

RESEARCH ARTICLE

OPEN ACCESS

MITIGATING ADVERSARIAL VULNERABILITIES IN DEEP LEARNING BASED FACE RECOGNITION USING STACKED ATTENTION RESIDUAL GAN AND FIRE HAWK OPTIMIZATION

Sheilla Ann Pacheco¹, Mahesh Goyani*², Jayzel P. Bangoy³ and Born Christian A. Isip⁴

^{1, 3, 4}North Eastern Mindanao State University, Tandag City, Surigao del Sur, 8300, Philippines.

²Government Engineering College, Modasa – 383315, India.

¹<http://orcid.org/0000-0002-5632-8772>, ²<https://orcid.org/0000-0002-7269-1176>

³<http://orcid.org/0009-0009-7853-9747>, ⁴<http://orcid.org/0000-0002-7847-3006>

E-mail: sabpacheco@nemsu.edu.ph, *mgoyani@gmail.com, jpbangoy@nemsu.edu.ph, bcaisip@nemsu.edu.ph

ARTICLE INFO

Article History

Received on: March 03, 2026

Reviewed on: April 03, 2026

Accepted on: May 01, 2026

Published on: July 31, 2026

Keywords:

Adversarial attack,
Face recognition,
SARGAN,
LCPC,
Enhanced Fire Hawk
Optimization.

ABSTRACT

In computer vision and pattern recognition tasks, deep learning models are widely used, especially in face recognition systems. Even with their excellent performance, these models are still susceptible to a variety of adversarial manipulations, such as blur, additive noise, translation, flipping, scaling, rotation, and changes in illumination. Furthermore, some architectures might experience optimization problems like vanishing gradients, which would further impair the stability of the model. In order to provide robust face verification under adversarial attack, Optimized Deep Learning-based Adversarial Defense Mechanism (ODL-ADM) is proposed in this work. It projects adversarial samples into an immune feature space. A Learnable Convolutional Principal Component Network (LCPCN) is incorporated into the framework to create a representation space that is both discriminative and resistant to perturbations. Adversarially corrupted facial images are suppressed and reconstructed using a Stacked Attention-based Residual Generative Adversarial Network (SARGAN). Accurate identity recognition is achieved by an Improved Cross-Triple MobileNetV1 architecture after perturbation removal. Enhanced Fire Hawk Optimization (EFHO) is used for performance maximization and parameter tuning to further improve recognition performance. Following image reconstruction and adversarial perturbation removal, the suggested model achieves a 98% face recognition accuracy.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Deep learning has become a revolutionary paradigm that makes it possible to solve complicated, high-dimensional problems that are difficult for conventional algorithms to handle. In a variety of fields, such as computer vision [1], speech processing [2], natural language processing (NLP) [3], mobile payment systems [4], autonomous driving [5], robotics [6], and medical diagnosis, artificial intelligence and deep learning techniques have recently produced state-of-the-art outcomes. Deep neural networks, in particular, are now essential to many applications that are security-critical. Smart city infrastructures are made possible by fields like self-driving cars, intelligent healthcare systems, and facial recognition. By enabling real-time data processing, lower latency, and scalable intelligent services, the combination of deep learning models with edge computing devices has further accelerated the development of smart cities.

Even though deep neural networks (DNNs) work very well, they are very easy to trick with adversarial examples. These adversarial examples are made by adding small, carefully planned changes to the input data that DNNs cannot see. This can cause the DNNs to make wrong predictions [7]. These kinds of changes are usually too small for the human eye to see, but they change the model's decision boundary a lot, which leads to confident but wrong outputs. Szegedy et al. [8] were the first to systematically show how weak deep neural networks are. They did this by showing that DNN-based image classifiers can fail when only small amounts of noise are added on purpose. This groundbreaking discovery exposed a critical security vulnerability in deep learning models, especially in image classification tasks. Even though a lot of research has been done, there is still no complete theoretical explanation or universally effective defense mechanism that can fully solve this security problem.

