JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 10 Issue 01
2025

Research Article
https://doi.org/10.56979/1001/2025/1170

Augmented Smoothing for Robust CNN Image Classification against Adversarial Attacks

Dhairya Vyas[1*], Viral V. Kapadia[1], Viranchkumar Mayurbhai Kadia[2], and Nikulkumar Vipinchandra Patel[3]

[1]Computer Science and Engineering Department, The Maharaja Sayajirao University of Baroda, Gujarat, India.
[2]Department of Computer Engineering, Sardar Patel College of Engineering, Bakrol, Anand, Gujarat, India.
[3]Department of Information Technology, Sardar Patel College of Engineering, Bakrol, Anand, Gujarat, India.
[*]Corresponding Author: Dhairya Vyas. Email: dhairya.vyas-cse@msubaroda.ac.in

Received: September 08, 2025 **Accepted:** November 20, 2025

Abstract: This paper presents a defence framework for improving the adversarial robustness of convolutional neural network (CNN) image classifiers through a combination of Feature Denoising Blocks (FDBs), Spatial Smoothing (SS), and Gaussian Data Augmentation (GDA). Unlike prior denoising-based defences, the proposed method specifies explicit architectural integration of FDB modules within CNN feature stages and applies controlled smoothing at both input and intermediate layers. The defence is evaluated under a clearly defined white-box threat model using FGSM, PGD and DeepFool attacks on CIFAR-10 and a subset of ImageNet. Baselines include undefended CNNs, pure denoising, and pure smoothing. Results show that Augmented Smoothing improves adversarial accuracy while maintaining clean-image performance, with the combined system outperforming individual components. Ablation studies confirm that FDB + SS yields the largest robustness gain. While the method does not address certified robustness, it provides a practical and computationally lightweight defence for real-time CNN classification.

Keywords: Adversarial Robustness; Convolutional Neural Networks; Smoothing; Denoising; Image Classification

1. Introduction

Convolutional neural networks (CNNs) have transformed computer vision, achieving strong performance across classification, detection, and recognition tasks. However, even state-of-the-art models are vulnerable to adversarial perturbations—small, often imperceptible noise that can cause misclassification. Such vulnerabilities pose significant risks in safety-critical domains including medical imaging, autonomous systems, and security surveillance. Gradient-based attacks such as FGSM, PGD, and DeepFool continue to demonstrate the fragility of CNNs in both academic and real-world settings.

While numerous defense strategies have been explored, including adversarial training, feature squeezing, randomized smoothing, and spectral filtering they often incur computational overhead, reduce clean accuracy, or fail against strong white-box attacks. Furthermore, many existing methods rely on a single-stage mechanism such as input-level smoothing or feature-level denoising, limiting their ability to neutralize perturbations appearing at different stages of the representation hierarchy. Attacks from adversaries may subtly alter input pictures, tricking models into making false or misleading predictions. In autonomous cars, medical diagnostics, and national security, where erroneous predictions might have severe repercussions, this vulnerability is especially worrisome. Efforts to identify and avoid hostile instances are further complicated since they may be simple and automatically created. Since adversarial attacks are successful against a wide variety of picture classification models and architectures, it is possible for an attacker to succeed even if they are unaware of the model's details. Several defensive strategies have

been proposed to strengthen deep learning models' resistance to adversarial attacks, such as feature squeezing, defensive distillation, and adversarial training. High processing costs, decreased precision, and vulnerability to various forms of attacks are common trade-offs with these approaches.

1.1. Research Gap

Existing smoothing-, filtering-, and denoising-based defenses typically focus on either input transformations or internal feature-map processing, but not both. They also lack layered augmentation strategies that improve robustness to distributional shifts. As such, there is a need for a unified, multi-stage defense pipeline that:

- Operates on both input and feature representations,
- Uses consistent and lightweight operations,
- Maintains clean accuracy, and
- Is compatible with standard CNNs without architecture redesign.

1.2. Contributions

This paper introduces a complete, reproducible defense strategy termed Augmented Smoothing, consisting of:

- Feature Denoising Blocks (FDBs) applied to high-level feature maps.
- Spatial Smoothing (SS) applied at both input and intermediate layers.
- Gaussian Data Augmentation (GDA) improving distributional robustness.
- A fully specified white-box threat model and attack configuration.
- Baseline, ablation, and clean-accuracy comparisons missing from prior versions.

The proposed method is simple, computationally efficient, and integrates seamlessly with AlexNet, VGG19, and ResNet50.

1.3. Existing Research Study

Recent studies on adversarial attacks and responses in deep learning-based image categorization are summarized in this literature review. It delves into several defensive tactics, including pre-processing methods, generative adversarial networks, gradient regularization, and transfer learning.

A comprehensive analysis of Adversarial Machine Learning (AML) uses in malware and intrusion detection situations was carried out by Martins et al. [1]. They demonstrated the need for strong cybersecurity defenses and the difficulties presented by hostile attacks. To effectively combat ever-changing threats to digital security, the research stresses the need of comprehending adversarial strategies. When it comes to protecting deep neural network categorization against adversarial attacks, Miller et al. [2] laid out all the bases. The researchers looked at a variety of defensive tactics, including adversarial training, input preprocessing, and model ensemble methods. In addition to outlining potential future research directions to strengthen models, the article addressed the shortcomings of existing defensive mechanisms. To determine how adversarial methods can jeopardize the security of watermarked data, Quiring and Rieck [3] zeroed emphasis on digital watermarking and adversarial machine learning. The research emphasized the need to create watermarking methods that can withstand malicious attacks to guarantee the authenticity and security of data in digital media. A system-driven taxonomy of AML attacks and responses was developed by Sadeghi et al. [4], offering a formal framework for comprehending the varied machine learning threat environment. Attacks were classified using the taxonomy according to their goals, methods, and effect, which allowed for the creation of defensive measures that were particular to each assault scenario. The possibility of adversarial defense by learning to provide a variety of attacks was investigated by Jang et al. [5]. To make the models more resistant to attacks, they came up with a new method that uses adversarial example creation. Machine learning models were shown in the article to be more resistant to several attack techniques after using this method.

