 0 & \text{if } i \neq c, \end{cases}$$

Up to the Nth power. Finally, the unofficial y^{SL},

$$y_i^{SL} = \begin{cases} 1 - \frac{N-1}{N}\alpha & \text{if } i = c \\ \frac{\alpha}{N} & \text{if } i \neq c, \end{cases}$$

Where α is a value for label smoothing that is near zero, for a predetermined number of iterations, each client trains the server-assigned model using soft labels and enhanced training examples. The model aggregation method used in this study was the Federated Average algorithm, also known as FedAvg. After the client-server communication cycle concludes, the server stores the local model parameters θ_k^t Across all employees. We average the weights of the local models according to their data contribution ratio to aggregate them. $\frac{|D_K|}{|D|}$It is seen in the subsequent equation.

$$\theta^{t+1} = \frac{1}{|D|}\sum_{k=1}^{K} |D_k|\theta_k^t,$$

where $|D_K|$ determines the client's dataset size, and $|D| = \sum_{K=1}^{K} |D_K|$. We then explain our federated adversarial training architecture, which is derived from the FedAvg AT algorithm proposed by Shah et al. When creating local model updates, Fed Avg AT considers adversarial examples, making it an extension of the Fed Avg algorithm in an adversarial federated setting. In our federated adversarial training strategy, the algorithm is fully described in Algorithm 1. Below, we outline the primary features of Algorithm 1. In the first step, the model is trained locally on each client's private data for E epochs. Subsequently, every client transmits the model weights that were trained with θ_k^t Through the server. The server then updates the weights by aggregating them using a fusion function F, like Fed Avg θ^{t+1} With the customers. One round of client-server communication has ended. To obtain the final global model, the process is repeated for R communication rounds.

When we want to assess the resilience of the global model, we add random noise to the test images. The model is resilient to slight random noise because it was trained using Gaussian noise. On the other hand, adversarial images will most likely have their intentionally planned perturbations distorted when noise is added to them.

The following two scenarios are also considered in this study: Clients have information that is uniquely theirs (IID), and clients do not (non-IID). Each client is anticipated to have an equal quantity of data with the same distribution of classes in the IID scenario. The non-IID scenario involves customers with non-uniform class distributions and varying amounts of data. We employed two methods outlined in Zhao et al. to generate client-side non-IID data. Initially, we used a method wherein every client was given data from a single class. As a result, there is a significant departure from the IID scenario in terms of the distribution of client classes. The second method involved randomly assigning two classes' worth of data to every customer. To reduce the impact of the non-IID data's skewness, we draw inspiration from Zhao et al.'s data sharing technique.

IV. EXPERIMENTAL RESULTS

In this section, we discuss how we implement the proposed adversarial (re)-training system into action and assess it using experimental data. We test our federated adversarial training method and centralized adversarial training method on IID and non-IID data to see which one performs better. Our goal is to demonstrate the practicality of both approaches by comparing their robustness and natural accuracy, and by bridging the performance gap between federated adversarial training and centralized training.

Experimental Setup and Computational Resources

Previous researchers employed the Adversarial Robustness Toolbox (ART), whereas our study used Python 3.7.6 and PyTorch 1.13.1 to create adversarial examples for training and testing. The work

employed ART with a more standardized and automated technique for creating and evaluating the (re)trained models, in contrast to our method, which enabled fine-grained control over the production of adversarial examples. Both studies seek to assess the model's resilience, despite these variations in methodology. We should note that our FL implementation is a simulation; to represent a central server and numerous client nodes, we used various variables in a Python program. As a result, the server and its clients did not exchange any data. We reduce a model's training time by utilizing a shared NVIDIA A40 GPU card to accelerate computations. But we switch to an NVIDIA GeForce GTX 1080 Ti while the A40 is in use by other academics.

Dataset and Modified Model Architecture

Our research utilizes the CIFAR-10 dataset, a widely recognized benchmark in adversarial machine learning, to evaluate the effectiveness of the proposed strategies. We use the ResNet18 architecture for experimentation with the CIFAR-10 dataset. However, we make two changes, as shown in Figure 3. We begin by adjusting the kernel size of the initial convolution layer. Second, except for the first Res Net block, which does not include down-sampling, we eliminate this step in all subsequent blocks. All experiments in this work utilized the retrained ResNet-18 architecture.