After the initial discovery, many studies, especially in the field of computer vision, have looked into how to make adversarial examples and suggested different ways to make models more robust [9], [10], [11], [12]. Face recognition is the process of finding or confirming a person's identity by looking at their facial features in a still image or video and comparing them to a database of known identities [13]. Face detection and face recognition are still important and technically difficult tasks in biometric-based security systems [14]. In the last twenty years, a lot of research has been done to make face recognition systems more accurate and reliable. Previous methods mainly used hand-made feature extraction methods. Commonly used traditional descriptors for facial representation and classification tasks include Gabor filters [15], Local Binary Pattern (LBP) [16], and Histogram of Oriented Gradients (HOG) [17]. Pose, illumination, expression, and occlusion all vary greatly in the human face, which is a complex, high-dimensional visual entity. Compact, low-dimensional feature representations are usually created from facial data to facilitate effective storage and computational processing. Researchers have widely used dimensionality reduction methods like Principal Component Analysis (PCA) [18] and Linear Discriminant Analysis (LDA) [19] to accomplish this. These techniques improve classification performance and computational efficiency by projecting high-dimensional facial data onto a lower-dimensional subspace while maintaining the most discriminative or variance-rich information [20].

Szegedy et al. [8] were the first to propose adversarial examples, arguing that such perturbations are inherent in real-world data distributions. However, deep models find it challenging to learn robust decision boundaries against them during training due to their incredibly low probability of occurrence. Classifiers frequently struggle to correctly identify adversarial samples once they are added to the test set, which significantly degrades performance. Later, Ian Goodfellow et al. [21] contended that deep models' high-dimensional linear behavior—especially when linear activation functions are used—is primarily responsible for their susceptibility to adversarial attacks. Small, well-designed perturbations can accumulate and significantly alter model predictions due to the linearity of high-dimensional feature spaces. Anil Rozsa et al. [22] introduced a technique called the Fast Flipping Attribute (FFA) attack to modify particular facial characteristics while maintaining the overall identity in the context of facial attribute manipulation. Similar to this, Vahid Mirjalili and Arun Ross [23] showed how to manipulate facial images specifically by flipping just the gender attribute while maintaining the same biometric features. Additionally, Anubha Chhabra et al. [24] created adversarial perturbations with the goal of changing specific facial characteristics while preserving visual realism and identity consistency.

The main contribution of the study is as follows:

- To lessen the effects of adversarial attacks, we suggest an auto encoder-based Learnable Convolutional Principal Component Network (LCPCN) that projects adversarial images into an immune feature space, which reduces the strength of attack.
- We also developed a Stacked Attention Residual Generative Adversarial Network (SARGAN) for effective image reconstruction. The model uses stacked attention mechanisms and residual learning to accurately rebuild facial images, which improves its performance on face verification tasks.
- For face verification, we employed MobileNetV1 as the baseline architecture. The model was further enhanced by incorporating Cross Triplet Loss to improve feature discrimination. Additionally, we proposed a novel optimization algorithm, termed Enhanced Fire Hawk Optimization, to effectively optimize the network weights and enhance overall performance.
- Lastly, an ablation study was carried out to quantify the incremental performance improvements in face recognition attained by the overall framework and to show the contribution of each individual component.

The remainder of the paper is structured as follows. Section II presents a comprehensive review of the related literature. Section III describes the proposed methodology in detail. Section IV discusses the experimental results and performance analysis. Finally, Section V concludes the paper and outlines potential future research directions.

II. THEORETICAL REFERENCE

Deep neural networks (DNNs) have been extensively used in computer vision, speech recognition, natural language processing (NLP), and many other application domains in recent years. Their remarkable empirical success has fueled rapid advancements in artificial intelligence, especially through Convolutional Neural Networks (CNNs), and has led to significant progress in both academic research and industrial systems. However, empirical data shows that even small changes to model parameters or input images can seriously impair deep learning systems' performance [8]. Attackers can introduce attacks at various points in the pipeline, such as by changing learnable model parameters (model-space attacks) or directly modifying input data (input-space attacks). The adversary has full knowledge of the model architecture, parameters, and gradients in a white-box attack scenario, which allows for accurate and efficient attack generation. On the other hand, in a black-box setting, the adversary must rely on query-based tactics or transferability properties to create adversarial examples because they have little or no access to internal model details. The majority of adversarial defense systems in use today are made to handle specific or frequently occurring attack scenarios.

