Optimized Adversarial Defense: Combating Adversarial Attacks with Denoising Autoencoders and Ensemble Learning

Tushar Bhatia*, Peeyush Kumar Singh, Kanishk Vikram Singh, Jayesh, Faisal Rais

Department of Computer Science & Engineering, HMR Institute of Technology and Management, Hamidpur, New Delhi, 110036, Delhi, India

* Corresponding author

doi: https://doi.org/10.21467/proceedings.178.18

ABSTRACT

Adversarial attacks pose a significant risk to machine learning models by introducing carefully crafted perturbations that can mislead the models into producing incorrect outputs. This research investigates the effectiveness of denoising autoencoders as a defense mechanism against adversarial attacks on image classification tasks. A strategy combining a denoising autoencoder with a convolutional neural network (CNN) classifier is proposed and evaluated on the Modified National Institute of Standards and Technology (MNIST) dataset. The ability of autoencoders to learn robust representations and reconstruct original images from noisy inputs is leveraged to mitigate the impact of adversarial perturbations generated by the Fast Gradient Sign Method (FGSM). A K fold cross validation ensemble technique was employed to ensure robust and generalizable results. Findings demonstrate the potential of autoencoder based defense in enhancing the robustness of classifiers against FGSM adversarial attacks, achieving significantly higher classification accuracy compared to the unprocessed adversarial set. However, due to the autoencoder being a lossy reconstruction technique, a trade off between robustness and overall classification performance is observed, with diminishing effectiveness for more severe adversarial perturbations. Despite these limitations, the research motivates further research into autoencoder based defense mechanisms, exploring more complex architectures, combining with other techniques such as ensemble learning, and extending to real world applications.

Keywords: Adversarial Attacks, Denoising Autoencoders, Convolutional Neural Networks, K fold Cross Validation, Fast Gradient Sign Method (FGSM)

1 Introduction

Deep neural network models have achieved great success across various applications, including image classification. However, recent studies have exposed a disconcerting vulnerability known as adversarial attacks [1]. An adversarial attack occurs by introducing small and specifically designed distortions of the input data that lead the machine learning model to produce dramatically incorrect outputs. Adversarial examples are often imperceptible to humans and can pose a serious security threat to the deployment of machine learning systems in safety critical applications like autonomous driving, medical diagnosis, and cybersecurity. For their effective utilization, many of the proposed and developed defense strategies have also been reported [2]–[6]. Denoising autoencoders are examples of one such useful technique. Autoencoders are a class of neural network architectures constructed to learn compact representations of input data through encoding and subsequent decoding [3]. Their ability to reconstruct original inputs from noisy versions has led to applications in anomaly detection and denoising tasks. This characteristic of autoencoders makes them potentially valuable tools for filtering out adversarial perturbations, which can be viewed as a specific type of noise introduced to mislead the classifier and reduce accuracy.

Proceedings DOI: 10.21467/proceedings.178; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-984081-8-1

1.1 Autoencoders for Denoising

Autoencoders are neural networks made up of two primary components which include an Encoder and a Decoder. The encoder takes the input data (an image, in this case) and maps it into latent space, which is a simply lower dimensional representation of an image. From this latent representation, the decoder reconstructs the original input, minimizing reconstruction error. An autoencoder is trained to minimize the reconstruction error of the input relative to its decoded version [3]. The schematic representation of the autoencoder shown in Figure 1, provides a brief overview of the encoding and decoding processes. For adversarial defense, denoising autoencoders are trained specifically on noisy versions of the input images. In image processing, an autoencoder learns how to recreate the original image from a corrupted version by emphasizing on the vital features of images and suppressing the noise. This characteristic of denoising autoencoders makes it a promising tool for combating adversarial attacks.

