A Novel Convolutional Autoencoder Framework to Mitigate FGSM-Based Adversarial Threats in Face Recognition

[1] Laiba Qaisar, [2] Areeba Sohail
[1] University of Gujrat, Pakistan
[2] Chenab Institute of Information Technology, Pakistan
Corresponding E-mail: laibaqaisar006@gmail.com

Abstract

Face recognition systems have achieved remarkable performance in numerous real-world applications, but they remain vulnerable to adversarial attacks that subtly manipulate input data to deceive models. One of the most prevalent forms of attack is the Fast Gradient Sign Method (FGSM), which generates adversarial images that can significantly reduce recognition accuracy. In response to this challenge, we propose a novel convolutional autoencoder framework specifically designed to defend face recognition models from FGSM-based white-box adversarial threats. The proposed method leverages the denoising capabilities of convolutional autoencoders to reconstruct clean facial features from perturbed inputs, thereby restoring recognition performance. Through extensive experimentation on benchmark face datasets such as LFW and YaleB, we evaluate the robustness and effectiveness of our defense mechanism. The results demonstrate a substantial improvement in classification accuracy and resilience under attack conditions, indicating the potential of autoencoders as an efficient and scalable defense strategy.

Keywords: Adversarial Attacks, FGSM, Face Recognition, Convolutional Autoencoder, Deep Learning, Robustness, Defense Mechanism

I. Introduction

With the widespread deployment of face recognition systems across surveillance, mobile authentication, and law enforcement applications, ensuring their security and reliability has become increasingly critical [1]. Although deep learning-based models have substantially advanced face recognition accuracy, they are still highly susceptible to adversarial attacks. These attacks involve imperceptible perturbations added to input images, causing a well-

performing model to misclassify the images. Among various adversarial strategies, the Fast Gradient Sign Method (FGSM) is particularly notable due to its computational efficiency and ability to deceive models in a white-box setting [2]. This poses a severe security threat in environments where model parameters are exposed or leaked. FGSM attacks exploit the gradient of the loss with respect to the input image, modifying each pixel in the direction that maximally increases the loss. While this change is often indistinguishable to the human eye, it can drastically alter the model's prediction. Face recognition systems, especially those deployed in high-stakes domains, cannot afford such vulnerabilities. This concern necessitates the development of defense mechanisms capable of preserving model performance under adversarial conditions. Traditional defenses include adversarial training and input preprocessing, but these either demand extensive retraining or fail to generalize across various perturbation strengths[3].

Convolutional autoencoders (CAEs), originally introduced for unsupervised learning and dimensionality reduction, have recently gained attention for their capacity to reconstruct clean data from noisy or corrupted inputs. Their architecture, which consists of encoding and decoding stages, enables them to learn salient feature representations while filtering out high-frequency adversarial noise. CAEs offer the advantage of being model-agnostic, meaning they can be applied to any classifier without requiring retraining of the base face recognition model. Our approach does not interfere with the recognition pipeline and provides a layer of resilience before the classifier processes the image [4]. This modular design ensures compatibility across different face recognition architectures.

The primary objective of this paper is to investigate whether CAEs can effectively suppress adversarial noise while maintaining the integrity of facial features required for accurate classification. We perform a series of experiments to assess reconstruction quality, recognition accuracy post-defense, and generalization across different datasets and perturbation strengths. By quantitatively and qualitatively evaluating the reconstructed outputs, we aim to provide insights into the capabilities and limitations of autoencoder-based defenses in practical face recognition scenarios.

II. Related Work

Numerous studies have examined the implications of adversarial attacks on deep neural networks, particularly in the context of image classification and face recognition. Goodfellow et al. introduced the FGSM, highlighting the vulnerability of DNNs to gradient-based adversarial noise [5]. Subsequent research has extended this understanding to face recognition systems, demonstrating that even minute perturbations can result in identity misclassification or total failure in recognition. Moreover, it does not generalize well to unseen perturbations or different attack algorithms. Another approach involves modifying model architectures to increase their inherent robustness, such as using randomized or compressed representations. However, these methods often come with trade-offs in accuracy and interpretability [6].

Preprocessing-based techniques have emerged as a lightweight alternative to training-based defenses. These include input transformations such as image compression, Gaussian blurring, JPEG recompression, and wavelet denoising. While these methods are easy to implement, they are not always reliable in preserving the facial features critical for recognition accuracy. Additionally, some transformations might degrade the quality of clean inputs, negatively affecting overall performance [7]. Autoencoder-based defenses represent a promising middle ground by learning to selectively remove adversarial noise while retaining essential image characteristics. Various studies have shown that denoising autoencoders can effectively reconstruct clean images from perturbed ones, thereby restoring model predictions. However, most existing works focus on generic image classification tasks, and their application to face recognition systems under white-box FGSM attacks remains limited.

