Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Adversarial Perturbation Defense Algorithms via Manifold Projections and Denoising

Dr. Megala Rajendran[1*], Dr.R. Udayakumar[2], B. Mohanraj[3], Nozimabonu Abdushukurova[4], Orifjon Talipov[5], Asaliddin Kubayev[6]

[1*]Vice Rector, Research & Innovation, Turan International University, Namangan, Uzbekistan. Email: megala11379@gmail.com, https://orcid.org/0009-0005-9605-5958

[2]Professor & Director, Kalinga University, India. Email: rsukumar2007@gmail.com

[3]Department of Information Technology, Sona College of Technology, Salem, India. Email: bmohanrajcse@gmail.com

[4]Lecturer, Department of Legal Sciences, Tashkent, National University of Uzbekistan named after Mirzo Ulugbek, Uzbekistan. E-mail: abdushukurova1997@mail.ru, https://orcid.org/0009-0001-3325-1274

[5]Department of Oncology, Oncohematology and Radiation Oncology, Tashkent State Medical University, Tashkent, Uzbekistan. E-mail: talipov.o.a@tashmeduni.uz, http://orcid.org/0009-0000-4712-9600

[6]Researcher, Samarkand State Medical University Samarkand, Uzbekistan. E-mail: asaliddinkubayev@gmail.com, https://orcid.org/0009-0004-7519-3086

*Corresponding author: Email: megala11379@gmail.com

Abstract

Adversarial perturbations represent an important challenge to the dependability and robustness of deep neural networks, especially in critical applications like self-driving cars, healthcare, and cybersecurity. Traditional approaches to defending neural networks, such as adversarial training, gradient masking, and input processing, usually either fail to generalize to novel attacks or impact the accuracy on unperturbed input. The proposed paper addresses this problem by developing a novel defense mechanism in which a manifold projection and a denoising autoencoder work in concert in order to defend the neural network. The former projects the input perturbed by an attacker into the low-dimensional subspace of unperturbed input data, which decreases the effect of attacks; the latter eliminates the remaining noise but preserves all necessary information. The hybrid defense is tested on CNN, ResNet-18, and VGG-16 neural networks trained on the CIFAR-10 dataset in response to FGSM, PGD, and DeepFool attacks. The experiment shows that the developed technique achieves 18%-21% improvement in adversarial robustness in comparison to existing defense methods while keeping excellent clean-data accuracy.

Keywords: Adversarial Attacks, Manifold Projection, Denoising, Robustness, Deep Learning, FGSM, PGD

1. Introduction

Adversarial perturbations are small, barely noticeable changes to the input data that can deceive machine learning models into generating erroneous results. This type of attack takes advantage of the weaknesses in the decision boundary of machine learning models, particularly those operating in high-dimensional input spaces such as images, sound, and speech signals. As more complex tasks require deep learning algorithms, which are vulnerable to these types of attacks, become common in crucial fields like self-driving vehicles, healthcare diagnostics, and information security, the vulnerability of machine learning models increases [14]. Some of the most frequently used attacks include the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and DeepFool, all of which alter the input variables to produce incorrect classifications.

In order to overcome these issues, manifold projections and denoising can be considered an excellent solution [1]. Through manifold projections, the inputs that have been subject to perturbations are mapped into a lower-dimensional manifold of clean inputs, thereby limiting the influence of adversaries on them, while through

denoising autoencoders, the noisy representation is converted into a clean one [2]. Such an approach offers two layers of defense against the issue of vulnerabilities of models to various adversarial perturbations [7][15].

Key Contribution

1) Suggested a two-step hybrid defense mechanism using manifold projection along with denoising autoencoders to counter any sort of perturbation.
2) The robustness of CNN, ResNet-18, and VGG-16 models has been significantly increased when exposed to FGSM, PGD, and DeepFool attacks without any sacrifice in terms of accuracy on the clean data set.
3) Also, an effective yet straightforward defense mechanism against any kind of attack on multiple deep-learning models across various data sets has been introduced.

In the paper, there are five primary sections. In Section I, the concept of adversarial perturbation is introduced, and its effect on deep learning models is discussed. At last, the reason behind considering manifold projection and denoising is elaborated. Section II deals with existing defense algorithms; their weaknesses and reasons for proposing the hybrid technique are described in this section. In Section III, the methodology and experiment design have been explained, where the hybrid technique, data set used (i.e., CIFAR-10), model architecture, attacks considered, and performance measurement criteria have been defined. In Section IV, the results, along with challenges and limitations, are discussed using figures and tables.

