Computer Fraud and Security
ISSN (online): 1873-7056

Securing AI Systems Against Adversarial Attacks: A Framework for Building Robust and Trustworthy Machine Learning Models

Hamza Afzal
Software Engineer
American Technology Group LLC
Baltimore Maryland
hafzal.student@wust.edu

Malik Huzaifa
System Administrator
FNR Solutions INC
Baltimore Maryland
Hmalik.student@wust.edu

***Abstract*—**Harnessing AI systems with adversarial attacks has turned out to be an issue of urgent concern as machine learning models continue to work in high-stakes settings. This research suggests an alternative and all-in-one adversarial defense, which incorporates safe data preprocessing, CNN-based feature extraction, and iterative adversarial retraining to improve the robustness and reliability of the model. The framework is tested on the MNIST data set and also includes adversarial samples created by using FGSM, PGD, BIM, and C&W. The retraining cycle allows the network to acquire more stable decision boundaries by repeatedly exposing the model to changing perturbations and helping to counter weaknesses both on white and gray-box threat conditions. Experimental findings indicate a great increment of robustness, accuracy upgrades between 7% and over 97% following retraining across all types of attack. The model is clean with an accuracy of 99.1 %, and it performs better than the current methods, including conventional adversarial training and AEDPL-DL. Comparative evaluation proves the fact that iterative retraining of adversarial models is more resilient to data poisoning, evasion, and gradient-based attacks. The suggested solution represents a promising avenue for creating secure, attack-deterrent, and trustworthy machine learning systems that are applicable in the real world.

Keywords*—Adversarial Attacks, Machine Learning, Convolutional Neural Networks (CNNs), Adversarial Defense, Data Poisoning, White-Box Attacks, Grey-Box Attacks, Model Robustness.*

I. INTRODUCTION

AI and ML have been integrated into the new digital ecosystem in the form of the backbone of the contemporary healthcare diagnostics application, financial analytics, cybersecurity monitoring, autonomous transportation, and massive enterprise-level systems [1]. Their ability to handle complex information, identify latent trends, and provide quick and automatic determination of information has made them valuable assets in a high-stakes operational setting. The stable and reliable operation of models as an organization leans more on AI-driven systems to guarantee accuracy, efficiency, and reliability, which has become a pressing need to ensure safety, economic stability, and continuity of services [2].

Although effective when the conditions are standard, ML models are extremely vulnerable to adversarial interference. Tiny, well-designed perturbations, which may be completely invisible to human perceptions, can cause large misclassification of otherwise correct models [3], [4]. The Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), Basic Iterative Method (BIM), and Carlini and Wagner (C&W) attack are examples of well-known adversarial attack strategies that exploit gradient-sensitive design flaws to achieve significant failures [5]. Medical imaging classifiers can be misled, or fraud detectors bypassed, or autonomous vehicles may be affected by these attacks to alter perceptions of traffic signs [6], [7]. These vulnerabilities are potentially very dangerous, showing that even state-of-the-art classifiers can be compromised in adversarial conditions.

The standard training processes have not been trained to be resistant to these targeted disorders [8]. There is a considerable difference in robustness of security-sensitive models when adversarial manipulated inputs are presented to models trained on a clean distribution only, showing a major gap in generalization [9]. Current methods offer partial defines but are often limited in scale, attack-modes, and attack modes, and more holistic and versatile defensive methods are required.

To solve this issue, the current article proposes a framework of adversarial defines that combines secure pre-processing, a CNN-based classification model, and adversarial retraining. The framework would reinforce decision boundaries through the creation of controlled adversarial samples and embed the samples into several training cycles, which makes it resilient to both white-box and Gray-box attacks [10]. It is expected to establish a defines mechanism that maintains high clean accuracy but with substantially greater resistance to a variety of adversarial perturbations, and so enables the use of AI systems in mission-critical settings with greater reliability and security.

A. Motivation and Contribution of the Paper

With AI-powered systems fully integrating into the stakes of various areas like healthcare diagnostics, financial safety, biometrics, and automated navigation, the issue of reliability in adverse settings has been a major concern. Even the slightest, indistinguishable changes can cause machine learning models to misclassify their inputs, undermine safety measures, and be maliciously exploited. Traditional defensive mechanisms tend to be reactive and insufficient to scale to new attack vectors, and leave the state-of-the-art ML pipeline vulnerable to advanced adversarial manipulation. This increasing weakness encourages the necessity of smart, adaptive, and strong defense systems that can withstand various threat conditions. To sustain the trust, guarantee the operational stability, and maintenance of safe deployment of AI systems in the environment where accuracy and security should be provided simultaneously, it would be vital to improve adversarial resilience. The major contributions of the paper are as follows:

- Use of a well-curated and structured dataset to simulate clean inputs and generate adversarial samples systematically by having a controlled assessment of vulnerabilities to the model in a variety of threat cases.
- Creating an end-to-end data processing pipeline with functionalities such as normalization, sensitive pre-processing, and generation of adversarial perturbations to prepare the data to be used to train a robust model.
- Combining a CNN-based classification model as the main learner and making it more resilient by retraining it via the adversarial method iteratively.
- Adopting an adaptive adversarial retraining mechanism, which involves the use of adversarial samples across several training steps, can achieve learning stability on decision boundaries and enhance resilience.
- Assessing model resilience to both simple and highly advanced perturbations by observing model performance on various adversarial attacks, including FGSM, PGD, BIM, and CW.
- Comparison of the suggested method to the existing techniques and proving better clean accuracy and much higher adversarial robustness.