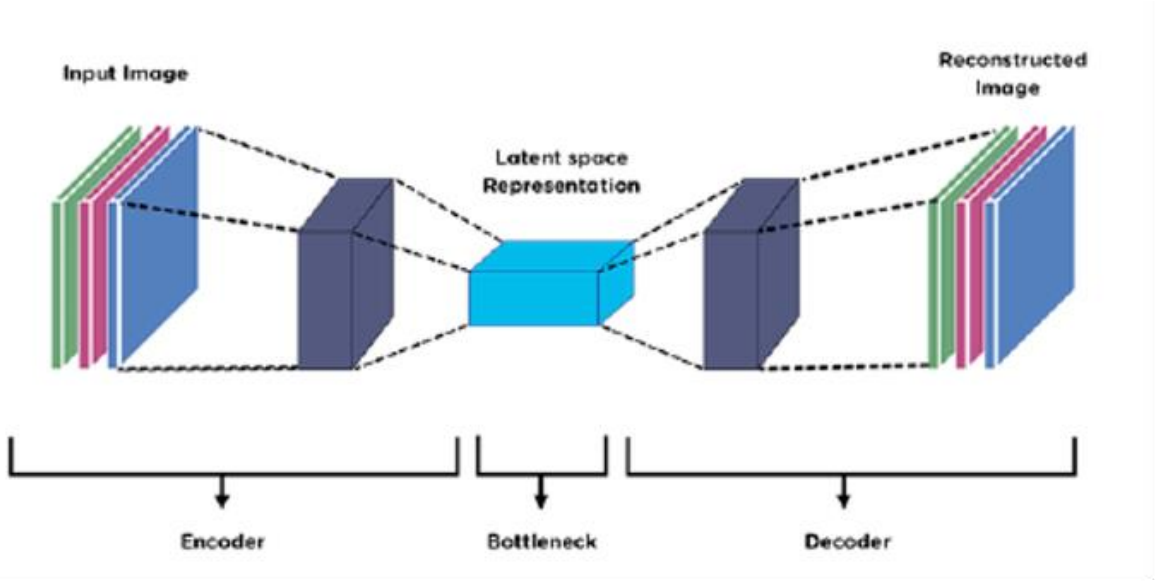

Figure 1: Autoencoder architecture illustrating different components involved in training process

1.2 The Fast Gradient Sign Method (FGSM)

The Fast Gradient Sign Method (FGSM) is a prominently used technique to produce adversarial samples. It functions by computing the loss function of the gradient concerning the input image. The sign of this gradient is then taken and multiplied by a small perturbation factor (epsilon). This modified gradient is added to the original image, thereby pushing the decision boundary of the model in a direction that increases the likelihood of misclassification. While computationally efficient, FGSM is considered a white box attack, meaning the attacker can access the internal model features and gradients.

1.3 K Fold Cross Validation

The K fold cross validation technique is employed to ensure the robustness and generalization of the autoencoder based defense strategy. This ensemble learning technique works by partitioning the dataset into K subsets called folds, and then it trains and evaluates the model iteratively on different combinations of these folds. Specifically, in each iteration, one of the folds is taken out as the validation set, while the remaining (K - 1) folds are used for training purposes. This iterative process is repeated K times, in which each fold serves as the validation set exactly once. More reliable estimates of performance and generalization

parameters are obtained by K fold cross validation and the problems associated with overfitting or bias which occur in a single train test split are almost eradicated. This enhances the generalizability of the results and allows for a more complete assessment to be made for enhancing the effectiveness of autoencoders in defending against adversarial attacks.

1.4 Research Focus

This research evaluates the efficacy of denoising autoencoders against FGSM adversarial attacks in an image classification task using the MNIST dataset. The primary goal is to assess the performance of autoencoders in reconstructing and denoising adversarial images so that the performance of the model on clean images is not affected while making it more resistant to adversarial perturbations. Specifically, the performance of the autoencoder based defense is examined in terms of classification performance, reconstruction loss and computational cost by applying the K fold cross validation ensemble technique for obtaining strong and generalizable conclusions from the experiment. Furthermore, the potential of combining the autoencoder with other defense mechanisms is explored to provide a more comprehensive defense against adversarial attacks.