However, model performance frequently deteriorates due to inadequate generalization capability when substantial variability is introduced in the input data distribution. Systematic data augmentation techniques increase dataset variability to improve robustness against a variety of input conditions, allowing the model to learn invariant feature representations. Overfitting problems in a number of earlier studies, such as [7], [25], and [26], restrict their applicability to new data. The suggested work uses dropout regularization and skip connections to reduce overfitting. Additionally, MobileNetV1 incorporates an Enhanced Fire Hawk Optimization (EFHO) algorithm to optimize parameter learning and further minimize overfitting in the face recognition framework. In contrast to earlier studies like [27], which mainly concentrate on linearly separable data, real-world datasets usually show high intra-class variability and are non-linearly separable. The proposed study uses 20 different classes with significant image diversity to better reflect real-world scenarios, which leads to a classification problem that is non-linearly separable. Furthermore, a large portion of the literature currently in publication (such as [12], [28], [29] focuses exclusively on the detection or elimination of adversarial attacks.

On the other hand, by assessing the downstream performance of face recognition following adversarial noise removal, the current work goes beyond attack mitigation. Furthermore, the suggested approach can be applied to more general pattern recognition tasks and is not limited to face recognition. Additionally, vanishing and exploding gradient issues are common with deep learning models [30]. The suggested SARGAN architecture uses residual connections, which transfer data from previous layers to later layers, to solve this problem. Residual learning keeps gradients from diminishing toward zero and stabilizes the training process by promoting gradient flow across layers. Although some previous studies suggest methods for evaluation, they frequently lack thorough analytical validation, extensibility, and accurate metrics. The current study uses a variety of evaluation metrics, such as accuracy, precision, recall, F1-score, Cohen's kappa, Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM), to guarantee robust performance assessment. White-box settings, where complete model knowledge is assumed, were the main focus of early research on adversarial attacks against face recognition systems. For example, by tricking local DNN-based face models, Mahmood Sharif et al. [31] created adversarial eyeglass frames that allow an attacker to pose as someone else. Adversarial patches of different shapes were designed to attack the public face recognition model ArcFace in order to further improve transferability [32].

For adversarial patches to continue to be effective in real-world situations, they must have strong robustness against environmental interference. Once static adversarial examples fail under changing circumstances, they are frequently hard to modify. Adversarial light-based attacks, on the other hand, are more adaptable to changing physical environments. By creating a wearable hat with infrared LEDs embedded to trick facial recognition software, Zhou et al. [33] suggested a non-invasive physical attack. More recently, Anh Nguyen et al. [34] presented a projection-based technique that does away with wearable accessories by projecting digital adversarial patches onto faces using commercial projectors. Black-box attacks are now more useful in real-world applications than white-box techniques. Application programming interfaces (APIs), which enable external queries without disclosing internal model details, are commonly used by commercial face recognition systems. Through repeated API queries and output observation, malicious actors can use this interface to create adversarial examples. For instance, assuming the target model is unknown, Mahmood Sharif et al. used iterative querying to update adversarial patterns. In addition to creating perturbations, identifying vulnerable facial regions is a crucial challenge in adversarial patch attacks.

Targeting weak areas reduces overfitting of perturbations to particular conditions and increases attack efficiency. In order to identify the most vulnerable facial regions for cartoon sticker attacks, Wei et al. [35] suggested a query-based method to find the best adversarial positions and rotation angles. Even though these techniques do not require explicit perturbation gradient calculations, system administrators may become suspicious of excessive querying. In order to solve this, Guo et al. [36] suggested a joint optimization strategy for both perturbation and patch location, which reduced the number of queries needed from thousands to just a few dozen. Additionally, Sun et al. [37] optimized ApaNet by introducing a novel loss function. The hyperbolic tangent (Tanh) activation function used by the framework may cause vanishing gradient problems, which could cause training to stall or slow down because of insignificant weight updates during backpropagation. ApaNet uses intermediate and high-level feature activations from FaceNet to optimize a loss function that measures the deviation between reconstructed outputs and supervisory signals in order to reduce imperceptible adversarial perturbations in source images. The suggested model works well in both white-box and black-box attack scenarios, according to experimental evaluations on benchmark datasets like LFW, YouTube Faces DB, and CASIA-FaceV5. Prominent adversarial attack mechanisms are summarized in Table 1.

Table 1: Comparative analysis of prominent adversarial attack mechanisms.

