A Robust Neural Network against Adversarial Attacks

Mohammad Barr
Department of Electrical Engineering, College of Engineering, Northern Border University, Saudi Arabia
mohammed.barr@nbu.edu.sa (corresponding author)

Received: 12 December 2024 | Revised: 30 December 2024 | Accepted: 12 January 2025

 DOI: https://doi.org/10.48084/etasr.9920

ABSTRACT

The security and dependability of neural network designs are increasingly jeopardized by adversarial attacks, which can cause false positives, degrade performance, and disrupt applications, particularly on resource-constrained Internet of Things (IoT) devices. This study adopts a two-step approach: first, designs a robust Convolutional Neural Network (CNN) that achieves high performance on the MNIST dataset, and second, evaluates and enhances its resilience against advanced adversarial techniques such as Deepfool and L-BFGS. Initial evaluations revealed that while the proposed CNN performs well on standard classification tasks, it is vulnerable to adversarial attacks. To mitigate this vulnerability, APE-GAN, an innovative adversarial training technique, was employed to re-train the proposed CNN, significantly improving its robustness against adversarial attacks while optimizing performance for embedded systems with limited computational resources. Systematic experimentation demonstrates the effectiveness of APE-GAN in enhancing both the accuracy and resilience of the proposed CNN, outperforming conventional methods and establishing it as a pioneering solution in adversarial machine learning. By integrating APE-GAN into the training process, this research ensures the secure and efficient operation of the proposed CNN in real-world IoT applications, marking a significant step forward in addressing the challenges posed by adversarial attacks.

Keywords-neural networks; adversarial attack; robustness; L-BFGS

I. INTRODUCTION

In recent years, there have been major shifts in the technological environment due to the extensive use of neural networks in embedded systems such as IoT devices. Image classification and natural language processing are only two of the many tasks that advanced computer models have proven to be exceptionally adept at. Due to this, complicated data analysis, decisions are now more efficient and accurate. A new age of invention and advancement has begun with the broad adoption of this technology, which has expanded across numerous sectors and industries [1]. With the introduction of deep learning, a subfield of machine learning, models can now learn from data and gain a systematic understanding of the world, completely transforming computing approaches. To better represent data through several layers, deep learning models use backpropagation algorithms to sift through massive datasets in search of complicated patterns. Advancements in high-performance hardware and refined architectures for Deep Neural Networks (DNNs) have allowed deep learning to make great strides in several domains. This development encompasses established fields, such as image classification and voice recognition, and cutting-edge ones, such as pharmacological chemical analysis and neural circuit rebuilding. Industry giants such as Google, Alibaba, Intel, and Nvidia have exerted a great influence in driving advances in AI-based services by presenting remarkably accurate models [2].

Concerns have arisen about the adversarial attack susceptibility of neural networks and deep learning, despite the significant advances in these technologies. Neural networks are vulnerable to adversarial attacks that alter the input data on purpose. The accuracy and dependability of neural networks can be compromised by carefully crafted disruptions that cause misclassifications or inaccurate predictions [4]. The implications of these vulnerabilities are more severe in safety-critical applications, such as security systems, medical diagnostics, and autonomous vehicles. Serious hazards, including loss of life or the theft of sensitive data, could result from seemingly insignificant errors [3]. Therefore, it is critical to address weaknesses in neural network designs and limit the dangers of adversarial attacks if neural network-based systems are to continue to be effective and reliable. Researchers have intensified their efforts across many domains to better understand adversarial attacks, the resilience of neural network architectures, and how to strengthen defensive techniques [5].

Research on neural network vulnerabilities has been extensive, with many studies shedding light on various parts of the problem and offering novel approaches to strengthening their security. However, concerns about the security and reliability of Deep Neural Networks (DNNs) have grown in importance as these systems expand their use beyond the laboratory. Although attackers can subtly alter legitimate inputs, changes that are frequently invisible to human observers can trick trained models into producing false outputs. In [6], it

was shown how easily attackers can target high-performing DNNs, demonstrating how important it is to have strong security measures in place. This security issue is prevalent, as weaknesses have been found in Automatic Speech Detection (ASR), Voice-Controllable Systems (VCS), and other applications [7-9]. In [2], attacks on autonomous cars were presented by manipulating traffic signs to trick learning systems. The adversarial learning of linear models has been supported by extensive analysis and experiments in supplementary studies [10]. In [11], the generalizability of adversarial scenarios was explored. In [12], different approaches were presented, reflecting the current state of affairs in neural network safety. Using distributed deep learning algorithms to protect the privacy of training data is at the heart of these methods.

