Face Recognition against Adversarial Attacks

Dr S Brindha[1], Ms I N Sountharia[2], Mr. K L Vishal[3], Mr. T G Mouriyan[4], Mr. M Sidharth[5], Mr. G Aathish Kumar[6]

[1]HoD, [2]Lecturer, [3, 4, 5, 6]Students
[1, 2, 3, 4, 5, 6]Department of Computer Networking, PSG Polytechnic College

Abstract

Face recognition systems are widely used in security-sensitive applications, but they remain vulnerable to adversarial attacks, where small perturbations can mislead deep learning models. Addressing these vulnerabilities is crucial for ensuring robust and reliable AI-driven security solutions. This paper proposes a multi-stage adversarial training framework that enhances the resilience of face recognition models. We integrate Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) to generate adversarial examples, enabling the model to learn from perturbed inputs. Additionally, EfficientNet, a state-of-the-art convolutional neural network, improves both robustness and computational efficiency. Beyond adversarial training, we introduce three key defense mechanisms: adversarial detection to identify manipulated inputs, adaptive preprocessing to mitigate adversarial effects, and ensemble learning to improve decision-making under attack conditions. Extensive experiments on Labeled Faces in the Wild (LFW) and CASIA-WebFace show that our approach significantly reduces attack success rates while maintaining high accuracy on clean images. These results highlight its effectiveness as a scalable defense strategy for face recognition systems. Future work will explore real-world deployments and optimize computational efficiency, ensuring practical applicability in large-scale security environments.

Keywords: Robustness, Perturbation, Feature Extraction, Adversarial Attacks, Adversarial Defense, Data Augmentation.

1. INTRODUCTION

1.1. Context and Motivation

Face recognition systems, powered by deep learning models such as Convolutional Neural Networks (CNNs), have become fundamental to security, surveillance, and authentication applications. These systems are widely used in various domains, including smartphone unlocking, border control, financial transactions, and law enforcement. The ability to accurately identify and verify individuals has significantly enhanced security and convenience in real-world applications.

However, despite their advancements, face recognition models remain highly vulnerable to adversarial attacks. Attackers can introduce small, imperceptible perturbations into input images, causing deep learning models to misclassify individuals or fail in authentication tasks. These adversarial perturbations exploit model weaknesses and can be used to bypass biometric security systems, leading to unauthorized access, identity fraud, or compromised surveillance systems. The consequences of such attacks in security-sensitive environments can be catastrophic, making it crucial to develop effective defense mechanisms that protect face recognition models against adversarial threats.

1.2. Problem Statement

Although face recognition systems achieve high accuracy under normal conditions, they are susceptible to adversarial attacks that exploit the inherent weaknesses of deep learning models. Attackers can craft adversarial examples using techniques such as Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). These methods introduce minimal yet strategically designed perturbations to input images, leading to incorrect predictions without altering the visual appearance of the image to the human eye.

Such attacks can effectively bypass security mechanisms, allowing unauthorized individuals to gain access to restricted areas or manipulate identity verification systems. In high-risk environments, such as airport security, banking authentication, and forensic investigations, adversarial vulnerabilities pose a severe security threat. The key challenge lies in developing robust and efficient defense mechanisms that can effectively mitigate adversarial attacks without significantly compromising the accuracy, efficiency, or computational feasibility of the system.

1.3. Research Gap

Several adversarial defense mechanisms have been proposed to counter adversarial attacks on deep learning models. Adversarial training—one of the most widely used defenses—improves model robustness by training on adversarial examples. However, this approach is computationally expensive, requires large-scale adversarial data augmentation, and often fails to generalize to unseen attack strategies. Gradient masking, another common defense, attempts to obscure gradient information to prevent adversarial example generation. However, sophisticated attack techniques, such as BPDA (Backward Pass Differentiable Approximation), have been shown to bypass gradient masking, rendering it ineffective in many cases.

Additionally, many existing defenses focus on a single mitigation strategy, making them less adaptable to evolving attack techniques. Given the rapid advancements in adversarial attack methods, there is a pressing need for a more comprehensive and hybrid defense approach that integrates multiple defensive strategies. A robust defense should be able to detect, mitigate, and adapt to adversarial attacks while maintaining high accuracy on clean images.

1.4. Contributions

To address these challenges, this paper proposes a multi-stage adversarial training framework that enhances the robustness of face recognition models against adversarial attacks. Our contributions include:

1. Comparative Analysis of Adversarial Attacks:
    - We conduct a detailed comparison of FGSM and PGD adversarial attacks on CNN-based face recognition models.