Adversarial Attacks	Attack Type	Attack Target	Attack Frequency	Perturbation Type	Perturbation Norm	Attack Strategy	Primary Strength
L-BFGS [8]	WB	T	OS	SP	l_∞	CO	Early optimization approach – finds small perturbations
FGSM [21]	WB	T	OS	SP	l_∞	GO	Extremely fast, simple
JSMA [38]	WB	T	ITR	SP	l_2	SA	Sparse, targeted attacks
C & W [9]	WB	T	ITR	SP	l_0, l_2, l_∞	CO	Strong baseline, low distortion
One-Pixel [39]	BB/WB	NT	ITR	SP	l_0	EDB	Minimal pixel changes
DeepFool [40]	WB	NT	ITR	SP	l_0, l_2, l_∞	GO	Finds minimal perturbation
ZOO [41]	BB	T	ITR	SP	l_2	MM	Black-box gradient estimation
UAP [42]	WB/BB	NT	ITR	UN	l_2, l_∞	GO	Generalizes across inputs

Source: Authors, (2026).

WB: White box, BB: Black box, T: Targeted, NT: Non Targeted, OS: One Shot, ITR: Iterative, SP: Specific, UN: Universal, CO: Constrained Optimization, GO: Gradient Optimization, MM: Migration Mechanism, SA: Sensitivity Analysis, EDB: Estimated Decision Boundary.

III. MATERIALS AND METHODS

By projecting adversarial samples into an immune space, a novel adversarial defense mechanism is put forth to neutralize adversarial perturbations and provide robust face verification in challenging and unrestricted scenarios. In contrast to other areas of the representation space, the immune space is conceived as a feature subspace where adversarial perturbations have a negligible impact on the recognition model. First, an in-house dataset is created. To guarantee variety in posture and expression, 20 representative frames are taken from each recorded video sequence. Systematic data augmentation techniques are used to increase variability even more. Facial images with different scale, rotation, translation, horizontal flipping, illumination, and other transformations are included in the augmented dataset. Data augmentation accomplishes two main goals: (1) it broadens the size of the training dataset, which enhances representation learning and lessens overfitting; and (2) it increases intra-class diversity, which allows the model to generalize to unknown test conditions.

The suggested framework then locates and builds the immune space, which allows potentially harmful inputs to be projected into a strong subspace where their negative effects on face verification performance are greatly reduced. The immune space is first determined from the input data using a novel framework called the Learnable Convolutional Principal Component Network (LCPCN). This method reduces dimensionality while maintaining the most important features needed for immune space estimation. In contrast to traditional PCA-based methods, LCPCN enables automatic and data-driven immune space discovery by directly learning discriminative projections from image data in an end-to-end fashion. LCPCN effectively processes high-resolution images and captures spatial correlations across local regions by utilizing convolutional neural network (CNN) architectures. This makes the framework appropriate for challenging immune space detection tasks found in real-world face verification systems and scalable for large-scale datasets. The projected data are then run through an Attention-Based Residual Generative Adversarial Network (SARGAN), which serves as an adversarial purification module, to suppress adversarial perturbations. The network can highlight prominent facial regions while attenuating irrelevant or noise-corrupted areas thanks to the attention mechanism. During reconstruction, residual connections maintain crucial identity information and further stabilize training. The Enhanced Cross-Triple MobileNetV1 model is used to process the purified images for face verification following the removal of adversarial noise.

MobileNetV1's lightweight and computationally efficient design makes it ideal for deployment on platforms with limited resources, like mobile devices and embedded systems. The improved architecture maintains these efficiency benefits, allowing for quick inference and minimal computational overhead. Furthermore, because the model can adjust to different image resolutions, it can be used in a variety of deployment scenarios, such as low-resolution security cameras and high-definition video conferencing systems. Lastly, an Enhanced Fire Hawk Optimization (EFHO) algorithm is used to minimize losses and fine-tune parameters. Because of EFHO's quick convergence, near-optimal solutions can be found in a small number of iterations, which lowers the computational cost of optimization. EFHO also includes adaptive mechanisms that dynamically balance exploitation and exploration according to the search space's properties. This flexibility ensures successful parameter tuning and enhanced overall model performance by enabling effective navigation of intricate and dynamic optimization landscapes. Figure 1 shows the conceptual framework for the suggested methodology. Data collection, data pipeline construction, data cleaning, data validation, adversarial attack injection, immune space generation, perturbation removal, and model development and classification are some of its sequential phases. To guarantee strong adversarial mitigation and precise face verification under various operating conditions, each phase is methodically designed.

