[1] Bhasha Anjaria*

[2] Jaimeel Shah

Assistive-GAN Based Adversarial Learning and Defence for Black-box And White-box Attacks

Abstract: - This research paper addresses the ongoing challenge of adversarial attacks in machine learning security by introducing an Assistive-GAN framework tailored to enhance adversarial learning and defence mechanisms against black-box and white-box attacks. The framework is designed to integrate seamlessly with existing defence strategies, augmenting model resilience while maintaining performance metrics. Utilizing a dual-phase training process, the Assistive-GAN generates assistive samples strategically to reinforce the model's ability to identify and withstand adversarial perturbations. Through comprehensive experiments evaluating diverse datasets and attack scenarios, including black-box and white-box attacks, the framework demonstrates significant improvements in model robustness and accuracy compared to state-of-the-art techniques. This research highlights the potential of the Assistive-GAN framework as an effective proactive defence mechanism in bolstering machine learning security against adversarial threats, contributing valuable insights to the cybersecurity domain.

Keywords: Machine Learning, Adversarial learning, Attack, Defence, Assistive GAN

1. INTRODUCTION

Machine learning (ML) has revolutionized various domains, from image recognition to cybersecurity. However, this progress has also brought about new challenges, particularly regarding security vulnerabilities. Adversarial attacks, wherein malicious inputs are crafted to deceive ML models, have emerged as a significant threat [1]. These attacks can have severe consequences, ranging from misclassifying images to compromising sensitive systems. In response, researchers have focused on developing robust defence mechanisms to safeguard ML models against such attacks [2].

This research paper delves into adversarial machine learning, specifically targeting the development and evaluation of an Assistive-GAN framework. The framework aims to enhance adversarial learning and defence mechanisms against black-and-white-box attacks [3]. By strategically generating assistive samples, the framework fortifies the model's ability to discern and withstand adversarial perturbations [4]. This proactive defence approach is crucial in ensuring the resilience of ML models, particularly in critical applications like intrusion detection and malware analysis [5].

Understanding the broader landscape of adversarial machine learning is essential to contextualize this study. Previous research has explored various attack and defence strategies, including image transformation-based defences [6], denoising and verification cross-layer ensembles [7], and reinforcement learning-based black-box adversarial attacks [8]. These efforts underscore the complexity of the adversarial landscape and the necessity for robust defence mechanisms [9]. Moreover, datasets such as CIFAR-10 [10] and ImageNet [11] have been instrumental in evaluating the efficacy of defence strategies against adversarial attacks. These datasets provide a diverse range of images, enabling researchers to assess the generalizability and effectiveness of defence mechanisms across different domains [12].

Given these advancements and challenges, this research paper contributes a novel Assistive-GAN framework to bolster model robustness and resilience against adversarial attacks. Through comprehensive experiments and benchmarking against existing defence strategies, this study aims to provide valuable insights and advancements in adversarial machine learning security.

2. LITERATURE STUDY

P. Yu, K. Song, and J. Lu [1] introduced a method using Conditional Generative Adversarial Networks (C.G.A.N.s) to generate adversarial examples. Their work delves into the intersection of generative models and adversarial attacks, showcasing the potential of C.G.A.N.s in crafting deceptive inputs for machine learning systems.

Almuflih et al. [2] proposed a feature-map-based detection approach to identify adversarial attacks, aiming to enhance the robustness of classification systems. By focusing on analyzing feature maps, their method improves the reliability of detecting and mitigating adversarial inputs. Bakhti et al. [3] presented D.D.S.A., a defence mechanism leveraging Deep Denoising Sparse Autoencoders to combat adversarial attacks. Their work emphasizes the significance of denoising techniques in fortifying machine learning models against crafted

[1,2] Department of Computer Engineering, Parul University, Vadodara, Gujarat, India
[1*] bhasha.anjaria21316@paruluniversity.ac.in
[2] jaimeel.shah@paruluniversity.ac.in

perturbations. Dasgupta and Gupta [4] introduced Dual-filtering (D.F.) schemes as a defence strategy to prevent adversarial attacks. By employing filtering methods in learning systems, their approach contributes to bolstering the resilience of models against adversarial perturbations.

Deldjoo et al. [5] conducted a comprehensive survey on Adversarial Recommender Systems, covering attack/defence strategies and the role of Generative Adversarial Networks (GANs) in recommendation systems. Their work provides valuable insights into the evolving landscape of adversarial techniques in recommendation algorithms. Despiegel and Despiegel [6] proposed Dynamic Dynamic Autoencoders (D.D.A.) as a defence mechanism against adversarial attacks. Their approach focuses on dynamic adaptation within autoencoder architectures, showcasing a proactive strategy to mitigate adversarial perturbations. Folz et al. [7] presented an adversarial defence approach based on Structure-to-Signal Autoencoders. By prioritizing structural information preservation, their method enhances models' robustness against adversarial inputs in computer vision tasks. Han et al. [8] reviewed interpreting adversarial examples in deep learning, providing insights into various methods and strategies for understanding and mitigating adversarial attacks. Their work is a valuable resource for researchers and practitioners aiming to enhance the security of deep learning models. Hu and Tan [9] explored the generation of adversarial malware examples for black-box attacks based on Generative Adversarial Networks (GANs). Their study sheds light on the challenges and implications of adversarial attacks in cybersecurity, particularly in the context of malware detection systems. Jin et al. [10] discussed Generative Adversarial Network (GAN) technologies and applications in computer vision, highlighting the versatility of GANs in generating realistic data samples. Their work showcases the potential of GANs beyond adversarial settings, contributing to advancements in computer vision tasks. Kuribayashi [11] focused on defence strategies against adversarial attacks, emphasizing the importance of robust mechanisms in fake media generation and detection. Their work contributes to the ongoing efforts in developing defensive measures to combat the proliferation of adversarial content in digital media.