Although there has been significant progress in discovering vulnerabilities in neural networks and developing solutions to mitigate them, there is still a need for more robust and efficient alternatives. Previous studies have frequently ignored the special difficulties that arise in neural network applications. Specific policies are required to handle resource-constrained situations and the potential for attacks. This study aimed to address this gap by thoroughly investigating adversarial threats in neural network applications. Specifically, the sensitivity of neural networks trained on the MNIST dataset to various attack strategies was investigated, including Fast Gradient Sign (FGSM), Deepfool, and L-BFGS, to illuminate multiple problems caused by unfavorable threats in reality-global distribution scenarios. Furthermore, this study presents a strategy that utilizes adversarial training approaches, specifically the APE-GAN method, to strengthen the resistance of neural network architectures to adversarial attacks. This research aims to enhance the knowledge of neural network security and aid in the creation of strong defense mechanisms by thoroughly evaluating and testing these methods.

As deep learning models continue to evolve, the demand for efficient model deployment in resource-constrained environments has led to the development of various compression and quantization techniques. These methods aim to reduce the size and complexity of neural networks while maintaining their accuracy, which is crucial for embedded system applications. Adversarial attacks are carefully designed to manipulate the decision-making process of a classifier and drastically affect accuracy, loss, and precision, which are important performance measures [13]. These attacks are based on computer vision principles. In [14], an assortment of attacks, including white-box and black-box attacks, as well as techniques such as FGSM, PGD, and MIM, were described. This study started by reviewing the ground rules that are critical to comprehending the patterns of adversarial attacks on neural networks, distinguishing between white-box and black-box adversarial attacks [15].

In a white box attack, the attacker knows all the architecture details, variables, and gradients of the attacked model, making it a very sophisticated way to apply adversarial methods [16]. This in-depth knowledge allows adversaries to painstakingly manipulate input data to expose weaknesses in the model's decision-making process. On the other hand, black box attacks are characterized by a lack of visibility into the target model's inner workings. Without understanding the model's structure or parameters, attackers in such cases depend entirely on its input-output behavior. Therefore, black-box attackers employ strategies such as transferability [17], where adversarial instances created for one model are applied to another that shows similar behavior. In contrast to black-box attacks, which operate within the constraints of restricted access to information, white-box attacks use all internal knowledge. This is the fundamental difference between these two attack types.

Many studies [3, 18, 19] have classified adversarial attacks into different types, with white-box attacks being the first. It is possible to further categorize these hostile attacks as targeted or non-targeted. An adversarial example is designed to induce a specific classification in targeted attacks, such as labeling all images as representing one person, whereas, in non-targeted attacks, the only goal is to cause misclassifications regardless of the resulting class. Iterative model access to compute gradients is used in adversarial example generation [20], and white-box attacks can be classified into two basic types: limited optimization and gradient-based optimization.

In [21], a novel defense mechanism was proposed against white-box adversarial attacks. To reduce the susceptibilities of neural networks to adversarial disruptions, this strategy incorporated randomized discretization as a robust countermeasure. This strategy was effective in strengthening the resistance of neural network designs against white-box attacks. In [22], a unique approach to generating white-box adversarial examples was presented, which was specifically customized for text classification tasks. The Hotflip approach aims to generate adversarial disruptions that are capable of fooling text classification machines while preserving semantic coherence. This study adds to the expanding body of knowledge on adversarial attacks in natural language processing, providing useful insights into the weaknesses of text classification algorithms. In [23], transfer learning approaches were used to increase the efficacy of adversarial attacks across a variety of models and domains. Incorporating a transfer-based method provides significant gains in producing hostile instances that are capable of fooling target models. This study provided crucial insights into the domain of adversarial machine learning.