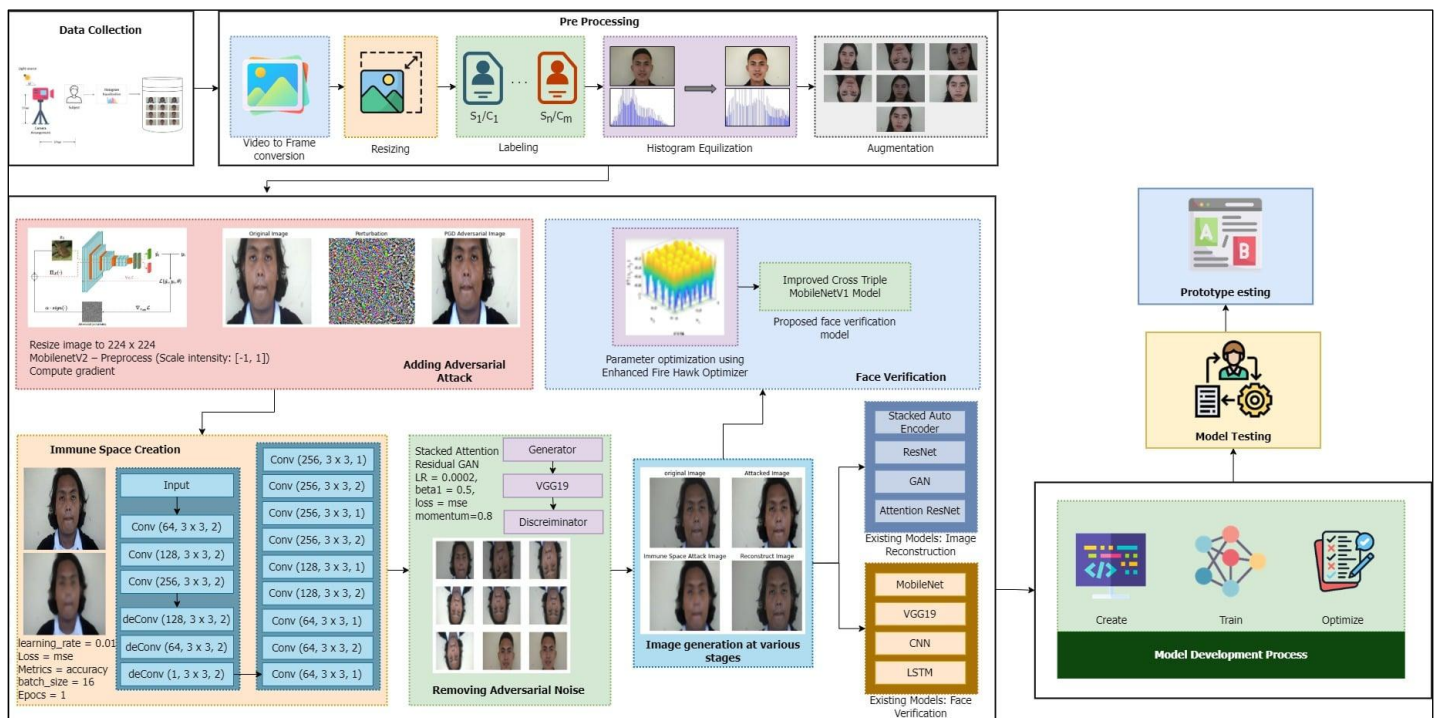


Figure 1: Conceptual framework of proposed work.
Source: Authors, (2026).

In order to deceive the recognition network, the adversarial perturbations are purposefully added to preprocessed (cleaned) images. A purification module based on a Generative Adversarial Network (GAN) is used to eliminate adversarial noise prior to classification in order to counteract these perturbations. After that, MobileNetV1 receives the cleaned photos for facial recognition. To strengthen resilience and improve generalization against such attacks, the network is rigorously trained with adversarial exposure. MobileNetV2 is combined with the Projected Gradient Descent (PGD) technique to generate adversarial examples. This method creates perturbations that push the model toward misclassification by repeatedly maximizing the loss function in relation to the input image.

