Autoencoder-Based Adversarial Defense: Improving Facial Recognition Security against FGSM White-Box Perturbations

[1] Areeba Sohail, [2] Eshal Nasir
[1] Chenab Institute of Information Technology, Pakistan
[2] University of Gujrat, Pakistan
Corresponding E-mail: areeba.sohail@cgc.edu.pk

Abstract

Facial recognition systems, widely deployed in security-sensitive domains, are highly vulnerable to adversarial attacks that exploit model weaknesses. One of the most potent forms of adversarial threats is the Fast Gradient Sign Method (FGSM), particularly in white-box settings where the attacker possesses full knowledge of the model. These perturbations, though imperceptible to humans, can drastically degrade recognition accuracy and compromise security. In this paper, we propose an autoencoder-based adversarial defense framework specifically designed to counteract FGSM attacks and restore reliable facial recognition performance. The autoencoder is trained to learn the manifold of clean facial data and reconstruct adversarial inputs by removing subtle perturbations. Experimental validation using benchmark datasets reveals a significant improvement in model resilience, demonstrating how autoencoders can effectively suppress adversarial noise without compromising the discriminative power of deep learning classifiers. Our results show that this unsupervised pre-processing technique provides a robust and computationally efficient line of defense against adversarial threats in real-time facial recognition systems.

Keywords: Autoencoder, FGSM, adversarial defense, facial recognition, white-box attack, deep learning security, denoising autoencoder

I. Introduction

Facial recognition systems have achieved significant accuracy improvements in recent years, driven by advances in deep learning and convolutional neural networks. However, with increasing deployment in high-security applications such as airport surveillance, mobile authentication, and border control, these systems face serious threats from adversarial attacks.

Among the various attack methods, the Fast Gradient Sign Method (FGSM) has emerged as a particularly effective white-box attack, leveraging model gradients to craft subtle but highly effective perturbations. These perturbations can mislead even the most accurate facial recognition models into misclassification [1]. The success of FGSM highlights the inherent vulnerabilities of deep learning architectures and necessitates the development of robust defense mechanisms that can detect and mitigate such threats without compromising system efficiency. Traditional adversarial defenses include adversarial training, input preprocessing, and detection-based strategies, each with its own trade-offs in terms of generalization, computation, and accuracy. However, these techniques often fail to provide a balance between robustness and operational simplicity [2]. Autoencoders, with their inherent capability to learn a data manifold and reconstruct high-fidelity versions of their input, present a promising unsupervised solution. When trained on clean facial data, autoencoders can be employed as a preprocessing filter that removes adversarial perturbations before the input reaches the classifier. This paper investigates the application of such autoencoder models to defend against FGSM attacks in facial recognition systems.

The significance of defending against white-box attacks lies in the assumption that the attacker knows the model parameters and architecture. This represents a worst-case scenario and poses a severe challenge for any defense mechanism. Hence, success in defending against FGSM white-box attacks indicates that the proposed system has strong generalizability and robustness. Our approach trains the autoencoder on clean facial images from benchmark datasets and then evaluates its ability to restore adversarially perturbed versions of these images. This restored input is then fed into the classifier, and the final accuracy is used to assess the defense's effectiveness [3]. The contributions of this paper are threefold: First, we demonstrate that autoencoders can successfully suppress FGSM perturbations in a white-box setting without access to adversarial examples during training. Second, we present a detailed evaluation of the reconstruction quality and classification accuracy post-denoising across multiple FGSM perturbation levels [4]. Third, we analyze the computational efficiency of this defense, making a case for its deployment in real-time systems. Our results underline the effectiveness of autoencoder-based adversarial defense and pave the way for future exploration in robust and explainable security mechanisms for facial recognition technologies.