In [24], a unique approach was presented to effectively distill hostile black-box attacks. This method aimed to enhance the effectiveness of black-box attacks by transferring information from white-box attacks. The results showed substantial gains in the efficiency and efficacy of creating adversarial samples that are capable of fooling target models. This study offered useful insights into the process of developing viable adversarial attack tactics. As shown in [25], this strategy is effective in enhancing the resistance of machine learning models to adversarial disruptions. This study made a significant contribution to ongoing attempts to design solid defenses against adversarial attacks in computer vision and pattern recognition applications. In [26], effective defensive strategies were studied and proposed to reduce the impact of adversarial attacks. This study provided useful insights and tactics to improve the resilience and security of artificial

intelligence systems against external disturbances. In [27], a unique training strategy, named bilateral adversarial training, was proposed to accelerate the process of building robust models to overcome the challenges of adversarial attacks. This study made significant contributions to the fields of computer vision and adversarial machine learning by introducing novel approaches to improve the robustness of machine learning models. In [28], the efficacy of the correlation power analysis approach in the presence of noise for side channel assaults was investigated. The findings provided valuable insights into the performance of side-channel attack strategies in noisy settings, shedding light on possible vulnerabilities and countermeasures in the field of information security.

This study presents a unique Convolutional Neural Network (CNN) design that is suitable for the MNIST dataset and assesses its robustness against adversarial attacks. This CNN architecture was put through rigorous testing by employing state-of-the-art adversarial attack approaches, such as Deepfool and L-BFGS to evaluate its resilience. Following that, a novel strategy was proposed to enhance the model's defenses against adversarial attacks. The adversary training process was incorporated with the APE-GAN approach to strengthen the defenses of the model.

II. PROPOSED APPROACH: STRENGTHENING CNN DEFENSES AGAINST MALICIOUS ATTACKS

This study aimed to improve the security of CNNs against offensive assaults. It begins by providing a comprehensive description of the process of developing a CNN architecture that is particularly designed for the MNIST dataset. Then an exhaustive assessment is performed to determine the resilience of the model in the face of a variety of adversarial attack strategies. This method incorporates a new training strategy that uses the APE-GAN technique in the adversarial training process of the proposed CNN to strengthen the defenses of the model and reduce the effect of adversarial disruptions. Exhaustive experiments and in-depth analyses were carried out to determine if the proposed technique is effective in strengthening CNN's security against adversarial threats. Figure 1 shows the approach suggested for the CNN, along with how it can be customized to withstand an adversarial attack.

A. MNIST Dataset

The MNIST dataset has an important position in the field of neural networks and is well-known for its function as a core benchmark in image classification tasks. MNIST is a basic resource for training and assessing a variety of algorithms [29]. It consists of a collection of grayscale pictures that are 28×28 pixels in size and display handwritten numbers ranging from 0 to 9. Due to its straightforwardness and the fact that it is clearly defined, it is an excellent starting place for researchers who want to investigate and improve machine learning models, especially CNNs. The MNIST offers a trustworthy framework for evaluating the robustness of CNN models against adversarial attacks due to its extensive application and uniform structure. Researchers can gain significant insights into the efficacy of protection mechanisms and methods by putting CNNs trained on MNIST to rigorous testing using adversarial approaches. MNIST continues to play an important role in the understanding and development of strong machine-learning models that can survive adversarial challenges in real-world applications.

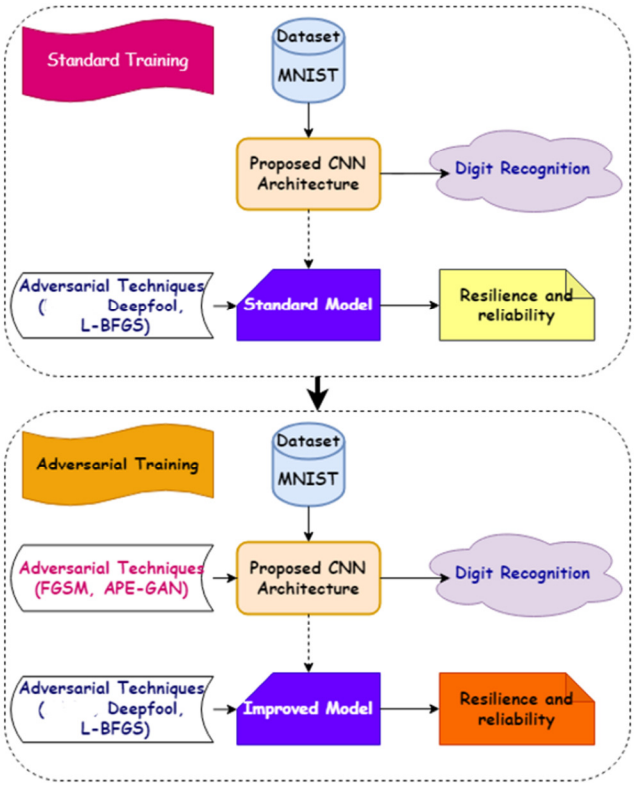

Fig. 1. The proposed method to improve a CNN against adversarial attacks.