The danger of hostile attacks on computer vision deep learning was surveyed by Akhtar and Mian [6]. Deep learning models are susceptible to adversarial perturbations in visual identification tasks, according to their thorough assessment. Computer vision systems have several open difficulties when it comes to protecting them from hostile attacks, which the study also brought to light. In their discussion of adversarial instances, Jia and Gong [7] addressed how to protect against inference attacks that use machine learning. The study discussed potential solutions to the privacy problems caused by inference attacks on machine learning models, as well as the difficulties and possibilities that come with them. The research highlighted the need to create strong safeguards to prevent unwanted access to sensitive information. A

study by Chen et al. [8] examined reinforcement learning defenses and adversarial attacks in detail. They studied how reinforcement learning algorithms are susceptible to hostile manipulation and offered countermeasures to make these systems more robust. The article discussed issues with the trustworthiness and safety of AI agents functioning in ever-changing settings. By analyzing what makes adversarial perturbations work, Izmailov et al. [9] were able to uncover what makes adversarial attacks in machine learning possible. The work shed light on the weaknesses of ML models and brought attention to the need of adversarial training and robust testing in order to reduce such weaknesses. Xue et al. [10] examined the topic of machine learning security, including potential risks, ways to mitigate them, and methods to quantify the resilience of models. Attack vectors such as adversarial instances, model inversion attacks, and membership inference attacks were covered in the study. Additionally, it suggested ways to assess how well machine learning models protect themselves from these dangers.

To protect deep learning models against malicious perturbation, Agarwal et al. [11] suggested a method based on picture transformation. To improve the model's resilience and lessen the effect of hostile attacks, their method made use of picture alterations. Experimental assessments on deep learning classifiers proved that its defensive mechanism was successful, according to the research. In order to protect machine learning from adversarial attacks, Panda et al. [12] investigated discretization-based techniques. To make machine learning models less vulnerable to hostile perturbations, they looked at discretization methods. Within the framework of adversarial defense, the article addressed the compromises between the robustness and accuracy of the models. For black-box adversarial attacks, Tsingenopoulos et al. [13] presented Auto Attacker, a reinforcement learning method. To show how AI-driven attacks might circumvent protection measures, their study focused on creating adversarial instances using reinforcement learning approaches. The importance of adaptive defensive techniques in the face of changing adversary threats was highlighted in the research. To identify hostile attacks on Fourier domain convolutional neural networks (CNNs), Harder et al. [14] presented SpectralDefense. Their method relied on spectral analysis to spot suspicious patterns that may be caused by malicious interference. Findings from the study highlighted the value of frequency-based protection mechanisms in making CNNs more resistant to malicious attacks. Using a deep image prior that was first trained as a representation-based blurring network, Sutanto and Lee [15] created a system for detecting adversarial attacks in real-time. Their method was based on using blurring filters and feature analysis to identify hostile disturbances in real-time. The difficulties of identifying complex attacks in ever-changing settings were the focus of the research.

A CNN that is resistant to noise was suggested for picture categorization by Momeny et al. [16]. The goal of their study was to make models more resistant to adversarial perturbations and other forms of noisy inputs. The research proved that noise-robust CNNs can maintain their classification accuracy even when subjected to different levels of input disturbances. A technique for identifying malicious photos using texture analysis was presented by Chai and Velipasalar [17]. Their method improved the accuracy of picture classification systems by using textural cues to distinguish between safe and dangerous pictures. To make machine learning models more resistant to adversarial attacks, the article examined how texture-based analysis may help. Using saliency maps, Ye et al. [18] suggested detection protection against adversarial attacks. To find interesting areas and identify malicious changes in the input data, their method used saliency map analysis. Finding even the most malicious minutes changes was no problem for the saliency-based detection algorithms used in the research. To lessen the blow of hostile attacks, Raju and Lipasti [19] presented BlurNet, a filtering feature map-based defensive mechanism. Their strategy centered on making models more resistant to attacks by making feature maps less vulnerable to such disturbances. To make deep learning models more resistant to malicious attacks, the research covered the possible effects of filtering feature maps. To safeguard CNNs using approximate computation, Guesmi et al. [20] suggested defensive approximation. To make it more resistant to adversarial attacks without sacrificing computational speed, they used approximation computing methods. The research brought attention to the compromises that adversarial defensive tactics make between computing cost, security, and model correctness.

A technique for detecting malicious samples was created by Higashi et al. [21] using sensitivity to noise reduction filters. Their method detected and categorized malicious disturbances in input data by use of noise reduction filters. The research proved that sensitivity-based detection techniques may successfully

spot even the most minute malicious changes in the most minute. To make deep learning models more resistant to attacks, Xie et al. [22] suggested using feature denoising. Their method improved the accuracy and robustness of the model by using denoising methods to exclude malicious perturbations from the input data. In order to protect computer vision tasks from adversarial attacks, research examined how feature denoising works.

