Adversarial Training Using FGSM Attack for Convolutional Neural Networks

Neha Mehra[1]*, Urjita Thakar[2], Vrinda Tokekar[3]

[1, 2] Shri Govindram Seksaria Institute of Technology and Science (SGSITS), Indore, India
[3]Institute of Engineering and Technology (IET-DAVV), Indore, India
mehra.neha40@gmail.com*, thakarurjita@gmail.com, vtokekar@ietdavv.edu.in

Abstract. Deep neural networks are highly vulnerable to adversarial perturbations, which can significantly reduce their classification performance. To address this vulnerability, this work applies Fast Gradient Sign Method (FGSM) based adversarial training to improve the robustness of convolutional neural networks. FGSM generates perturbed inputs through a single gradient-based step, making it an efficient method for exposing model weaknesses. FGSM generates adversarial perturbation examples by applying a one-step perturbation in the direction of the gradient sign, making it fast and efficient attack generation method. In this study, FGSM-crafted samples are incorporated during training, and the effect of varying epsilon values and clean–adversarial data ratios is examined on MNIST (Modified National Institute of Standards and Technology) dataset consists of handwritten digit images and CIFAR-10 (Canadian Institute for Advanced Research) dataset contains color images across 10 classes. Experimental results show that adversarial training enhances resilience against FGSM attacks while maintaining acceptable accuracy on clean inputs, highlighting its effectiveness as a practical defense strategy for secure deep learning systems.

Keywords: Adversarial Training, FGSM, Convolutional Neural Networks, Machine Learning Security, Gradient-based Methods, Robustness

1. Introduction

Machine learning, especially deep neural networks, has become a widely adopted approach in computer science and engineering applications, including in domains such as cybersecurity, image analysis, natural language processing, and pattern recognition. With increasing use in security-critical tasks, ensuring the reliability of training data and the learned model has become essential [1]. It has been shown that crafted or manipulated training inputs can lead to incorrect model behavior, making the protection of datasets a crucial aspect of secure machine learning. Adversarial machine learning is a technique that show how model will confuse from mislead perturbations and how defensive mechanisms can be used to counter different attacks. A modified input that provides a change in perturbation and causes the model to produce an incorrect prediction is known as an adversarial example[2]. These perturbations expose model flaws in deep neural network decisions particularly convolutional neural networks (CNNs)[3].

Among various adversarial attack techniques, the Fast Gradient Sign Method (FGSM) is a basic type of adversarial attack method as it is simple and fast method based on gradient. By changing the input in the direction of the loss of gradient, FGSM [4] creates adversarial samples that allows important areas which are more sensitive in convolutional neural networks (CNNs).

S. Bhalerao et al. (eds.), *Proceedings of the 2nd International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2026)*, Advances in Intelligent Systems Research 207,
https://doi.org/10.2991/978-94-6239-678-4_13

To overcome the problem, several defensive methods have been studied, and adversarial training is one of the most effective techniques used to solve this problem. During adversarial training, examples that are modified are added to the learning process so that the model can gradually get used to the adversarial examples, enabling the model to learn more robust feature representations[5][6]. This will help CNN model gradually adapts to adversarial noise, improving its stability and generalization.

In this paper, we studied how FGSM gradient-based adversarial training method used to make convolutional neural networks more robust to the noisy data[7]. FGSM-generated adversarial examples are added into the training process with clean samples, and the effects of varying epsilon values and different clean-to-adversarial data ratios are analyzed. Experiments are conducted on the MNIST and CIFAR-10 benchmark datasets to evaluate the trade-off between adversarial robustness and clean accuracy.

2. Related Work

It has been studied that deep neural networks are highly vulnerable to suspicious inputs. It also studied how small and often modified perturbations can cause significant degradation in model performance. This section presents a systematic literature review on adversarial attack techniques and corresponding defense mechanisms.

2.1 Adversarial Attacks

The Fast Gradient Sign Method (FGSM), is a basic and a single-step gradient-based technique that perturbs inputs in the direction of the loss gradient. It is simple and computationally efficient, it was first presented by Goodfellow et al. [8] and used by many researchers to do their work. In contrast, a technique DeepFool, which is proposed by Moosavi-Dezfooli et al. [9], it works by using iterative type of work that minimally perturbs inputs by finding the closest decision boundary. It is more effective but slower, when it is used for generating samples it takes lots of time and need GPU for execution.