The dataset was divided into 60,000 shapes for the training set and 10,000 shapes for the test set. Figure 2 presents a sample of the MNIST dataset.

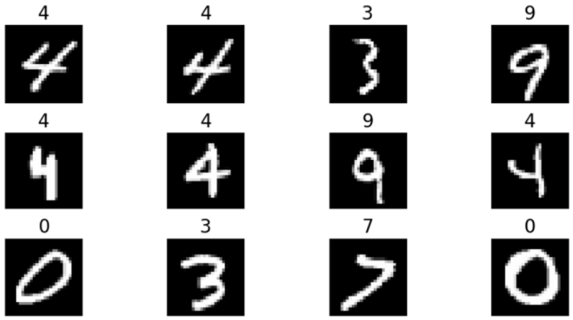

Fig. 2. MNIST dataset sample.

B. Proposed CNN Architecture

As shown in Figure 3, the proposed CNN design is built with a series of layers that are optimized to handle and extract features from input data in a streamlined manner. It employs three convolutional layers to extract detailed geometric patterns and characteristics from the data. The training process is

enhanced by deliberately integrating three layers of batch normalization to stabilize and expedite convergence. The design also includes three dropout layers to prevent overfitting by randomly turning off some neurons while training. A one-dimensional matrix is created by flattening the output of the convolutional layers. This allows for a smooth transition to fully linked layers. Lastly, the architecture provides a strong framework, appropriate for tasks such as image identification and classification, which comprises two dense layers that are responsible for classification using learned features.

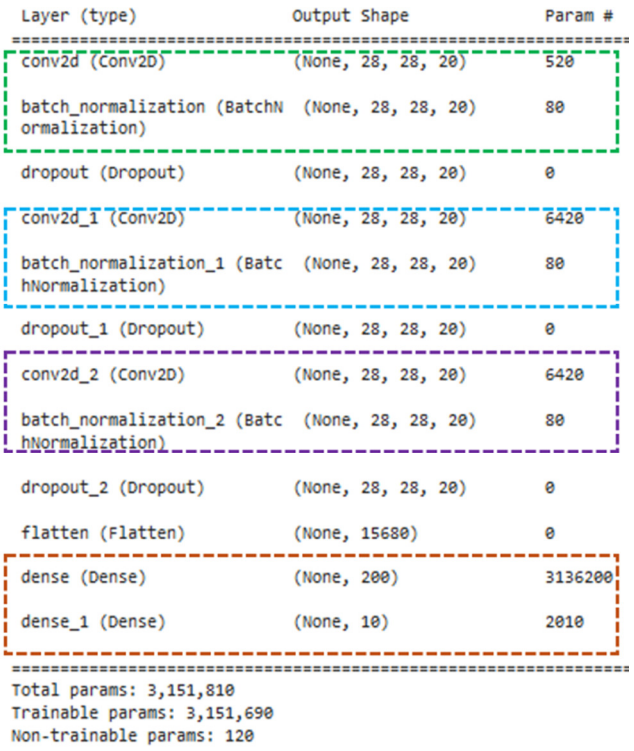

Fig. 3. Proposed CNN architecture.

C. Deepfool Technique

The Deepfool method describes an iterative strategy for constructing adversarial instances taking advantage of the linear character of neural networks [30]. This method seeks to identify the smallest amount of disturbance that is necessary to incorrectly categorize an input sample. Mathematically, the disrupted input x' is calculated as follows:

$$x' = x + \delta \cdot \frac{f(x)-f(x')}{\|\nabla f(x)\|}.\nabla f(x) \quad (2)$$

where x' denotes the disrupted input data, x designates the original input data, δ denotes the small scalar value controlling the magnitude of the disruption, $f(x)$ denotes the output of the neural network for the input x, and finally, $\nabla f(x)$ is the gradient of the neural network's output for the input data. This method makes adjustments to the disruption in the direction of the gradient in a repeated manner until the neural network incorrectly classifies the disrupted input. This iterative process will continue until the required degree of misclassification is attained.