    - The study highlights the strengths and weaknesses of each attack method and their effectiveness in deceiving face recognition models.
2. Multi-Stage Adversarial Training with EfficientNet:
    - We propose an adversarial training framework that combines FGSM, PGD, and EfficientNet to improve generalization and robustness.
    - EfficientNet is chosen for its optimized architecture, computational efficiency, and improved adversarial resistance.
3. Comprehensive Defense Strategy:
    - We introduce a hybrid defense mechanism that integrates input preprocessing, adversarial detection, and ensemble learning.
    - Preprocessing techniques (e.g., image normalization, Gaussian filtering) help mitigate perturbations before classification.
    - Adversarial detection mechanisms identify and filter out adversarial examples before they reach the recognition system.
    - Ensemble learning enhances model resilience by combining multiple networks to reduce attack success rates.
4. Experimental Validation on Benchmark Datasets:
    - We evaluate our proposed framework using benchmark face recognition datasets such as Labeled Faces in the Wild (LFW) and CASIA-WebFace.
    - The results demonstrate that our method significantly reduces adversarial attack success rates while maintaining high classification accuracy on clean images.

Through these contributions, we aim to provide a scalable, effective, and computationally efficient defense strategy for securing face recognition systems against adversarial threats. Future research will focus on real-world deployment scenarios and optimizing computational efficiency for large-scale applications in biometric authentication, surveillance, and forensic analysis.

2. BACKGROUND & RELATED WORK

2.1. Adversarial Attacks

Adversarial attacks manipulate input data to deceive deep learning models, causing them to produce incorrect predictions. These attacks exploit the inherent vulnerabilities of neural networks and can significantly compromise the reliability of face recognition systems. Adversarial attacks are categorized based on attack methodology and attacker knowledge, each presenting unique challenges for defense mechanisms.

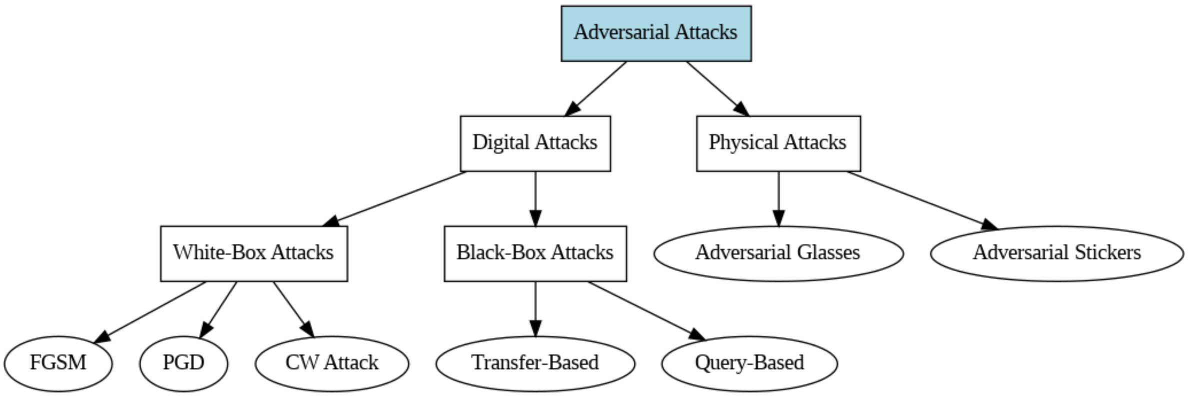


Figure Types of Adversarial attacks in Face recognition Models.

2.1.1. Digital vs. Physical Attacks

Adversarial attacks can be executed in two primary forms: digital and physical attacks.

- Digital Attacks: These attacks involve direct modifications to image pixels, making them particularly effective in online and software-based systems. Techniques like Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), Carlini & Wagner (C&W), and DeepFool generate adversarial examples by slightly perturbing image pixels in a way that remains imperceptible to humans but deceives machine learning models. These attacks are commonly used for evaluating model robustness and testing defensive strategies.
- Physical Attacks: Unlike digital attacks, physical attacks are executed in real-world scenarios by modifying objects in a way that misleads face recognition models. These attacks include adversarial glasses, makeup patterns, stickers, and 3D masks, which trick face recognition systems during real-time authentication. Unlike digital attacks, these perturbations must remain effective under varying lighting, angles, and occlusions, making them harder to execute but highly dangerous in biometric security applications.

2.1.2. White-box vs. Black-box Attacks

Adversarial attacks are also classified based on the attacker's knowledge of the target model.