Several studies have done on the transfer learning of method and how to make model robust when these attacks comes. Kurakin et al. [10] uses FGSM but modified version which uses iterative manner. Madry et al.[11] express the method that deals with adversarial methods that show how model robustness will be done using a saddle-point optimization problem and introduced a method Projected Gradient Descent (PGD), which is a multi-step extended version of FGSM which uses a repeation of gradient updates and some constraints[4]. Carlini and Wagner [12] developed an adversarial attack based optimization method that targets neural networks trained with defensive distillation. The attack minimizes the L2, L∞, or L0 norm of perturbations while maximizing the adversarial objective using a custom loss function. Papernot et al. [13] developed the Jacobian-based Saliency Map Attack (JSMA), which selectively modifies a limited number of highly influential features identified through the network's Jacobian. Kurakin et al. [14] presented FGSM-based data augmentation in boosting adversarial resilience. Their experiments use the

CIFAR-10 and TinyImageNet datasets, and they find that integrating weak FGSM examples into the training pipeline helps the model generalize better under gradient-based attacks, especially when combined with label smoothing and batch normalization. Carlini et al. [15] conducted a rigorous analysis of DeepFool attack, PGD attack, and Carlini & Wagner (C&W) attack which was evaluated on CIFAR-10 dataset, SVHN dataset, and ImageNet-subset datasets. Their findings showed that DeepFool, though computationally lighter than PGD and C&W, generated perturbations that aligned closely with local decision boundary shifts. Croce and Hein [16] proposed a reliable evaluation framework using an ensemble of diverse parameter-free adversarial attacks. Their work demonstrated that strong boundary-based adversaries can effectively expose robustness weaknesses in convolutional neural networks and highlighted the importance of using multiple adversarial attack methods for accurate robustness evaluation.

2.2 Adversarial Defenses and Training Strategies

Beyond attack generation, other researchers also give their study of models in different a situation which works on security and on sensitive data. Eleftheriadis et al. [17] presented different ways to make DNNs improve model robustness. They said that defenses method that use a single method are not enough and suggested that multiple attack methodologies be explored for adversarial data. Rong et al. [18] proposed multi-resolution training, which gives downsampling and upsampling during training to manage the effects of adversarial changes. Although it is not a hybrid attack-based method, but it suggests that modifying the way the model learns might help it better protect from different kinds of attacks. Kafali et al. [19] discussed DATNS framework with Non-Sequential Adversarial Epochs, which adversarial data are added to some training phase only. This approach help training process most effective and keep models stable and fast as generating samples itself taken lots of resources and need GPU. Villegas-Ch et al. [20] tested FGSM, PGD, and C&W against different defense strategies and came to the conclusion that hybrid and multi-strategy methods are needed to make systems more robust in practical applications.

In the overall literature review we saw that CNN based model is affected by even small changes. FGSM is basic method[21] for generating samples as it is easy to use and fast, which make it a base for adversarial training. In contrast to other attacks DeepFool, PGD, and C&W attack method are iterative based method. They are very optimized method which uncover the weaknesses in neural networks but have higher computational costs[22]. These study highlight the basic adversarial training methodologies, especially FGSM-based methods which is used in practical applications in the field of security of ML based model.

3. Adversarial Training Methodology

As the security of machine learning models is a crucial concern, many critical decisions rely on machine learning models, so the main motive of this study is to enhance capacity of convolutional neural networks to learn patterns against gradient-based adversarial attacks. To do this, FGSM-based adversarial training is very effective way to do this. FGSM method make changes as per the direction of the

gradient of the loss function with respect to the input. This make adversarial samples to utilize computing power required.

The proposed methodology integrates FGSM-generated adversary samples into the training process with clean data. Exposing the model to both types of data enables it to develop feature representations that are less sensitive. Different clean-to-adversarial data ratios and multiple epsilon values are explored to analyze their impact on model robustness and generalization. This FGSM-based adversarial training approach improves resistance against gradient-based attacks while maintaining competitive performance on clean data.

3.1 Dataset Preparation

Two benchmark datasets are used to evaluate the impact of FGSM-based adversarial training.

- MNIST consists of 60,000 training images and 10,000 test images of handwritten digits (28×28 grayscale, 10 classes)[23].
- CIFAR-10 includes 50,000 training and 10,000 testing color images (32×32×3, 10 classes)[24].

All images are normalized to [0, 1] range to ensure stable gradient computation during FGSM generation. No additional augmentation is applied to isolate the effects of adversarial training.

3.2 CNN Model Architecture

A standard convolutional neural network (CNN) is used for training of both the datasets, consisting of:

- Optimizer: Adam [25]
- Convolutional layers activated with ReLU to introduce nonlinear transformations[26]
- Max-pooling layers used for spatial downsampling for feature maps.
- Dense layers is used for representation learning next flattening the data.
- A final fully connected layer producing logits for classification.

The architecture differs slightly for MNIST and CIFAR-10, adjusted to match input dimensionality, but the overall design and training principles remain consistent.