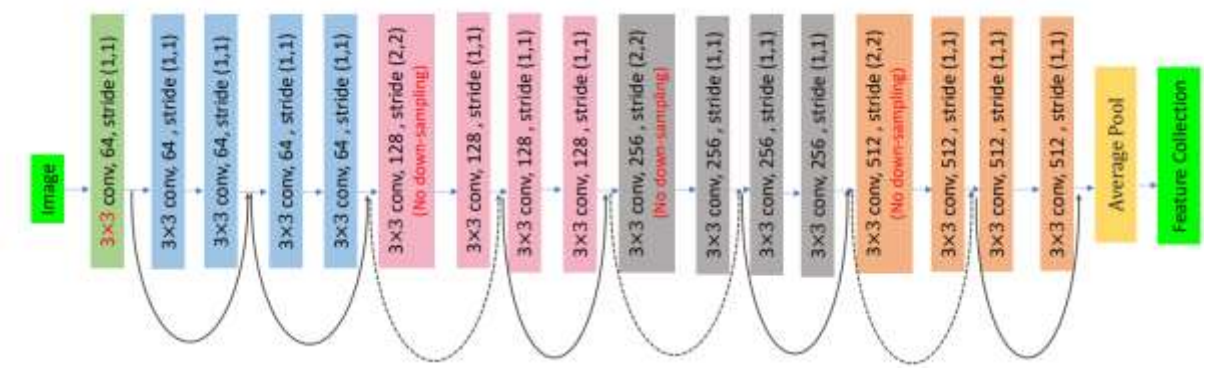

Figure 3: Modified ResNet18 Architecture of a Deep Convolutional Neural Network for Feature Extraction.

Centralized Adversarial Training Strategy

Following the methodology, we tested the resilience of our suggested centralized model on the CIFAR-10 dataset. We used adversarial examples based on PGD to train the model, with the following parameters: iteration number = 7, perturbation magnitude = 8/255, step size = 2/255. Furthermore, we supplemented the training set with 0.1 standard deviation Gaussian noise, with a mean of 0. In the methodology section, we covered the specifics of creating the adversarial images. We used the SGD optimizer to conduct adversarial training on the centralized model, setting the learning rate to 0.001, momentum to 0.9, and weight decay to 0.0002. During training, the learning rate was adjusted using a function that, after 100 and 150 epochs, respectively, reduced the initial learning rate by a factor of ten. Specifically, 0.1 is the initial learning rate. In cases where the present epoch is 100 or above, the learning rate is divided by 10. The learning rate is divided by 10 once again if it is equal to or greater than 150. As a loss function, we selected categorical cross-entropy. We assess our model's resilience to adversarial attacks, such as FGSM, C&W, DeepFool, and PGD, during the testing phase. We preprocessed the test photos by adding 0.1 standard deviation Gaussian noise with a mean of 0. Table 2 presents the experimental data used to compare our method with the existing process.

Federated Adversarial Training with IID and Non-IID Data

Within the framework of our proposed decentralized adversarial training strategy, we will address how to prepare for IID and non-IID data in the sections that follow. In all of our experiments, we used K workers, where K is an integer between 5 and 10. To generate adversarial examples, we employ the same hyperparameters used in the centralized adversarial training case (Section V-C), along with the PGD method and Gaussian noise. We do our experiments with E = 1, 3, and 5.

1) IID Data: Partitioning the training set into separate subsets at random, with each subgroup distributing data uniformly across the ten classes of the CIFAR-10 dataset, is the IID CIFAR-10 data split. The number of subsets is directly proportional to the number of clients, and each client is given their subset. A key

assumption in training machine learning models is that each client's training set follows an IID data distribution. The data partition achieves this. After that, each client uses its private data to train its model. With five and ten clients, respectively, we will evaluate our federated adversarial learning with IID data in Section V-E2.