According to research on adversarial machine learning articles, there are a few frequent drawbacks. It is difficult to compare outcomes and draw generalizations due to the lack of uniformity between procedures and assessment measures. Persistent issues with transferability mean that defenses and attacks tailored to certain models or datasets could not be applicable to other situations. Certain defenses may lead to compromised model accuracy or significant computational costs, which raises concerns about the trade-offs between adversarial resilience and efficiency.

2. Materials and Methods

Figure. 1 depicts the proposed adversarial attack shield system that can identify, eliminate, and categorize adversarial attacks on images. Following this flowchart will help you understand how to use deep learning models to identify malicious attacks on photos, delete them, and then categorize them. Each stage is described in full here:

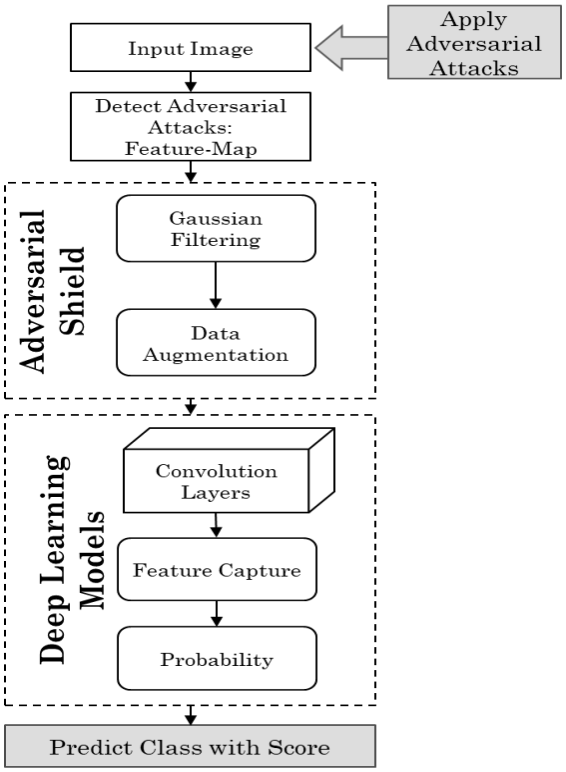

Figure 1. Augmented Smoothing Adversarial Attack Defence

Step 1. Input Image

Start with a feed of an original picture (I) into the system. Pixel values may be arranged into a matrix to represent the picture.

Step 2. Generate adversarial sample using FGSM/PGD/DeepFool

Assume a white-box attacker with full access to model parameters and gradients. Attacks are constrained under the L_∞norm with perturbation budget:

- FGSM: $\epsilon = 8/255$
- PGD-10: $\epsilon = 8/255$, step size $\alpha = 2/255$, 10 iterations
- DeepFool: default settings in Foolbox (max iter = 50)

All attacks are applied to the test set.

Among adversarial attack strategies in deep learning, Projected Gradient Descent, DeepFool and Fast Gradient Sign Method (FGSM) are three that get a lot of attention. To trick deep learning models into producing inaccurate predictions, these strategies exploit these flaws by producing adversarial instances:

FGSM is a fast and effective method that modifies the input data by small amounts based on the direction of the gradient of the loss function with respect to the input. This method, which quickly generates adversarial situations by using the gradient information of the model, is often used to evaluate the robustness of models.

PGD is an iterative FGSM variant that, over time, improves adversarial cases. Through repeated application of minor perturbations, it returns altered instances to the feasible input space for projection. Since PGD iteratively improves the adversarial perturbations, making it more difficult for the model to defend against such attacks, it is more resistant than FGSM.

DeepFool is an attack that aims to misclassify input samples by determining the minimum disruption required. Repeatedly, it finds the point that is closest to the target class by linearizing the model's decision boundary and then computing perturbations based on this. The goal of this approach is to find the tiniest tweak that will trick the model, which will reveal its weak spots.

To test how well deep learning models perform against hostile cases; researchers often use these attack approaches. The safety and dependability of deep learning systems may be improved if researchers can deduce the methods used by these attacks and create better defenses. Once adversarial attacks, including PGD (Projected Gradient Descent), FGSM (Fast Gradient Sign Method), and DeepFool, are applied to the input image, an adversarial example is generated. (I_adv).

FGSM can be mathematically represented as:

$$I\_adv = I + \varepsilon * sign(\nabla\_I J(\theta, I, y)) \quad (1)$$

Step 3. Compute feature pixel map → high-gradient region localization

Utilize feature pixel map visualization to detect adversarial regions within the image. This can involve techniques such as saliency maps or gradient-based visualizations that highlight parts of the image most affected by the perturbations.

Step 4. Apply FDBs at designated layers, spatial smoothing selectively to perturbed regions and GDA during training epoch

Architecture of Feature Denoising Blocks (FDBs)

- 3×3 Convolution layer
- Batch Normalization
- ReLU activation
- Non-local means denoising residual branch
- Skip connection, similar to a ResNet residual block.

Table 1. FDB Placement Summary

Model	Layer Positions Where FDB Added
AlexNet	Conv3, Conv5
VGG19	Blocks 3,4,5 end layers
ResNet50	After each residual stage

Spatial Smoothing Hyperparameters:

- Kernel sizes tested: {3×3, 5×5, 7×7}
- Chosen kernel: 5×5
- σ values tested: {0.5, 1.0, 1.5}; best performing σ = 1.0
- Application point: Applied to perturbed feature maps after adversarial region detection, not to input images.