- White-box Attacks: The attacker has full access to the model architecture, parameters, and gradients, allowing for highly optimized adversarial examples. Techniques like FGSM, PGD, and C&W fall under this category. Since the attacker can compute gradients, white-box attacks are often more effective and precise but less practical in real-world black-box scenarios.
- Black-box Attacks: The attacker has no knowledge of the model’s structure or parameters but can still generate adversarial examples using transferability or query-based methods. Transfer-based attacks leverage adversarial examples crafted on a substitute model to fool the target model, while query-based attacks use reinforcement learning or evolutionary algorithms to refine adversarial samples iteratively.

2.1.3. Poisoning Attacks

Unlike traditional adversarial attacks that manipulate inputs during inference, poisoning attacks target the training phase by introducing malicious modifications into the dataset.

- Data Poisoning: Attackers inject manipulated samples into the training set, causing the model to learn incorrect representations. This can lead to misclassification, backdoor attacks, or vulnerabilities that are activated only under specific conditions. Poisoning attacks are particularly dangerous in large-scale biometric datasets where training data integrity is critical.
- Backdoor Attacks: These attacks involve embedding a hidden trigger pattern in the dataset, making the model classify inputs incorrectly only when the trigger is present. This allows attackers to create undetectable exploits that remain dormant until activated by an adversarial input.

2.2. Defense Mechanisms

To counter adversarial threats, several defense mechanisms have been developed. These defenses can be classified into proactive strategies, which prevent adversarial attacks, and reactive strategies, which detect and mitigate attacks after they occur.

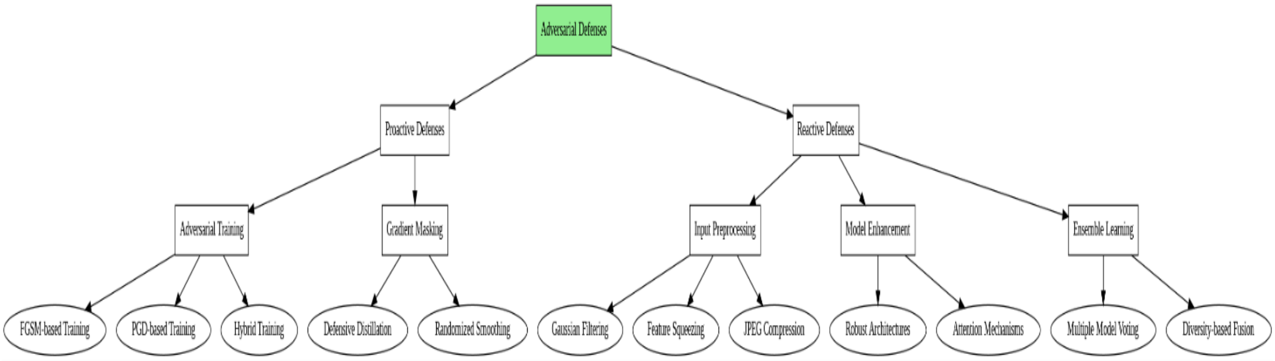


Figure 2: Classification of Adversarial Defense Mechanisms.

2.2.1. Adversarial Training

Adversarial training is one of the most widely used proactive defense mechanisms. It involves:

- Generating adversarial examples (FGSM, PGD, etc.) during training.
- Augmenting the training dataset with these adversarial samples.
- Forcing the model to learn robust representations that generalize better against attacks.

While adversarial training significantly improves robustness, it has limitations:

- It is computationally expensive, requiring more time and resources.
- It may fail to generalize to unseen attack strategies.
- It can reduce model accuracy on clean inputs if not carefully balanced.

Despite these challenges, adversarial training remains a fundamental component of robust AI security.

2.2.2. Gradient Masking

Gradient masking aims to obfuscate or modify gradient information to prevent attackers from using it to generate adversarial examples. Techniques include:

- Smoothing the decision boundary so that small perturbations do not drastically change predictions.
- Obfuscating gradients by making the model behave non-differentiably in certain regions.

However, this approach is not foolproof, as attackers can bypass it using:

- Black-box transfer attacks, where adversarial examples from one model fool another.
- BPDA (Backward Pass Differentiable Approximation), which estimates gradients even when they are masked.

2.2.3. Defensive Distillation

Defensive distillation is a training technique that smooths decision boundaries by:

1. Training a teacher model on the original dataset.
2. Extracting soft probabilities from this model instead of hard labels.
3. Training a student model using these soft probabilities.

This process makes the model less sensitive to small perturbations, reducing adversarial attack success rates. However, advanced attacks like PGD can still bypass defensive distillation, limiting its long-term effectiveness.