B. Justification and Novelty of the Paper

The article is supported by the fact that quick, reliable AI systems that resist advanced, adaptive, adversarial threats are urgently needed. This study presents a dynamic iterative retraining process as opposed to the traditional training processes, which depend on fixed adversarial samples, and as such, new adversarial samples emerge and are included throughout, preserving robustness in a continuously evolving fashion. The novelty lies in the fact that clean, noisy, and multi-attack samples are combined into one training loop that has a robustness threshold, which results in a stable performance regardless of the type of attack. This multi-layered upgrading of the existing defense strategies offers greater generalization and enhanced resilience.

C. Structure of the Paper

The paper will be organized in the following way: Section II will provide a review of the current studies. Section III describes the suggested methodology, which consists of the dataset preparation, secure preprocessing, and architecture design. Section IV includes the experimental configuration, assessments, and comparison outcomes in several scenarios of adversarial attacks. The final section of Section V finishes with an overview of the main findings and future research prospects.

II. LITERATURE REVIEW

Adversarial attacks threaten the AI system's reliability. Recent research examines various defense mechanisms, exploiting machine learning methods to identify, reduce, and respond to cyber threats to prove dependable and reliable smart systems.

Sanapala, Lavanya (2024) introduces the ML Filter, which is a new method that incorporates the aspects of security in machine learning so as to detect and eliminate the known and unknown threats efficiently. The Statistical Perturbation Bounds Identification Algorithm and the ML-Filter Detection Algorithm are used to assess if a dataset is contaminated. The data is divided using DBSCAN so that an algorithm can analyze it. The research considers its performance through the True Positive Rate, significance test accuracy, which the performance of detection depends on the perturbation size and not on the dataset or the ML models used. ML Filter is also able to detect known attacks with a rate of 99.63% and a generalized rate of 98%, which shows that there is a lot of progress in machine learning and deep learning algorithms [11].

Villegas-Ch et al. (2024) evaluated the VGG16 image classification model's resistance and adverse example creation. To assess the impact on the classification accuracy, they employed techniques to attack the original images, such as the Carlini and Wagner attack, the Projected Gradient Descent method, and the Fast Gradient Sign method. Because the average accuracy decreased by 25% when attacked by the Fast Gradient Sign and Projected Gradient Descent assaults and by 35% when attacked using the Carlini and Wagner approach, the study found that the VGG16 model was vulnerable to adversarial cases. As potential defences against these hostile threats, picture manipulation techniques like noise reduction, image compression, and Gaussian blurring were also investigated [12].

Wibawa (2023) examines the security concerns and attacks on AI systems as adversarial threats. The study is based on AI models simulating and testing their robustness using Python programming and libraries like Clever Hans, which is suited to evaluate AI security. The authors point out that any little tampering in the input data can have severe impacts on AI

predictions, as the FGSM model shows such a sharp decline in accuracy of approximately 66% at epsilon of 0.1 when attacked. These vulnerabilities are important to understand in order to protect the progress of AI technology. Through a thorough insight into the essence of AI attacks and the security issues that accompany them [13].

Zhu, Zhang, and Chen (2023) provide AI-Guardian, a novel approach to defeating adversarial attacks that employs purposefully placed backdoors to crash the adversarial disturbance while maintaining the functionality of the initial primary job. They evaluate AI-Guardian severely using five popular adversarial example generation techniques, and the experiment's outcomes demonstrate how well it withstands adversarial attacks. Specifically, compared to the state-of-the-art works, the attack's success rate drops by 30.9% rather than 97.3%, but only the correctness of clean data drops by 0.9%. Moreover, AI-Guardian adds minimum overhead to the model prediction time; it is 0.36% of the model prediction time, which is nearly insignificant in the majority of instances [14].

Anastasiou et al. (2022) provide a novel AI architecture that will enhance AI security and dependability by integrating defense algorithms and adversarial cases. Enhancing deep neural network (DNN) classifiers—primarily convolutional neural networks (CNNs)—under challenging manufacturing settings that include noise, vibrations, and data transfer mistakes is its focus. The architecture promotes the dynamic process between the attacker and the defender by training adversarial, evaluating defense algorithms, and a multiclass discriminator to distinguish between the attacked and the non-attacked data. As shown in the experimental findings, the defense algorithms and multiclass discriminator work well together to rejuvenate the weakened models and to build a stronger DNN classifier [15].

