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

Peeyush Kumar Singh[1], Tushar Bhatia[2], Jayesh[3], Kanishk Vikram Singh [4]

[1], Student, Department of Computer Science & Engineering, HMR Institute of Technology and Management
[2], Student, Department of Computer Science & Engineering, HMR Institute of Technology and Management
[3], Student, Department of Computer Science & Engineering, HMR Institute of Technology and Management
[4], Student, Department of Computer Science & Engineering, HMR Institute of Technology and Management

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 study investigates the effectiveness of denoising autoencoders as a defense mechanism against adversarial attacks on image classification tasks. We propose a strategy that combines a denoising autoencoder with a convolutional neural network (CNN) model classifier, evaluated on the MNIST dataset. The autoencoder's ability 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). The K-fold cross-validation ensemble technique is employed to ensure robust and generalizable results. Our findings demonstrate the potential of the autoencoder-based defense in enhancing the classifier's robustness against FGSM adversarial attacks, achieving significantly higher classification accuracy compared to the unprocessed adversarial set. However, due to 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, our study 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

Machine learning (ML) models, particularly deep neural network models, have obtained great success across various aspects, including image classification. However, recent studies have exposed a disconcerting vulnerability: Adversarial attacks. These meticulously crafted small perturbations to input data can mislead ML models into producing drastically incorrect outputs. Adversarial examples, while often imperceptible to humans, pose a serious security threat to the deployment of ML systems in applications like autonomous driving or medical diagnosis which are safety-critical.

To combat their impact, various defense strategies have been proposed and developed. One promising avenue lies in denoising autoencoders. Autoencoders are neural network architectures which are constructed to learn compact representations of input data through encoding and subsequent decoding. Their ability to reconstruct original inputs from noisy versions has led to applications in anomaly detection and denoising tasks. This characteristic makes autoencoders potentially valuable tools for filtering out adversarial perturbations, which can be viewed as a specific type of noise introduced to mislead the classifier.

1.1Autoencoders for Denoising

Autoencoders have two primary components: an encoder and a decoder. The encoder works by mapping the input data (an image in this case) to a lower-dimensional representation called the "latent space," which captures the image's main properties. The decoder works by attempting to reproduce the original image from this representation. The training objective of an autoencoder being a lossy technique, is to minimize the reconstruction error between the input and its decoded counterpart.

In the context of adversarial defense, denoising autoencoders are trained specifically on noisy versions of input images. By learning to reconstruct the original image from a corrupted version, the autoencoder is forced to focus on the core features of the image while filtering out the noise. This characteristic makes denoising autoencoders a promising tool for combating adversarial attacks, which essentially introduce a specific type of noise to mislead the classifier.

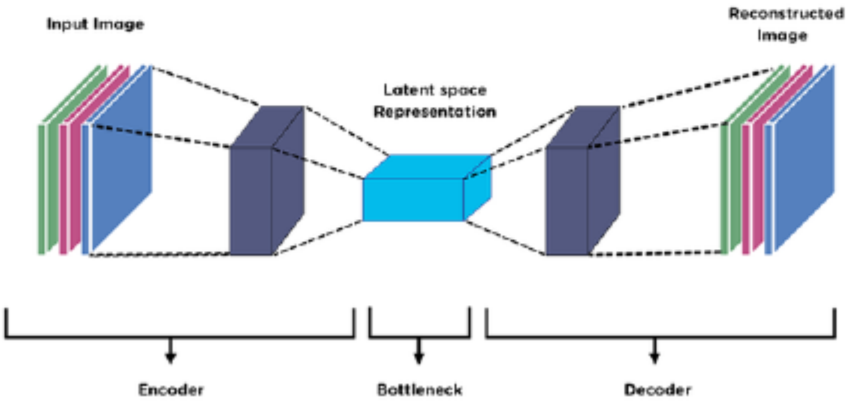

Fig.1. Autoencoder architecture

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 model's decision boundary 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

To ensure the robustness and generalization of our autoencoder-based defense strategy, we employ the K-fold cross-validation technique. This ensemble learning technique works by partitioning the dataset into K subsets which are called "folds," and then trains and evaluates the model iteratively on different combinations of these folds. Specifically, in each iteration, one of the fold is taken out as the validation set, while the remaining K-1 folds are used for training purpose. This iterative process is repeated K times, in which each fold serves as the validation set exactly once.