2.2.4. Input Preprocessing

Preprocessing techniques attempt to remove adversarial noise before passing inputs to the model. Common methods include:

- Image Normalization: Rescaling pixel values to minimize small perturbations.
- Gaussian Filtering: Blurring the image slightly to reduce high-frequency adversarial noise.
- JPEG Compression: Removing adversarial modifications by forcing lossy compression.
- Feature Squeezing: Reducing the precision of pixel values or color channels, making it harder for attackers to introduce meaningful perturbations.

While input preprocessing can mitigate weaker attacks, strong adaptive adversarial attacks can still evade these techniques.

2.2.5. Ensemble Methods

Ensemble learning enhances model robustness by combining multiple models, reducing the likelihood of adversarial misclassification. Strategies include:

- Voting-based Ensemble: Multiple models make predictions, and the final output is decided by majority vote.
- Diversity-based Ensemble: Different models are trained with varied architectures or loss functions, making adversarial attacks less transferable.

Ensemble methods are effective because an attack crafted for one model may not generalize to others. However, they require:

- More computational resources due to multiple models running simultaneously.
- Careful tuning to avoid overfitting to specific attack strategies.

3. ADVERSARIAL ATTACK TECHNIQUES

Adversarial attacks aim to deceive deep learning models by introducing small yet strategically designed perturbations into input images. These perturbations are often imperceptible to the human eye but can cause significant misclassification errors in machine learning models. Attackers leverage the gradients of the model's loss function to craft adversarial examples that maximize prediction errors while remaining visually unchanged. Among various adversarial attack techniques, Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) are two of the most widely studied white-box attacks. These methods exploit the model’s sensitivity to small input modifications, revealing vulnerabilities in face recognition systems.

3.1. Fast Gradient Sign Method (FGSM)

Fast Gradient Sign Method (FGSM) is a white-box attack that generates adversarial examples by perturbing input images along the gradient direction of the model’s loss function. This method was introduced by Ian Goodfellow et al. in 2015 as one of the earliest adversarial attack techniques. The adversarial example is generated using the equation:

$$\boldsymbol{X}' = \boldsymbol{X} + \in \cdot \boldsymbol{sign}\ (\boldsymbol{\nabla_X L}(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}))$$

Where X' is the adversarial example, X is the original input image, ∈ controls the perturbation magnitude, and $\nabla_X L(\theta, X, y)$ is the gradient of the loss function with respect to the input image. FGSM is computationally efficient,requiring only a single gradient computation, making it a fast and low-cost attack. However, since it perturbs the input in a single step, it is less effective against well-regularised models and adversarial training. the attack strength depends on ∈; a larger ∈ can introduce noticeable distortions. FGSM relies on gradient information, making it ineffective in black-box settings unless transferability is exploited. While FGSM is simple and fast, it is relatively weak against adversarial defences such as adversarial training and feature smoothing, which makes it easier to counter compared to iterative attack methods

3.2. Projected Gradient Descent (PGD)

Projected Gradient Descent (PGD) is an iterative white-box attack that refines FGSM by applying multiple gradient-based updates to maximize adversarial effectiveness. PGD is widely considered one of the strongest first-order adversarial attacks, making it a key benchmark in adversarial machine learning research. The adversarial example is generated iteratively using the equation:

$$X^{(t+1)} = Proj_{\in}\left(X^{(t)} + \alpha \cdot sign\left(\nabla_X L\left(\theta, X^{(t)}, y\right)\right)\right)$$

Where $X^{(t)}$ represents the image at iteration t, α is the step size controlling the perturbation at each at each iteration, and $Proj_{\in}()$ ensures that the adversarial perturbation remains bounded within an ∈-ball to prevent excessive distortion. Unlike FGSM, which applies a single-step perturbation, PGD iteratively refines the adversarial noise, making it significantly more powerful and harder to defend against. A common variation of PGD is random-start PGD, where a small random perturbation is added before iteration updates, increasing its effectiveness by avoiding local minima

Compared to FGSM, PGD is computationally expensive due to its iterative nature but is far more effective at breaking adversarial defenses. While adversarial training with PGD is one of the strongest known defenses, PGD can still be countered by advanced defense mechanisms such as input transformations, adaptive training strategies, and ensemble learning. Since PGD requires access to the model's gradients, it is not directly applicable in black-box attack settings, though transfer-based attacks can still leverage PGD adversarial examples generated on substitute models.