Lin (2022) explores ways of making AI models more resilient to data poisoning by focusing on threat taxonomies like clean-label poisoning, backdoor insertion, and gradient-based adversarial contamination. The paper has assessed defensive practices in three phases: data sanitization and anomaly detection in the pre-training phase, robust optimization in the model development phase, and runtime monitoring in the post-training phase. It proposes hybrid solutions as a combination of effective statistical learning with uncertainty estimation and federated data verification to remove points of failure. Also, it highlights the value of explainability, accountability, and constant validation to promote trust in settings prone to poisoned data. It concludes that a multi-layered approach of interventions in all stages is imperative to reduce risks [16].

Table I describes the comparison between existing studies on adversarial attacks in AI systems, based on approaches, findings, advantages, limitations, and future work.

TABLE I. COMPARATIVE SUMMARY OF EXISTING ADVERSARIAL ATTACK AND DEFENSE RESEARCH

References	Approach	Key Findings	Advantages	Limitations	Future Work
Sanapala, Lavanya (2024)	A proposed ML-Filter for detecting gradient-based data poisoning in industrial ML systems using DBSCAN, a detection algorithm, and a statistical perturbation bounds identification algorithm.	Achieved 99.63% detection for known and 98% generalized detection for unknown attacks; effectiveness depends on perturbation size rather than dataset or ML model.	High detection accuracy; generalized detection capability; statistically grounded method.	Limited scalability across large, real-time data streams; performance under dynamic attack scenarios untested.	Extend ML-Filter to real-time adaptive detection frameworks and explore integration with federated learning security.
Villegas-Ch et al., (2024)	Evaluated FGSM, PGD, and Carlini & Wagner attacks on image classification models (VGG16) and tested image manipulation defenses (noise reduction, compression, blurring).	Observed 25–35% accuracy drops under attacks; image manipulation defenses partially recovered accuracy.	Demonstrated comparative robustness of image manipulation techniques; practical evaluation across common attack methods.	Defenses are not model-agnostic; limited to the image domain; performance trade-off between defense strength and accuracy.	Develop adaptive multimodal defense techniques applicable across image, text, and signal domains.
Wibawa (2023)	Simulated adversarial attacks using the CleverHans library with FGSM ($\varepsilon = 0.1$) to assess AI model vulnerability.	Found a 66% accuracy drop under attack, emphasizing neural network vulnerability to small perturbations.	Demonstrated effectiveness of simple attacks; emphasized awareness of AI security threats.	Focused on basic attacks; lacked comprehensive defense evaluation and quantitative	Design automated defense frameworks integrated into model training pipelines to

				mitigation strategies.	mitigate gradient-based attacks.
Zhu, Zhang & Chen (2023)	Introduced AI-Guardian, embedding intentional backdoors to neutralize adversarial perturbations while preserving main task performance.	Reduced attack success rate from 97.3% to 3.2% with a negligible 0.9% accuracy loss on clean data.	High defense efficacy with minimal computational overhead; novel use of intentional backdoors.	Potential ethical and security risks with embedded backdoors; evaluation limited to specific attack models.	Investigate secure backdoor embedding protocols, ensuring transparency and preventing misuse.
Anastasiou et al. (2022)	To make convolutional neural networks (CNNs) more resilient in noisy industrial settings, they built an AI architecture using adversarial examples and defense mechanisms.	Hybrid defense and multiclass discriminator improved model robustness and accuracy under adversarial conditions.	Practical validation in real manufacturing data; dual attacker–defender simulation.	Limited generalization beyond industrial settings; computationally intensive; requires domain-specific tuning.	Extend hybrid defense mechanisms to cross-domain applications and improve computational efficiency.
Lin (2022)	Comprehensive review of data poisoning and adversarial resilience across pre-, in-, and post-processing stages; focused on hybrid defense integration.	Layered defense strategies combining robust learning, uncertainty estimation, and federated validation effectively reduce risks.	Provides a holistic resilience framework combining technical and governance measures.	Lacks empirical implementation; primarily conceptual; limited quantitative validation.	Develop empirically validated hybrid frameworks integrating governance, explainability, and real-world testing.