Laykaviriyakul and Phaisangittisagul [12] proposed a Collaborative Defence-GAN approach to protect against adversarial attacks on classification systems. By leveraging collaborative defence mechanisms, their method aims to enhance the resilience of classifiers in the face of crafted adversarial inputs. Li et al. [13] presented a defence method based on multiple Filtering and image rotation to mitigate adversarial attacks. Their approach highlights the effectiveness of image transformations and filtering techniques in improving model robustness against adversarial perturbations. Liang et al. [14] surveyed adversarial attack and defence techniques, providing insights into various methodologies and advancements in adversarial robustness. Their comprehensive review is a valuable resource for researchers and practitioners navigating the landscape of adversarial machine learning. Liu et al. [15] proposed an adversarial sample defence method based on saliency information, focusing on leveraging salient features to detect and mitigate adversarial inputs. Their work enhances the interpretability and robustness of machine learning models against adversarial attacks. Ren et al. [16] explored adversarial attacks and defences in deep learning, highlighting the evolving strategies and challenges in safeguarding deep learning models. Their study provides insights into the dynamic nature of adversarial threats and the need for adaptive defence mechanisms. Ryu and Choi [17] introduced a hybrid adversarial training approach for deep learning models, combining adversarial training with denoising networks to enhance robustness against adversarial examples. Their method strengthens the resilience of deep learning models in real-world scenarios. Samangouei et al. [18] proposed Defence-GAN, a technique to protect classifiers against adversarial attacks using generative models. By leveraging generative adversarial networks, their approach aims to improve the generalization and robustness of classifiers in the presence of adversarial inputs. Singh et al. [19] presented a Defence Against Adversarial Attacks Using Chained Dual-GAN Approach, highlighting the role of chained generative models in mitigating adversarial threats. Their method contributes to enhancing the security and reliability of machine learning systems. Taheri et al. [20] developed a robust defensive system against adversarial examples using generative adversarial networks, emphasizing the importance of generative models in adversarial defence. Their work addresses the challenges of adversarial attacks in machine learning applications. Wang et al. [21] proposed Immune Defence, a novel adversarial defence mechanism aimed at preventing the generation of adversarial examples. Their approach draws inspiration from immune system principles to enhance the robustness of machine learning models. Wang et al. [22] conducted a contemporary survey on adversarial attacks and defences in machine learning-powered networks, providing an overview of the current landscape of adversarial machine learning research. Their comprehensive survey serves as a roadmap for understanding and addressing adversarial threats. Yadav et al. [23] introduced an integrated Auto Encoder-Block Switching defence approach to prevent adversarial attacks. By combining autoencoders with block-switching techniques, their method aims to improve the resilience of machine learning models against adversarial perturbations.

Yang et al. [24] explored the generation of adversarial samples by manipulating image features with autoencoders. Their work showcases the potential of autoencoder-based techniques in crafting adversarial examples and understanding model vulnerabilities. Yu et al. [25] proposed a defence mechanism for adversarial examples using conditional generative adversarial networks, highlighting the role of conditional generative models in adversarial defence. Their approach improves the robustness of machine learning models against adversarial inputs. Zhao et al. [26] evaluated a GAN-based model for adversarial training, exploring the effectiveness of generative adversarial networks in enhancing model robustness. Their work provides insights into the practical implications of GANs in adversarial defence strategies. Zhou et al. [27] investigated adversarial ranking attacks and defences, focusing on techniques to protect ranking algorithms against adversarial manipulations. Their study contributes to securing ranking systems in the face of adversarial threats. The limitations across the range of papers in adversarial attacks and defences in machine learning and deep learning are varied. Many papers focus on specific defence mechanisms or attack strategies, which may not generalize well across different datasets or model architectures. Some papers lack extensive empirical validation or real-world deployment scenarios, limiting the practical applicability of their proposed methods. Additionally, most papers do not address the challenges of adaptive adversaries that can circumvent static defence mechanisms. There is also a lack of consensus on standardized evaluation metrics for comparing different defence approaches, making it challenging to assess their relative effectiveness comprehensively. Moreover, the rapid evolution of adversarial techniques necessitates continuous updates and improvements to defence strategies, highlighting the ongoing nature of research in this area.

3. PROPOSED METHODOLOGY

This proposed work aims to advance the field of human activity recognition by developing a customized CNN & L.S.T.M. architecture with a tracker capable of robust feature extraction. Fig. 1 depicts a suggested system that utilizes a primary generator, an assistant generator, a discriminator, and a classifier to generate and categorize images. The process commences with an authentic image inputted into the primary and assistance generators. The main generator creates an image, whereas the helper generator creates an image generated by the assistant. Subsequently, the image produced by the primary generator and the image generated by the assistant is forwarded to a discriminator and a classifier. The discriminator assesses the authenticity of the images, determining whether they are genuine or artificially created. The system analyzes the actual, generated, and assistant-generated images and produces labels indicating whether each image is authentic or artificially generated. At the same time, the classifier analyzes the generated images to calculate their confidence ratings for each class. The system evaluates the genuineness and accuracy of categorizing the produced photos by providing class confidence values and adversarial class values as output. These values aid in comprehending the degree of similarity between the generated images and the anticipated classes. In summary, this system improves the image creation by using an assistance generator to enhance the output and increase categorization accuracy. It also utilizes the discriminator to ensure the generated images' realistic appearance.