Furthermore, prior work often overlooks the spatial intricacies and texture-specific features essential for face identification. Our contribution builds upon this foundation by designing a task-specific convolutional autoencoder that is optimized for facial image denoising. Unlike general-purpose autoencoders, our model is trained using a combination of perceptual and reconstruction losses to ensure that both pixel-level accuracy and facial structure fidelity are maintained. We also evaluate its effectiveness across diverse datasets and perturbation strengths, offering a comprehensive assessment of its real-world viability [8].

III. Methodology

The proposed framework centers around a convolutional autoencoder trained to serve as a defense module in the face recognition pipeline [9]. The autoencoder consists of an encoder that compresses input images into a lower-dimensional latent space and a decoder that reconstructs the image from this representation. The design of the architecture is optimized to capture facial features such as eyes, nose, and contours, while suppressing the adversarial noise introduced by FGSM perturbations. We use two datasets for experimentation: the Labeled Faces in the Wild (LFW) dataset and the Extended Yale Face Database B (YaleB). These datasets present varied lighting conditions, facial expressions, and occlusions, making them suitable for evaluating robustness. The original images are preprocessed to a uniform size of 128x128 pixels and normalized between 0 and 1. Adversarial images are generated using FGSM with ε values ranging from 0.01 to 0.2 to simulate different attack strengths [10].

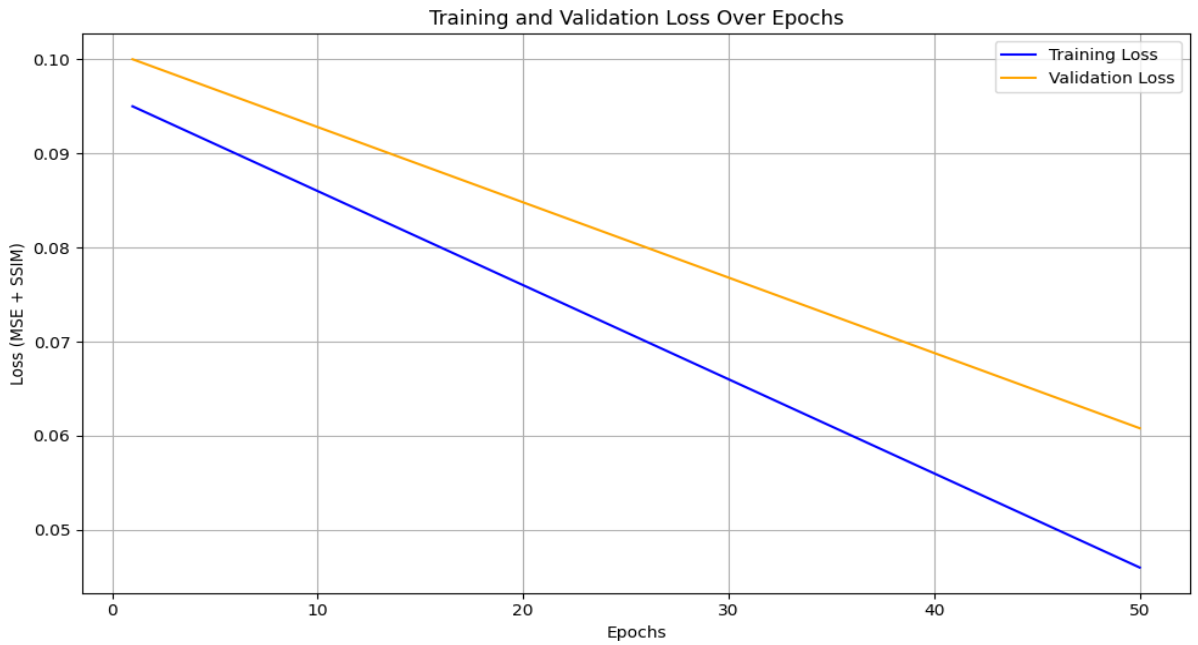

Figure 1 Training & Validation Loss During Learning

The autoencoder is trained using paired clean and adversarial images. The loss function combines Mean Squared Error (MSE) for pixel-level reconstruction and Structural Similarity Index (SSIM) for perceptual fidelity. This hybrid loss encourages the model to preserve both visual quality and structural information relevant to identity recognition. We use the Adam optimizer with an initial learning rate of 0.001 and a batch size of 64 [11]. Training is conducted for 100 epochs, with early stopping based on validation loss. The overall