III. Methodology

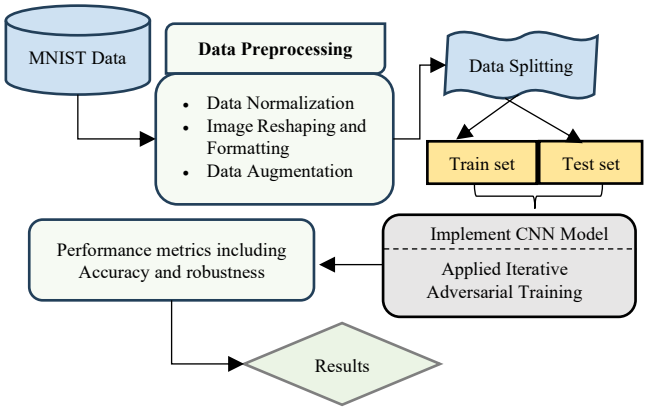

Fig. 1. Methodological Process for the Proposed Framework with Integrated Adversarial Defense Mechanisms

The methodology describes a systematic approach that aims to improve the machine learning model's robustness to adversarial attacks with secure pre-processing, adversarial training, and strict evaluation. The suggested framework starts with the MNIST dataset that is subjected to systematic data pre-processing in order to enhance the quality of inputs and resilience. Pre-processing involves data normalization, image reshaping and formatting, and focused data augmentation to improve generalization. The dataset is then split into training and testing portions after it is prepared so that performance may be evaluated fairly and reliably. The main architecture to be used in image classification is the Convolutional Neural Network (CNN) model. In order to make the models robust, they use iterative adversarial training, whereby adversarial examples are created and added to the training loop. This procedure subjects the model to malicious Peruke, which allows the model to be trained to learn more robust decision boundaries. The structure also calculates important performance measures like accuracy, robustness scores, and stability when subjected to adversarial influence to determine the reliability of the model. The findings eventually show that the suggested secure learning pipeline is effective in creating attack-resistant and trustworthy AI systems—Fig. 1 Proposed Framework of Integrated Adversarial Defense Mechanisms.

A. Data Collection

The MNIST dataset[1] consists of a total of 70,000 grayscale images of handwritten digits (between 0 and 9), each image is 28x28 pixels in size. Each image is classifiable into one of ten different numeric classes, with the value of each pixel intensity falling between 0 and 255, which are different degrees of gray. The dataset is balanced with almost an equal representation of samples in each digit category, so there was an equal distribution of samples per category. All the images have been scaled to be of equal size and positioned in the middle to ensure uniformity. Fig. 2 displays the Data samples:

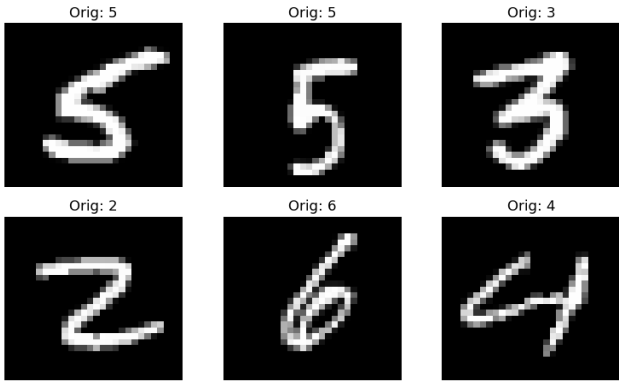

Fig. 2. Data Samples

B. Data Preprocessing

The pre-processing step aims at cleaning the raw data, enhancing its clarity, consistency, and formatting to make sure that the model receives cleaner data that is more reliable, and it brings about better training and higher robustness.

C. Normalization

The min-max normalization technique begins with data normalization. The smallest value for each characteristic or pixel is converted to 0, the greatest value to 1 [17], and all other values are converted to a decimal between 0 and 1. Equation (1) is used to implement the min–max normalization procedure:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

D. Image Reshaping and Formatting

The purpose of this step is to alter the form of the input data to suit the form needed by the neural network. The images of MNIST are originally in a 28x28 pixel format, which is converted to a channel format to give the height of the image 28x28x1. This is to give the grayscale channel a precise definition so that it can be used with convolutional layers. Also, the digit labels are transformed to one-hot coded vectors so that the model can easily do multi-class classification among the ten digit categories.

E. Data Augmentation

In machine learning, data augmentation is a significant approach to artificially increasing the size and variety of a dataset and does not require the collection of new data. It entails the use of different transformations on an existing data sample to generate a new, slightly different sample, but without changing its original label or meaning [18]. This can assist in generalization of the model, decrease overfitting as well, and lead to improved performance on unseen data. Geometric transformations (like rotation, flipping, scaling, and cropping), color space manipulation, and noise-based manipulation are some of the common techniques of data augmentation [19]. A commonly used method of noise is Gaussian Noise Augmentation, in which the input data is corrupted with random noise that has a Gaussian (normal) distribution. This can be mathematically written as shown in Equation (2)

$$x' = x + \mathcal{N}(0, \sigma^2) \quad (2)$$

In this case, x is a given pixel value, and $\mathcal{N}(0, \sigma^2)$As a Gaussian noise whose mean is 0 and the variance is s 2. And after the data samples have some noise, which is presented in Fig. 3:

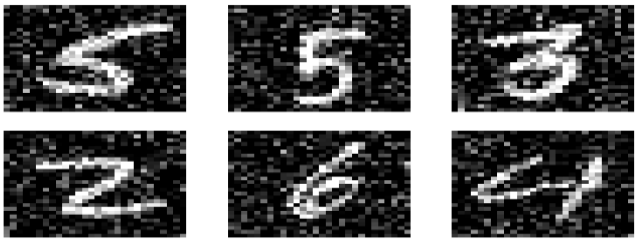

Fig. 3. Data Samples After Augmentation

F. Data Splitting

The dataset is separated into subsets for testing and training, with 20% set aside for performance and generalization accuracy evaluation, and 80% of the data used for model training.