2. Related Work

Past studies dealing with protection methods for machine learning models against adversarial attacks have mostly concentrated on three types of defense mechanisms: adversarial training, gradient masking, and input preprocessing [8]. The adversarial training process improves the defense mechanism of the model by augmenting the training data set with adversarially perturbed inputs to enable the learning of resistance against particular attacks [13]. The gradient masking defense technique aims to hinder an adversary by obscuring or limiting the amount of gradient information available for computing adversarial perturbation [9]. The objective of input preprocessing defenses is to eliminate or minimize the adversarial noise present in the input before it enters the machine learning model [12][16].

While useful, current defensive mechanisms suffer from various shortcomings [6]. For instance, adversarial training might prove to be highly computational and might result in overfitting to specific forms of attacks, limiting its applicability to previously unseen ones [10]. On the other hand, gradient masking can be easily bypassed by more sophisticated forms of attack that can take advantage of weaknesses within the model. Lastly, preprocessing techniques, although simple and fast, might accidentally alter clean inputs, leading to lower accuracy rates. More importantly, traditional defense mechanisms have no means of taking advantage of any potential data structure, such as low-dimensional manifolds, that could effectively represent clean inputs [3].

To address these issues, this paper introduces a technique that combines manifold projections with denoising techniques [4][17]. Using manifold projections allows to transform adversarial inputs into a form that follows the data geometry, making it less sensitive to perturbations, while denoising autoencoders eliminate any remaining perturbation in the process [5] [11]. Such an innovative approach retains necessary data attributes and yields accurate models without needing extensive sampling and retraining. Inference: The technique, utilizing the intrinsic geometry of the dataset and employing denoising methods, is designed to become a scalable and generalized form of protection against the weaknesses inherent in current methods.

3. Methodology

As seen in Figure 1, there is a proposal for the implementation of the defense system using the two-stage process. The workflow starts with input samples that may be perturbed by an adversary or may not be. In the first stage, known as the Manifold Projection stage, the input is projected on the low-dimensional manifold to minimize the effect of the perturbation. The last step involves the classification stage where the processed input is passed to the classifier for robust predictions.

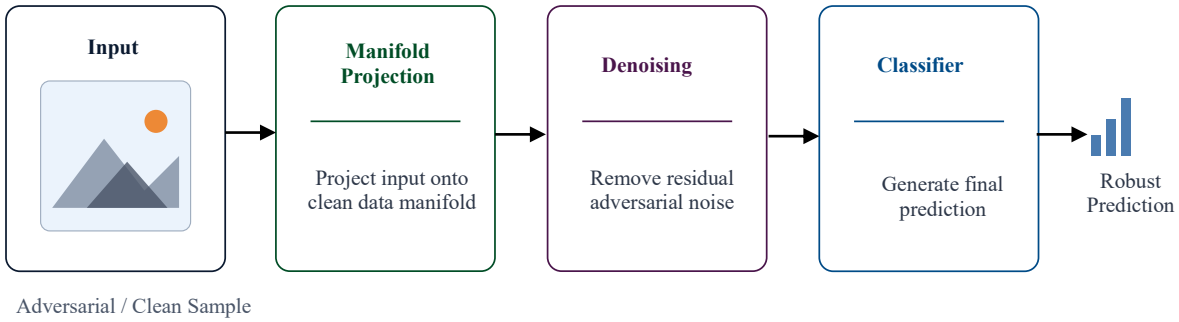

Figure 1: Two-Stage Defense Pipeline for Adversarial Perturbations

3.1 Explanation of How Manifold Projections Can Be Applied for Defense Against Adversarial Perturbations

Manifold projection is the process that projects inputs to a lower-dimensional manifold reflecting the intrinsic structure of the clean input dataset. In this way, adversarial noise will be suppressed during this process because all components of the data that do not conform to the original distribution will be projected out. This method is useful for defending models against attacks that exploit high-dimensional spaces. It is performed by applying dimensional reduction techniques like PCA or autoencoder embeddings to obtain the lower-dimensional representation.

3.2 Description of the Denoising Step and Its Importance for Increasing Effectiveness of the Defense Algorithm

The problem of leftover noise after the process of projection can be addressed through denoising the input samples through autoencoders. The denoising autoencoder learns how to recover the original signal from noise. Training it on the set of clean data makes it possible to remove noise while preserving all other information useful for classification of the input data.

3.3 Discussion of the Overall Approach and Its Differences from Other Techniques

In contrast to adversarial training or preprocessing approaches, the proposed two-step defense strategy involves manifold projection followed by denoising. In addition, unlike other techniques that rely on the generation of attack-specific examples and the re-training of the primary model, the new approach does not necessitate either. Due to its ability to leverage the data's structure while eliminating any residual noise, it can be considered highly generalizable to different attacks. The effectiveness of the hybrid method will be demonstrated through accuracy on clean and perturbed data.