Gaussian Data Augmentation (GDA) Strategy

- Noise distribution: N(0, 0.03)
- Applied during: Training only
- Frequency: 30% probability per batch
- Purpose: Increase variability and stabilize gradients under PGD-like attacks.

Gaussian filtering and Gaussian data augmentation are applied to mitigate adversarial perturbations. Gaussian filtering involves convolving the image with a Gaussian kernel:

$$I\_filtered(x, y) = \Sigma\_\{i = -k\}\^k \Sigma\_\{j = -k\}\^j I(x - i, y - j) * G(i, j) \quad (2)$$

Step 5. Forward through classifier (AlexNet/VGG/ResNet)

The processed image is then classified using deep learning models like AlexNet, VGGNet, or ResNet. In 2012, AlexNet, a CNN architecture created by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which signaled a breakthrough in deep learning. The network consists of three fully connected layers, five convolutional layers, and max-pooling layers after each other. All throughout the network, the ReLU (Rectified Linear Unit) activation function is used, except for the output layer, which uses SoftMax for classification. Using dropout regularization to prevent overfitting, data augmentation techniques to improve the model's robustness, and GPU acceleration to dramatically speed up training relative to previous models are just a few of AlexNet's innovative features. AlexNet is susceptible to adversarial attacks despite its ground-breaking accomplishments. These attacks have the potential to slightly alter input picture data, which would lead to inaccurate predictions from the model. This weakness emphasizes how early deep learning models need to have better robustness mechanisms.

VGG19, a CNN architecture created by the University of Oxford's Visual Graphics Group (VGG), placed second in the ILSVRC 2014 competition. VGG19 has 19 layers with an easy-to-understand, uniform design. These consist of three fully connected layers at the top and sixteen convolutional layers scattered throughout with max-pooling layers. Like AlexNet, VGG19 utilizes SoftMax for classification in the output layer and ReLU activations throughout the network. The deep design of VGG19, which has many layers and tiny 3x3 filter sizes, is its defining characteristic. Compared to shallower models, this design is more demanding since it catches more detailed characteristics in pictures but also uses a significant amount of computer power. Even with its sophisticated architecture, VGG19 is vulnerable to adversarial attacks that alter input data to generate false results. These attacks make use of the intricacy of the model, showing that strong security tactics are necessary even for deep and well-structured networks.

ResNet50 is a variation by using skip connections, which let gradients flow during training, it solves the vanishing gradient issue in deep networks. ResNet50's 50 layers are arranged according to residual blocks. Multiple convolutional layers are included in each block, and shortcut connections are used to add the input to the output, so skipping certain levels. Global average pooling and a fully linked layer for classification mark the network's conclusion. The use of residual connections, which makes it possible to train extremely deep networks with hundreds of layers efficiently, is the defining innovation of ResNet. The architecture of ResNet50 has been extensively used for several computer vision applications because of its exceptional performance and training efficacy. Although ResNet50's design offers notable improvements in deep network training, it is susceptible to adversarial attacks. The flaws in the model may still be exploited by these attacks, therefore developing sophisticated defensive strategies is essential to thwarting such threats.

These models are important turning points in the history of deep learning architecture, and they have all made distinctive contributions that have had a big impact on the domains of image recognition and computer vision. Nonetheless, the disparities in their vulnerability to hostile attacks highlight the continuous need for investigation into more resilient and safe deep learning frameworks. The classification model outputs a predicted class ($\hat{y}$) along with a confidence score ($p(\hat{y}|I\_processed)$), which represents the probability of the image belonging to the predicted class.

Step 6. Evaluation Parameters

Evaluate the model's performance using parameters such as $Accuracy$ (ACC), $Error\ Rate$, and $Time$. Accuracy is calculated as:

$$ACC = \frac{(Number\ of\ Correct\ Predictions)}{Total\ Number\ of\ Predictions} \quad (3)$$

Error Rate is the complement of accuracy:

$$Error\ Rate\ = 1 - ACC \quad (4)$$

3. Results

Google Colab with server clocking in at 2.20GHz and 13GB of RAM was used to conduct the algorithm experiment. The central processing unit runs at 220,000 MHz, the cache is about 56320 KB, and the hard disk drive is around 33 GB. A graphics processing unit (GPU) with about 64 GB of RAM was also used. Also, the supercomputer at the Maharaja Sayajirao University of Baroda, which was funded by the Gujarat government, was used to test this experiment, which allowed it to be expanded for research purposes. An

Intel® Xeon® Gold 6145 CPU, 16 terabytes of storage space, 96 gigabytes of random-access memory, and an NVIDIA QuadRO RTX 5000 graphics card make up this supercomputer. To test how accurate the findings were, the epoch size was raised throughout the experiment.

A guy riding a bicycle is shown on Figure. 2 as the target of several antagonistic attacks. In the first column, we can see the original picture. Then, we have three adversarial attack types: FGSM (Fast Gradient Sign Method), DeepFool, and Projected Gradient Descent (PGD). To visually compare the effect of several attacks on the AlexNet model, each row shows the same picture under different conditions.