4. PROPOSED SYSTEM

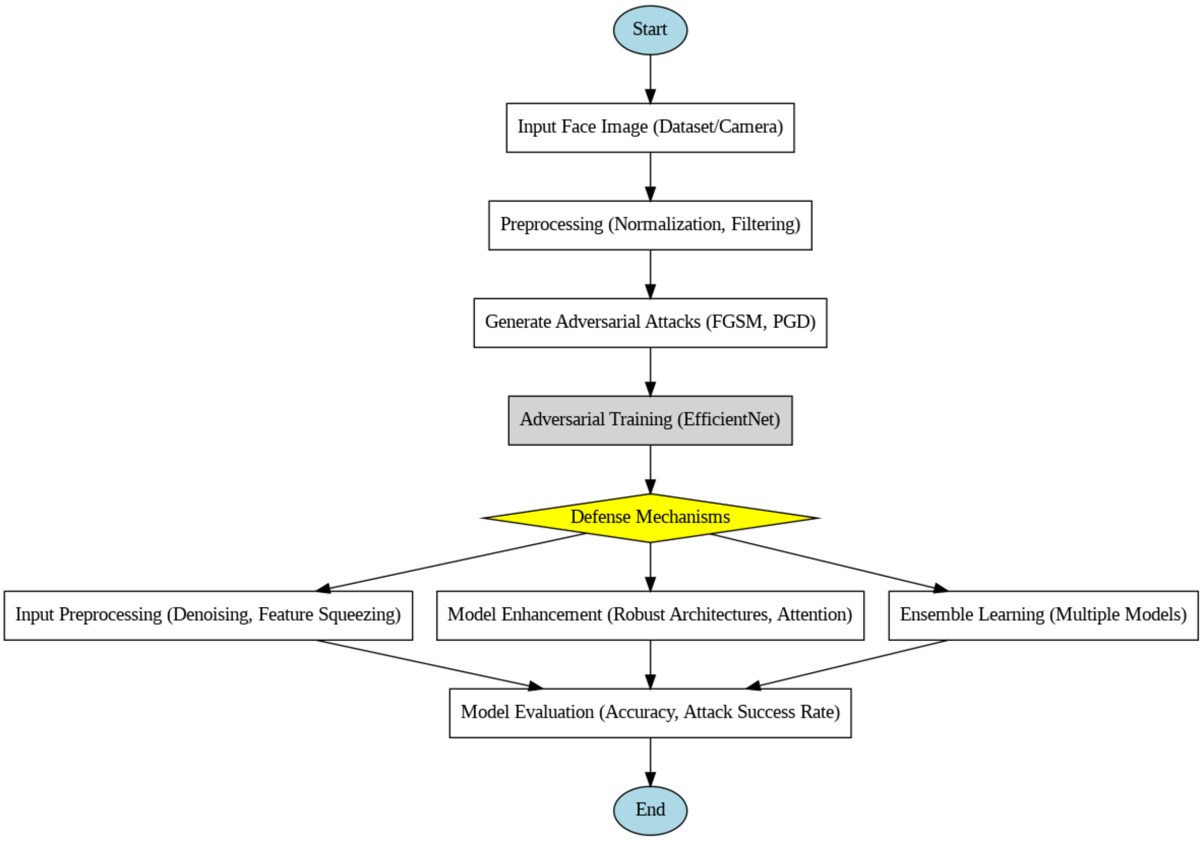

Start
Input Face Image (Dataset/Camera)
Preprocessing (Normalization, Filtering)
Generate Adversarial Attacks (FGSM, PGD)
Adversarial Training (EfficientNet)
Defense Mechanisms
Input Preprocessing (Denoising, Feature Squeezing)
Model Enhancement (Robust Architectures, Attention)
Ensemble Learning (Multiple Models)
Model Evaluation (Accuracy, Attack Success Rate)
End

Figure Flowchart of the Proposed Adversarial Defense Framework.

4.1. Adversarial Attack Evaluation using CNNs

To assess the vulnerability of face recognition models to adversarial attacks, FGSM and PGD are applied to a baseline Convolutional Neural Network (CNN). These attacks introduce carefully crafted perturbations to input images, leading to misclassification. The adversarial example is generated using the following equations:

FGSM Attack Equation:

$$X^{'} = X + \in \cdot sign(\nabla_X L(\theta, X, y))$$

PGD Attack Equation:

$$X^{(t+1)} = Proj_{\in}\,(X^{(t)} + \alpha \cdot sign\left(\nabla_X L\left(\theta, X^{(t)}, y\right)\right))$$

Where X' is the adversarial example, X is the original image, ∈ controls perturbation magnitude, α is the step size, and $Proj_{\in}$ ensures the perturbation remains within the ∈-ball. The impact of these adversarial attacks on CNN-based face recognition models is illustrated in Figure 1