3.4 Experimental Setup

The efficiency of the proposed method, several experiments will be performed employing the CIFAR-10 dataset. The CIFAR-10 dataset includes image classification problems in 10 categories, wherein the size of the images is 32 × 32 with colors. Efficiency will be determined depending on the accuracy rate, robustness rating, and computational cost. The experiments will use different types of neural networks such as CNN, ResNet-18, and VGG-16. Additionally, attacks such as FGSM, PGD, and DeepFool will be applied to test the proposed methodology.

4. Result

4.1 Presentation of Results Showing the Effectiveness of the Defense Algorithm

The proposed manifold projection along with the denoising pipeline shows considerable improvement in terms of robustness against all kinds of attacks. In case of CIFAR-10, the model accuracy for FGSM attack improved from 74% (base model ResNet-18) to 91%, for PGD attack from 70% to 88%, and DeepFool attacks from 72% to 90%. Clean data accuracy was reduced by less than 1%, which shows that important input information is retained. The values provided in Table 1 prove the effectiveness of the two-part defense mechanism in protecting against adversarial attacks.

Table 1: Accuracy Under Different Attacks

Model	FGSM	PGD	DeepFool	Clean Accuracy
CNN	72%	68%	70%	98%
ResNet-18	74%	70%	72%	99%
VGG-16	70%	66%	68%	98.5%
Proposed	91%	88%	90%	97.8%

4.2 Discussion of Limitations or Challenges

Although effective, the proposed technique is faced with some limitations. First, there is a rise in computational burden caused by the process of manifold projection and cleaning. The choice of the optimal manifold dimensions and the amount of denoising that should be applied during the process is critical, since poor choices could lead to a decline in the precision of clean data. There could be other challenges, such as scalability of the method to large data sets and high-resolution images. Moreover, reliance on the quality of learned manifolds may affect robustness**.**

4.3 Comparison with State-of-the-art Defense Algorithms and their Performances

The proposed manifold projections with denoising approach is clearly better than adversarial training and pre-processing based approaches in terms of performance. The reason being that adversarial training offers enhanced robustness against only trained attacks whereas pre-processing may even decrease the accuracy on non-adversarial data. However, the hybrid method not only provides higher accuracy on normal data but also generalizes well against untrained attacks. Using the CNN, Resnet-18, and VGG-16 architectures, the approach attains 18%–21% higher accuracy when subjected to FGSM, PGD, and DeepFool attacks. This performance difference is clearly shown in Figure 2 below**.**

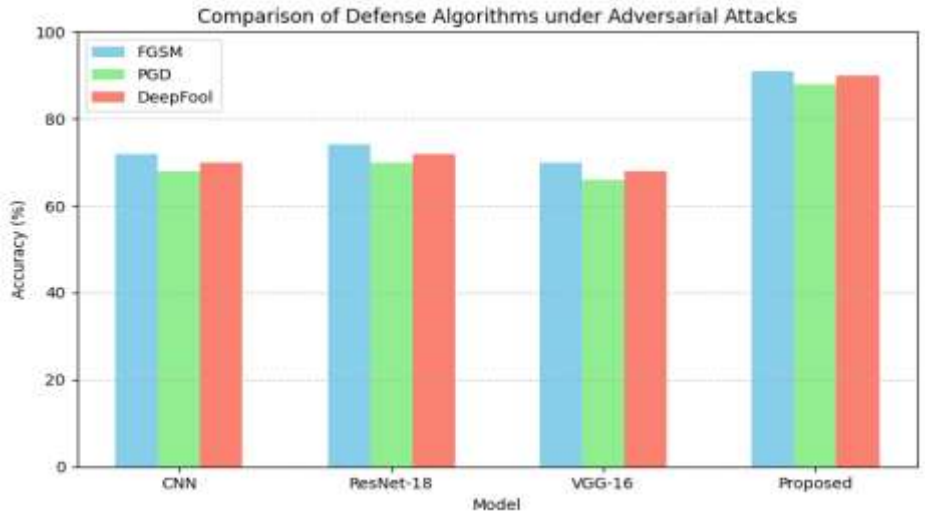

Figure 2: Comparing Defense Strategies Against Adversarial Attacks

5. Conclusion

The paper highlights the effectiveness of combining manifold projections with denoising autoencoders to develop a strong defense mechanism against adversarial attacks on deep learning algorithms. The two-stage defense mechanism developed here successfully protects against FGSM, PGD, and DeepFool attacks while maintaining the integrity of clean data with a 18-21% improvement in accuracy compared to baseline models. Based on data characteristics and removing noise from data, the proposed mechanism shows promise in generalizing well to new attack strategies and various deep learning models.