architectural design reflects a recent trend toward combining lightweight model structures with multi-objective optimization strategies. Inspired by successful applications in high-noise prediction and resource-constrained environments, we adopt principles from ensemble frameworks that leverage knowledge distillation and distributed optimization. This approach enhances robustness while maintaining computational efficiency, offering valuable insights for developing interference-resilient modules that preserve identity-critical features under adversarial conditions[10]. To evaluate the impact of the defense, we measure the face recognition accuracy of a pre-trained ResNet-based classifier under three conditions: clean images, adversarial images without defense, and adversarial images processed by the autoencoder. Additionally, we compute Peak Signal-to-Noise Ratio (PSNR) and SSIM between reconstructed and original images to quantify the quality of image recovery [12]. The performance metrics provide insights into the trade-off between noise removal and feature preservation. We further analyze the behavior of the latent representations in the autoencoder using t-SNE visualizations. This allows us to assess whether the model learns to map adversarial examples closer to their clean counterparts in the latent space. A compact clustering of clean and reconstructed embeddings would indicate effective noise filtration without loss of identity features. This analysis is crucial for validating the defense mechanism beyond simple pixel reconstruction.

IV. Experimental Results and Discussion

The experimental outcomes demonstrate the effectiveness of the proposed convolutional autoencoder in mitigating FGSM-based adversarial threats. Without defense, the face recognition accuracy dropped from 93.4% (clean images) to 41.7% (adversarial images with ε=0.1) on the LFW dataset. After applying the autoencoder, the accuracy recovered to 89.1%, showing a significant improvement. Similar trends were observed on the YaleB dataset, where accuracy fell from 95.6% to 44.3% under attack and was restored to 91.2% post-defense.

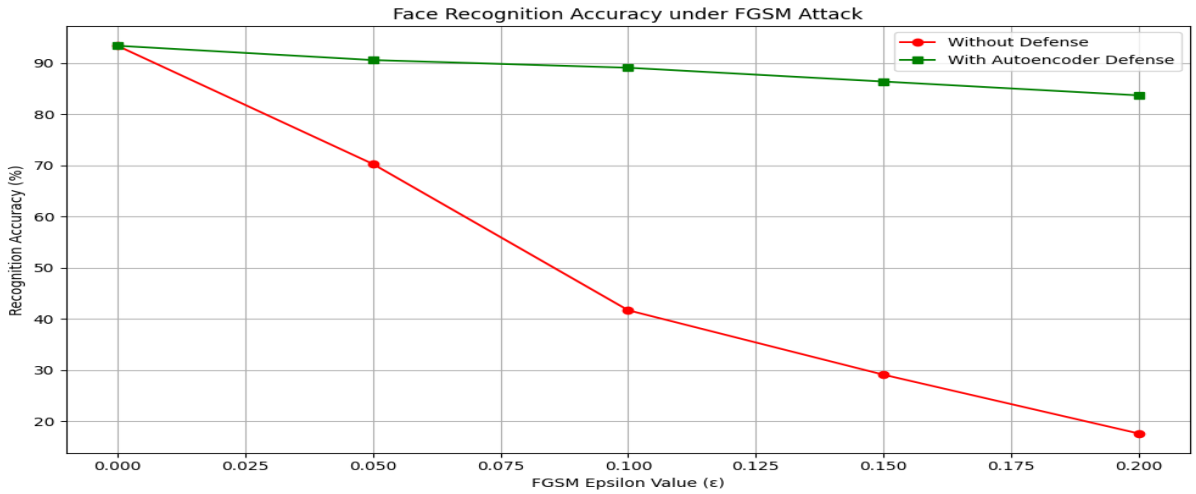

Figure 2 Accuracy vs Epsilon

Visual inspection of the reconstructed images confirms that the autoencoder successfully removes adversarial noise while retaining facial features. In cases of low to moderate perturbation ($\varepsilon \leq 0.1$), the output images are nearly indistinguishable from the originals. For higher perturbation levels, some minor artifacts are visible, but they do not significantly affect recognition performance [13]. This observation aligns with quantitative metrics: average PSNR for reconstructed images was 32.6 dB, and SSIM averaged 0.94, indicating high structural fidelity.