3.3 FGSM Attack Generation

FGSM used method generates adversarial examples using[2][27].

$$x\_adversarial = x\_input + \epsilon \ . sign\left(\nabla_{x\_input}\ Loss(\theta, x\_input, y\_t)\right)$$ Equation(1)

where:

- x_input is the original sample
- y_t denotes the true class label
- Loss represents the loss function

- θ stands for the model's parameters
- The perturbation intensity is controlled by ϵ.

FGSM is used to produce adversarial instances for multiple ε values.
ϵ ={0.05,0.08,0.1,0.12,0.15,0.18,0.2,0.3,0.4,0.5}
Evaluation of both weak and extreme attack scenarios is made possible by stronger disturbances produced by higher ε.

3.4 Evaluation Metrics

Three points are chosen to assess the impact of adversarial training: 1. Clean Accuracy on test images that have not been manipulated. 2. Accuracy of FGSM against suspicious test images produced with the same ε. 3. Combined Accuracy (Primary Metric) Accuracy on suspicious and clean data using the same ratio as training. Additionally, the Attack Success Rate (ASR) is also calculated. It shows the proportion of adversarial inputs that are successful in forcing incorrect classification. An attack is more successful if the ASR is higher .

3.5 Experimental Setup

- Training is conducted for 5–10 epochs depending on dataset.
- Batch size is 128.
- Optimizer: Adam
- Loss function: Sparse Categorical Crossentropy (from logits)
- All experiments use fixed random seed.

This given experimental setup allows a detailed comparison across multiple ε values and given training ratios.

4. Experiments and Results

The experimental analysis focuses on evaluating how FGSM attack based adversarial training influences the performance of convolutional neural networks on the MNIST dataset and CIFAR-10 dataset[28]. FGSM-based adversarial training is implemented by mixing clean samples and FGSM-perturbed samples during training. The following clean-to-adversarial ratios are explored:

- 50% Clean – 50% FGSM
- 30% Clean – 70% FGSM
- 25% Clean – 75% FGSM
- 60% Clean – 40% FGSM

For each mini-batch during training, a subset of clean samples is selected, and an equal-sized subset is perturbed using FGSM with chosen perturbation strength (ε). The clean and adversarial samples are then combined to form a mixed batch, which is used to update the CNN parameters.

The experimentation was conducted on two benchmark datasets. CIFAR-10 contains 50,000 colored training images and 10,000 test images across ten object categories. MNIST consists of 60,000 grayscale digit images for training and 10,000 for testing.

A simple three-layer convolutional neural network is implemented for both datasets. The model is trained for 5–10 epochs using the Adam optimizer and cross-entropy loss, depending on dataset complexity.

4.1 FGSM Vulnerability Analysis (MNIST)

Experiments were conducted for a wide range of perturbation strengths (ϵ) and multiple clean-to-adversarial training ratios. Metrics such as Clean Accuracy, FGSM Accuracy, and Attack Success Rate (ASR) were computed for CNN model as well as after adversarial training.

The Clean model achieved 97.35% accuracy on MNIST. Before adversarial training, FGSM attacks were applied to the clean model using ε ϵ{0.05, 0.5}. As shown in Fig. 1, accuracy showed a sharp decline under FGSM perturbations.