2 Related Works

Adversarial attacks have been recognized as a significant threat to machine learning models, particularly deep neural networks [1], [2]. Initial observations showed that adding imperceptible perturbations to input data can cause neural networks to produce incorrect outputs [1]. Following this, Fast Gradient Sign Method (FGSM) was proposed as an efficient technique to generate adversarial examples, highlighting the vulnerability of neural networks to such attacks [2]. To overcome these vulnerabilities, various defense mechanisms have been developed, including adversarial training, input transformations, and model modifications [4]–[6]. Adversarial training involves augmenting the training data with adversarial examples to improve model robustness [4], while input transformations, such as denoising or compression, aim to remove adversarial perturbations from the input data [5]. Model modifications encompass architectural changes or regularization techniques designed to enhance resilience against attacks [6].

Denoising autoencoders have emerged as a promising input transformation technique for adversarial defense [3]. Their ability to reconstruct clean inputs from corrupted versions has been explored across various domains, including image classification [7], speech recognition [8], and malware detection [9]. This versatility demonstrates the potential of autoencoders as a generalizable defense mechanism against adversarial attacks in diverse applications. Additionally, innovative approaches like Feature Squeezing [10] and Defense GAN [11] aim to reduce the attack surface available to adversaries by coalescing samples within small spatial regions or leveraging generative models to project inputs onto safe manifolds before classification. Recent comparative studies, such as [12], have further contributed to understanding adversarial attacks by analyzing their effects across different pre trained models, which is crucial for identifying common vulnerabilities and informing the development of more robust defense strategies. Despite these advancements, challenges remain, including the trade off between model robustness and overall performance [13]. This highlights the ongoing need for defense mechanisms that maintain high performance on unperturbed inputs while effectively mitigating the impact of adversarial attacks. This research builds upon these insights by investigating the effectiveness of denoising autoencoders coupled with ensemble learning techniques like K fold cross validation to develop a more robust and generalizable defense strategy against FGSM attacks in image classification tasks.

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

3 Methodology

3.1 Dataset

Experiments were conducted using the MNIST dataset, which is a prominently used benchmark for image classification tasks. The MNIST dataset consists of 70,000 grayscale images of handwritten digits, each of 28x28 pixels. A training set comprising 60,000 examples and a test set of 10,000 examples is used, with each example belonging to one of ten classes. The simplicity and well understood nature of the MNIST dataset makes it ideal for exploring adversarial attacks and defense mechanisms. Though the low complexity of the dataset may not fully reflect real world scenarios, the insights gained from these experiments can serve as a foundation for future research on more complex datasets and applications.

3.2 Image Classification Model

The classification task utilizes a Convolutional Neural Network (CNN) architecture consisting of two convolutional layers, each followed by a max pooling layer, and a fully connected layer with a softmax activation function for multiclass classification. The first convolutional layer has 32 filters with a kernel size of 3x3, followed by a 2x2 max pooling layer. The second convolutional layer has 64 filters with a kernel size of 3x3, followed by another 2x2 max pooling layer. The flattened output is then fed into a fully connected layer with 10 units, representing the 10 digit classes. The CNN model is trained on the MNIST training set using the Adam optimizer and categorical cross entropy loss function. A batch size of 128 is used to train the model for approximately 20 epochs. This model serves as the baseline for evaluating the impact of FGSM attack and the effectiveness of the autoencoder based defense strategy.

3.3 Adversarial Attack Generation

The adversarial pattern generation is a crucial component of the methodology process which focuses on creating perturbations in input data to mislead neural network models using the Fast Gradient Sign Method (FGSM) [2]. FGSM operates by perturbing input data based on the gradient information of the loss function for the input calculating perturbations to maximize loss and lead to misclassifications. The perturbed image X adversarial is mathematically generated using the following formula

$$X(adv) = X + \epsilon\Sigma(\nabla X J\,(X, Y(true))) \quad (1)$$

where X represents the original input image, Y(true) is the true label, J is the loss function, ∇X denotes the gradient with respect to X, Σ indicates the signum function, and ϵ controls the magnitude of the perturbation. By adjusting ϵ, the strength of the attack is controlled. Smaller ϵ values result in subtle perturbations, while larger values lead to more pronounced changes.