D. L-BFGS Technique

The Limited-memory BFGS (L-BFGS) is another technique that is often used for adversarial attacks [2]. However, in contrast to previous methods, L-BFGS explicitly optimizes the disruption to maximize the loss and cause misclassification. The disrupted input x' is iteratively calculated as follows:

$$x' = x + \delta \cdot sign(\nabla_{adv} L(x', y_{true})) \quad (3)$$

where x' is the perturbed input data, x represents the original input data, and δ denotes the small scalar value controlling the magnitude of the perturbation.

III. RESULTS AND DISCUSSIONS

A. Standard Training Results of the Proposed CNN

Table I presents the performance evaluation results of the proposed CNN trained on the MNIST dataset, showcasing its classification metrics across all 10-digit classes (0–9). The model achieves near-perfect results, with precision, recall, and F1-score values consistently close to 1.00 for each class. The overall accuracy is 100%, indicating exceptional performance on the dataset. The macro average (0.99) and weighted average (1.00) further demonstrate the model's robustness and balanced performance across varying class distributions. These results reflect the effectiveness of the standard training process applied to the proposed CNN on the MNIST dataset

TABLE I. PERFORMANCE EVALUATION

Classes	Precision	Recall	F1 score	Support
0	1.00	1.00	1.00	420
1	0.99	0.99	0.99	471
2	1.00	0.99	0.99	410
3	1.00	1.00	0.99	423
4	0.99	0.99	0.99	427
5	1.00	1.00	0.99	382
6	1.00	1.00	0.99	413
7	0.99	0.99	0.99	470
8	0.99	1.00	0.99	386
9	1.00	0.99	0.99	395
Accuracy			1.00	4210
Macro avg	0.99	1.00	1.00	4210
Weighted avg	1.00	1.00	1.00	4210

B. Attack Results on the Standard-Trained CNN Model

1) Deepfool Attack Results

Meticulous experimentation and analysis elucidate the repercussions of this adversarial attack method on the model's performance and resilience. By varying parameters such as the number of iterations and the maximum perturbation, the impact of the Deepfool attack was evaluated on the model's ability to accurately classify digits. Additionally, the resulting misclassifications induced by the attack were scrutinized, shedding light on the specific patterns and characteristics of adversarial perturbations that lead to erroneous predictions. This examination allows gaining deeper insights into the vulnerabilities of the proposed model when subjected to sophisticated adversarial attacks such as Deepfool and underscores the importance of developing robust defense mechanisms to protect against such threats. Figure 4 illustrates the Deepfool attack on the standard CNN.

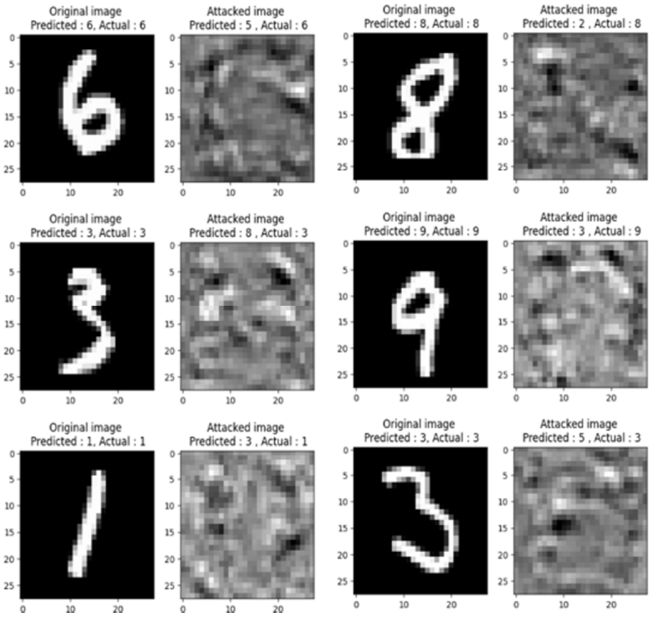

Fig. 4. Deepfool attack.