G. Proposed Framework Architecture

Convolutional Neural Network (CNN) with the ability to generate meaningful spatial features of the data. It is characterized by the two convolutional layers, which have 32 and 64 filters (3x3), along with the ReLU activation and max-pooling layer (2x2) to minimize the spatial dimension and calculations. The dropout (0.25) is added in a way to avoid overfitting, and a 1024 neuron fully connected layer with ReLU activation is used to learn the features [20] further. The output layer has 11 neurons (10 digit classes and 1 adversarial class), which use the softmax activation to produce class probabilities. To minimize the categorical cross-entropy loss function, the network parameters are optimized, which is represented by Equation (3)

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log\left(\hat{y}_{ic}\right) \qquad (3)$$

y_{ic}Is the actual label, $\hat{y}_{ic}$He forecasted probability, and C is the overall count of classes [21]. Softmax activation used in the output layer transforms the logits into normalized probabilities. Its mathematical form is given by Equation (4)

$$\hat{y}_{ic} = \frac{e^{z_{ic}}}{\sum_{k=1}^{C} e^{z_{ik}}} \qquad (4)$$

Assuming that z_{ic}Is the result of the final fully connected neuron on class c. The design guarantees its effectiveness in extracting features, classification, and adversarial retraining.

H. Model Training

The training stage is aimed at optimization of constructed CNN with clean and noisy samples, together with adversarial samples, to increase the accuracy and resilience of the classification. The model is assembled using the Adam optimizer and trained in mini-batches comprising an equal representation of all data types. One-hot labels are also applied, but with soft labels that allow uncertainty and minimize overfitting. New adversarial samples are generated with FGSM, PGD, BIM, and C&W attacks with every training iteration and concatenated with the training set to undergo iterative adversarial retraining. The goal function reduces the weighted average of clean and adversarial losses, which enables the model to increase its resistance to perturbations automatically. The training process will continue until the robustness threshold ($\rho \geq 0.1$) or the upper iteration limit (kmax = 3000) is met, such that the final model can be highly accurate and have strong resistance to attack-related conditions.

I. Evaluation Metrics

The proposed model is evaluated in terms of standard classification and adversarial robustness. The percentage of accurately classified cases, both (benign) and (adversarial), with respect to all instances, as shown in Equation (5):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \qquad (5)$$

The values TP, TN, FP, and FN stand for the following: true positives, true negatives, false positives [22], and false negatives, respectively. The measure of robustness in the presence of adversarial perturbation is the robustness metric (r), which calculates the difference between model accuracy (pre-adversarial retraining) and that under adversarial retraining. The mathematical form represented by Equation (6).

$$\rho = Accuracy_{after} - Accuracy_{before} \qquad (6)$$

The larger the ρvalues, the more resistance to adversarial attacks there is. There are also two evaluation scenarios defined, including White-Box Attacks, where the attacker is fully aware of the parameters and gradients of the model[23], as well as Gray-Box Attacks, where the attacker only knows partial information about the model. These types of attacks assist in testing the model concerning its strength against various levels of threats.

IV. Result and Discussion

The proposed CNN-based adversarial robust framework was evaluated on the MNIST dataset through experimental evaluation. The system was trained and tested on a GPU-enabled platform through repeated adversarial retraining on FGSM, PGD, BIM, and C&W attacks. The findings indicate that adversarial retraining significantly increases the accuracy and the robustness of the classification.

TABLE II. Classification Accuracy Under White-Box Attacks

Attack Type	Before Retraining (%)	After Retraining (%)
FGSM	28.7	98.6
PGD	32.4	98.1
BIM	26.5	98.3
C&W	7.0	96.9