By leveraging K-fold cross-validation, we can achieve a more reliable approximation of the model's performance and mitigate the risk of overfitting or bias introduced by a single train-test split. This approach improves the generalizability of our findings and offers a more comprehensive evaluation of the autoencoder's effectiveness in defending against adversarial attacks.

1.4 Research Focus

In this study, we estimate denoising autoencoder's efficacy in defending against FGSM adversarial attacks on an image classification task involving the MNIST dataset. Our primary objective is to assess the autoencoder's ability to reconstruct and denoise adversarial images, thereby enhancing the model's robustness against adversarial perturbations while maintaining its performance on clean, unperturbed images. Specifically, we evaluate the autoencoder-based defense strategy in terms of classification accuracy, reconstruction error, and computational efficiency, utilizing the K-fold cross-validation ensemble technique to ensure robust and generalizable results. Furthermore, we explore the potential of combining the autoencoder with other defense mechanisms to provide a more comprehensive defense against adversarial attacks.

2. Related Works

Szegedy et al. [1] pioneered the concept of adversarial examples, which present subtle modifications to input data that can mislead well-trained models into inaccurate predictions. Since then, other research have looked into various sorts of adversarial attacks and offered defense techniques to lessen their effects.

Shekhar et al. [2] provide a survey of various attacks and defense mechanisms in machine learning, covering a wide range of techniques and applications. They categorize adversarial attacks based on different criteria, such as the attack methodology (e.g., gradient-based, optimization-based, or transfer-based), the attacker's knowledge (e.g., white-box or black-box), and the attack objective (e.g., misclassification, targeted misclassification, or source-target misclassification).

The Fast Gradient Sign Method (FGSM) [3], introduced by Goodfellow et al., is widely used for its computational efficiency. It generates adversarial examples by varying the input data in the direction of the sign of the loss function's gradient with respect to the input. Our study focuses on defending against FGSM attacks using denoising autoencoders.

Shekhar et al. [2] categorize defense mechanisms into three main categories: adversarial training, input transformations, and model modifications. Adversarial training entails augmenting the training dataset with adversarial examples to improve the model's robustness [4]. Input transformations, such as denoising or compression, aim to remove or reduce the adversarial perturbations from the input data [5]. Model modifications involve architectural changes or regularization techniques to enhance the model's resilience against adversarial attacks [6].

Denoising autoencoders fall under the category of input transformations for adversarial defense. These unsupervised neural networks are trained to reconstruct the original input from a corrupted or noisy version, effectively learning robust representations of the data [7]. Several studies have explored the use of denoising autoencoders as a defensive mechanism against adversarial attacks in various domains, including image classification [8], speech recognition [9], and malware detection [10].

In the context of intrusion detection systems (IDS), Alotaibi and Rassam [11] provide a comprehensive survey on adversarial machine learning attacks and defense strategies. They highlight the importance of defending IDS against adversarial attacks, as these systems play a crucial role in cyber security. The authors discuss various defense mechanisms, including input transformations, adversarial training, and model modifications, and their application in securing IDS against adversarial threats.

Our study builds upon the existing literature by investigating the effectiveness of denoising autoencoders in defending against FGSM adversarial attacks on the MNIST image classification task. We employ the K-fold cross-validation ensemble method to ensure robust and generalizable results, which has not been extensively explored in previous studies. Additionally, we provide insights into the trade-offs between robustness to adversarial attacks and overall classification performance, as well as the autoencoder's behavior and reconstruction capabilities through visualization and analysis of reconstruction errors.

3. Methodology

This section outlines the dataset, models, and experimental setup employed in our research to assess the efficacy of denoising autoencoders in protecting against adversarial attacks on image classification tasks.

3.1 Dataset

We conducted our experiments using the MNIST dataset, a prominently used benchmark for image classification jobs. The MNIST dataset consists of 70,000 grayscale images of handwritten digits, each of 28x28 pixels. A training set comprised of 60,000 examples and test set of 10,000 examples is used which each example belonging to one of ten classes (digits 0-9).

The simplicity and well-understood nature of the MNIST dataset make it an ideal testbed for exploring adversarial attacks and defense mechanisms. While the dataset's low complexity may not fully reflect real-world scenarios, insights gained from these experiments can serve as a junction for future research on much more complex datasets and applications.