2) *LBFGS Attack Results*

The model's performance was analyzed in the face of adversarial disruptions induced by the LBFGS Attack. By examining the resulting misclassifications, specific patterns and characteristics of adversarial perturbations that lead to incorrect predictions can be identified. This analysis provides valuable insights into the vulnerabilities of the proposed model when subjected to sophisticated adversarial attacks such as LBFGS, highlighting the need for robust defense mechanisms to enhance its resilience against such threats. Figure 5 presents the LBFGS attack on the standard CNN model.

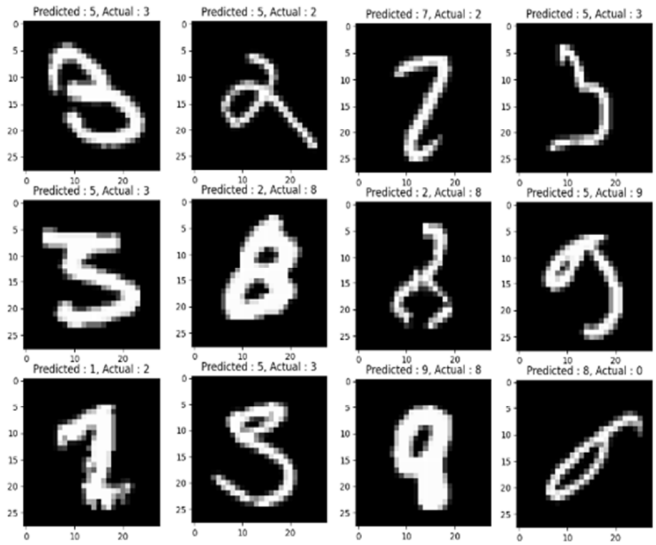

Fig. 5. LBFGS attack.

C. *CNN-based Adversarial Training using APE-GAN*

The results presented in Table II highlight the effectiveness of the APE-GAN framework in mitigating the impact of adversarial attacks on the CNN trained on the MNIST dataset. The original CNN achieves perfect accuracy (100%) on standard data but suffers severe degradation under adversarial scenarios. For instance, the accuracy drops dramatically to 25.6% and 30.8% under the Deepfool and L-BFGS attacks, respectively, underscoring the vulnerability of CNN designs to adversarial perturbation. However, with adversarial training using APE-GAN, the accuracy of the proposed CNN under these attacks improves significantly to 60.3% (Deepfool) and 69.4% (L-BFGS), reflecting improvements of 34.7% and 38.6%, respectively. This shows that APE-GAN effectively enhances the robustness of the CNN against attack scenarios while maintaining high performance.

TABLE II. PERFORMANCE OF ADVERSARIAL TRAINED CNN USING APE-GAN

Scenario Test	**CNN accuracy (%) standard**	**CNN accuracy (%) using APE-GAN**	**Improvement (%)**
Original Data	100	99.6	-0.4
Adversarial attacks (mean CNN accuracy after varying the introduced error ε at the input)			
Deepfool attack	25.6	60.3	34.7
LBFGS attack	30.8	69.4	38.6

D. *Comparative Study*

Table III compares the proposed APE-GAN approach with a related method that employs federated learning [30]. While both techniques target adversarial robustness, APE-GAN achieves superior accuracy (99.6%) on the MNIST dataset compared to 97.2% for the federated learning approach. Additionally, the CNN based on APE-GAN shows notable improvement in robustness, effectively mitigating the L-BFGS attack ($\varepsilon = 0.1$) with an improvement of 85.7%, while the federated learning approach reports a 58.3% robustness against label-flipping and Gaussian attacks. This comparison further highlights the superiority of APE-GAN in adversarial training, particularly in scenarios involving highly sophisticated attack techniques. The results underscore the contribution of this study in addressing critical security challenges in neural network-based IoT applications. By leveraging APE-GAN, the proposed CNN not only improves adversarial robustness but also ensures computational efficiency, making it particularly suited for resource-constrained environments. This research sets a benchmark for the development of secure and reliable neural network designs, paving the way for more resilient IoT systems against evolving adversarial threats.