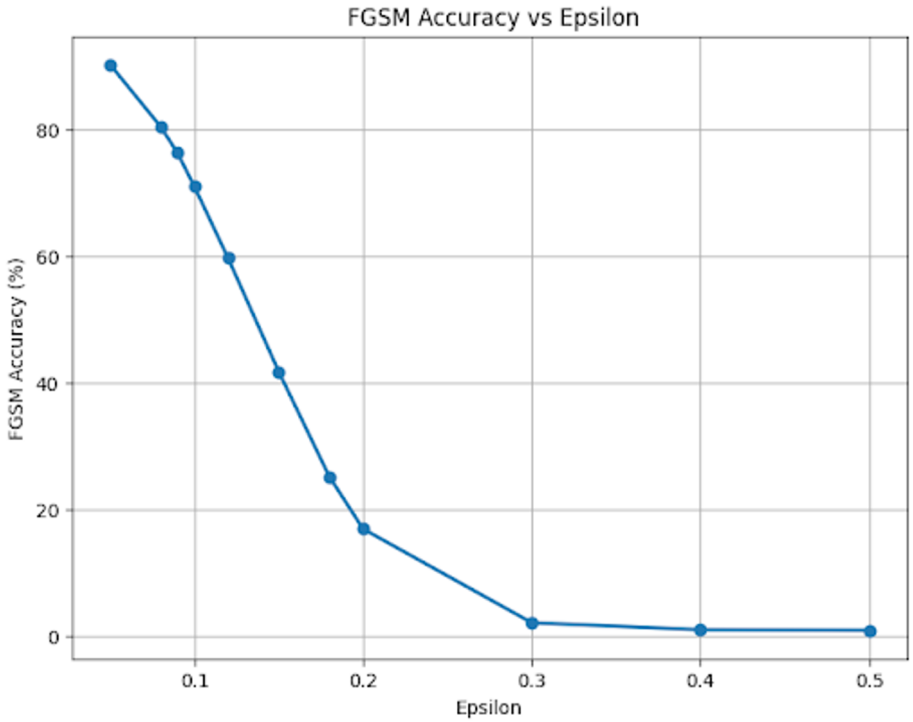

Fig.1: FGSM Accuracy vs Epsilon

The Attack Success Rate shown in Fig. 2 curve clearly shows that even small perturbations significantly degrade accuracy, confirming that MNIST models remain highly vulnerable to FGSM without defenses technique.

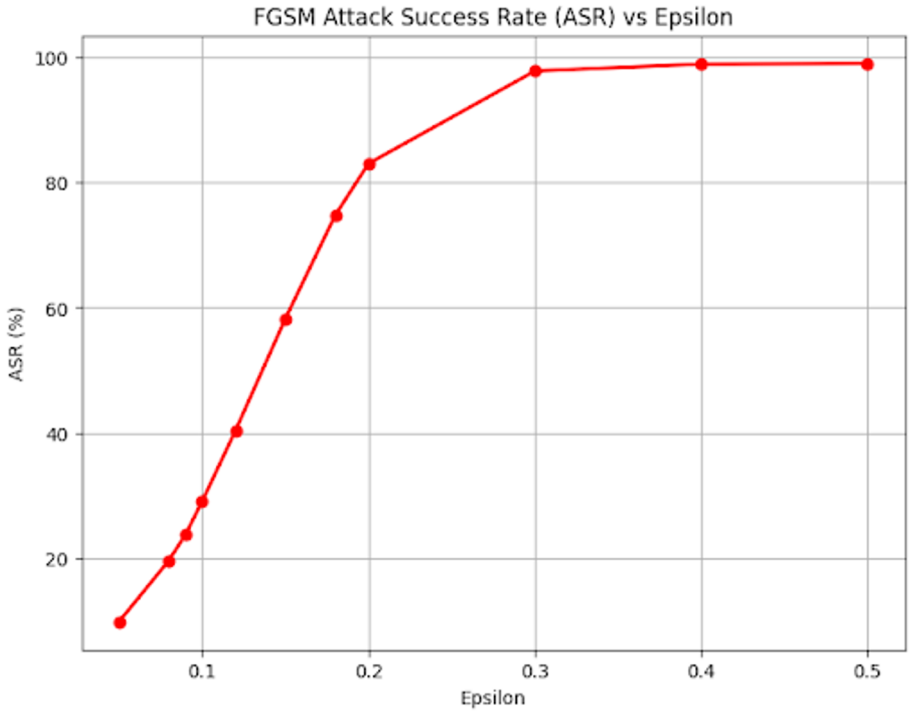

Fig. 2: FGSM Attack Success Rate vs Epsilon

4.2 FGSM Adversarial Training (MNIST)

Adversarial training was performed for each epsilon across the following mixing ratios:

- 50% Clean – 50% FGSM
- 30% Clean – 70% FGSM
- 25% Clean – 75% FGSM
- 60% Clean – 40% FGSM

The Combined Accuracy (mixed on clean and FGSM test data) is shown in Fig.3.

Key Observations

1. Lower epsilons (0.05–0.1) significantly improves from adversarial training

Combined accuracy remains 85–97% depending on ratio.

At $\epsilon = 0.05$

- 50/50 ratio → 96.05% FGSM accuracy
- 30/70 ratio → 95.90%
- 25/75 ratio → 95.95%

Training on more adversarial samples does not drastically reduce clean accuracy for low ϵ.