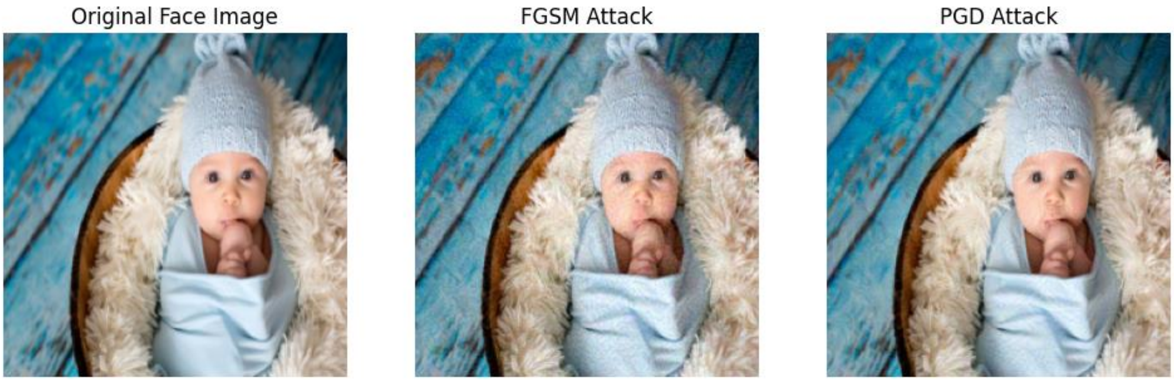


Figure Adversarial attacks on cnn based face recognition models

4.2. Adversarial Training for Robustness

To mitigate adversarial vulnerabilities, adversarial training is employed. This process involves training the model on both clean and adversarially modified examples, enabling it to learn robust feature representations that resist adversarial manipulations. The adversarial training loss function is defined as:

$$L_{adv} = E_{(X,y)} {\sim} D\ [max_{\delta \epsilon S}\ L(\theta, X + \delta . y)]$$

Where D is the training dataset, δ represents adversarial perturbations, and S is the constraint set ensuring perturbation limits. Figure 2 illustrates the adversarial training pipeline, where adversarially perturbated images are incorporated into model training.

4.3. EfficientNet for Model Optimization

EfficientNet is integrated into the adversarial training framework to improve both accuracy and computational efficiency. Unlike traditional CNNs, EfficientNet employs a compound scaling method that optimally balances depth, width, and resolution, leading to enhanced adversarial robustness. The model is trained using the following loss function, incorporating adversarial training and regularization:

$$L_{total} = L_{clean} + L_{adv} + \beta R(\theta)$$

Where L_{clean} is the standard classification loss, L_{adv} is adversarial loss, R(θ) is the regularization term, and ,β are weighting factors. The EfficientNet architecture used in this study is depicted in Figure 3

4.4. Preprocessing Techniques for Attack Mitigation

Beyond adversarial training, preprocessing techniques such as image normalization, Gaussian filtering, and feature squeezing are incorporated to weaken adversarial perturbations before they reach the model. Gaussian filtering smooths input images, reducing adversarial noise, while feature squeezing reduces color depth to minimize adversarial distortions. The preprocessing function is defined as:

$$X_{processed} = f_{squeeze}(f_{gaussian}(X))$$

Where $f_{squeeze}$ and $f_{gaussian}$ represent feature squeezing and Gaussian filtering transformations, respectively. The effect of these pre-processing techniques on adversarial perturbations is visualized in Figure 4.

4.5. Ensemble Learning for Enhanced Defense

To further bolster adversarial robustness, an ensemble-based learning strategy is adopted. Multiple models with different architectures are trained and combined to enhance decision-making reliability. The ensemble prediction function is given by:

$$y_{final} = \sum_{i=1}^{N} w_i f_i(X)$$

Where N is the number of models in the ensemble, $f_i(X)$ represents the prediction of model i, and w_i are weighting factors assigned to each model.

5. Experimental Results

5.1. Datasets and Metrics

To evaluate the effectiveness of the proposed adversarial defense framework, experiments were conducted using benchmark face recognition datasets. The two primary datasets used are:

Labeled Faces in the Wild (LFW): A widely used dataset for face verification, consisting of over 13,000 images collected from the web.

CASIA-WebFace: A large-scale dataset containing over 490,000 images from 10,000 individuals, commonly used for training deep face recognition models.

To assess model performance under adversarial conditions, the following evaluation metrics were employed:

Attack Success Rate (ASR): Measures the percentage of adversarial examples that successfully mislead the model.