TABLE III. COMPARATIVE STUDY

Study	**Training techniques**	**Dataset**	**Accuracy (%)**	**Robustness improvement (%)**
This work (APE-GAN)	APE-GAN	MNIST	99.98	85.7 L-BFGS attack ($\varepsilon = 0.1$)
[30]	Federated learning	MNIST	97.2	58.3 Label-flipping, Gaussian

IV. CONCLUSION AND FUTURE WORK

This study introduces a novel perspective to enhance the resilience of neural network designs against adversarial attacks, particularly in resource-constrained embedded systems such as IoT devices. The APE-GAN adversarial training method achieved substantial improvements in both accuracy and robustness compared to conventional approaches such as FGSM and similar methods. Unlike previous strategies, APE-GAN not only mitigates adversarial vulnerabilities but also achieves these results with a level of efficiency suitable for deployment on devices with limited computational resources. These findings underscore the importance of integrating advanced adversarial training techniques into the design process of machine learning systems to protect against evolving attack strategies. The systematic evaluation of APE-GAN against sophisticated attacks such as Deepfool and L-BFGS highlights its unique ability to address both the accuracy and reliability challenges posed by adversarial inputs. This represents a significant step forward in the field of adversarial machine learning, particularly in ensuring the robustness of neural networks in real-world applications. The novelty of this work lies in its dual focus on enhancing security and optimizing performance for embedded systems, paving the way for future research. Expanding the scalability and applicability of APE-GAN across diverse datasets and domains could unlock new pathways to create resilient machine-learning solutions that meet the demands of increasingly sophisticated adversarial environments. Thus, this study establishes APE-GAN as a promising foundation for the next generation of secure and reliable neural network designs.

ACKNOWLEDGEMENT

The author extends his appreciation to the Deanship of Scientific Research at the Northern Border University, Arar, KSA for funding this research work through project number NBU-FFR-2025-XXXX-01.

REFERENCES

[1] H. Jiang, J. Lin, and H. Kang, "FGMD: A robust detector against adversarial attacks in the IoT network," *Future Generation Computer Systems*, vol. 132, pp. 194–210, Jul. 2022, https://doi.org/10.1016/j.future.2022.02.019.

[2] Y. Wang, Y. Tan, W. Zhang, Y. Zhao, and X. Kuang, "An adversarial attack on DNN-based black-box object detectors," *Journal of Network and Computer Applications*, vol. 161, Jul. 2020, Art. no. 102634, https://doi.org/10.1016/j.jnca.2020.102634.

[3] A. Aldahdooh, W. Hamidouche, S. A. Fezza, and O. Déforges, "Adversarial example detection for DNN models: a review and experimental comparison," *Artificial Intelligence Review*, vol. 55, no. 6, pp. 4403–4462, Aug. 2022, https://doi.org/10.1007/s10462-021-10125-w.

[4] L. Liu, Y. Guo, Y. Cheng, Y. Zhang, and J. Yang, "Generating Robust DNN With Resistance to Bit-Flip Based Adversarial Weight Attack," *IEEE Transactions on Computers*, vol. 72, no. 2, pp. 401–413, Feb. 2023, https://doi.org/10.1109/TC.2022.3211411.

[5] R. Duan, Y. Chen, D. Niu, Y. Yang, A. K. Qin, and Y. He, "AdvDrop: Adversarial Attack to DNNs by Dropping Information." arXiv, 2021, https://doi.org/10.48550/arXiv.2108.09034.

[6] C. Szegedy *et al.*, "Intriguing properties of neural networks." arXiv, 2013, https://doi.org/10.48550/arXiv.1312.6199.

[7] N. Carlini *et al.*, "Hidden Voice Commands," presented at the 25th USENIX Security Symposium (USENIX Security 16), 2016, Art. no. 513–530, [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini.

[8] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "DolphinAttack: Inaudible Voice Commands," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, Dallas Texas USA, Oct. 2017, pp. 103–117, https://doi.org/10.1145/3133956.3134052.

[9] M. Ben Ammar, R. Ghodhbani, and T. Saidani, "Enhancing Neural Network Resilence against Adversarial Attacks based on FGSM Technique," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14634–14639, Jun. 2024, https://doi.org/10.48084/etasr.7479.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples." arXiv, Mar. 20, 2015, https://doi.org/10.48550/arXiv.1412.6572.

[11] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples." arXiv, May 24, 2016, https://doi.org/10.48550/arXiv.1605.07277.

[12] M. Abadi *et al.*, "Deep Learning with Differential Privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Vienna, Austria, Oct. 2016, pp. 308–318, https://doi.org/10.1145/2976749.2978318.