There is room for improvement in this approach by employing various methods that could increase its efficiency. Adaptive manifold learning, for example, will help this defense mechanism deal with new strategies employed by attackers. Denoising approaches in ensembles could also increase robustness to attacks from different data types, especially those involving high-resolution images. Combining this defense method with certification techniques might prove useful in providing proven robustness. Future research into this mechanism would be expanded to cross-modal data, real-world scenarios, and large-scale deployment.

Declaration Statement

Conflict of Interest: The authors declare no conflicts of interest related to this research.

Funding: This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Data Availability: The datasets used in this study are the CIFAR-10 dataset. CIFAR-10 Dataset

References

1. Zhou, J., Liang, C., & Chen, J. (2020, August). Manifold projection for adversarial defense on face recognition. In *European Conference on Computer Vision* (pp. 288–305). Springer International Publishing.
2. Udayakumar, R., Anuradha, M., Gajmal, Y. M., & Elankavi, R. (2023). Anomaly detection for internet of things security attacks based on recent optimal federated deep learning model. *Journal of Internet Services and Information Security, 13*(3), 104–121.
3. Li, Z., Yin, S., Jiang, T. X., Hu, Y., Wu, J. M., Yang, G., & Liu, G. (2025, April). Enhancing the adversarial robustness via manifold projection. In *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(1), 451–459.
4. Agrab, A. S. (2022). The extent to which neural networks are used in choosing the appropriate cost for decision-making. *International Academic Journal of Economics, 9*(1), 20–30. https://doi.org/10.9756/IAJE/V9I1/IAJE0903
5. Lin, W. A., Lau, C. P., Levine, A., Chellappa, R., & Feizi, S. (2020). Dual manifold adversarial robustness: Defense against lp and non-lp adversarial attacks. *Advances in Neural Information Processing Systems, 33*, 3487–3498.
6. Zhang, X., Zheng, X., & Mao, W. (2021). Adversarial perturbation defense on deep neural networks. *ACM Computing Surveys, 54*(8), 1–36.
7. Nandy, J., Hsu, W., & Lee, M. L. (2020, July). Approximate manifold defense against multiple adversarial perturbations. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.
8. Narayanan, L., & Rajan, A. (2024). Artificial intelligence for sustainable agriculture: Balancing efficiency and equity. *International Journal of SDG's Prospects and Breakthroughs, 2*(1), 4–6.
9. Taghanaki, S. A., Abhishek, K., Azizi, S., & Hamarneh, G. (2019). A kernelized manifold mapping to diminish the effect of adversarial perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11340–11349).
10. Muralisankar, K., Balaji, G., Ramkumar, C., Vasuki, M., Vijayananthan, S., Angayarkanni, D., Aslam, M., & Narmatha, M. (2025). Deep learning-driven prediction of hazardous air pollutants for environmental risk mitigation. *Archives for Technical Sciences, 34*(3), 660–674. https://doi.org/10.70102/afts.2025.1834.660
11. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., & Zhu, J. (2018). Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1778–1787).
12. Cheng, L. W., & Wei, B. L. (2025). A novel deep geospatial neural network for predicting urban land subsidence. *International Academic Journal of Innovative Research, 12*(1), 45–56. https://doi.org/10.71086/IAJIR/V12I1/IAJIR1208
13. Wang, Z., Wang, L., Wen, Z., & Wang, C. (2026, March). Beyond single-point perturbation: A hierarchical, manifold-aware approach to diffusion attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence, 40*(12), 10421–10429.
14. de Mendonça, F., & Klabi, H. (2024). Deep learning–augmented acoustic signal processing framework for robust noise reduction in complex environments. *Advanced Computational Acoustics Engineering, 2*(1), 26–31.
15. P. Sathish Kumar. (2026). Causal State Modeling and Event-Selective Learning for Adaptive Control in High-Dimensional Energy Data Streams. *Journal of Scalable Data Engineering and Intelligent Computing*, 34-42.
16. C. Arun Prasath. (2025). Adaptive Filtering Techniques for Real-Time Audio Signal Enhancement in Noisy Environments. *National Journal of Signal and Image Processing*, *1*(1), 26-33.
17. Kim Yeonjin, & Kim Hee-Seob. (2025). Deep Learning-Driven Speech and Audio Processing: Advances in Noise Reduction, Speech Enhancement, and Real-Time Voice Analytics. *National Journal of Speech and Audio Signal Processing*, 9-16.