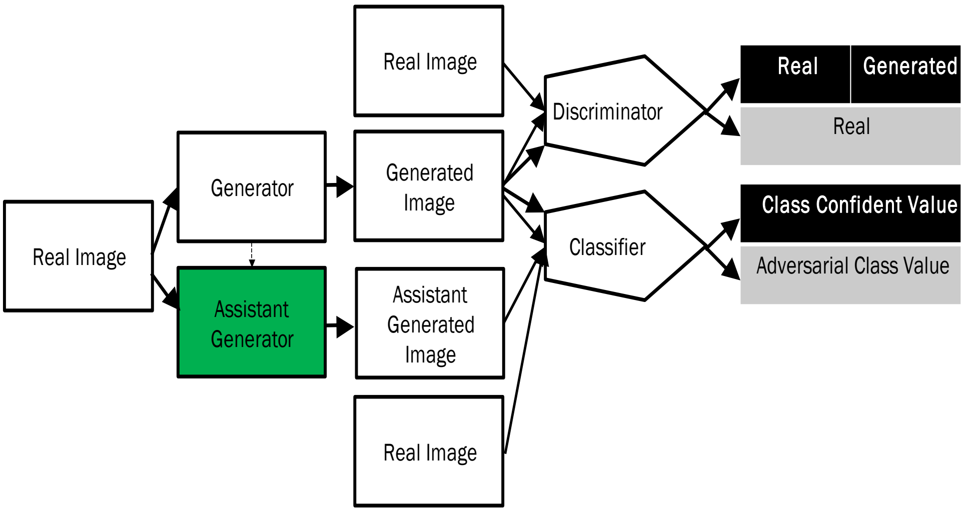


Figure 1. Proposed System Block Diagram

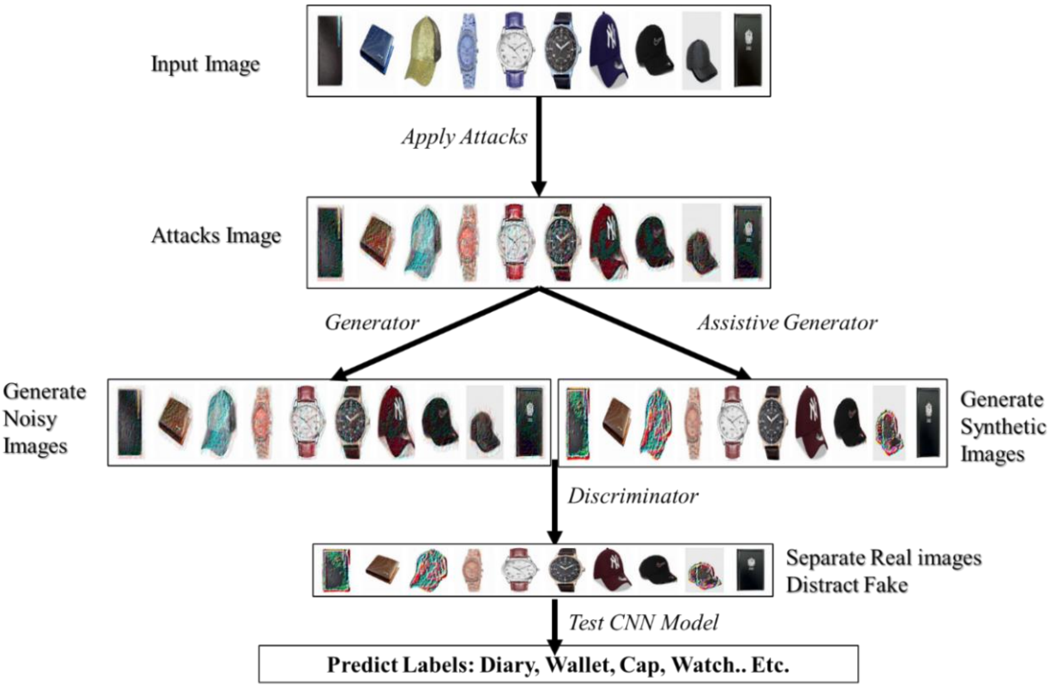


Figure 2. Proposed System Process

Figure 2 depicts the suggested system method for producing and categorizing images, specifically in the presence of adversarial attacks. The method commences with an input image subjected to a series of assaults to generate an "Attacks Image." These attacks mimic noise and distortions to assess the system's resilience. Subsequently, the "Attacks Image" is inputted into two concurrent generators: the main generator and the auxiliary generator. The main generator produces images with much noise, whereas the auxiliary generator generates synthetic images. Both sets of generated images aim to replicate the input images while incorporating alterations caused by the attacks. Subsequently, the results from both generators are transmitted to a discriminator. The function of the discriminator is to differentiate between authentic and counterfeit images, effectively segregating genuine images from artificially created ones. This stage guarantees that only the most authentic produced photos are kept for subsequent processing. Ultimately, the output of the discriminator is evaluated by employing a Convolutional Neural Network (CNN) model. The CNN approach utilizes image classification to predict labels for various objects, categorizing them into specific categories such as diaries, wallets, caps, watches, etc. The system's objective is to precisely categorize images even in the face of adversarial noise, improving image recognition models' resilience and dependability. This process utilizes a combination of adversarial training and classification techniques to enhance the performance of image recognition systems in difficult circumstances.

3.1. Pseeudo Code Input:

```
- Original image: X
- Target model: M
- Assistant Generator: G
- Adversarial attack: A

Output:
- Defended image: X_defended

Procedure:
  1. Initialize hyperparameters and model parameters for M and G.
  2. Train M and G on a dataset of clean images:
  a. Train M using a standard dataset with labels.
  b. Train G as a generative model to assist M. It learns to generate images similar to clean images.
  3. Generate assistant-generated 25 images for X:
  a. Use G to generate an assistant-generated image X_assistant = G(X).
  4. Apply the adversarial attack A to X_assistant to create an example X_adversarial.
  5. Defend against the adversarial attack using the assistant-generated image:
  a. Pass X_adversarial through M to obtain the predicted class probabilities P_before = M(X_adversarial).
```

b. Calculate the difference between the predicted class probabilities for X_adversarial and X_assistant:
Diff = |P_before - M(X_assistant)|.
c. If Diff is below a certain threshold or meets a specific criterion, consider X_adversarial a failed attack.
d. Otherwise, revert the adversarial perturbations using the assistant-generated image:
X_defended = X_adversarial - (X_adversarial - X_assistant) * α, where α is a small regularization parameter.
6. Evaluate the defended image:
a. Pass X_defended through M to obtain the final predicted class probabilities P_after = M(X_defended).
b. If P_after is similar to the predicted class probabilities for the clean image X, consider the defence successful.
c. Otherwise, adjust hyperparameters, model architectures, or retrain G and M to improve defence performance.
7. Repeat steps 3-6 for a batch of images to improve robustness.
8. Output the defended image X_defended.
End.