2) Non-IID Data: A non-IID data split entails dividing the data in such a way that each subset has its distinct data distribution, as opposed to the IID data split that randomly divides the training set into separate subsets with a uniform distribution. Using this approach, the training data is distributed non-uniformly among all clients. Specifically, a non-IID training subset is generated by randomly assigning one or two data classes to each client from the training sets in a highly heterogeneous manner. It causes the data distribution of all clients to be biased towards the allocated subset of classes and different from each other. Part V-E3 detailed our methodology and the effects of non-IID data on it.

3) Local Adversarial Training: To protect against adversarial attacks, we customize our centralized training process for each client. For both the centralized training case and the local model training across all clients, we stick to the same set of hyperparameters (section V-C) for creating adversarial instances. In line with the strategy employed by Shah et al., our experimental design assumes that each client is selected to participate in each communication cycle.

Experimental Results

This section presents the experimental findings of our proposed model's operation in federated adversarial training, as well as in natural and adversarial example settings. The model's generalizability and resistance to adversarial attacks during testing are part of our evaluation. Both federated adversarial training and centralized training with IID and non-IID data show significant improvements when using our strategy.

1) Centralized Training: On the CIFAR-10 test set, the ResNet-18 that was trained on the dataset achieves a robust accuracy of 3.37% and a normal accuracy of 99.26%. Centralized adversarial training with PGD examples yields a natural accuracy of 78.17% and a robust accuracy of 47.05% for the modified ResNet-18 model. Adding samples of Gaussian noise to the training dataset also leads to a considerable improvement in the robust accuracy (65.41%). When FGSM instances with F = 8/255 are used instead of PGD adversarial examples, the robust accuracy drops from 47.05% to 27.27%. Evidence from PGDs produced by an iterative local

Table 1: Comparison of Adversarial Training Strategies Using Customized and Official ResNet18 Models under FGSM Attack Conditions

Training Set	Model	Learning Rate	Test Set	Natural Accuracy	Robust Accuracy
Natural examples	Customized ResNet18	Varied	FGSM	99.26%	3.37%
PGD examples	Customized ResNet18	Varied	FGSM	78.17%	47.05%
FGSM examples	Customized ResNet18	Varied	FGSM	78.59%	27.27%
PGD examples + Gaussian	Customized ResNet18	Varied	FGSM + Gaussian	78.65%	65.41%
PGD examples + Gaussian	Official ResNet18	Varied	FGSM + Gaussian	70.17%	44.82%
PGD examples + Gaussian	Customized ResNet18	Fixed	FGSM + Gaussian	79.87%	55.95%

Search methods that focus on the immediate area around the primary examples yield better results. In terms of natural accuracy (78.65% vs 70.17%) and robust accuracy (65.41% vs 44.82%), our modified ResNet-18 outperformed the official ResNet-18. A slight improvement in natural accuracy (from 78.17% to 79.87%) and a significant drop in robust accuracy (from 65.41% to 55.95%) are observed when the learning rate is fixed to 0.1 rather than varied. In Table 1, the primary outcome is presented. Table 2 compares the existing technique to the customized ResNet18 model trained with varied learning rates.

Table 2 illustrates how our centralized adversarial training method outperforms existing methods and how effectively it defends against various white-box attacks. When tested against FGSM, C&W, and DeepFool attacks, our technique demonstrates superior resilient accuracy.

Table 2: Centralized adversarial training - Robust accuracy (%) on the CIFAR-10 test dataset under various white-box attacks.

Training	FGSM	C&W	Deep Fool
Existing method	47%	78%	36%
Our method	65.41%	81%	83%

2) Federated Adversarial Training with IID Data: Here, we evaluate federated adversarial training using IID data and contrast it with centralized adversarial training. Our goal is to demonstrate that federated approaches can be equally resilient to malicious attacks. For C&W, the robust accuracy in the federated scenario with five clients is comparable to that of the centralized model (Table 3), while for Deep Fool assaults, it is within 4% of the centralized model. When it comes to the PGD assault, the federated method with five clients is marginally less effective than the centralized scenario. When we examine ten clients, we observe the same pattern. The performance of the federated technique for the C&W attack is within 5% of the results of centralized adversarial training, and for the Deep Fool attack, it is within 7%. Comparing the centralized and federated approaches, the latter, with 10 clients, is 5% more effective against the PGD attack. Our experiments show that federated adversarial training using IID data can achieve robustness levels similar to the centralized scenario, particularly when all participants are involved in every communication round.