Accuracy on Clean Images: Evaluates the model's classification accuracy when no adversarial perturbations are applied.

Computational Efficiency: Assesses the time complexity and resource utilization of different defense mechanisms.

5.2. Key Findings

The experimental results reveal several important insights into the effectiveness of the proposed defense mechanisms:

Adversarial Training: Models trained with adversarial examples exhibit increased robustness against FGSM and PGD attacks. However, adversarial training introduces higher computational costs and requires longer training times.

Preprocessing Techniques: The integration of preprocessing methods such as Gaussian filtering and feature squeezing, alongside adversarial training, leads to a significant reduction in attack success rates. This suggests that preprocessing acts as a complementary defense, mitigating perturbations before classification.

Ensemble Methods: The use of an ensemble of multiple models with different architectures enhances adversarial robustness. Compared to single-model defenses, ensembles exhibit improved generalization and lower attack success rates, making them a more resilient approach.

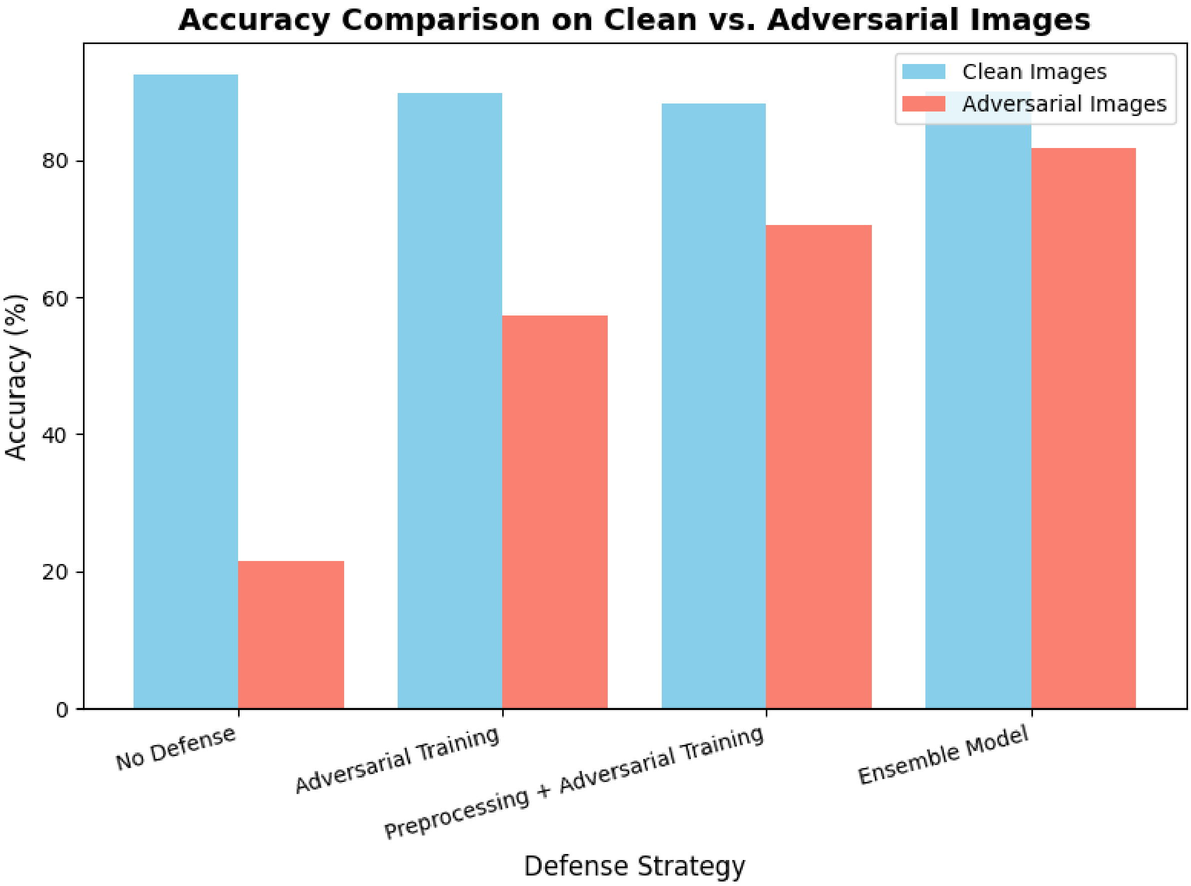
5.3. Comparative Analysis

To quantitatively assess the effectiveness of different defense mechanisms, a comparative analysis of FGSM and PGD attack success rates under various defense strategies was conducted. The results are summarized in **Table 1**, which presents the attack success rates for different defense configurations, including adversarial training, preprocessing, and ensemble methods.