3.2. GAN Working

1. **Initialization**:

1. Initialize the parameters of the generator (G) and the discriminator (D) networks. Weights and biases in neural networks typically represent these parameters.

2. **Generate Fake Data**:

1. "The generator takes random noise samples (usually from a simple distribution like Gaussian) and produces synthetic data."
2. "Mathematically, this is represented as $G(z;\ \theta\_g)$, where z is the random noise vector, and θ_g represents the generator's parameters."

3. **Training Discriminator**:

1. "The discriminator takes a batch of real data samples and a batch of synthetic (fake) data samples."
2. "It is trained to correctly classify real data as real (output close to 1) and fake data as fake (output close to 0)."
3. "The loss function for the discriminator is typically the binary cross-entropy loss, such as $-\log(D(x)) - \log(1 - D(G(z)))$."
4. "Update the parameters of the discriminator to minimize this loss."

4. **Generate More Fake Data**:

1. "After training the discriminator, generate another batch of synthetic data using the updated generator."

5. **Training Generator**:

1. "The generator aims to generate data indistinguishable from real data."
2. "It does this by maximizing the probability that the discriminator classifies its generated data as real."
3. "The generator's loss function is typically $-\log(D(G(z)))$."
4. "Update the parameters of the generator to minimize this loss."

6. **Iterate**:

1. "Repeat steps 3-5 for a fixed number of iterations or until convergence."
2. "The adversarial training process involves the continual back-and-forth training between the generator and the discriminator."

7. **Convergence**:

1. "Ideally, the GAN training converges to a point where the generator produces data very similar to the real data, and the discriminator cannot reliably distinguish between real and generated data."

8. **Result**:

"Once the training is complete, you can use the trained generator to generate new data samples that resemble the training data.

4. RESULTS AND DISCUSSION

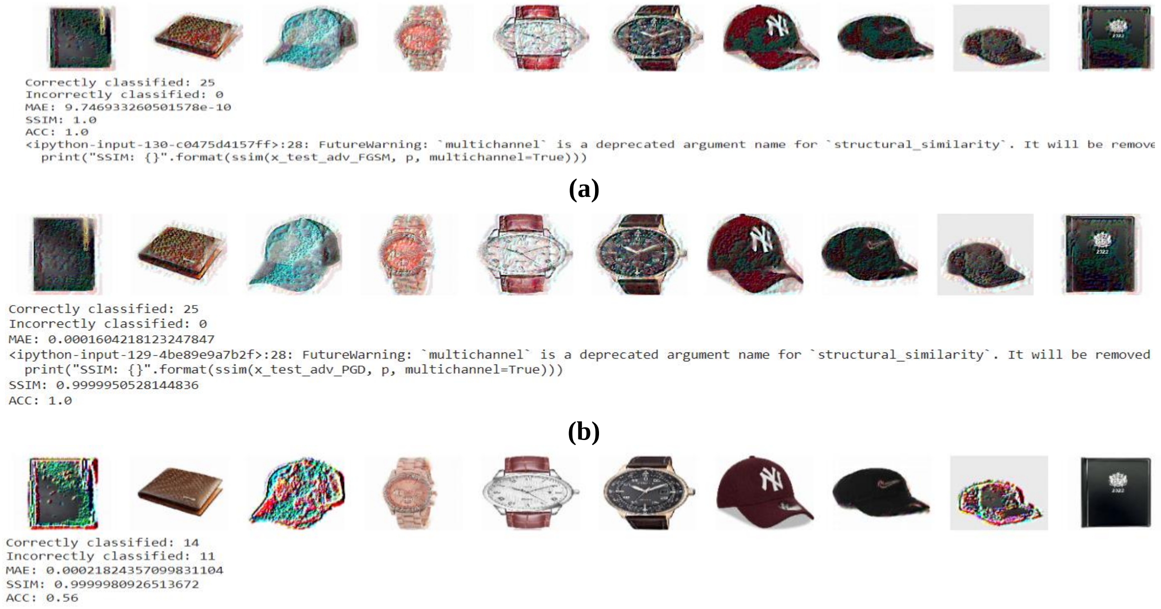
4.1. Parameters

Table 1 displays three fundamental metrics for assessing the efficacy of image processing models, particularly for noise elimination and classification tasks. The Mean Absolute Error (MAE) is used in regression tasks to calculate the average absolute discrepancies between predicted and actual values. It is employed to evaluate the quality of an image after removing noise by calculating the mean of the absolute differences between the actual and expected values. The Structural Similarity Index (SSIM) is a metric that assesses the perceived similarity of two images by considering their luminance, contrast, and structure. This measure guarantees the preservation of the structural integrity of the images following processing, employing a sophisticated formula that incorporates the means, variances, and covariances of the image intensities. Accuracy (ACC) is a metric that measures the proportion of adequately predicted cases to the total instances in classification tasks. It aids in assessing the efficacy of noise removal by measuring the model's ability to categorize the images accurately. Every statistic offers a distinct viewpoint on the model's performance, providing a thorough evaluation framework to ensure excellent image processing results. This table is essential for comprehending and implementing these criteria in real-life situations.