3.4 Denoising Autoencoder Architecture

The denoising autoencoder architecture employed in this research study consists of an encoder and a decoder network, inspired by techniques outlined in the Keras blog [14]. The encoder network is a CNN structure that maps the input image (28x28x1) to a lower dimensional latent representation. It includes a convolutional layer with 16 filters, a 3x3 kernel size, ReLU activation, and a stride of 2, followed by another convolutional layer with 8 filters, a 3x3 kernel size, ReLU activation, and a stride of 2. The decoder network, a transpose convolutional network, uses the latent representation learned by the encoder to attempt recreating the original input image. It comprises a transpose convolutional layer with 8 filters, a 3x3 kernel size, ReLU activation, and a stride of 2, followed by another transpose convolutional layer with 16 filters, a

3x3 kernel size, stride of 2, and ReLU activation. The final layer is a convolutional layer with 1 filter, a 3x3 kernel size, and sigmoid activation to produce the reconstructed 28x28x1 image. The autoencoder is trained on the MNIST training set with added Gaussian noise with noise factors of 0.1, 0.2, and 0.3 to simulate noisy input conditions. The objective is to minimize the mean squared error (MSE) between the input images and their reconstructed counterparts using the Adam optimizer.

3.5 Ensemble Learning Implementation

The K fold cross validation technique is utilized to ensure the autoencoder based defense mechanism is robust and generalizable. By dividing the dataset into K equal sized subsets or folds, this strategy trains and assesses the model repeatedly using various fold combinations. In each of the iterations, one fold is taken as the validation set, while the remaining (K - 1) folds are used for training. The entire process is executed K times, resulting in each fold being used as a validation set once. The entire process is executed K times, resulting in each fold being used as a validation set once. The final performance metric is computed as the average across all K iterations. In the experiments, K is set to 10, resulting in a 10 fold cross validation process. This approach helps mitigate the risk of overfitting or bias introduced by a single train test split and provides a more reliable estimate of performance of the model on unseen data.

3.6 Experimental Setup

The primary goal of the experiments was to evaluate the effectiveness of the denoising autoencoder in reconstructing and denoising adversarial images generated by the Fast Gradient Sign Method (FGSM) attack, thereby enhancing the robustness of the image classification model against adversarial perturbations. The experiments followed a systematic approach, which began with training a Convolutional Neural Network (CNN) image classification model on the MNIST training set. Adversarial examples were then generated from the MNIST test set using the FGSM attack. Subsequently, a denoising autoencoder was trained on the MNIST training set with added Gaussian noise. The adversarial examples were passed through the trained autoencoder to obtain reconstructed and denoised images. For each fold of the K fold cross validation, the dataset was split into training and validation sets based on the current fold assignment. A new CNN classifier was trained on the training set of the current fold, and its performance was evaluated using the denoised images as the test set. Classification accuracy and loss were recorded for each fold. Ultimately, the performance of the CNN classifier was comprehensively assessed on three distinct sets: the clean test set, containing original MNIST test images, the adversarial test set, consisting of adversarial examples generated by FGSM, and the reconstructed test set, comprising denoised adversarial examples from the autoencoder. Classification accuracy and loss were meticulously measured and compared across these sets to assess the effectiveness of the autoencoder based defense strategy. Furthermore, original, adversarial, and reconstructed images were meticulously visualized and compared to qualitatively analyze the ability of autoencoders to denoise and recover the original image content from adversarial perturbations. The experiments were conducted on a system configured with an Intel Core i7 processor, 16 GB RAM, and an NVIDIA RTX 3080 GPU. TensorFlow 2.13 was employed for model development, and the Adversarial Robustness Toolbox library facilitated adversarial sample generation and evaluation.