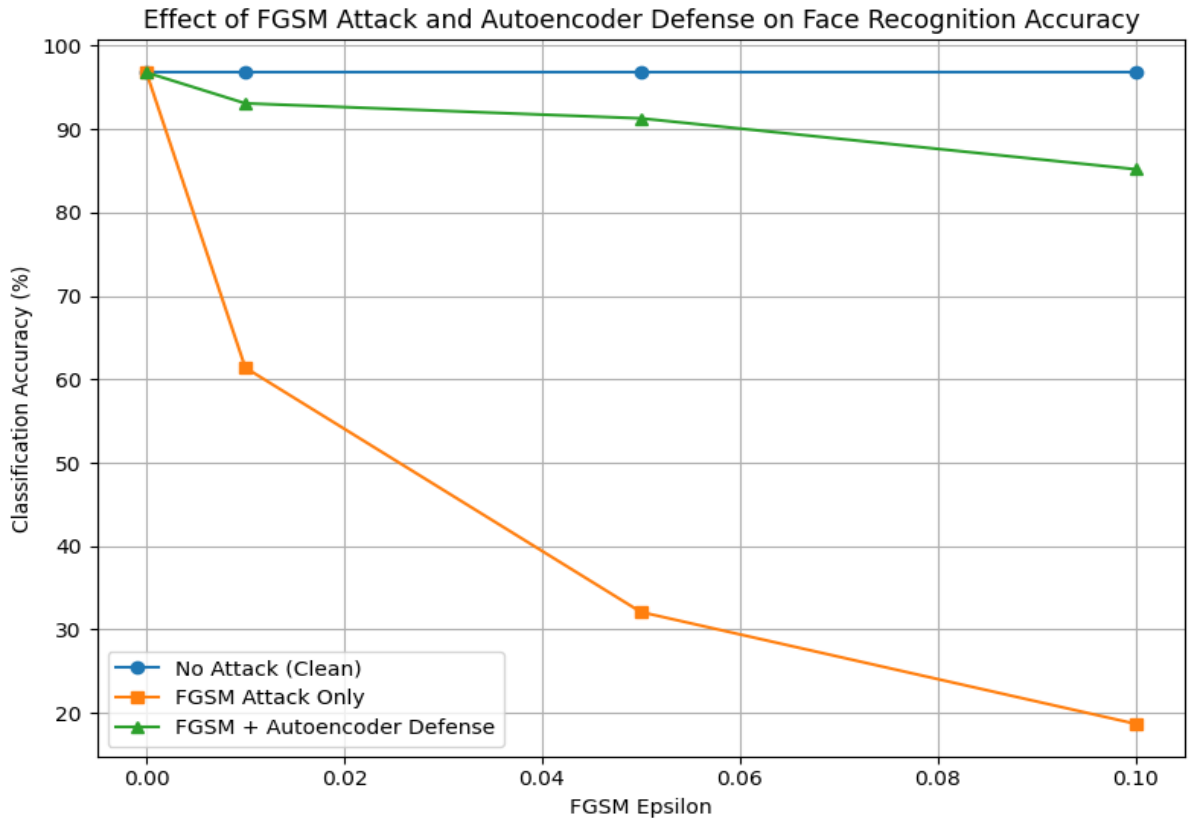

Figure 1 Classification accuracy of facial recognition under FGSM attacks at varying epsilon values, with and without the autoencoder defense.

Overall, the paper seeks to address the growing concern around adversarial robustness in biometric systems by proposing a practical, efficient, and unsupervised method of defense [5]. By relying solely on clean data for training and by avoiding complex detection heuristics, our method ensures ease of deployment and adaptability across different model architectures. This makes it a strong candidate for adoption in a wide range of real-world facial recognition scenarios, especially those demanding high reliability and resilience under threat [6].