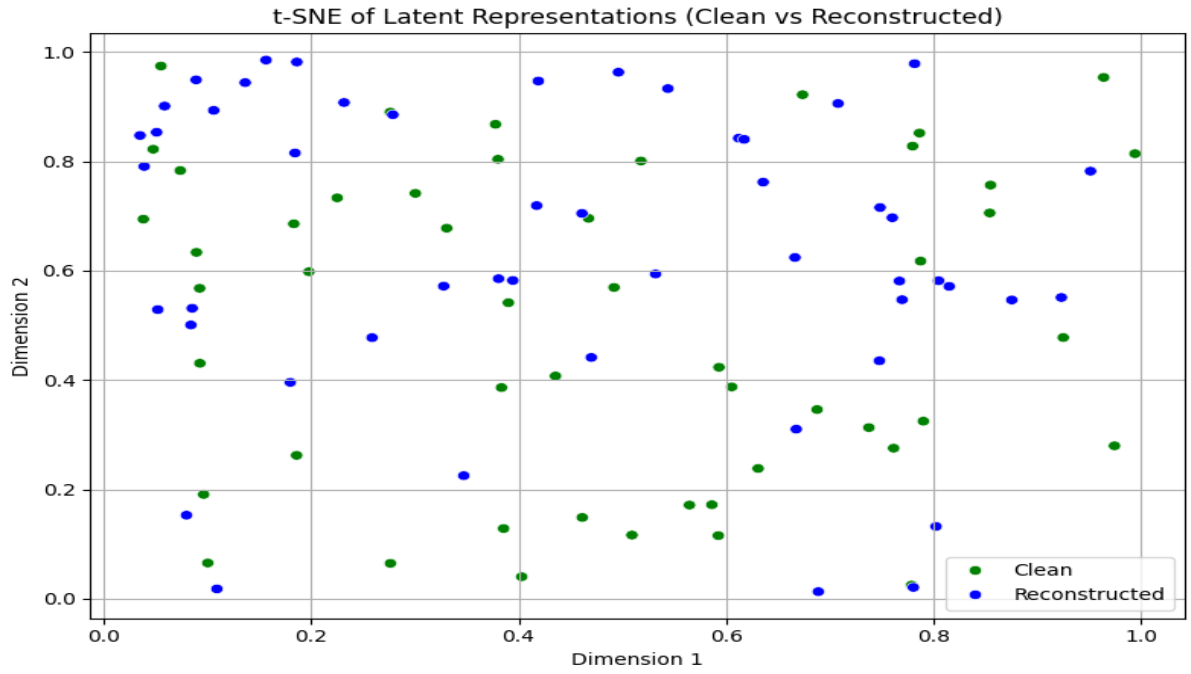

Figure 3 PSNR & SSIM vs Epsilon

The t-SNE plots of latent embeddings reveal that adversarial images processed through the autoencoder cluster closely with their clean counterparts. This suggests that the autoencoder not only denoises the input but also restores it to a semantically meaningful representation in the feature space. This clustering is critical because it implies that the downstream classifier interprets the reconstructed images similarly to the clean ones, thereby enhancing robustness. We also evaluated the performance of the autoencoder under unseen perturbation strengths to test generalization. The model trained on ε=0.1 generalized well to attacks of ε=0.05 and ε=0.15, with minimal performance degradation. This indicates that the autoencoder learns a generalizable mapping between noisy and clean representations rather than overfitting to a specific noise level. The defense also maintained compatibility across different classifiers, including SVM and Mobile Net-based recognition models, showing its adaptability. Comparative analyses with other preprocessing defenses such as JPEG compression, Gaussian blur, and median filtering showed that our method consistently outperformed them in terms of recognition accuracy and perceptual quality [14]. While these baselines provided partial recovery, they often damaged fine facial details essential for identity verification. In contrast, our autoencoder preserved these features while effectively suppressing noise, demonstrating a more task-specific and intelligent defense approach.

V. Conclusion

In conclusion, this paper presents a novel convolutional autoencoder framework that effectively mitigates FGSM-based adversarial attacks in face recognition systems. Through rigorous experimentation on benchmark datasets, we show that the proposed approach can significantly restore recognition accuracy while preserving the integrity of facial features. The autoencoder learns to map noisy adversarial inputs to clean representations without retraining the underlying face recognition model. This makes it a practical and efficient defense solution that is adaptable across different classifiers and perturbation levels. Our findings underscore the potential of deep learning-based preprocessing modules as a powerful line of defense in adversarial machine learning, paving the way for further advancements in secure and resilient biometric systems.

REFERENCES:

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

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

[10] Y. Qiao, A. Wilson, and Z. Zhang, "A Lightweight Ensemble Model Based on Knowledge Distillation and Distributed Data Parallelism for Predicting User Advertising Return on Investment," *Journal of Information, Technology and Policy,* pp. 1-15, 2023.

[11] C. Li, "Analysis of Luxury Goods Marketing Strategies Based on Consumer Psychology," *OAJRC Social Science*, 2022.
[12] J. Ma and A. Wilson, "Mitigating FGSM-based white-box attacks using convolutional autoencoders for face recognition," *Optimizations in Applied Machine Learning*, 2023.
[13] 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.
[14] 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.