4 Results and Discussion

4.1 Visualizing Adversarial Examples and Reconstructions

To qualitatively assess the impact of adversarial perturbations and the ability of autoencoders to reconstruct and denoise the images, a sample of original, adversarial, and reconstructed images from the MNIST test set were visualized.

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

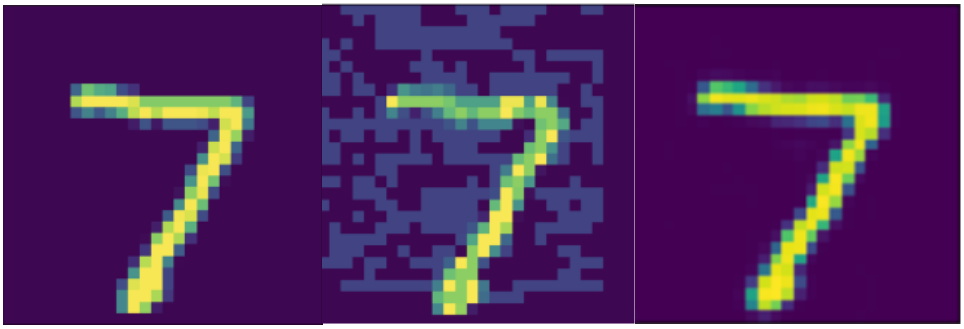

Figure 2: Sample of original, adversarial, and reconstructed image from the MNIST test set

Figure 2 illustrates comparisons between original images, adversarial images, and reconstructed images. The pristine MNIST digit serves as the baseline for comparison. The subtle noise introduced by the FGSM attack, although not visible to the human eye, significantly impacts the performance of classifiers, highlighting the deceptive nature of adversarial attacks. The denoising autoencoder demonstrates its ability to filter adversarial perturbations and recover the overall structure of digits. While not a perfect restoration, this reconstruction is significantly closer to the original image, improving the chances of accurate classification.

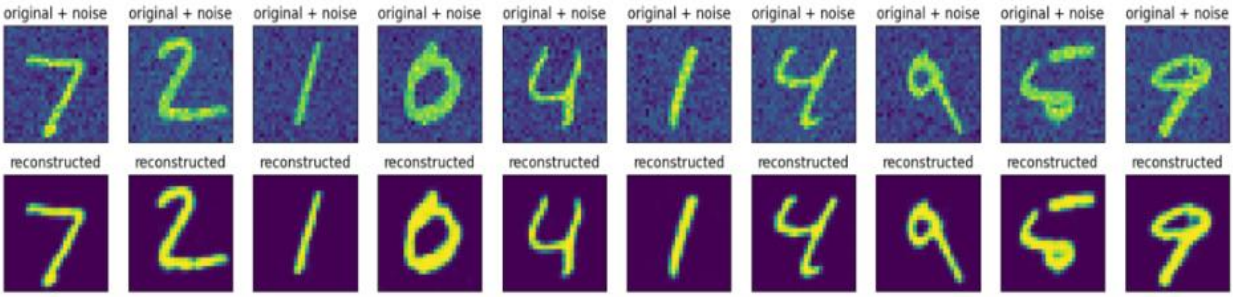

Figure 3: Comparative illustration of noisy images and images reconstructed with autoencoder training process

Figure 3 illustrates the core concept behind the autoencoder based defense. By training the autoencoder on noisy variations of MNIST digits, it learns to reconstruct the underlying clean images. Even though the reconstructions may exhibit some minor artifacts, they preserve the essential features necessary for successful digit classification. This training strategy enables the autoencoder to denoise perturbed images effectively.

4.2 Classification Performance on Clean and Adversarial Examples

10.3 The effectiveness of the autoencoder based defense was quantitatively evaluated by comparing the classification performance of CNN model under the following different conditions

1. **Clean Test Set:** Original MNIST test images.
2. **Adversarial Test Set:** Generated by FGSM attacks with varying epsilon (ϵ) values.
3. **Reconstructed Test Set:** Obtained by passing adversarial images through a denoising autoencoder (which may be trained on varying noise factors).
4. **K Fold + Reconstructed Test Set:** The same process as above, but with the addition of K fold cross validation for training and evaluation.