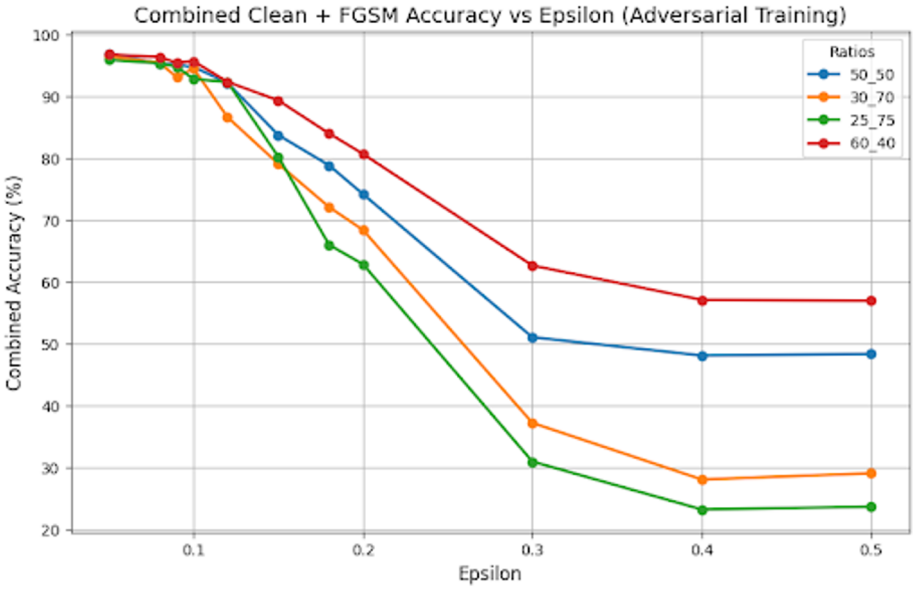

Fig. 3: Accuracy vs Epsilon of Adversarial Training on MNIST Dataset

2. Mid-level epsilons (0.12–0.18) show partial robustness

At $\epsilon = 0.15$

- FGSM accuracy ranges from 61% to 75%, depending on ratio.

Overly adversarial ratios (25/75) degrade clean accuracy more drastically.

3. High epsilons (0.3–0.5) remain extremely challenging

- FGSM accuracy falls below 10% for nearly all ratios.
- Even after adversarial training, ASR ≈ 90–99%.

This confirms that FGSM at very large epsilons steps outside the perceptually acceptable region, making classification a challenge.

4.3 CIFAR-10 FGSM Analysis

The CIFAR-10 dataset on FGSM evaluation shows much higher vulnerability due to the high-dimensional, colorful inputs:

As shown in Fig. 4, even an epsilon of 0.01 reduces accuracy by 50%, showing CIFAR is far more sensitive to adversarial noise than MNIST.

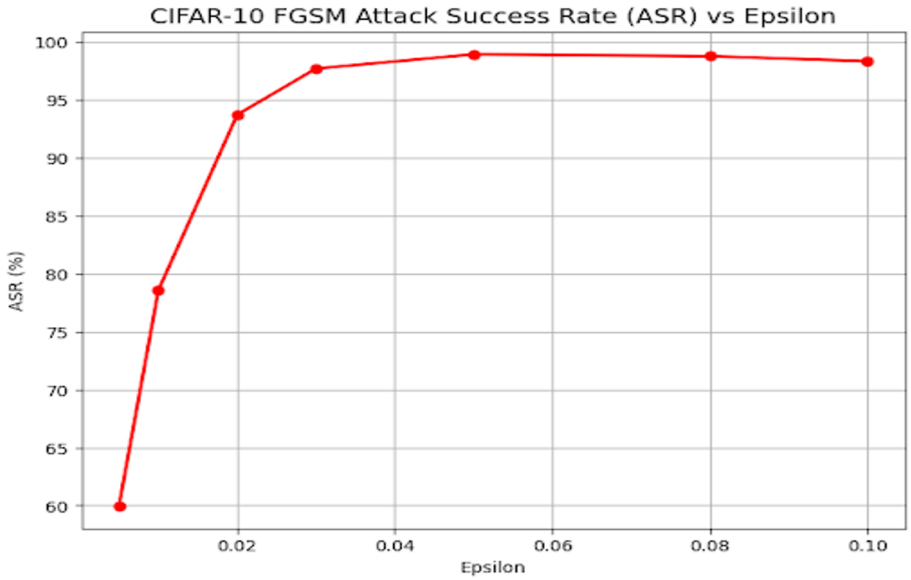

Fig. 4 : Attack Success rate on CIFAR-10

4.4 CIFAR-10 Adversarial Training Performance Analysis

The combined accuracy on Clean and FGSM for CIFAR-10 across different adversarial training ratios (50/50, 30/70, 25/75 and 60/40) and a wide range of FGSM perturbation strengths (ϵ = 0.05–0.5) shown in Fig. 5. The combined accuracy reflects how well the model performs when evaluated on combination of clean samples and FGSM generated adversarial samples using the same ratio as used for training.

Key findings include:

1. CIFAR-10 behaves differently from MNIST

Unlike MNIST, where accuracy generally decreases with higher epsilon, CIFAR-10 shows a non-monotonic behavior. For small epsilons (ϵ = 0.05–0.1), accuracy is low across all ratios (10–20%), indicating extremely high vulnerability. However, accuracy increases significantly for medium to high epsilons (ϵ = 0.2–0.5).