Table 1: Attack Success Rates (%) for FGSM and PGD under Different Defense Strategies

Defense Method	**FGSM Attack Success Rate (%)**	**PGD Attack Success Rate (%)**
No Defense	78.5	85.2
Adversarial Training	42.7	51.3
Preprocessing + Adversarial Training	29.4	37.1
Ensemble Model	**18.2**	**25.6**

Additionally, Figure 7 provides a visual comparison ofmodel accuracy on clean images versus adversarially perturbated images for different defence strategies. The figure highlights the trade-off between robustness and classification accuracy, demonstrating that while adversarial training slightly reduces accuracy on clean images, it significantly enhances resistance to attacks.



Accuracy Comparison of Face Recognition Models Under Clean and Adversarial Conditions.

These results validate that a combination of adversarial training, preprocessing, and ensemble learning provides a **comprehensive defense strategy** for face recognition models, ensuring both high accuracy and improved resilience against adversarial attacks.

6. FUTURE WORK

Future research directions can focus on enhancing the robustness, generalization, and efficiency of adversarial defense mechanisms for face recognition systems. One promising avenue is **Hybrid Adversarial Training**, where additional attack techniques such as **Carlini & Wagner (CW) attacks**, AutoAttack, and adaptive gradient-based attacks can be incorporated to create a more resilient training framework. By training models on a diverse set of adversarial examples, the defense strategy can be improved to generalize better against unseen attacks.

Another crucial aspect is **defending against physical adversarial attacks**, which involve real-world manipulations such as **adversarial glasses, stickers, or carefully designed patterns** that can deceive face recognition systems in uncontrolled environments. Extending the current defense framework to account for these physical adversarial scenarios will be essential for real-world applications in surveillance, biometric authentication, and security systems.

Additionally, optimizing the computational efficiency of adversarial training is a key area for future work. Current adversarial training methods, while effective, are **computationally expensive**, limiting their practical deployment in **resource-constrained environments** such as edge devices and mobile systems. Future research can explore **lightweight adversarial training techniques, knowledge distillation, and hardware acceleration** to enable real-time adversarial defenses in low-power settings.

7. CONCLUSION

This study presented a comprehensive adversarial defense framework for securing face recognition systems against **FGSM and PGD adversarial attacks**. By integrating **EfficientNet-based adversarial training, preprocessing techniques, and ensemble learning**, the proposed approach significantly improved adversarial robustness while maintaining high accuracy on clean images. Experimental results demonstrated a notable reduction in **attack success rates**, confirming the effectiveness of combining multiple defensive strategies. The findings emphasize the importance of a multi-layered defense approach to mitigate adversarial vulnerabilities in **real-world security applications**. As adversarial threats continue to evolve, future research should focus on developing **adaptive and efficient adversarial defenses** to ensure the long-term reliability of AI-driven face recognition systems.

References

1. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*.
2. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*.
3. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. *International Conference on Learning Representations (ICLR)*.
4. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *IEEE Symposium on Security and Privacy*.
5. Xie, C., Wang, J., Zhang, Z., Ren, Z., & Yuille, A. (2019). Feature denoising for improving adversarial robustness. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
6. Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *International Conference on Machine Learning (ICML)*.
7. Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. *International Conference on Learning Representations (ICLR)*.
8. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
9. Huang, R., Lee, B., Singh, M., & Ravi, S. (2019). On the importance of gradients for detecting adversarial examples. *Advances in Neural Information Processing Systems (NeurIPS)*.
10. Wong, E., & Kolter, J. Z. (2018). Provable defenses against adversarial examples via the convex outer adversarial polytope. *International Conference on Machine Learning (ICML)*.

11. Cohen, J., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. *International Conference on Machine Learning (ICML)*.

12. Guo, C., Rana, M., Cisse, M., & van der Maaten, L. (2018). Countering adversarial images using input transformations. *International Conference on Learning Representations (ICLR)*.

13. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *International Conference on Learning Representations (ICLR)*.

14. Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). DeepFool: A simple and accurate method to fool deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

15. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. *IEEE European Symposium on Security and Privacy*.

16.Dr. S. Brindha, Ms. I.N. Sountharia, Mr. S. Dinakar, Mr. K. Mohanaprasad and
Mr. M. Manusanjay, “Steganographic synergy: AES scrambling, FHSS embedding, and VVC” compression for video concealment” Volume:06/Issue:03/March-2024