3.2 Image Classification Model

For classification task, we employed 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 multi-class 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 was 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 20 epochs. This model serves as the baseline for evaluating the impact of FGSM attack and the effectiveness of the autoencoder-based defense strategy.

3.2 Adversarial Attack Generation

Adversarial Pattern Generation is a crucial component of our methodology, focusing on creating perturbations in input data to mislead neural network models. We employ the Fast Gradient Sign Method (FGSM), a well-established technique in adversarial machine learning.

FGSM operates by perturbing input data based on the gradient information of the loss function for the input. The perturbation is calculated to maximize the loss, leading to misclassifications. Mathematically, the perturbed image, $X(\text{adv})$, is generated as follows:

$$X(\text{adv})=X+\epsilon\cdot\mathtt{sign}\,(\triangledown XJ\,(X,\,Y(true)))$$

Here:

- X represents the clean input image.
- $Y(\text{true})$ is the true label of the clean image.
- $J\,(X,\,Y(\text{true}))$ is the loss function based on the true label.
- $\triangledown X$ denotes the gradient with respect to the input.
- ϵ controls the magnitude of perturbation.

The sign function indicates that the perturbation is added in the direction of increased loss, aiming to induce misclassification. By adjusting ϵ, we control the strength of the attack. 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 our study consists of an encoder and a decoder network.

Encoder Network:

It is a CNN structure that works by mapping the input image (28x28x1) to a lower-dimensional latent representation. It comprises the following layers:

1. Convolutional layer with 16 filters, 3x3 kernel size, ReLU activation and stride of 2.
2. Convolutional layer with 8 filters, 3x3 kernel size, ReLU activation and stride of 2.

Decoder Network:

The decoder network, a transpose convolutional network, uses the latent representation that the encoder has learnt to attempt and recreate the original input image. It is made up of the following layers:

1. Transpose convolutional layer with 8 filters, 3x3 kernel size, ReLU activation and stride of 2.
2. Transpose convolutional layer with 16 filters, 3x3 kernel size, stride of 2, ReLU activation.

3. Convolutional layer with 1 filter, 3x3 kernel size and sigmoid activation to produce the reconstructed 28x28x1 image.

The autoencoder was trained on the MNIST training set with added Gaussian noise (noise factor of 0.1,0.2,0.3) to simulate noisy input conditions. The objective was to minimize the mean squared error (MSE) between the input images and their reconstructed counterparts using the Adam optimizer.

3.5 Ensemble Learning: K-Fold Cross-Validation

We used the K-fold cross-validation technique to make sure our autoencoder-based defense mechanism was 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 utilized as the validation set, while the remaining K-1 folds are used for training. The entire process is done K times. Hence, each fold is used as a validation set once. The final performance metric is then computed as the average across all K iterations.

In our experiments, we set K=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 the model's performance on unseen data.

3.6 Experimental Setup

The primary goal of our experiments was to evaluate the effectiveness of the denoising autoencoder in reconstructing and denoising adversarial images generated by the FGSM attack, thereby improving the robustness of the image classification model against adversarial perturbations.

We followed these steps:

1. Train the CNN image classification model on the MNIST training set.

2. Generate adversarial examples from the MNIST test set using the FGSM attack.

3. Train the denoising autoencoder on the MNIST training set with added Gaussian noise.

4. Pass the adversarial examples through the trained autoencoder to obtain reconstructed and denoised images.

5. For each fold of the K-fold cross-validation:

   - Split the dataset into training and validation sets based on current fold assignment

   - Train a new CNN classifier on the training set of the current fold.

   - Evaluate the performance of CNN classifier using the denoised images as test set.

- Record the classification accuracy and loss for the current fold.

6. Evaluate the performance of the CNN classifier on the following sets:

    - Clean test set (original MNIST test images)
    - Adversarial test set (adversarial examples generated by FGSM)
    - Reconstructed test set (denoised adversarial examples from the autoencoder)

We measured and compared the classification accuracy and loss on these three sets to assess the effectiveness of the autoencoder-based defense strategy.

Additionally, we visualized and compared the original, adversarial, and reconstructed images to qualitatively analyze the autoencoder's ability to denoise and recover the original image content from adversarial perturbations.

The experiments were implemented using TensorFlow 2.13 and the Adversarial Robustness Toolbox (ART) library for generating adversarial samples and assessing the model's robustness.