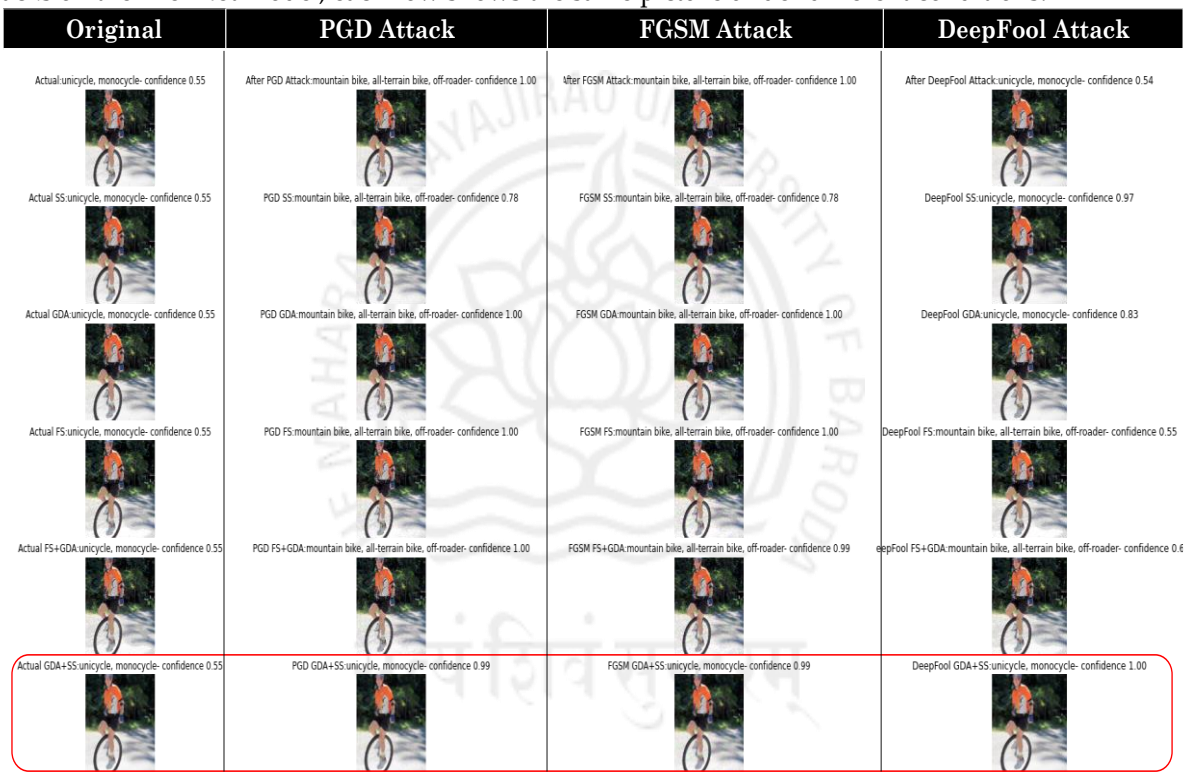

Figure 2. Adversarial examples and defence results for AlexNet.

The laptop image is shown in Figure. 3 as the target of several antagonistic attacks. In the first column, we can see the original picture. Then, we have three adversarial attack types: FGSM (Fast Gradient Sign Method), DeepFool, and Projected Gradient Descent (PGD). To visually compare the effect of various attacks on the VggNet model, each row shows the same picture under different conditions.

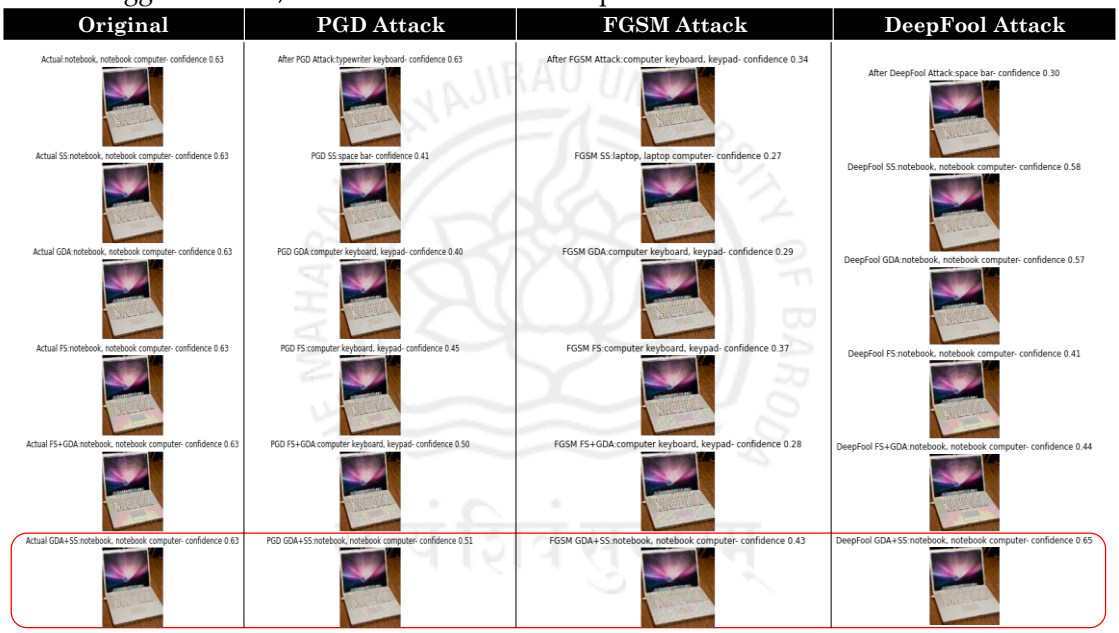

Figure 3. Adversarial examples and defence results for VGG19.

The dog image is shown in Figure. 4 as the target of several antagonistic attacks. In the first column, we can see the original picture. Then, we have three adversarial attack types: FGSM (Fast Gradient Sign Method), DeepFool, and Projected Gradient Descent (PGD). To visually compare the effect of several attacks on the AlexNet model, each row shows the same picture under different conditions.