2. 50/50 Ratio Performs the Most Consistently

The blue curve (50/50) shows:

- Strong improvement from 46% → 63% (ϵ = 0.05 → 0.1)
- Peak performance at ~71% accuracy for ϵ = 0.5

- Smooth, stable trend across all epsilons
- No sudden drops compared to 25/75

This makes 50/50 the most balanced and stable ratio for CIFAR-10 adversarial training.

3. 30/70 Ratio Peaks at Medium Epsilons

The orange curve (30/70):

- Low accuracy at small epsilons (17–20%)
- Sharp rise at medium epsilons (ϵ = 0.15–0.25)
- Reaches 70% combined accuracy at ϵ = 0.2
- Slight drop at high epsilon (65–68% at ϵ = 0.4–0.5)

This suggests that training with more adversarial samples improves robustness for mid-range perturbation.

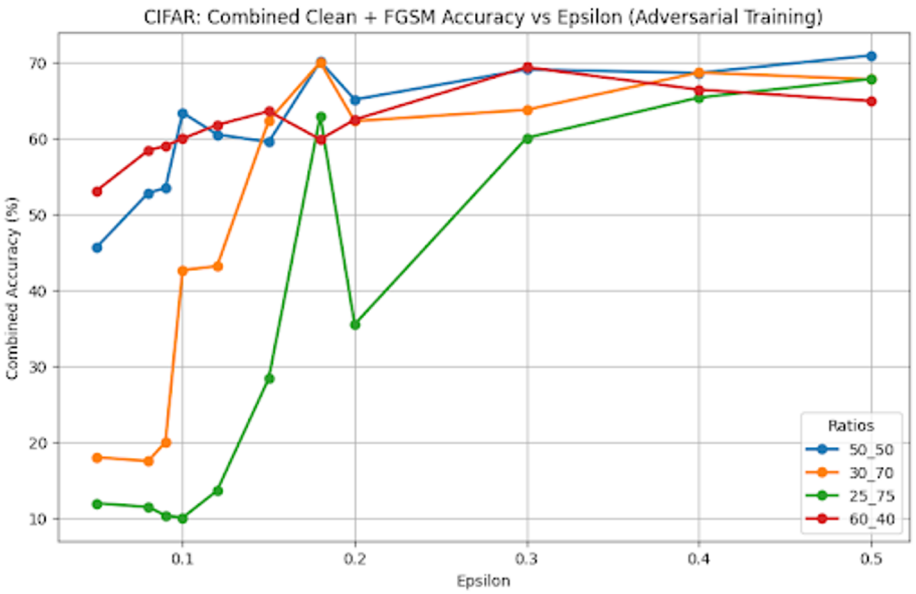

Fig. 5: Accuracy vs Epsilon of Adversarial Training on CIFAR-10 Dataset

4. 25/75 Ratio is highly Unstable

The green curve (25/75):

- Extremely low accuracy (10–12%) at ϵ = 0.05–0.1
- Sudden spike (63%) at ϵ = 0.18
- Drops again at ϵ = 0.2 (35%)
- Then rises to 60–67% at high epsilons

5. 60/40 Ratio Preserves Clean Accuracy but Has Slightly Lower FGSM Robustness

The red curve (60/40):

- Best performance at small epsilons (53–60%)
- Smooth rising trend
- Peaks at ~69% around $\epsilon = 0.3–0.4$
- Slight decline at $\epsilon = 0.5$

4.5 CIFAR-10 FGSM Attack Success Rate (ASR) Analysis

The Attack Success Rate (ASR) of FGSM was trained on CIFAR-10 dataset on model across a wide range of perturbation strengths (ϵ = 0.05 to 0.5) and four training ratios (50/50, 30/70, 25/75, 60/40) shown in Fig. 6.