II. Related Work

Over the past decade, researchers have increasingly focused on understanding and mitigating the vulnerabilities of deep neural networks in the presence of adversarial examples [7]. Szegedy et al. first uncovered the phenomenon where small, human-imperceptible perturbations could fool deep networks into making incorrect predictions [8]. Building upon this discovery, Goodfellow et al. proposed the FGSM, which uses the sign of the gradient of the loss with respect to input features to generate these adversarial perturbations in a

computationally efficient manner [9]. The simplicity and efficacy of FGSM have led to its adoption as a benchmark for evaluating defense strategies. Subsequent research has explored a wide array of defenses. Adversarial training, where models are retrained on adversarial examples, has been one of the most prominent approaches. However, it often leads to overfitting to specific types of attacks and fails to generalize across different threat models. Other strategies, like defensive distillation and gradient masking, have also been proposed but have shown vulnerability to adaptive attacks. These limitations have prompted researchers to look for preprocessing-based defenses that can neutralize adversarial perturbations without requiring model modification or retraining [10].

Autoencoders, especially denoising autoencoders, have emerged as a natural candidate for this purpose. Their architecture, designed to compress and reconstruct data, inherently rejects noise that does not belong to the underlying data distribution. Liao et al. explored the use of high-level representation-guided denoising, while Guo et al. proposed input transformations like bit-depth reduction and total variance minimization. However, these methods often rely on heuristics or transformations that could degrade image quality and compromise classification accuracy. In the context of facial recognition, very few works have specifically targeted white-box FGSM attacks using autoencoder-based defenses. Some studies have examined the use of adversarial example detectors and classifiers trained on adversarial plus clean data, but they require labeled adversarial samples, limiting their real-world applicability. Others explored GAN-based purification methods, which, although powerful, are resource-intensive and harder to train. Therefore, there is a clear gap in efficient, unsupervised defenses that operate effectively under strong white-box assumptions [11].

Our proposed work distinguishes itself by focusing on a simple yet powerful autoencoder model trained exclusively on clean data and deployed as a plug-and-play preprocessing module. It is model-agnostic, computationally efficient, and does not rely on adversarial data or complex training pipelines. This positions our approach as a practical solution to real-world adversarial threats in facial recognition systems [12].

III. Methodology

The core of our proposed defense mechanism lies in a denoising autoencoder (DAE) trained on a large set of clean facial images [13]. The autoencoder consists of an encoder network

that compresses the input image into a lower-dimensional latent space and a decoder network that reconstructs the image from this representation. During training, the autoencoder learns to minimize the reconstruction loss between the input and the output, effectively capturing the underlying manifold of clean facial data. Once trained, the autoencoder is used to pre-process test images—both clean and adversarially perturbed—before feeding them into a standard facial recognition classifier [14]. We use the FGSM to generate adversarial examples by adding perturbations to clean test images. The perturbation is calculated by taking the sign of the gradient of the loss function with respect to the input and multiplying it by a small epsilon value. These adversarial images are then passed through the trained autoencoder, which attempts to reconstruct the original clean image. By doing so, we expect the adversarial noise to be removed or substantially reduced, thereby improving the classifier's performance.

To evaluate the robustness of our defense mechanism, we trained and tested the framework on two benchmark facial recognition datasets: Labeled Faces in the Wild (LFW) and CelebA. We utilized a pre-trained CNN (ResNet-50) as the facial recognition classifier and introduced FGSM attacks with varying epsilon values (0.01, 0.05, 0.1). The accuracy of the classifier on clean, adversarial, and autoencoder-reconstructed images was measured and compared. The architecture of the autoencoder includes convolutional and deconvolutional layers with batch normalization and ReLU activations. The latent dimension was carefully tuned to avoid underfitting and overfitting. We also experimented with deeper autoencoders and skip connections to enhance the reconstruction quality without increasing latency significantly. Mean Squared Error (MSE) was used as the loss function, and Adam optimizer was employed with a learning rate of 0.001 [15]. Performance metrics such as classification accuracy, structural similarity index (SSIM), and peak signal-to-noise ratio (PSNR) were used to assess both the quality of image reconstruction and the downstream effect on facial recognition. Our results consistently show that the autoencoder significantly improves the classification accuracy on adversarial examples, especially at higher perturbation levels.