Table 1: Parameters Description

Metric	Parameter Name	Description	Used When	Equation
MAE	Mean Absolute Error	Measures the average absolute differences between predicted and actual values in regression tasks.	Parameter are used to check the image quality after removal of a noise.	MAE= (1/n) ∗ Σ |y_actual− y_predicted|
SSIM	Structural Similarity Index	Assesses the perceptual similarity between two images, considering luminance, contrast, and structure.		SSIM = (2 ∗ μ_x ∗ μ_y + C1) ∗ (2 ∗ σ_xy + C2) / (μ_x^2 + μ_y^2 + C1) ∗ (σ_x^2 + σ_y^2 + C2)"
ACC	Accuracy	Measures the ratio of correctly predicted instances to the total number of instances in classification tasks.	To check whether the noise from the image is removed or not.	Accuracy = (Number of Correct Predictions) / (Total Number of Predictions)

4.2. Results



```
Correctly classified: 25
Incorrectly classified: 0
MAE: 9.746933260501578e-10
SSIM: 1.0
ACC: 1.0
<ipython-input-130-c0475d4157ff>:28: FutureWarning: `multichannel` is a deprecated argument name for `structural_similarity`. It will be remove
  print("SSIM: {}".format(ssim(x_test_adv_FGSM, p, multichannel=True)))
```

(a)

```
Correctly classified: 25
Incorrectly classified: 0
MAE: 0.0001604218123247847
<ipython-input-129-4be89e9a7b2f>:28: FutureWarning: `multichannel` is a deprecated argument name for `structural_similarity`. It will be removed
  print("SSIM: {}".format(ssim(x_test_adv_PGD, p, multichannel=True)))
SSIM: 0.9999950528144836
ACC: 1.0
```

(b)

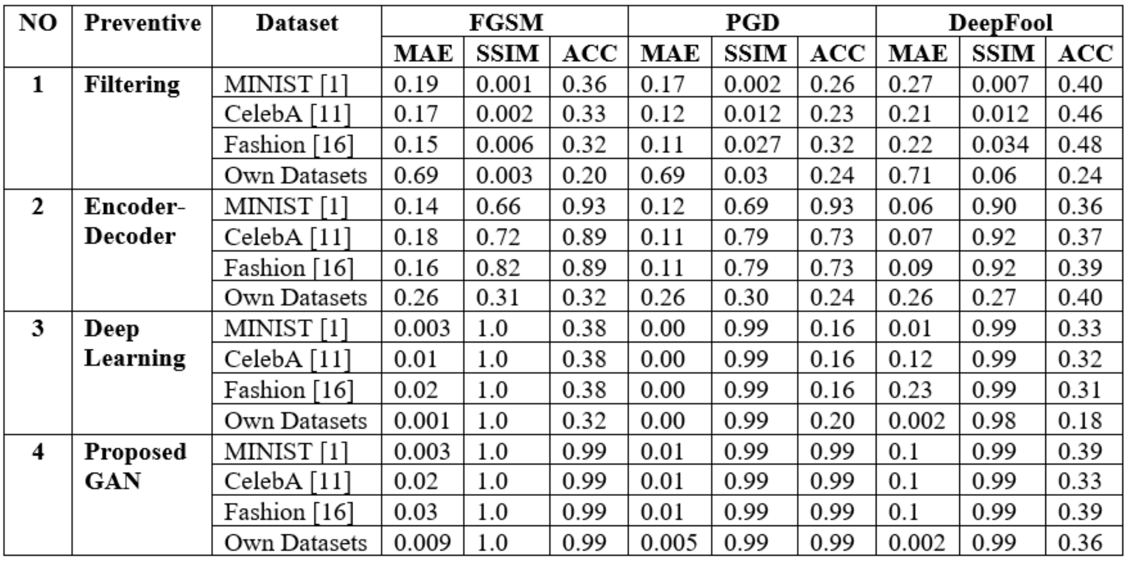
```
Correctly classified: 14
Incorrectly classified: 11
MAE: 0.00021824357099831104
SSIM: 0.9999980926513672
ACC: 0.56
```

(c)

Figure 3: (a) FGSM (b) PGD, and (c) DeepFoo Attacks Defence by Proposed Assistive GAN