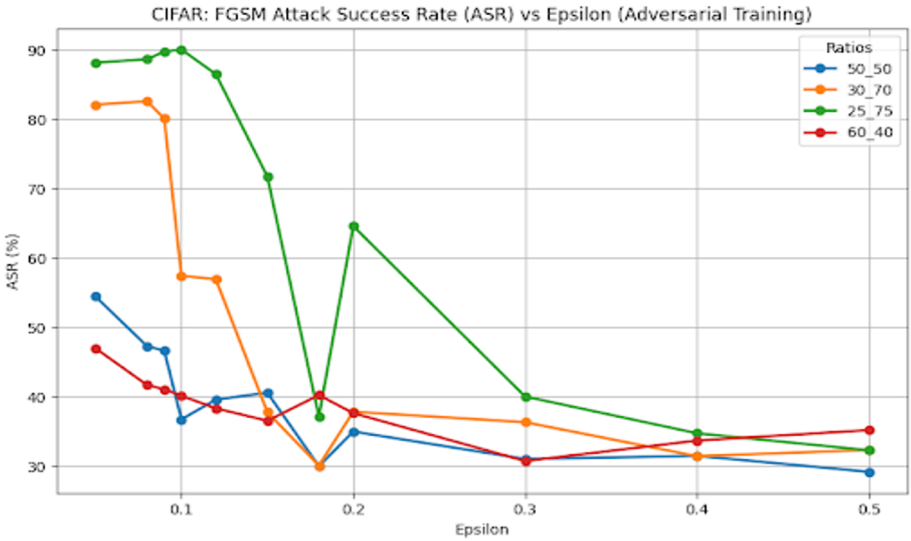

Fig. 6: ASR vs Epsilon of Adversarial Training on CIFAR-10 Dataset

Key Observations from the ASR Curves:

1. All ratios show extremely high ASR at low epsilons (0.05–0.1)

- 50/50 → 47–55% ASR
- 30/70 → 80–82% ASR
- 25/75 → 88–90% ASR
- 60/40 → 40–47% ASR

This confirms that CIFAR-10 models remain highly vulnerable to FGSM even after adversarial training, especially for shallow perturbations. This is consistent with prior

work showing that single-step attacks exploit linearity in high-dimensional space, making CIFAR significantly more vulnerable to adversarial noise than MNIST.

2. ASR drops sharply for mid-range epsilons (0.15–0.3)

All curves show a distinct downward crash, indicating that: FGSM adversarial training becomes more effective when epsilon increases.

For example:

- For 50/50, ASR drops from 39% → 30%
- For 30/70, ASR drops from 57% → 37%
- For 25/75, ASR plummets from 87% → 40

Thus, higher ASR means stronger attack success, while lower ASR means better model robustness.

5. Conclusion

This study demonstrated the effectiveness of FGSM based adversarial training in enhancing robustness. It also demonstrates how convolutional neural networks using gradient based adversarial attacks method evaluated on the MNIST and CIFAR-10 datasets improve model robust. Training with a mixture of clean and FGSM-perturbed samples across different ε values, the proposed approach enables CNNs to learn more stable and robust feature representations with consistent performance. The experimental results also show that the choice of clean-to-adversarial data ratio plays a critical role in achieving balance between adversarial robustness and clean accuracy. Ratios such as 50/50 and 60/40 were particularly effective in balancing clean accuracy with adversarial accuracy. Although FGSM is a gradient-based attack strategy, adversarial training with FGSM significantly reduces vulnerability of the model, as its computational efficiency makes it well suited for practical applications. From the observed study, FGSM-based adversarial training remains limited in its ability to defend against stronger multi-step adversarial attacks. Future work can involve exploring multi-step adversarial training and exploring additional defense techniques to further enhance model accuracy and robustness.

References

[1] Abomakhelb, A., Jalil, K.A., Buja, A.G., Alhammadi, A., Alenezi, A.M.: A comprehensive review of adversarial attacks and defense strategies in deep neural networks. In: MDPI *Technologies* 13(5), Article 202 (2025). https://doi.org/10.3390/technologies13050202.

[2] Goodfellow, I.J., Shlens, J., and Szegedy, C.: Explaining and harnessing adversarial examples. In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)* (2015). https://doi.org/10.48550/arXiv.1412.6572.

[3] Biggio, B., Roli, F.: Wild patterns: Ten years after the rise of adversarial machine learning. In: *Proceedings of the 25th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 2154–2156 (2018). https://doi.org/10.48550/arXiv.1712.03141.

[4] Wen, X., Danso, E., Danso, S.: Improving security-sensitive deep learning models through adversarial training and hybrid defense mechanisms. In: *Journal of Cybersecurity* 7(1), 45–69 (2025). https://doi.org/10.32604/jcs.2025.063606

[5] Michael Gallagher, Nikolaos Pitropakis, Christos Chrysoulas, Pavlos Papadopoulos, Alexios Mylonas and Sokratis Katsikas.: Investigating machine learning attacks on financial