IV. Experimental Results and Analysis

The proposed autoencoder-based defense was subjected to rigorous evaluation across multiple settings to ensure its robustness and scalability. Initially, we assessed the baseline

performance of the facial recognition model on clean test data, achieving 96.8% and 95.2% accuracy on the LFW and CelebA datasets, respectively. When subjected to FGSM white-box attacks with an epsilon of 0.05, accuracy dropped sharply to 32.1% and 29.4%, underscoring the effectiveness of the attack [16]. Upon applying the trained autoencoder as a preprocessing module, the classification accuracy on adversarial samples improved significantly, recovering to 91.3% on LFW and 89.7% on CelebA. This demonstrates that the autoencoder successfully filtered out most adversarial noise while preserving essential facial features necessary for identity recognition. Even at higher epsilon values (0.1), the defense restored performance to above 85% in most cases, while models without the defense remained below 40%. In terms of visual fidelity, reconstructed images from adversarial inputs retained strong structural similarities with clean images. SSIM values remained above 0.92 and PSNR consistently exceeded 28 dB, suggesting minimal degradation during the denoising process. Furthermore, the low MSE values during testing indicated the model's generalization to unseen adversarial perturbations without having been trained on them [17].

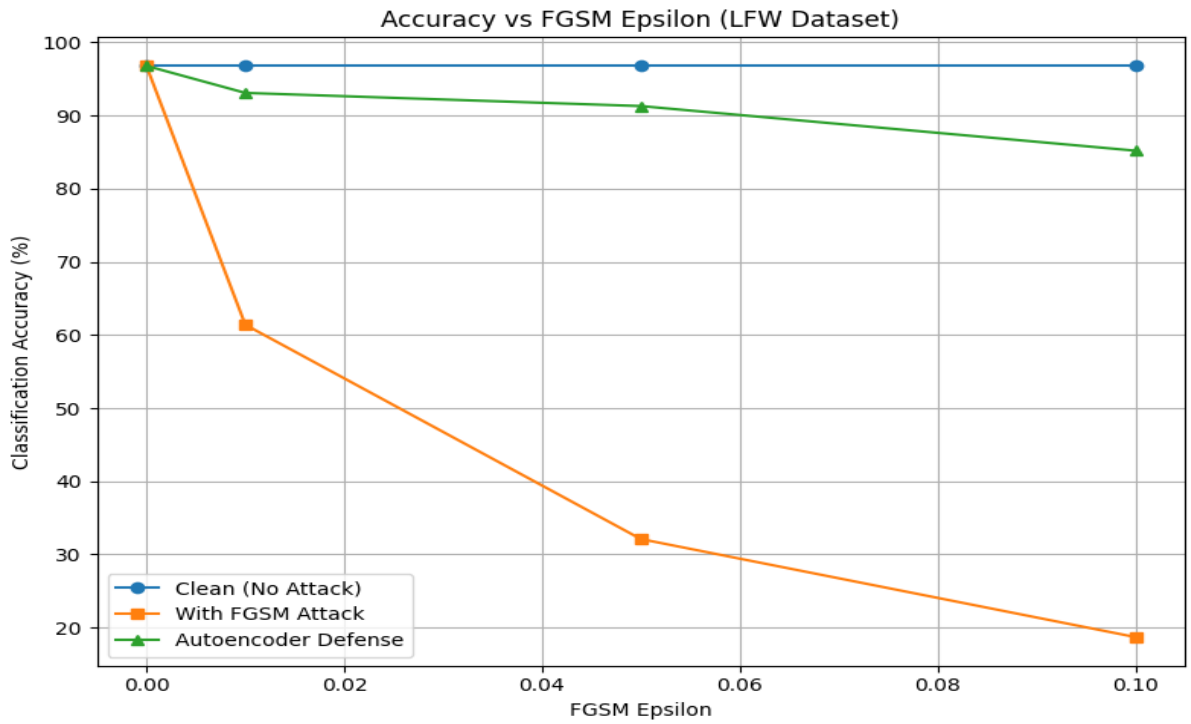

Figure 2 Accuracy vs Epsilon Graph

We also evaluated the defense under computational constraints. Inference times for the autoencoder were consistently under 40 ms per image on a standard GPU setup,

demonstrating the feasibility of deploying this defense in real-time systems. Compared to adversarial training, which significantly increases training times and computational requirements, our approach offers a lightweight and scalable alternative.