Figure 3 illustrates the performance of the proposed Assistive GAN system in defending against three types of adversarial attacks: FGSM (Fast Gradient Sign Method), PGD (Projected Gradient Descent), and DeepFoo. Each subsection of the figure (a, b, and c) corresponds to one type of attack and displays a series of images along with crucial metrics for evaluating the system's performance. In Fig. 3 (a), FGSM attacks are shown. The images demonstrate how the system handles noise introduced by FGSM The metrics indicate that 25 images were correctly classified and 0 were incorrectly classified. The Mean Absolute Error (MAE) is 0.746932366501578e-18, the Structural Similarity Index (SSIM) is 1.0, and the Accuracy (ACC) is 1.0. Fig. 3 (b) shows the results under PGD attacks. Like FGSM, all 25 images are correctly classified with an MAE of 0.00016014281232487487, an SSIM of 0.9999950528144836, and an ACC of 1.0. It indicates robust performance against PGD attacks. Fig. 3 (c) presents the system's response to DeepFoo attacks. Here, 14 images are correctly classified, and 11 are incorrectly classified. The MAE is significantly lower at 0.00021842357909831104, and the SSIM remains high at 0.999998926513672, but the accuracy drops to 0.56, indicating more difficulty in handling DeepFoo attacks. Overall, the figure demonstrates the proposed system's varying effectiveness across different adversarial attacks, with the highest robustness against FGSM and PGD and a noticeable drop in performance against DeepFoo attacks. Table 2 shows the efficacy of four preventive methods—Filtering, Encoder-Decoder, Deep Learning, and the Proposed GAN—across four datasets (MNIST., CelebA, Fashion, and Own Datasets) under three types of adversarial attacks (FGSM, PGD, and DeepFool). The metrics used are MAE, SSIM, and ACC. The conclusive observation is that the Proposed GAN consistently outperforms other methods, achieving the lowest MAE, highest SSIM, and highest ACC across all datasets and attack types. It indicates that the proposed GAN is the most effective method for maintaining image quality and classification accuracy under adversarial conditions. Figure 4 displays the Mean Absolute Error (MAE), with the Filtering method exhibiting a very high MAE. This suggests that the Filtering method performs less in reducing noise, particularly when applied to the Own Datasets. Encoder-decoder techniques decrease Mean Absolute Error (MAE) compared to Filtering, suggesting superior handling of noise. Deep Learning techniques demonstrate a low Mean Absolute Error (MAE) when applied to the MNIST. and CelebA datasets exhibit a greater MAE when applied to the Fashion and Own Datasets. The Proposed GAN demonstrates the lowest Mean Absolute Error (MAE) across all datasets and attack modes, indicating its superior capability to decrease noise successfully. Figure 5 examines the Structural Similarity Index (SSIM), which shows that Filtering produces low SSIM values, indicating a subpar preservation of image quality. Encoder-Decoder techniques demonstrate superior structural similarity index (SSIM) values, especially for MNIST. and CelebA datasets, but to a lesser extent for Fashion and Own Datasets. Deep Learning techniques demonstrate high performance on the MNIST. and CelebA datasets exhibit lower effectiveness when compared to the Fashion and Own datasets. The Proposed GAN exhibits superior SSIM values across all datasets and attack types, showcasing its capacity to preserve excellent picture quality despite adversarial attacks. The Proposed GAN generally surpasses the other approaches by a large margin, attaining the highest accuracy, lowest MAE, and SSIM. It demonstrates its resilience and efficacy in countering adversarial attacks while maintaining picture quality and classification accuracy. The extensive investigation highlights the exceptional performance of the Proposed GAN in several difficult situations and datasets, establishing it as the most efficient approach among the ones assessed. Figure 6 demonstrates the precision of the techniques, indicating that Filtering has a modest level of accuracy with notable fluctuations, particularly when dealing with DeepFool attacks. The Encoder-Decoder model demonstrates superior performance, attaining high accuracy for the MNIST. and CelebA datasets, whereas its performance is lower for the Fashion and Own datasets. Deep Learning techniques provide significant accuracy when applied to the MNIST. and CelebA datasets, regardless of the type of attack. However, their performance is comparatively poorer when dealing with the Fashion and Own Datasets, particularly when subjected to the DeepFool assault. On the other hand, the Proposed GAN consistently achieves the maximum accuracy across all datasets and attack types, showcasing its greater resilience and efficacy. The figures present a comprehensive graphical overview of the performance of four distinct methods—Filtering, Encoder-Decoder, Deep Learning, and the Proposed GAN—in the face of adversarial attacks across many datasets (MNIST., CelebA, Fashion, and Own Datasets). The success of each technique is assessed using three metrics: accuracy (ACC), mean absolute error (MAE), and structural similarity index (SSIM) when subjected to FGSM, PGD, and DeepFool assaults.

Table 1: Comparative Analysis



NO	Preventive	Dataset	FGSM			PGD			DeepFool		
			MAE	SSIM	ACC	MAE	SSIM	ACC	MAE	SSIM	ACC
1	**Filtering**	MINIST [1]	0.19	0.001	0.36	0.17	0.002	0.26	0.27	0.007	0.40
		CelebA [11]	0.17	0.002	0.33	0.12	0.012	0.23	0.21	0.012	0.46
		Fashion [16]	0.15	0.006	0.32	0.11	0.027	0.32	0.22	0.034	0.48
		Own Datasets	0.69	0.003	0.20	0.69	0.03	0.24	0.71	0.06	0.24
2	**Encoder-Decoder**	MINIST [1]	0.14	0.66	0.93	0.12	0.69	0.93	0.06	0.90	0.36
		CelebA [11]	0.18	0.72	0.89	0.11	0.79	0.73	0.07	0.92	0.37
		Fashion [16]	0.16	0.82	0.89	0.11	0.79	0.73	0.09	0.92	0.39
		Own Datasets	0.26	0.31	0.32	0.26	0.30	0.24	0.26	0.27	0.40
3	**Deep Learning**	MINIST [1]	0.003	1.0	0.38	0.00	0.99	0.16	0.01	0.99	0.33
		CelebA [11]	0.01	1.0	0.38	0.00	0.99	0.16	0.12	0.99	0.32
		Fashion [16]	0.02	1.0	0.38	0.00	0.99	0.16	0.23	0.99	0.31
		Own Datasets	0.001	1.0	0.32	0.00	0.99	0.20	0.002	0.98	0.18
4	**Proposed GAN**	MINIST [1]	0.003	1.0	0.99	0.01	0.99	0.99	0.1	0.99	0.39
		CelebA [11]	0.02	1.0	0.99	0.01	0.99	0.99	0.1	0.99	0.33
		Fashion [16]	0.03	1.0	0.99	0.01	0.99	0.99	0.1	0.99	0.39
		Own Datasets	0.009	1.0	0.99	0.005	0.99	0.99	0.002	0.99	0.36