Table 3: Performance of Federated Adversarial Training Across Varying Client Counts (K) Under Different Attack Methods

# Clients (K)	Natural	FGSM	C&W	DeepFool	PGD
K = 5	80.76%	63.07%	81%	79%	71%
K = 10	66.23%	51.51%	76%	76%	77%

3) Federated Adversarial Training with Non-IID Data: Here, we investigate how data heterogeneity affects the efficacy of our federated adversarial training approach. We focus on what happens to the global model's robust accuracy and its natural accuracy when non-IID data is either one-class or two-class. To further reduce the influence of data heterogeneity, we also assess how well the data sharing method works, which allows customers to contribute a small amount of their private data. Tables 4 and 5 display the outcomes of the experiments. We assess the robust accuracy of our federated adversarial training approach on the CIFAR-10 test dataset, both with and without employing the data sharing method for local training, in the tables below. To build the CIFAR-10 global shared training set, we follow the procedure outlined in.

Out of the training images, we select 1,000 for each class at random, for a total of 10,000 images that will be shared. To minimize the effect of data heterogeneity, we randomly choose 500 photos from the worldwide shared dataset for each class in our study. Thus, the other five thousand pictures are rendered useless. We split the unselected training samples into multiple parts after creating the global training subset. A client is given each partition so they can practice local adversarial tactics. In Section V-D2, we covered the specifics of generating non-IID data for each client.

We create the non-IID data for every client after making the global shared training set, using the steps outlined in Section V-D2. We found that on both one-class and two-class non-IID data, the federated adversarial training framework with data sharing achieved higher robust accuracy than the non-sharing framework in our experiments. In particular, the global model trained without utilizing the global shared training subset performs poorly in terms of both robust accuracy and natural accuracy when applied to the one-class non-IID dataset, which presents a more challenging and extreme example of data heterogeneity compared to the two-class non-IID samples. The specific accuracy rates against FGSM, C&W, DeepFool, and PGD assaults are 11%, 12%, 11%, and 1.61%, respectively, in the robust accuracy test. Table 4 shows that, in contrast to the non-data-sharing situation, the one-class non-IID federated adversarial training approach with data sharing produced substantially greater robust accuracy and natural accuracy. Even in the two-class non-IID scenario, we see the same patterns; for example, when comparing the versions with and without data sharing, the former achieves far better natural and robust accuracy.

Table 4: Accuracy Comparison of One-Class Non-IID Federated Adversarial Training with and Without Data Sharing Across Different Attack Methods

Training	Natural	FGSM	C&W	DeepFool	PGD
One class of non-IID federated AT **without** data sharing	10.97%	1.61%	11%	12%	11%
One class of non-IID federated AT **with** data sharing	67.42%	41.18%	53%	47%	48%

Table 5 contains more detailed results. The findings suggest that utilizing private data, even in small quantities, can help mitigate the effects of data heterogeneity and enhance the resilience of the federated learning model to adversarial attacks.

Table 5: Accuracy Comparison of Federated Adversarial Training with and without Data Sharing Against Various Adversarial Attacks

Training	Natural	FGSM	C&W	Deep Fool	PGD
Two-class non-IID federated AT without data sharing.	57.82%	54%	57%	62%	59%
Two-class non-IID federated AT with data sharing	85.04%	63.97%	72%	71%	67%

V. CONCLUSION

To ensure the dependability and credibility of machine learning models for use in real-world scenarios, it is crucial to assess their resilience against adversarial attacks. Researchers can successfully uncover weaknesses in deep learning models by utilizing systematic evaluation approaches. These methods include creating adversarial instances using white-box and black-box attack techniques (e.g., FGSM, PGD, Carlini-Wagner). These assaults demonstrate how vulnerable even well-performing models can be to modest input changes, as they simulate worst-case scenarios with tiny, often undetectable perturbations. One can quantify the model's resilience by measuring its response to these perturbations; standard measures for this include attack success rate, robust accuracy, and perturbation norms (L∐, L2, etc.).