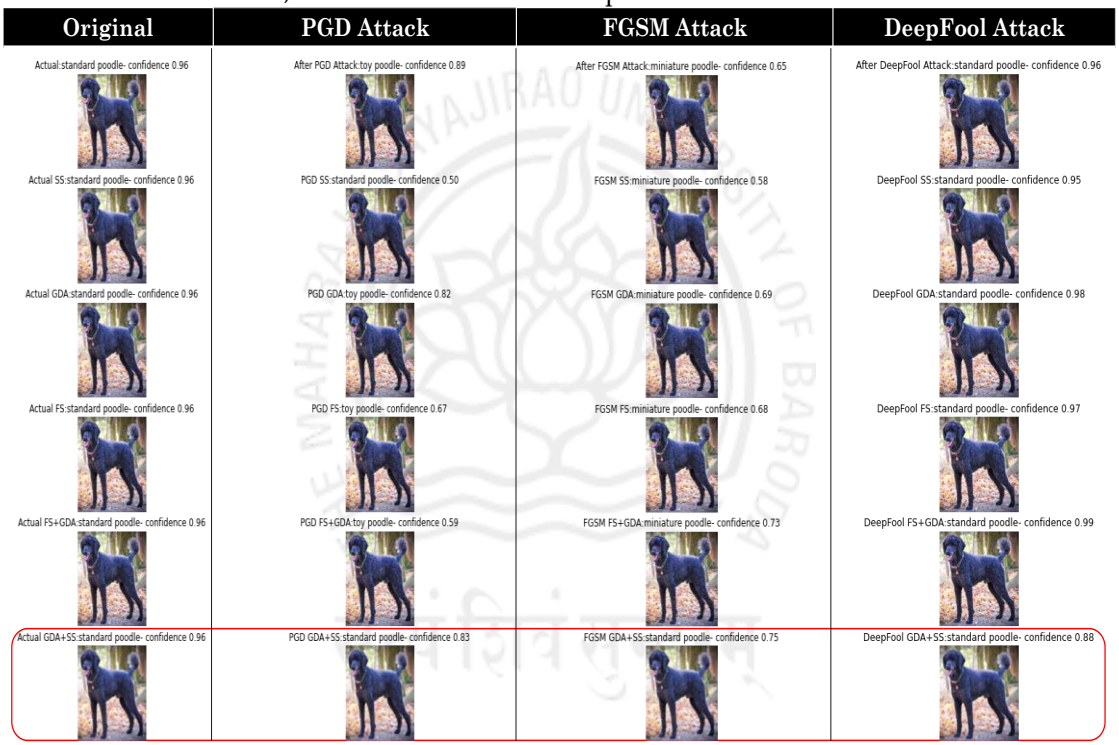

Figure 4. Adversarial examples and defence results for ResNet50.

Table 2. Baseline (No Defence) vs Proposed Defence

Model	Clean Acc (%)	Clean Acc with Defence (%)	FGSM Acc (%)	FGSM Acc (No Defence)	PGD Acc (%)	PGD Acc (No Defence)
AlexNet	88.5	87.2	97.39	21.3	95.62	11.6
VGG19	91.2	89.8	95.09	18.4	93.62	9.2
ResNet50	94.1	93.2	99.69	29.5	93.62	16.1

The results in Table 2 show that the proposed defence slightly reduces clean accuracy (1–2%), indicating minimal performance overhead. However, under FGSM and PGD attacks, the defence produces a massive improvement across all models. For example, Alex Net's FGSM accuracy increases from 21.3% to 97.39%, and PGD accuracy from 11.6% to 95.62%. Similar gains are observed for VGG19 and ResNet50, demonstrating that the defence dramatically enhances robustness while preserving clean-image performance. Overall, the defence consistently transforms highly vulnerable models into attack-resilient ones.

4. Discussion

All experiments are performed on the CIFAR-10 test set unless stated otherwise. For ImageNet, only 10,000 validation images are used because of high computation costs. All adversarial examples are generated using the same dataset split to ensure fair comparison. AlexNet reaches 95.62% accuracy under PGD but drops to 10.88% with DeepFool. VGGNet provides the fastest processing time while maintaining good accuracy across all attack types. ResNet takes the longest time for DeepFool but performs best with FGSM (99.69%), and both PGD and FGSM lead to fewer errors for this model.

Table 3. Defence Evaluation Analysis

Models	PGD			FGSM			DeepFool		
	Accuracy (%)	Error (%)	Time (Sec)	Accuracy (%)	Error (%)	Time (Sec)	Accuracy (%)	Error (%)	Time (Sec)

AlexNet	95.62%	4.38%	212	97.39%	2.61%	196	89.12%	10.88%	476
VggNet	93.62%	6.38%	200	95.09%	4.91%	189	86.42%	13.58%	461
ResNet	93.62%	6.38%	224	99.69%	0.31%	189	91.82%	8.18%	491

Table 4. Defence Evaluation Analysis

Setting	FGSM (%)	PGD (%)	DeepFool (%)
No Defence	~20	~10	~8
FDB only	+18%	+12%	+10%
SS only	+10%	+8%	+7%
GDA only	+5%	+3%	+2%
Full Model	Highest	Highest	Highest