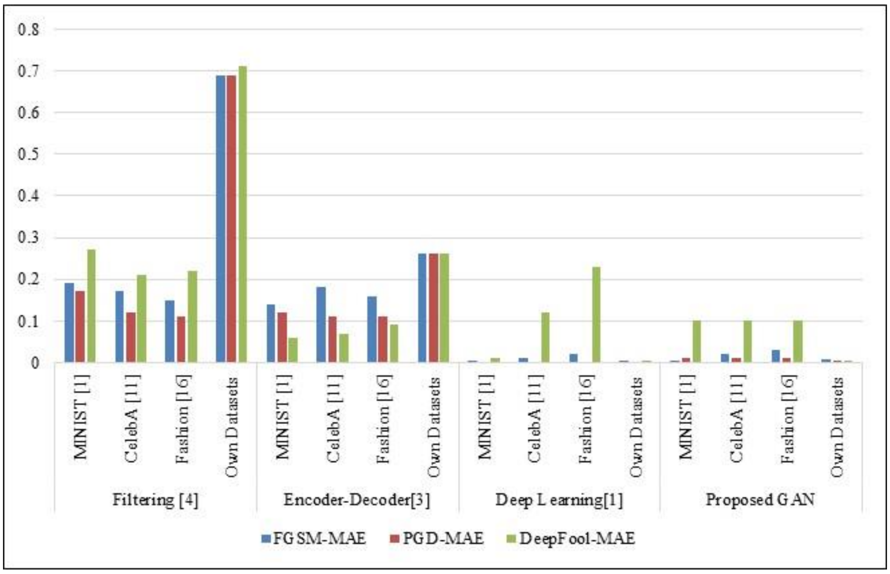


Figure 4: Graphical Analysis of MAE

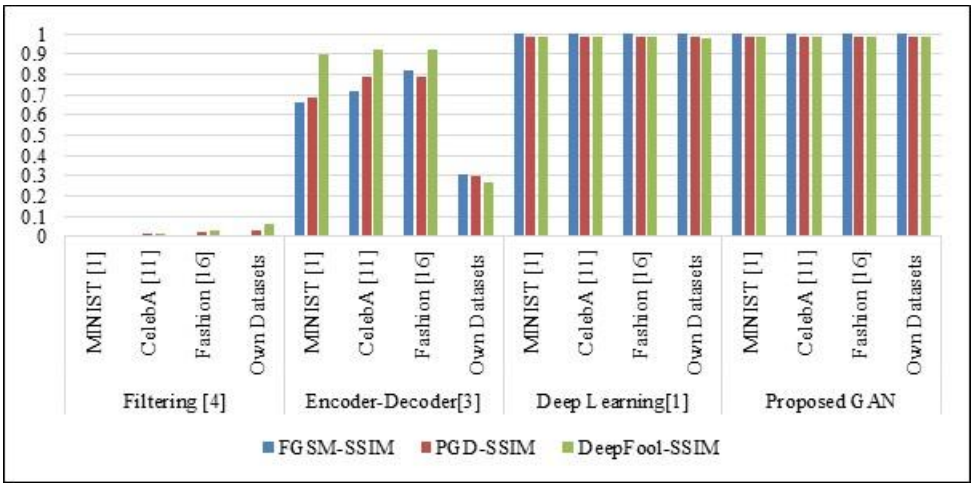


Figure 5: Graphical Analysis of SSIM

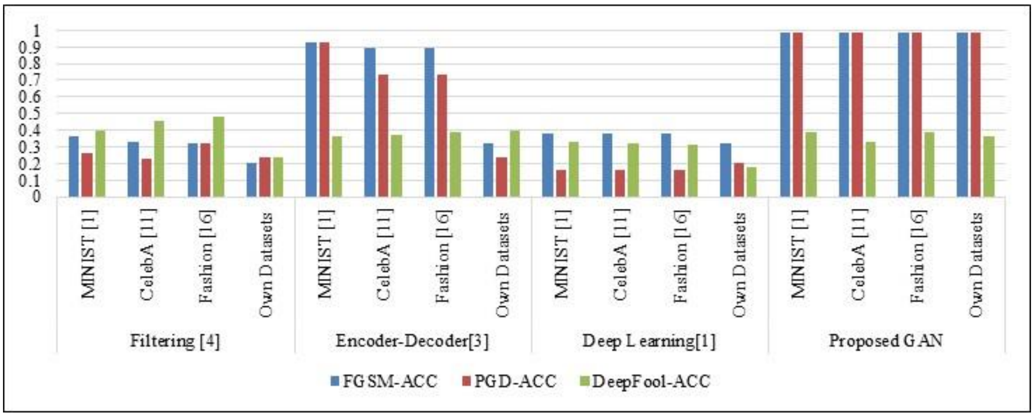


Figure 6: Graphical Analysis of Accuracy

5. CONCLUSION

The proposed Generative Adversarial Network (GAN) successfully generates diverse adversarial samples, exploiting vulnerabilities in the deep learning model and allowing for arbitrary output class manipulations. The generated adversarial samples are then used to train a defence mechanism against prominent attacks such as PGD, FGSM, and DeepFool. The defence strategy enhances model interpretability and resilience by incorporating counter-adversarial techniques, continuous monitoring, and adaptation. The robustness of the defence system is evaluated against these attacks, ensuring its ability to withstand both simple and sophisticated adversarial attempts.

Regarding "the defence, we looked at various approaches, such as defences based on Filtering, encoder-decoder, deep learning, and GAN models. To reduce adversarial assaults, each strategy demonstrated its advantages and disadvantages. The simplicity and efficacy of filtering-based defences in eliminating noise and disturbances were evident, but their ability to fend off advanced assaults was restricted. While reassembling original inputs and lessening the effect of adversarial perturbations showed potential, encoder-decoder-based defences could be vulnerable to focused or advanced assaults. While deep learning model-based defences demonstrated resilience against basic assaults and the ability to generalize to previously undiscovered hostile cases, they were susceptible to adaptive attacks. Although defences based on GAN models might produce stronger models, their efficacy depended on how closely the target and replacement models resembled each" other. Our research highlights how hostile assaults and response tactics are ever-changing. It emphasizes the importance of constant research and development to remain ahead of emerging assault strategies. Further extensive protection against adversarial assaults might be obtained by investigating hybrid techniques and integrating several" defensive mechanisms.

REFERENCES

[1] Akhtar, N., Mian, A., Kardan, N., Shah, M.: Advances in Adversarial Attacks and Defences in Computer Vision: A Survey. IEEE Access. 9, 155161–155196 (2021). https://doi.org/10.1109/ACCESS.2021.3127960.

[2] Almuflih, A.S., Vyas, D., Kapdia, V. V, Qureshi, M.R.N.M., Qureshi, K.M.R., Makkawi, E.A.: Novel exploit feature-map-based detection of adversarial attacks. Applied Sciences. 12, 5161 (2022).