Table 1: *Comparison of classification accuracy across clean, adversarial, and reconstructed test sets.*

Strategy	Classification Accuracy
Base Model (Clean)	99.06%
Adversarial (ε=0.1)	81.48%
Autoencoder (Noise factor=0.1)	96.21%
K Fold + Autoencoder	97.07%
Adversarial (ε=0.2)	30.30%
Autoencoder (Noise factor = 0.2)	89.88%
K Fold + Autoencoder	93.01%
Adversarial (ε=0.3)	8.38%
Autoencoder (Noise factor = 0.3)	68.04%
K Fold + Autoencoder	77.35%

Table 1 presents the classification accuracy values for various experimental setups. On the clean test set, the base learner (CNN model without any defense) achieves an accuracy of 99.06%, indicating its high performance on unperturbed images. However, when evaluated on the adversarial test sets generated by FGSM with increasing perturbation factors (ε = 0.1, 0.2, and 0.3), the accuracy of the model drops significantly, demonstrating its vulnerability to adversarial attacks. After applying the autoencoder based defense strategy, we observe a substantial improvement in classification accuracy on the reconstructed test sets. For instance, with a perturbation factor of 0.2, the autoencoder (trained on noise factor 0.2) improves the accuracy from 30.30% on the adversarial set to 89.88% on the reconstructed set. Furthermore, the K fold cross validation technique (k=10) with the autoencoder provides a more robust and generalizable estimate of performance of the model, achieving an accuracy of 93.01% on the reconstructed set.

It is important to note that while the autoencoder based defense strategy significantly enhances the robustness of the model against adversarial perturbations, the level of improvement diminishes as the perturbation factor increases. For example, with a perturbation factor of 0.3, the autoencoder (trained on noise factor 0.3) and the K fold cross validation setup achieve accuracies of 68.04% and 77.35%, respectively, on the reconstructed set, which are lower than the accuracies achieved for smaller perturbation factors. It is to be noted that using K fold ensemble learning, the accuracy of classification has improved thus cementing the fact that ensemble learning in conjunction with autoencoder can provide a significant jump in the robustness of neural networks against FGSM adversarial attack.

4.3 Discussions and Implications

The experiments demonstrate the potential of denoising autoencoders in defending against adversarial attacks on image classification tasks. The ability of autoencoders to learn robust representations and reconstruct original images from noisy inputs leads to significant improvements in the robustness of CNN classifier FGSM adversarial attacks. Crucially, the results highlight the importance of aligning the noise training level of autoencoder with the anticipated attack strength. For optimal protection, the noise factor used to train the autoencoder should be similar to the epsilon value employed in the FGSM attack. While the autoencoder based defense strategy enhances performance against adversarial examples, it generally

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

does not fully restore the original level of accuracy achieved on clean images. This suggests a potential trade off between adversarial robustness and overall classification performance. The denoising process may inadvertently alter some features critical for accurate classification. Furthermore, the experiments focused specifically on the FGSM attack. More sophisticated attack techniques, such as iterative or targeted attacks, could pose greater challenges for denoising autoencoders. In order to ensure generalizability, it is vital to evaluate the effectiveness of this defense strategy against a diverse range of adversarial attacks. The benefits of integrating K fold cross validation into the autoencoder based defense are also observed. The results suggest that K fold helps the autoencoder learn even more robust representations. Statistical significance tests would help determine the reliability of these observed improvements. If deemed significant, K fold integration could provide an additional layer of resilience in adversarial defense strategies.