[13] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey." arXiv, 2019, https://doi.org/10.48550/arXiv.1901.06796.

[14] R. R. Wiyatno, A. Xu, O. Dia, and A. de Berker, "Adversarial Examples in Modern Machine Learning: A Review." arXiv, 2019, https://doi.org/10.48550/arXiv.1911.05268.

[15] G. B. Ingle and M. V. Kulkarni, "Adversarial Deep Learning Attacks—A Review," in *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*, vol. 190, M. S. Kaiser, J. Xie, and V. S. Rathore, Eds. Springer Nature Singapore, 2021, pp. 311–323.

[16] Y. Lin, H. Zhao, Y. Tu, S. Mao, and Z. Dou, "Threats of Adversarial Attacks in DNN-Based Modulation Recognition," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, Toronto, Canada, Jul. 2020, pp. 2469–2478, https://doi.org/10.1109/INFOCOM41043.2020.9155389.

[17] H. Xu *et al.*, "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review." arXiv, 2019, https://doi.org/10.48550/ArXiv.1909.08072.

[18] A. Aldahdooh, W. Hamidouche, S. A. Fezza, and O. Déforges, "Adversarial example detection for DNN models: a review and experimental comparison," *Artificial Intelligence Review*, vol. 55, no. 6, pp. 4403–4462, Aug. 2022, https://doi.org/10.1007/s10462-021-10125-w.

[19] S. Y. Khamaiseh, D. Bagagem, A. Al-Alaj, M. Mancino, and H. W. Alomari, "Adversarial Deep Learning: A Survey on Adversarial Attacks and Defense Mechanisms on Image Classification," *IEEE Access*, vol. 10, pp. 102266–102291, 2022, https://doi.org/10.1109/ACCESS.2022.3208131.

[20] Y. Zhang and P. Liang, "Defending against Whitebox Adversarial Attacks via Randomized Discretization." arXiv, 2019, https://doi.org/10.48550/arXiv.1903.10586.

[21] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "HotFlip: White-Box Adversarial Examples for Text Classification," 2017, https://doi.org/10.48550/arXiv.1712.06751.

[22] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu, "Improving Black-box Adversarial Attacks with a Transfer-based Prior." arXiv, 2019, https://doi.org/10.48550/arXiv.1906.06919.

[23] Y. Gil, Y. Chai, O. Gorodissky, and J. Berant, "White-to-Black: Efficient Distillation of Black-Box Adversarial Attacks." arXiv, Apr. 04, 2019, https://doi.org/10.48550/arXiv.1904.02405.

[24] Y. Ding, L. Wang, H. Zhang, J. Yi, D. Fan, and B. Gong, "Defending Against Adversarial Attacks Using Random Forest," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019, pp. 105–114, https://doi.org/10.1109/CVPRW.2019.00019.

[25] V. Zantedeschi, M. I. Nicolae, and A. Rawat, "Efficient Defenses Against Adversarial Attacks," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, Dallas, TX, USA, Nov. 2017, pp. 39–49, https://doi.org/10.1145/3128572.3140449.

[26] J. Wang and H. Zhang, "Bilateral Adversarial Training: Towards Fast Training of More Robust Models Against Adversarial Attacks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 6628–6637, https://doi.org/10.1109/ICCV.2019.00673.

[27] F. Garipay, K. Uzgören, İ. Kaya, "Side channel attack performance of correlation power analysis method in noise," *International Journal on Information Technologies and Security*, vol. 15, no. 1, pp. 101–111, Mar. 2023, https://doi.org/10.59035/UWCB7557.

[28] R. Dixit, R. Kushwah, and S. Pashine, "Handwritten Digit Recognition using Machine and Deep Learning Algorithms," *International Journal of Computer Applications*, vol. 176, no. 42, pp. 27–33, Jul. 2020, https://doi.org/10.5120/ijca2020920550.

[29] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks." arXiv, 2015, https://doi.org/10.48550/arXiv.1511.04599.

[30] F. Hu, W. Zhou, K. Liao, H. Li, and D. Tong, "Toward Federated Learning Models Resistant to Adversarial Attacks," *IEEE Internet of Things Journal*, vol. 10, no. 19, pp. 16917–16930, Oct. 2023, https://doi.org/10.1109/JIOT.2023.3272334.