[3] Bakhti, Y., Fezza, S.A., Hamidouche, W., Deforges, O.: D.D.S.A.: A Defence against Adversarial Attacks Using Deep Denoising Sparse Autoencoder. IEEE Access. 7, 160397–160407 (2019). https://doi.org/10.1109/ACCESS.2019.2951526.

[4] Dasgupta, D., Gupta, K.D.: Dual-filtering (D.F.) schemes for learning systems to prevent adversarial attacks. Complex and Intelligent Systems. (2022). https://doi.org/10.1007/s40747-022-00649-1.

[5] Deldjoo, Y., Noia, T. Di, Merra, F.A.: A Survey on Adversarial Recommender Systems: From Attack/Defence Strategies to Generative Adversarial Networks. A.C.M. Comput. Surv. 54, (2021). https://doi.org/10.1145/3439729.

[6] Despiegel, V., Despiegel, V.: ScienceDirect Dynamic Dynamic Autoencoders Autoencoders Against Against Adversarial Adversarial Attacks Attacks. 00, (2023). https://doi.org/10.1016/j.procs.2023.03.104.

[7] Folz, J., Palacio, S., Hees, J., Dengel, A.: Adversarial defence based on structure-to-signal autoencoders. Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, W.A.C.V. 2020. 3568–3577 (2020). https://doi.org/10.1109/WACV45572.2020.9093310.

[8] Han, S., Lin, C., Shen, C., Wang, Q., Guan, X.: Interpreting Adversarial Examples in Deep Learning : A Review. (2023). https://doi.org/10.1145/3594869.

[9] Hu, W., Tan, Y.: Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN BT - Data Mining and Big Data. Presented at the (2022).

[10] Jin, L., Tan, F., Jiang, S.: Generative Adversarial Network Technologies and Applications in Computer Vision. Computational Intelligence and Neuroscience. 2020, (2020). https://doi.org/10.1155/2020/1459107.

[11] Kuribayashi, M.: Defence Against Adversarial Attacks B.T. - Frontiers in Fake Media Generation and Detection. Presented at the (2022). https://doi.org/10.1007/978-981-19-1524-6_6.

[12] Laykaviriyakul, P., Phaisangittisagul, E.: Collaborative Defence-GAN for protecting adversarial attacks on classification system. Expert Systems with Applications. 214, 118957 (2023). https://doi.org/https://doi.org/10.1016/j.eswa.2022.118957.

[13] Li, F., Du, X., Zhang, L.: Adversarial Attacks Defence Method Based on Multiple Filtering and Image Rotation. Discrete Dynamics in Nature and Society. 2022, (2022). https://doi.org/10.1155/2022/6124895.

[14] Liang, H., He, E., Zhao, Y., Jia, Z., Li, H.: Adversarial Attack and Defence: A Survey. Electronics (Switzerland). 11, 1–19 (2022). https://doi.org/10.3390/electronics11081283.

[15] Liu, S., Zhuang, Y., Ma, X., Wang, H., Cao, D.: An Adversarial Sample Defence Method Based on Saliency Information B.T. - Ubiquitous Security. Presented at the (2023).

[16] Ren, K., Zheng, T., Qin, Z., Liu, X.: Adversarial Attacks and Defences in Deep Learning. Engineering. 6, 346–360 (2020). https://doi.org/10.1016/j.eng.2019.12.012.

[17] Ryu, G., Choi, D.: A hybrid adversarial training for deep learning model and denoising network resistant to adversarial examples. Applied Intelligence. 9174–9187 (2022). https://doi.org/10.1007/s10489-022-03991-6.

[18] Samangouei, P., Kabkab, M., Chellappa, R.: Defence-Gan: Protecting classifiers against adversarial attacks using generative models. 6th International Conference on Learning Representations, I.C.L.R. 2018 - Conference Track Proceedings. (2018).

[19] Singh, A.B., Awasthi, L.K., Urvashi: Defence Against Adversarial Attacks Using Chained Dual-GAN Approach B.T. - Smart Data Intelligence. Presented at the (2022).

[20] Taheri, S., Khormali, A., Salem, M., Yuan, J.S.: Developing a robust defensive system against adversarial examples using generative adversarial networks. Big Data and Cognitive Computing. 4, 1–15 (2020). https://doi.org/10.3390/bdcc4020011.

[21] Wang, J., Wu, H., Wang, H., Zhang, J., Luo, X., Ma, B.: Immune Defence: A Novel Adversarial Defence Mechanism for Preventing the Generation of Adversarial Examples. (2023).

[22] Wang, Y., Sun, T., Li, S., Yuan, X., Ni, W., Hossain, E., Poor, H.V.: Adversarial Attacks and Defences in Machine Learning-Powered Networks: A Contemporary Survey. 1–46 (2023).

[23] Yadav, A., Upadhyay, A., Sharanya, S.: An integrated Auto Encoder-Block Switching defence approach to prevent adversarial attacks. (2022).

[24] Yang, J., Shao, M., Liu, H., Zhuang, X.: Generating adversarial samples by manipulating image features with auto-encoder. International Journal of Machine Learning and Cybernetics. 14, 2499–2509 (2023). https://doi.org/10.1007/s13042-023-01778-w.

[25] Yu, F., Wang, L., Fang, X., Zhang, Y.: The defence of adversarial example with conditional generative adversarial networks. Security and Communication Networks. 2020, (2020). https://doi.org/10.1155/2020/3932584.

[26] Zhao, W., Mahmoud, Q.H., Alwidian, S.: Evaluation of GAN-Based Model for Adversarial Training. Sensors. 23, (2023). https://doi.org/10.3390/s23052697.

[27] Zhou, M., Niu, Z., Wang, L., Zhang, Q., Hua, G.: Adversarial Ranking Attack and Defence. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 12359 LNCS, 781–799 (2020). https://doi.org/10.1007/978-3-030-58568-6_46.