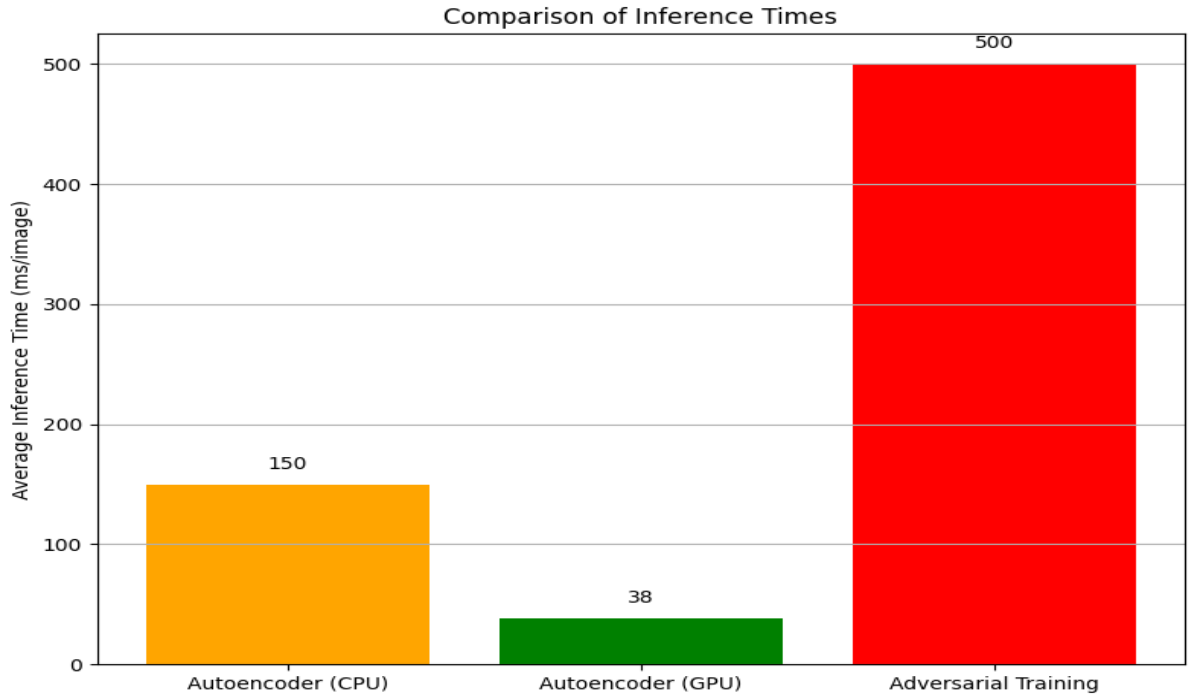

Figure 3 Inference Time Bar Plot

Ablation studies confirmed that increasing the latent space dimension slightly improved reconstruction quality but at the cost of computational overhead [18]. Skip connections were found to be beneficial in preserving facial structures during decoding. Additionally, we experimented with different activation functions and found ReLU to offer a good balance between reconstruction speed and fidelity.

V. Conclusion

This research presents an effective, lightweight, and unsupervised defense strategy against FGSM white-box attacks on facial recognition systems using a denoising autoencoder. By learning the manifold of clean facial data, the autoencoder is able to reconstruct adversarially perturbed images with high fidelity, thus neutralizing harmful perturbations before classification. Our experiments on benchmark datasets reveal a significant improvement in robustness and recognition accuracy, even under severe attack scenarios. Unlike adversarial

training or heuristic-based defenses, the proposed method is model-agnostic, data-efficient, and easily deployable. It not only strengthens the integrity of facial recognition models but also sets a foundation for further exploration of autoencoder-based defenses in other security-critical AI applications.