To further understand how well a model can resist adversarial manipulations, it is essential to incorporate adversarial training and several defense mechanisms into the robustness analysis. These include input preprocessing, defensive distillation, and certified defences. To ensure a fair comparison of robustness across models and architectures, these defences should be evaluated under powerful, standardized assault circumstances. To strengthen the validity of robustness assertions, frameworks such as Robust Bench and libraries like Clever Hans and Foolbox make it easier to create scalable and reproducible evaluation pipelines.

Ultimately, when deploying AI systems in sensitive industries such as healthcare, banking, and autonomous systems, creating and implementing rigorous techniques to test ML models against adversarial threats is essential. Research like this should inspire more secure and performance-oriented training paradigms and designs. For AI systems to be more resilient and accountable in the future, robustness evaluation has to be a standard component of the development lifecycle for machine learning models.

REFERENCES: -

[1] K. S. Divya, P. Bhargavi, and S. Jyothi, "Machine learning algorithms in big data analytics," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 1, pp. 63–70, 2018.

[2] R. K. Perumallapalli, "Machine Learning Approaches for Improving Supply Chain Efficiency and Demand Prediction," *IJSAT*, vol. 1, no. 2, pp. 1–7, Apr.–Jun. 2010.

[3] N. Gupta, G. Tur, D. Hakkani-Tür, and L. Heck, "Dialog state tracking using long short-term memory neural networks," in *Proc. Interspeech*, 2018.

[4] R. K. Perumallapalli, "AI-Enhanced Cybersecurity for Large-Scale Network Protection," *IJIRMPS*, vol. 2, no. 1, pp. 1–5, Jan.–Feb. 2014.

[5] R. K. Perumallapalli, "Deep Reinforcement Learning for Cloud Resource Provisioning," *IJIRMPS*, vol. 4, no. 1, pp. 1–6, Jan.–Feb. 2016.

[6] R. K. Perumallapalli, "SAP Cloud Integration with AI for Real-Time Data-Driven Decision Making," *IJIRCT*, vol. 5, no. 1, pp. 1–8, 2019. [Online]. Available: https://www.ijirct.org/viewPaper.php?paperId=2411015

[7] C. Szegedy et al., "Intriguing properties of neural networks," *arXiv preprint*, arXiv:1312.6199, 2014.
[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Int. Conf. Learn. Represent. (ICLR)*, 2015.
[9] J. Ajayi, "Topic: Robustness of Machine Learning Models Against Adversarial Attacks in Autonomous Systems," 2025.
[10] S. Bayani, I. A. Mohamed, and S. Venkatasubbu, "Robustness and Interpretability of Machine Learning Models in Financial Forecasting," *Eur. J. Technol.*, vol. 8, pp. 54–66, 2024, doi: 10.47672/ejt.2005.
[11] F. Avilov, R. Eremin, S. Budennyy, and I. Humonen, "On the Robustness of Machine Learning Models in Predicting Thermodynamic Properties: a Case of Searching for New Quasicrystal Approximants," *arXiv preprint*, arXiv:2410.13873, 2024.
[12] J. Henry, H. Lloyd, M. Turner, and C. Kendrick, "On the robustness of machine learning models for stress and anxiety recognition from heart activity signals," *TechRxiv*, 2022, doi: 10.36227/techrxiv 21688352.
[13] T. Freiesleben and T. Grote, "Beyond generalization: a theory of robustness in machine learning," *Synthese*, vol. 202, 2023, doi: 10.1007/s11229-023-04334-9.
[14] S. Divya, L. P. Suresh, and A. John, "Hybrid optimization algorithm-based generative adversarial network for change detection using pre-operative and post-operative MRI," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 36, no. 07, p. 2251007, 2022.
[15] A. Kapadnis, "Brain Tumor Detection using Transfer Learning Technique with AlexNet and CNN," Ph.D. dissertation, Natl. Coll. of Ireland, Dublin, 2021.
[16] B. Mahesh, "Machine learning algorithms—a review," *Int. J. Sci. Res. (IJSR)*, vol. 9, no. 1, pp. 381–386, 2020.