4. Results and Discussion

This section outlines the results obtained from our experiments on the ability of the denoising autoencoder to fend off FGSM adversarial attacks on the MNIST image classification task. We analyze and discuss the findings, highlighting the strengths and limitations of the proposed approach.

4.1 Visualizing Adversarial Examples and Reconstructions

To qualitatively assess the impact of adversarial perturbations and the autoencoder's ability to reconstruct and denoise the images, we visualize a sample of original, adversarial, and reconstructed image from the MNIST test set.

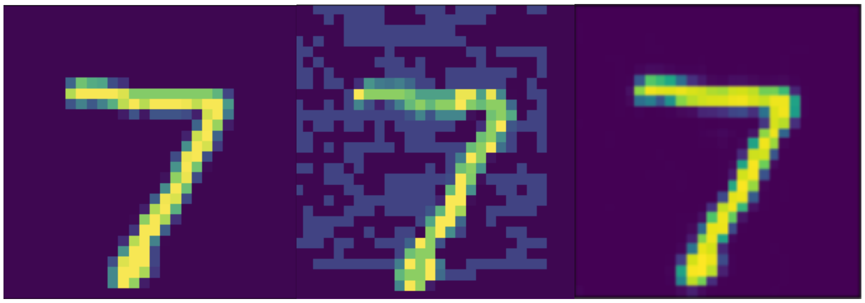

Fig.2. Sample of original, adversarial, and reconstructed image

Figure 2 illustrates the following:

- **Original Image:** The pristine MNIST digit '7' serves as the baseline for comparison.

- **Adversarial Image:** The subtle noise introduced by the FGSM attack, although imperceptible to the human eye, drastically impacts the classifier's performance. This highlights the deceptive nature of adversarial attacks.
- **Reconstructed Image:** The denoising autoencoder demonstrates its ability to filter adversarial perturbations and recover the digit's overall structure. While not a perfect restoration, this reconstruction is significantly closer to the original image, improving the chances of accurate classification.

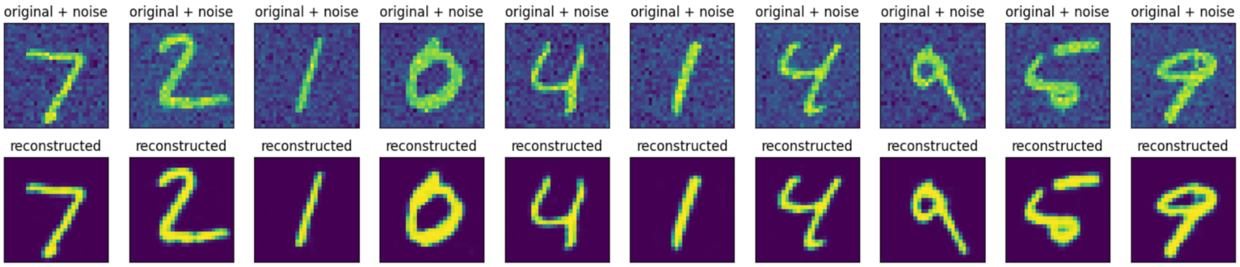

Fig.3. Training the Autoencoder for Robustness

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 is what enables the autoencoder to denoise perturbed images.

4.2 Classification Performance on Clean and Adversarial Examples

We quantitatively evaluate the effectiveness of the autoencoder-based defense by comparing the CNN model's classification performance under 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. Classification accuracy and loss on clean, adversarial, and reconstructed, kfold 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 model's accuracy 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 the model's performance, 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 model's robustness 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.

4.4 Discussion and Implications

Our experiments demonstrate the potential of denoising autoencoders in defending against adversarial attacks on image classification tasks. The autoencoder's ability to learn robust representations and reconstruct original images from noisy inputs led to significant improvements in the CNN classifier's robustness against FGSM adversarial attacks.

Crucially, our results highlight the importance of aligning the autoencoder's noise training level 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 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, our experiments focused specifically on the FGSM attack. More sophisticated attack techniques, such as iterative or targeted attacks, could pose greater challenges for denoising autoencoders. To ensure generalizability, it's vital to evaluate the effectiveness of this defense strategy against a diverse range of adversarial attacks.