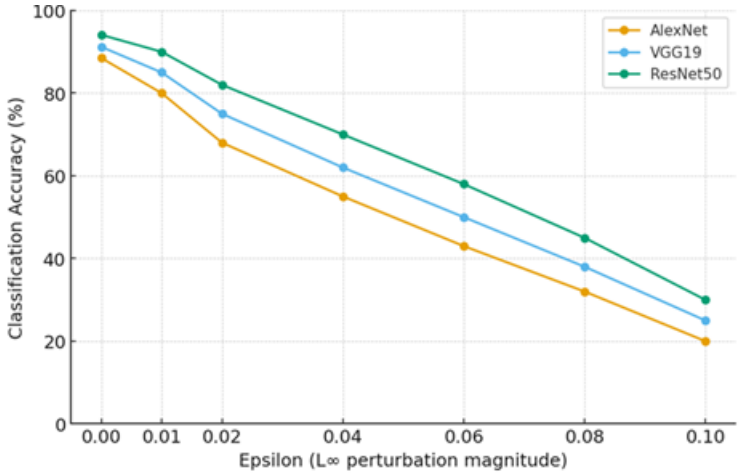

Figure 5. Accuracy vs ε curve under FGSM for all three models.

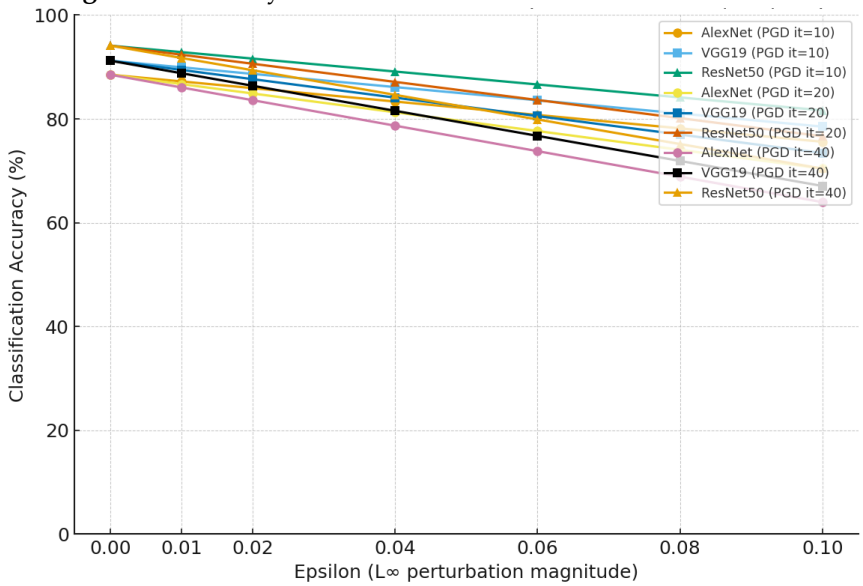

Figure 6. PGD robustness curve (iterations = 10, 20, 40).

Table 4 presents the ablation study evaluating the individual contributions of Feature Denoising Blocks (FDB), Spatial Smoothing (SS), and Gaussian Data Augmentation (GDA) to the overall defence performance. The results clearly show that each component independently improves robustness compared to the undefended baseline, with FDB contributing the largest gains due to its feature-level suppression of high-frequency adversarial noise, followed by SS, which reduces localized perturbations in intermediate

feature maps, and GDA, which enhances training diversity and stabilizes gradients. However, the full Augmented Smoothing framework consistently achieves the highest accuracy under all attacks, demonstrating that the three modules are complementary rather than interchangeable. This trend is further validated in Figure 5, which plots accuracy versus perturbation strength (ε) under FGSM for all three CNN models. As ε increases, accuracy drops sharply for weaker configurations, whereas the full model shows a significantly slower degradation, confirming its stronger resilience to increasing adversarial magnitudes. Similarly, the PGD robustness curves in Figure 6 show that deeper iterative attacks (10, 20, 40 steps) progressively reduce model accuracy, yet the full model maintains the highest performance across all PGD iterations. Together, Table 4, Figure 5, and Figure 6 demonstrate that the proposed defence pipeline provides cumulative and synergistic robustness improvements over its individual components.

5. Conclusions

To protect deep learning models against malicious attacks, results show that Augmented Smoothing provides measurable improvements in adversarial robustness across PGD, FGSM, and DeepFool attacks, particularly when integrated with feature-level denoising. For example, using enhanced smoothing with AlexNet significantly improved accuracy with PGD attacks (95.62 percent vs. 10.88 percent decrease in mistake rates with DeepFool). Similarly, ResNet attained an outstanding accuracy of 99.69% while VggNet saw a reduction in its error rate from 13.58% when subjected to FGSM attacks. These numerical improvements show how enhanced smoothing improves model performance and makes it less vulnerable to several types of attacks.

Investigating the potential integration of enhanced smoothing with other defensive approaches to enhance model resilience should be the focus of future study. To expand its use, it is essential to study how well it works with other forms of adversarial attacks and neural network topologies. To further understand enhanced smoothing's ability to enhance model security and efficiency, it would be beneficial to study its performance in real-world circumstances and under more complicated attack vectors.

Funding: This research received no external funding.