time series models. In: *Elsevier Computers & Security*, Volume 123, 2022. https://doi.org/10.1016/j.cose.2022.102933.
[6] Arthur Dantas Mangussi, Ricardo Cardoso Pereira, Ana Carolina Lorena, Miriam Seoane Santos, Pedro Henriques Abreu.: Studying the robustness of data imputation methodologies against adversarial attacks. In: *Elsevier Computers & Security*, Volume 157, Issue C, 2025, https://doi.org/10.1016/j.cose.2025.104574.
[7] Grosse, K., Papernot, N., Manoharan, P., Backes, M., and McDaniel, P.: On the (Statistical) detection of adversarial examples. In: *arXiv preprint* arXiv:1702.06280 (2017). https://doi.org/10.48550/arXiv.1702.06280.
[8] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R.: Intriguing properties of neural networks. In: *arXiv preprint* arXiv:1312.6199 (2013). https://doi.org/10.48550/arXiv.1312.6199.
[9] Moosavi-Dezfooli, S.M., Fawzi, A., and Frossard, P.: DeepFool: A simple and accurate method to fool deep neural networks. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2574–2582 (2016). https://doi.org/10.1109/CVPR.2016.282.
[10] Kurakin, A., Goodfellow, I., and Bengio, S.: Adversarial Machine Learning at Scale. In: *Proceedings of the 5th International Conference on Learning Representations (ICLR)* (2017). https://doi.org/10.48550/arXiv.1611.01236.
[11] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)* (2018). https://doi.org/10.48550/arXiv.1706.06083.
[12] Carlini, N., and Wagner, D.: Towards evaluating the robustness of neural networks. In: *Proceedings of the 2017 IEEE Symposium on Security and Privacy*, pp. 39–57 (2017). https://doi.org/10.1109/SP.2017.49.
[13] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., and Swami, A.: The limitations of deep learning in adversarial settings. In: *Proceedings of the 2016 IEEE European Symposium on Security and Privacy*, pp. 372–387 (2016). https://doi.org/10.1109/EuroSP.2016.36
[14] Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses. In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)* (2018). https://doi.org/10.48550/arXiv.1705.07204
[15] Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*, PMLR, vol. 80, pp. 274–283 (2018). https://doi.org/10.48550/arXiv.1802.00420.
[16] Croce, F., and Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: *Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, vol. 119, pp. 2206–2216 (2020). https://doi.org/10.48550/arXiv.2003.01690.*
[17] Eleftheriadis, C., Symeonidis, A., Katsaros, P.: Adversarial robustness improvement for deep neural networks. In: ACM Digital Library *Machine Vision and Applications* 35(3), Article 35 (2024). https://doi.org/10.1007/s00138-024-01519-1.
[18] Rong, S., Hamdan, E., Cetin, A.E.: Multi-resolution training improves robustness against adversarial attacks. In: Springer *Signal, Image and Video Processing* 19, Article 481 (2025). https://doi.org/10.1007/s11760-025-04038-2.
[19] Kafali, E., Semertzidis, T., Daras, P.: Dynamic trade-offs in adversarial training: Exploring efficiency, robustness, forgetting, and interpretability. In: Springer *Neural Processing Letters* 57(3), Article 47 (2025). https://doi.org/10.1007/s11063-025-11751-z.
[20] Villegas-Ch, W., Jaramillo-Alcázar, A., Luján-Mora, S.: Evaluating the robustness of deep learning models against adversarial attacks: An analysis with FGSM, PGD and CW. In:

MDPI *Big Data and Cognitive Computing* 8(1), Article 8 (2024). https://doi.org/10.3390/bdcc8010008.
[21] Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. In: *Proceedings of the ICLR 2017 Workshop* (2017). https://doi.org/10.48550/arXiv.1607.02533.
[22] Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey. In: *IEEE Access* 6, 14410–14430 (2018). https://doi.org/10.1109/ACCESS.2018.2807385
[23] Wong, E., Rice, L., and Kolter, J.Z.: Fast is better than free: Revisiting adversarial training. In: *Proceedings of the 8th International Conference on Learning Representations (ICLR)* (2020). https://doi.org/10.48550/arXiv.2001.03994.
[24] Burkule, M.: Robustness of adversarial defenses in computer vision. In: *Technical Report*, ResearchGate (2022). https://doi.org/10.13140/RG.2.2.21521.61288.
[25] Kingma, D.P., and Ba, J.: Adam: A method for stochastic optimization. In: *Proceedings of the 2nd International Conference on Learning Representations (ICLR)* (2015). https://doi.org/10.48550/arXiv.1412.6980
[26] Nair, V., and Hinton, G.: Rectified linear units improve restricted Boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 807–814 (2010). https://doi.org/10.5555/3104322.3104425
[27] Wang, S., Wang, X., Zhao, P., Wen, W., Kaeli, D., Chin, P., Lin, X.: Defensive dropout for hardening deep neural networks under adversarial attacks. In: *Proceedings of the 2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Article 71, pp. 1–8 (2018). https://doi.org/10.1145/3240765.3264699
[28] Ding, J., Zhao, J.-C., Sun, Y.-Z., Tan, P., Ma, J.-E., and Fang, Y.-T.: Improving the robustness of deep convolutional neural networks through feature learning. In: *arXiv preprint* arXiv:2303.06425 (2023). https://doi.org/10.48550/arXiv.2303.06425.