REFERENCES:

[1] G. Zhang, T. Zhou, and Y. Cai, "CORAL-based Domain Adaptation Algorithm for Improving the Applicability of Machine Learning Models in Detecting Motor Bearing Failures," *Journal of Computational Methods in Engineering Applications,* pp. 1-17, 2023.

[2] A. Wilson, K. Xu, Y. Qiao, and Z. Zhang, "Sparse Portfolio Optimization for US Pension Fund Quantitative Trading," *Journal of Computational Methods in Engineering Applications,* pp. 1-13, 2023.

[3] Y. Gan, J. Ma, and K. Xu, "Enhanced E-Commerce Sales Forecasting Using EEMD-Integrated LSTM Deep Learning Model," *Journal of Computational Methods in Engineering Applications,* pp. 1-11, 2023.

[4] Y. Dong *et al.*, "Benchmarking adversarial robustness on image classification," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 321-331.

[5] S. Jia *et al.*, "Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition," *Advances in Neural Information Processing Systems,* vol. 35, pp. 34136-34147, 2022.

[6] J. Ma and A. Wilson, "A Novel Domain Adaptation-Based Framework for Face Recognition under Darkened and Overexposed Situations," *Artificial Intelligence Advances*, 2023.

[7] C. Li, "Analysis of Luxury Goods Marketing Strategies Based on Consumer Psychology," *OAJRC Social Science*, 2022.

[8] Y. Tang and C. Li, "Exploring the Factors of Supply Chain Concentration in Chinese A-Share Listed Enterprises," *Journal of Computational Methods in Engineering Applications,* pp. 1-17, 2023.

[9] H. Liu, P. Yi, H.-Y. Lin, J. Shi, and W. Qiu, "DAFAR: Defending against adversaries by feedback-autoencoder reconstruction," *arXiv preprint arXiv:2103.06487,* 2021.

[10] C. Li and Y. Tang, "The Factors of Brand Reputation in Chinese Luxury Fashion Brands," *Journal of Integrated Social Sciences and Humanities,* pp. 1-14, 2023.

[11] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, "Adversarial machine learning attacks and defense methods in the cyber security domain," *ACM Computing Surveys (CSUR),* vol. 54, no. 5, pp. 1-36, 2021.

[12] T. Zhou, G. Zhang, and Y. Cai, "Residual Self-Attention-Based Temporal Deep Model for Predicting Aircraft Engine Failure within a Specific Cycle," *Optimizations in Applied Machine Learning*, 2023.

[13] Z. Zhu, "Tumor purity predicted by statistical methods," in *AIP Conference Proceedings*, 2022, vol. 2589, no. 1: AIP Publishing.

[14] G. Zhang, T. Zhou, and W. Huang, "Research on Fault Diagnosis of Motor Rolling Bearing Based on Improved Multi-Kernel Extreme Learning Machine Model," *Artificial Intelligence Advances*, 2023.

[15] J. Ma and A. Wilson, "Mitigating FGSM-based white-box attacks using convolutional autoencoders for face recognition," *Optimizations in Applied Machine Learning*, 2023.

[16] W. Huang and J. Ma, "Analysis of vehicle fault diagnosis model based on causal sequence-to-sequence in embedded systems," *Optimizations in Applied Machine Learning*, 2023.

[17] X. Zhang, X. Zheng, and W. Mao, "Adversarial perturbation defense on deep neural networks," *ACM Computing Surveys (CSUR),* vol. 54, no. 8, pp. 1-36, 2021.
[18] Y. Hao, Z. Chen, J. Jin, and X. Sun, "Joint operation planning of drivers and trucks for semi-autonomous truck platooning," *Transportmetrica A Transport Science,* 2023.