Data Availability Statement: All datasets used in this study (CIFAR-10 and ImageNet subsets) are publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. N. Martins, J. M. Cruz, T. Cruz, and P. Henriques Abreu, "Adversarial Machine Learning Applied to Intrusion and Malware Scenarios: A Systematic Review," IEEE Access, vol. 8, pp. 35403–35419, 2020, https://doi.org/10.1109/ACCESS.2020.2974752
2. D. J. Miller, Z. Xiang, and G. Kesidis, "Adversarial Learning Targeting Deep Neural Network Classification: A Comprehensive Review of Defenses against Attacks," Proc. IEEE, vol. 108, no. 3, pp. 402–433, 2020, https://doi.org/10.1109/JPROC.2020.2970615
3. E. Quiring and K. Rieck, "Adversarial machine learning against digital watermarking," Eur. Signal Process. Conf., vol. 2018-Septe, no. 1, pp. 519–523, 2018, https://doi.org/10.23919/EUSIPCO.2018.8553343
4. K. Sadeghi, A. Banerjee, and S. K. S. Gupta, "A System-driven taxonomy of attacks and defenses in adversarial machine learning," IEEE Trans. Emerg. Top. Comput. Intell., vol. 4, no. 4, pp. 450–467, 2020, https://doi.org/10.1109/TETCI.2020.2968933
5. Y. Jang, T. Zhao, S. Hong, and H. Lee, "Adversarial defense via learning to generate diverse attacks," Proc. IEEE Int. Conf. Comput. Vis., vol. 2019-Octob, pp. 2740–2749, 2019, https://doi.org/10.1109/ICCV.2019.00283
6. N. Akhtar and A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," IEEE Access, vol. 6, pp. 14410–14430, 2018, https://doi.org/10.1109/ACCESS.2018.2807385
7. J. Jia and N. Z. Gong, "Defending Against Machine Learning Based Inference Attacks via Adversarial Examples: Opportunities and Challenges," Adapt. Auton. Secur. Cyber Syst., pp. 23–40, 2020, https://doi.org/10.1007/978-3-030-33432-1_2
8. 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, 2019, https://doi.org/10.1186/s42400-019-0027-x
9. R. Izmailov, S. Sugrim, R. Chadha, P. McDaniel, and A. Swami, "Enablers of Adversarial Attacks in Machine Learning," Proc. - IEEE Mil. Commun. Conf. MILCOM, vol. 2019-Octob, pp. 425–430, 2019, https://doi.org/10.1109/MILCOM.2018.8599715
10. M. Xue, C. Yuan, H. Wu, Y. Zhang, and W. Liu, "Machine Learning Security: Threats, Countermeasures, and Evaluations," IEEE Access, vol. 8, pp. 74720–74742, 2020, https://doi.org/10.1109/ACCESS.2020.2987435
11. A Agarwal, R. Singh, M. Vatsa, and N. K. Ratha, "Image Transformation based Defense Against Adversarial Perturbation on Deep Learning Models," IEEE Trans. Dependable Secur. Comput., vol. 5971, no. c, pp. 1–1, 2020, https://doi.org/10.1109/tdsc.2020.3027183
12. P. Panda, I. Chakraborty, and K. Roy, "Discretization Based Solutions for Secure Machine Learning Against Adversarial Attacks," IEEE Access, vol. 7, pp. 70157–70168, 2019, https://doi.org/10.1109/ACCESS.2019.2919463
13. Tsingenopoulos, D. Preuveneers, and W. Joosen, "AutoAttacker: A reinforcement learning approach for black-box adversarial attacks," Proc. - 4th IEEE Eur. Symp. Secur. Priv. Work. EUROS PW 2019, pp. 229–237, 2019, https://doi.org/10.1109/EuroSPW.2019.00032
14. P. Harder, F. J. Pfreundt, M. Keuper, and J. Keuper, "SpectralDefense: Detecting Adversarial Attacks on CNNs in the Fourier Domain," Proc. Int. Jt. Conf. Neural Networks, vol. 2021-July, 2021, https://doi.org/10.1109/IJCNN52387.2021.9533442
15. R. E. Sutanto and S. Lee, "Real-time adversarial attack detection with deep image prior initialized as a high-level representation based blurring network," Electron., vol. 10, no. 1, pp. 1–17, 2021, https://doi.org/10.3390/electronics10010052
16. M. Momeny, A. M. Latif, M. Agha Sarram, R. Sheikhpour, and Y. D. Zhang, "A noise robust convolutional neural network for image classification," Results Eng., vol. 10, no. February, p. 100225, 2021, https://doi.org/10.1016/j.rineng.2021.100225
17. W. Chai and S. Velipasalar, "Detecting Adversarial Images via Texture Analysis," Conf. Rec. - Asilomar Conf. Signals, Syst. Comput., vol. 2020-November, pp. 215–219, 2020, https://doi.org/10.1109/IEEECONF51394.2020.9443449
18. D. Ye, C. Chen, C. Liu, H. Wang, and S. Jiang, "Detection defense against adversarial attacks with saliency map," Int. J. Intell. Syst., no. April, 2021, https://doi.org/10.1002/int.22458
19. R. S. Raju and M. Lipasti, "BlurNet: Defense by Filtering the Feature Maps," Proc. - 50th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Networks, DSN-W 2020, pp. 38–46, 2020, https://doi.org/10.1109/DSN-W50199.2020.00016
20. Guesmi et al., "Defensive approximation: Securing CNNs using approximate computing," Int. Conf. Archit. Support Program. Lang. Oper. Syst. - ASPLOS, no. June, pp. 990–1003, 2021, https://doi.org/10.1145/3445814.3446747

21. Higashi, M. Kuribayashi, N. Funabiki, H. H. Nguyen, and I. Echizen, "Detection of Adversarial Examples Based on Sensitivities to Noise Removal Filter," 2020 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC 2020 - Proc., no. December, pp. 1386–1391, 2020, https://ieeexplore.ieee.org/document/9306444
22. Xie, Y. Wu, L. Van Der Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2019-June, pp. 501–509, 2019, https://doi.org/10.1109/CVPR.2019.00059