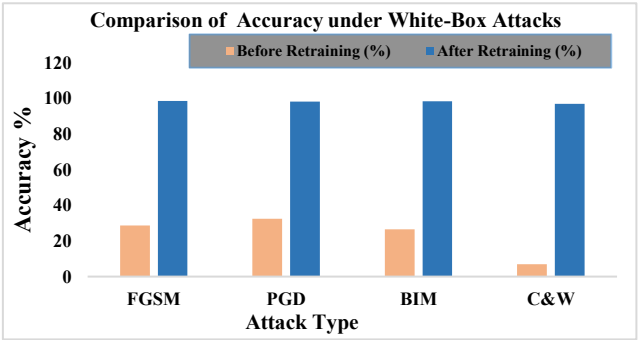

Fig. 4. Classification Accuracy under White-Box Attacks

Fig. 4 and Table II show that the proposed iterative adversarial retraining framework has contributed to a major improvement in model robustness. The chart presents the accuracy of the model in percentage across four types of white-box adversarial attack(where the attacker is fully aware of the model): FGSM, PGD, BIM, and C&W. The model is highly vulnerable. The accuracy of the model before retraining (light orange bars) is much lower, i.e., around 7% (C&W) and 33% (PGD). Nonetheless, the accuracy of the model increases significantly to almost 100% against all four attacks by implementing the suggested retraining method (dark blue bars), which proves the usefulness of the framework in ensuring AI systems are resistant to adversarial attacks and construct sound machine learning models.

TABLE III. Classification Accuracy Under Gray-Box Attacks

Attack Type	Before Retraining (%)	After Retraining (%)
FGSM	11.2	97.9
PGD	9.3	97.9
BIM	4.5	98.5
C&W	8.8	97.0

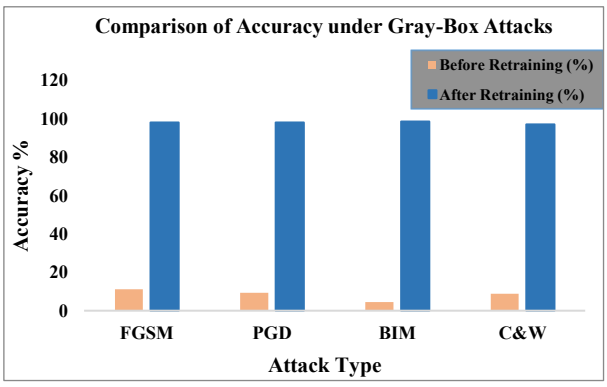

Fig. 5. Classification Accuracy under Gray-Box Attacks

Fig. 5 and Table III successfully illustrate the enhanced resiliency of the suggested CNN-based iterative adversarial retraining system to gray-box attacks, whereby the attacker partially knows about the model. The accuracy percentage of the

model is demonstrated on four attacks, namely FGSM, PGD, BIM, and C&W. The model, when retrained (light orange bars), has a very low accuracy, approximately 5% to 10% in all types of attacks, meaning that there is a high vulnerability to it. After the application of the proposed retraining method (dark blue bars), the accuracy of the model becomes almost 100% in all four gray-box attacks. This demonstration shows that the framework is quite efficient in protecting AI systems, as well as establishing strong and reliable machine learning models, even in the case when the adversary partially knows the system.

TABLE IV. PERFORMANCE EVALUATION OF THE PROPOSED FRAMEWORK

Metric	Value
Clean Accuracy	99.1
Average Accuracy (White-Box)	97.98
Average Accuracy (Gray-Box)	97.83
Robustness (ρ)	0.12

Table IV is a summary of the proposed defense framework in various evaluation environments. It claims a high clean accuracy of 99.1, and high resilience of both white and gray box attack situations. The score of robustness (0.12) also demonstrates how the model is robust to adversarial perturbations. The accuracy and loss curve depicted by Fig. 6:

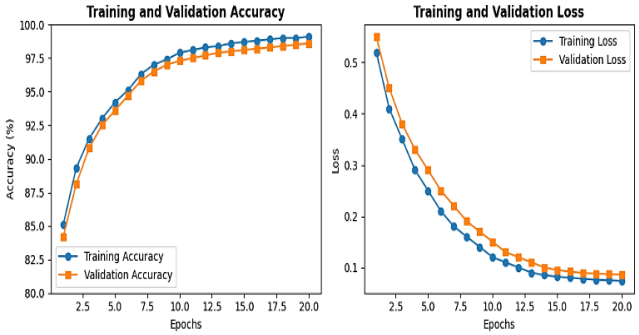

Fig. 6. Training and Validation Accuracy and Loss of the Proposed Framework

Fig. 6 presents evidence of the efficiency and sustainability of the suggested iterative adversarial retraining framework. The charts demonstrate the training process of 20 epochs: the Training and validation accuracy graph depicts a steady and consistent rise in both metrics, to almost 99%, which means that the model is learning without much overfitting. In the same vein, the Training and Validation Loss graph has a converging downward trend of smooth decreasing curves, to a low point. This great convergence is further substantiated by the quantitative results table, which indicates a high Clean Accuracy of 99.1%. Most importantly, the model operates at a high level even in adversarial environments with an average accuracy of 97.98% in White-Box and 97.83% in Gray-Box attacks. The low Robustness (0.12) value upholds the fact that the model is not sensitive to the adversarial perturbations, which leads to its high robustness and reliability to wide adversarial attacks.