We also observed the benefits of incorporating K-fold cross-validation into the autoencoder-based defense. 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. Conclusion and Future Scope

5.1 Conclusion

In this study, we investigated the effectiveness of denoising autoencoders as a defense mechanism against adversarial attacks, specifically the FGSM attack, on the MNIST image classification task. Our research aimed to leverage the autoencoder's ability to learn robust representations and reconstruct original images from noisy inputs, mitigating the effect of adversarial perturbations. To ensure the robustness and generalization of our approach, we employed the K-fold cross-validation ensemble technique.

The experimental results demonstrated the potential of the proposed autoencoder-based defense strategy in enhancing the robustness of the convolutional neural network (CNN) classifier against FGSM adversarial attacks. By passing the adversarial examples through the denoising autoencoder, we were able to reconstruct and denoise the images, enabling the classifier to achieve significantly higher accuracy compared to its performance on the unprocessed adversarial set. The K-fold cross-validation process generated a more trustworthy assessment of the model's performance, mitigating the risk of bias or overfitting caused by a single train-test split.

However, our findings also revealed a trade-off between robustness to adversarial attacks and overall classification performance. While the autoencoder-based defense strategy improved the model's accuracy on adversarial examples, it did not fully recover the original level of accuracy achieved on clean, unperturbed images. This suggests that the denoising process may inadvertently remove or alter some discriminative features crucial for accurate classification.

Furthermore, we observed that the effectiveness of the proposed approach diminished as the perturbation factor used in the FGSM attack increased. For more severe adversarial perturbations, the autoencoder's ability to reconstruct and denoise the images were hindered, resulting in lower classification accuracies on the reconstructed test sets.

Despite these limitations, the promising results obtained in this study motivate further research into autoencoder-based defense mechanisms against adversarial attacks. Potential avenues for future work include exploring more complex autoencoder architectures, combining autoencoders with other defense techniques for a more comprehensive defense strategy, and extending the approach to more complex datasets and real-world applications.

Overall, our findings highlight the importance of addressing the sensitivity of machine learning models towards adversarial attacks and the need for robust defense mechanisms to ensure the safe and reliable deployment of these systems in critical applications. The incorporation of ensemble techniques, like the K-fold cross-validation, can further strengthen the robustness and generalization of these defense strategies.

5.2 Future Scope

Although the results of this study are encouraging, there are several areas for future research and improvements:

1. Exploring Advanced Autoencoder Architectures: Our experiments employed a relatively simple autoencoder architecture. Investigating more complex and specialized architectures, such as variational autoencoders or adversarial autoencoders, may lead to improved denoising and reconstruction capabilities, potentially enhancing the defense against adversarial attacks.

2. Evaluating Against Diverse Adversarial Attacks: In this study, we focused on the FGSM attack, which is a relatively simple and computationally efficient adversarial attack method. Future work should evaluate the autoencoder-based defense strategy against more advanced and sophisticated attack techniques, such as iterative or targeted attacks, to assess its robustness and generalizability.

3. Extending to Complex Datasets and Real-World Applications: Our experiments were performed on the MNIST dataset, which is relatively simple and well-understood. Future research should explore the applicability and performance of the autoencoder-based defense approach on more complex datasets and real-world applications, such as image detection, natural language processing, or cybersecurity.

REFERENCES

[1] Szegedy, C., et al. (2014). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

[2] Shekhar, S., et al. (2023). A Comprehensive Survey on Adversarial Attacks and Defenses in Machine Learning. arXiv preprint arXiv:2312.03520.

[3] Goodfellow, I. J., et al. (2015). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[4] Madry, A., et al. (2018). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

[5] Guo, C., et al. (2018). Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117.

[6] Ross, A. S., & Doshi-Velez, F. (2018). Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. arXiv preprint arXiv:1711.09404.

[7] Vincent, P., et al. (2008). Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning (pp. 1096-1103).

[8] Gu, S., & Rigazio, L. (2015). Towards deep neural network architectures robust to adversarial examples. arXiv preprint arXiv:1412.5068.

[9] Du, J., et al. (2019). Robust speech recognition with denoising adversarial autoencoders. arXiv preprint arXiv:1902.07955.

[10] Al-Dujaili, A., et al. (2018). Denoising autoencoder adversarial attacks against deep malware classifiers. In Proceedings of the 2018 IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1233-1238).

[11] Alotaibi, A., & Rassam, M. A. (2023). Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense. Future Internet, 15(2), 62.https://doi.org/10.3390/fi15020062

[12] https://blog.keras.io/building-autoencoders-in-keras.html