5 Conclusions

This research investigates the effectiveness of denoising autoencoders as a defense mechanism against adversarial attacks on MNIST image classification. The findings demonstrate the potential of autoencoder based defenses in enhancing the robustness of convolutional neural network (CNN) classifiers against adversarial perturbations. By leveraging the ability of autoencoders to reconstruct original images from noisy inputs, significant improvements in classification accuracy on adversarial examples are achieved as compared to unprocessed adversarial sets. The integration of K fold cross validation further improves the robustness and generalizability of the results. While the proposed approach significantly reduced the impact of Fast Gradient Sign Method (FGSM) adversarial attacks, a trade off between adversarial robustness and overall classification performance was observed, as the denoising process may inadvertently alter some features critical for accurate classification. The effectiveness of the proposed approach diminishes with increasing perturbation strength, highlighting a limitation of this defense strategy that could be addressed by exploring adaptive autoencoder architectures. Future work should focus on investigating more complex autoencoder architectures, combining autoencoders with other advanced defense techniques to develop more comprehensive solutions, and extending the approach to diverse datasets and real world applications. Additionally, evaluating the defense strategy against a wider range of adversarial attack methods would provide a more comprehensive assessment of its effectiveness. Overall, this research contributes to the ongoing efforts to develop more resilient ML models which are capable of withstanding adversarial attacks in critical application domains such as autonomous systems and cybersecurity.

Declarations

Competing Interests

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note

AIJR remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

How to Cite

Tushar Bhatia, Peeyush Kumar Singh, Kanishk Vikram Singh, Jayesh, Faisal Rais (2025). Optimized Adversarial Defense: Combating Adversarial Attacks with Denoising Autoencoders and Ensemble Learning. *AIJR Proceedings*, 150-158. https://doi.org/10.21467/proceedings.178.18

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, et al., "Intriguing properties of neural networks," *arXiv preprint*, vol. 1312, pp. 6199, Feb. 2014.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint*, vol. 1412, pp. 6572, Mar. 2015.

[3] P. Vincent, H. Larochelle, I. Lajoie, et al., "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 1096–1103.

[4] A. Madry, A. Makelov, L. Schmidt, et al., "Towards deep learning models resistant to adversarial attacks," *arXiv preprint*, vol. 1706, pp. 6083, Sep. 2018.

[5] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," *arXiv preprint*, vol. 1711, pp. 00117, Jan. 2018.

[6] A. S. Ross and F. Doshi Velez, "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients," *arXiv preprint*, vol. 1711, pp. 09404, Apr. 2018.

[7] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," *arXiv preprint*, vol. 1412, pp. 5068, Sep. 2015.

[8] J. Du, X. Zhu, X. Shang, et al., "Robust speech recognition with denoising adversarial autoencoders," *arXiv preprint*, vol. 1902, pp. 07955, Feb. 2019.

[9] A. Al Dujaili, L. Huang, E. Hemberg, et al., "Denoising autoencoder adversarial attacks against deep malware classifiers," in *Proc. 2018 IEEE International Conference of Machine Learning and Applications (ICMLA)*, 2018, pp. 1233–1238.

[10] W. Xu, D. Evans, and Y. Qi, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks," *arXiv preprint*, vol. 1704, pp. 01155, Dec. 2017.

[11] P. Samangouei, M. Kabkab, and R. Chellappa, "DefenseGAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models," *arXiv preprint*, vol. 1805, pp. 06605, May. 2018.

[12] R. Kumari, T. Bhatia, P. K. Singh, and K. V. Singh, "Dissecting Adversarial Attacks: A Comparative Analysis of Adversarial Perturbation Effects on Pre-Trained Deep Learning Models," *International Journal of Scientific Research in Engineering and Management* (IJSREM), vol. 7, no. 12, pp. 1–6, Dec. 2023.

[13] A. Alotaibi and M. A. Rassam, "Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense," *Future Internet*, vol. 15, no. 2, p. 62, Jan. 2023.

[14] F. Chollet, "Building autoencoders in Keras," *Keras Blog, 2023.* [Online]. Available: https://blog.keras.io/buildingautoencodersinkeras.html [Accessed: Apr. 5, 2023].

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922