A. Comparative Analysis and Discussion

The comparative analysis indicates that the suggested framework of Iterative Adversarial Retraining is valuable in increasing model robustness to adversarial threats. As demonstrated in Table V, the proposed method had the best classification accuracy of 99.1%, which was higher than Adversarial Training Technique (98.5%) and AEDPL-DL (98.62%). This gain shows the robustness of the iterative retraining approach, in which the model is presented with freshly drawn adversarial samples until it reaches some specified robustness level. Training loop with clean, noisy, and adversarial data guarantees high generalization and stability both in a white-box and gray-box attack adversarial regime. The low but steady improvement in performance over current defense mechanisms validates that retraining can not only counter gradient-based attacks but also increase robustness to data poisoning and data evasion attacks, which makes retraining a more reliable and resilient AI model to apply AI in secure deep learning usages.

TABLE V. COMPARATIVE PERFORMANCE OF DIFFERENT ADVERSARIAL DEFENSE TECHNIQUES

Techniques	Accuracy
Iterative adversarial Retraining	99.1
Adversarial Training Technique[24]	98.5
AEDPL-DL[25]	98.62

The findings clearly show that the iterative adversarial retraining model is very strong in improving model trustworthiness during both white- and grey-box assaults. The radical change of accuracy, reducing below 30% to over 97%,

demonstrates that multiple exposures to changing adversarial samples help the model to internalize more fixed decision boundaries. In contrast to single-pass adversarial training, the iterative strategy is updated to a new perturbation in each cycle and improves its robustness and clean data performance. The results of the comparison with existing defenses prove that the suggested approach is more efficient and requires less time to be deployed to solve the problem of creating secure and credible AI systems.

V. Conclusion and Future Study

Enhancing machine learning systems towards adversarial resistance is crucial to the reliability and trust of AI-driven systems. The suggested framework shows that secure preprocessing, CNN-based classification, and iterative retraining of adversaries can contribute to a great level of resilience to various attacks. The experimental findings affirm gains of high strength, with model accuracy increasing to more than 97% with FGSM, PGD, BIM, and C&W attacks. An accuracy of 99.1%, clean, and much better performance than the current methods of defense are indicative of the results of continuously adding adversarial samples to the training cycle. The flexibility of retraining makes the model set reliable decision limits and maintain high accuracy despite extremely severe perturbations, confirming the potential of the framework to promote trustworthy AI. It is possible to consider future work extending this framework to more complex and high-dimensional data sets in order to test all domains, including healthcare, autonomous systems, defence, and finance. This might be incorporating explainable AI methods that would aid in the understanding of adversarial behaviors and enhance transparency. The distributed and federated learning scenarios can provide an extra level of protection due to the decentralization of the attack surfaces. Other studies can also focus on lightweight architectures and adaptive adversarial generators to be deployed in real-time. The further development of these directions will help to establish more robust and secure AI systems of the next generation.

References

[1] S. K. Chintagunta, “AI in Code, Testing, and Deployment: A Survey on Productivity Enhancement in Modern Software Engineering,” *Int. J. Res. Anal. Rev.*, vol. 10, no. 4, pp. 747–752, 2023.

[2] G. Modalavalasa and H. Kali, “Exploring Big Data Role in Modern Business Strategies: A Survey with Techniques and Tools,” *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 3, no. 3, pp. 431–441, Jan. 2023, doi: 10.48175/IJARSCT-11900B.

[3] S. Thangavel, S. Srinivasan, S. B. V. Naga, and K. Narukulla, “Distributed Machine Learning for Big Data Analytics: Challenges, Architectures, and Optimizations,” *Int. J. Artif. Intell. Data Sci. Mach. Learn.*, vol. 4, no. 3, pp. 18–30, Oct. 2023, doi: 10.63282/3050-9262.IJAIDSML-V4I3P103.

[4] B. R. Cherukuri, “Serverless revolution: Redefining application scalability and cost efficiency,” *World J. Adv. Res. Rev.*, vol. 2, no. 30, pp. 039–053, Jun. 2019, doi: 10.30574/wjarr.2019.2.3.0093.

[5] T. Chen, J. Liu, Y. Xiang, W. Niu, E. Tong, and Z. Han, “Adversarial attack and defense in reinforcement learning-from AI security view,” *Cybersecurity*, vol. 2, no. 1, p. 11, 2019.

[6] U. A. Korat and A. Alimohammad, “A Reconfigurable Hardware Architecture for Principal Component Analysis,” *Circuits, Syst. Signal Process.*, vol. 38, no. 5, pp. 2097–2113, 2019, doi: 10.1007/s00034-018-0953-y.

[7] S. K. Tiwari, “Integration of AI and Machine Learning with Automation Testing in Digital Transformation,” *Int. J. Appl. Eng. Technol.*, vol. 5, no. 1, pp. 95–96, 2023.

[8] C. Chang, J. Hung, C. Tien, C. Tien, and S. Kuo, “Evaluating Robustness of AI Models against Adversarial Attacks,” pp. 47–54, 2020.

[9] Z. Kong, J. Xue, Y. Wang, L. Huang, Z. Niu, and F. Li, “A survey on adversarial attack in the age of artificial intelligence,” *Wirel. Commun. Mob. Comput.*, vol. 2021, no. 1, p. 4907754, 2021, doi: 10.1155/2021/4907754.

[10] E. Shayegani, M. A. Al Mamun, Y. Fu, P. Zaree, Y. Dong, and N. Abu-Ghazaleh, “Survey of vulnerabilities in large language models revealed by adversarial attacks,” *arXiv Prepr. arXiv2310.10844*, 2023.

[11] L. G. Sanapala, Lavanya, “Mitigating Gradient-Based Data Poisoning Attacks on Machine Learning Models : A Statistical Detection Method,” *Indian J. Sci. Technol.*, vol. 17, no. 21, pp. 2218–2231, 2024.

[12] W. Villegas-Ch, A. Jaramillo-Alcázar, and S. Luján-Mora, “Evaluating the Robustness of Deep Learning Models against Adversarial Attacks: An Analysis with FGSM, PGD and CW,” *Big Data Cogn. Comput.*, vol. 8, no. 1, 2024, doi: 10.3390/bdcc8010008.

[13] S. Wibawa, “Analysis of Adversarial Attacks on AI-based With Fast Gradient Sign Method,” *Int. J. Eng. Contin.*, vol. 2, no. 2, pp. 72–79, 2023.

[14] H. Zhu, S. Zhang, and K. Chen, “AI-Guardian: Defeating Adversarial Attacks using Backdoors,” in *2023 IEEE Symposium on Security and Privacy (SP)*, 2023, pp. 701–718. doi: 10.1109/SP46215.2023.10179473.

[15] T. Anastasiou *et al.*, “Towards Robustifying Image Classifiers against the Perils of Adversarial Attacks on Artificial Intelligence Systems,” *Sensors*, vol. 22, no. 18, 2022, doi: 10.3390/s22186905.

[16] C.-J. Lin, “Building Resilient AI Models Against Data Poisoning Attacks,” *Multidiscip. Stud. Innov. Res.*, vol. 3, no. 4, pp. 1–16, 2022.

[17] H. Xu, X. Liu, Y. Wan, and J. Tang, "Towards Fair Classification against Poisoning Attacks," *arXiv Prepr. arXiv2210.09503*, 2022.

[18] R. F. Kharal, "Towards augmentation based defense strategies against adversarial attacks," in *2023 International Conference on Machine Learning and Applications (ICMLA)*, 2023, pp. 1430–1437. doi: 10.1109/ICMLA58977.2023.00216.

[19] Y. Liu, X. Yuan, R. Zhao, C. Wang, D. Niyato, and Y. Zheng, "Poisoning semi-supervised federated learning via unlabeled data: Attacks and defenses," *arXiv Prepr. arXiv2012.04432*, 2020.

[20] D. Gragnaniello, F. Marra, G. Poggi, and L. Verdoliva, "Analysis of adversarial attacks against CNN-based image forgery detectors," in *2018 26th European signal processing conference (EUSIPCO)*, 2018, pp. 967–971. doi: 10.23919/EUSIPCO.2018.8553560.

[21] Y. Li, D. Tian, M.-C. Chang, X. Bian, and S. Lyu, "Robust adversarial perturbation on deep proposal-based models," *arXiv Prepr. arXiv1809.05962*, 2018.

[22] S. Dhesi, L. Fontes, P. Machado, I. K. Ihianle, F. F. Tash, and D. A. Adama, "Mitigating adversarial attacks in deepfake detection: An exploration of perturbation and AI techniques," *arXiv Prepr. arXiv2302.11704*, 2023.

[23] Z. Qin, G. Liu, and X. Lin, "Enhancing Model Robustness Against Adversarial Attacks with an Anti-adversarial Module," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 2024, pp. 66–78. doi: 10.1007/978-981-99-8546-3_6.

[24] G. Apruzzese, M. Colajanni, L. Ferretti, and M. Marchetti, "Addressing Adversarial Attacks Against Security Systems Based on Machine Learning," in *2019 11th International Conference on Cyber Conflict (CyCon)*, IEEE, May 2019, pp. 1–18. doi: 10.23919/CYCON.2019.8756865.

[25] M. N. Al-Andoli, S. C. Tan, K. S. Sim, P. Y. Goh, and C. P. Lim, "A Framework for Robust Deep Learning Models Against Adversarial Attacks Based on a Protection Layer Approach," *IEEE Access*, vol. 12, pp. 17522–17540, Jan. 2024, doi: 10.1109/ACCESS.2024.3354699.