Algorithm PGD_ATTACK()

 $x_i \in X, K: \text{no of iterations}, \epsilon$
 $x_{adv} \leftarrow x_i$ **for** $k \in \{1, \dots, K\}$ **do** x_{adv} $\leftarrow x_{adv} + \alpha \cdot \text{sign}(\nabla_{x_i} \mathcal{L}(x_i, y_i, \theta))$ $x_{adv} \leftarrow \Pi_X x_{adv}$ **if** $\|x_{adv} - x_i\| > \epsilon$ **then** $\|x_{adv} - x_i\| \leftarrow \epsilon$ **end if****end for**

Here, x_i is i -th sample in dataset X , x_{adv} is adversarial sample, $\Pi_X x_{adv}$ is projection of adversarial sample. ϵ is maximum allowable difference. So it turns out to be constrained optimization problem. Figure 2 shows the original image, added perturbation and PGD adversarial image. To identify the immune space, the suggested framework uses a Learnable Convolutional Principal Component Network (LCPCN). There are two distinct phases to the LCPCN architecture's operation. In order to learn compact and noise-resilient feature representations, a denoising auto encoder is first trained using convolutional and transpose convolutional blocks. During this stage, the network can suppress unimportant variations while capturing dominant structural patterns. In the second stage, adversarially perturbed images are projected onto the learned immune space using an Eigenvector Selection Network (ESN). By keeping only, the most discriminative elements, this projection method reduces the impact of adversarial noise. An example of an adversarially attacked image and its corresponding projection onto the immune space are shown in Figure 3.

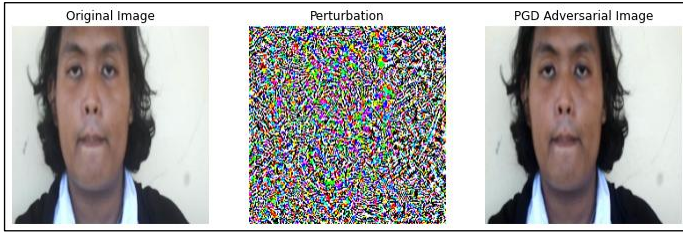


Figure 2: Original image, added noise and perturbed image.
Source: Authors, (2026).



Figure 3: Attacked image after projection on immune space.
Source: Authors, (2026).

The visual representation of the model is depicted in Figure 4.

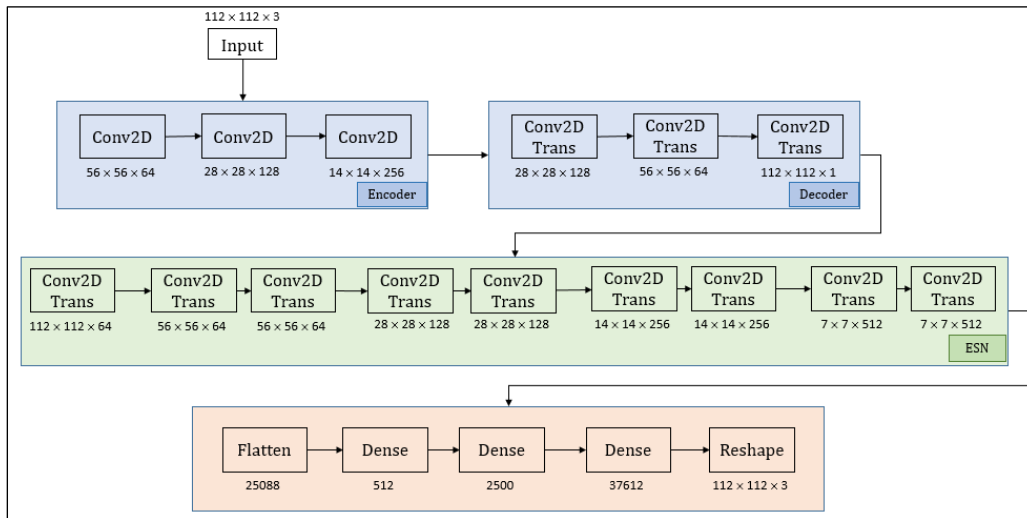


Figure 4: Visual representation of LCPCN model.
Source: Authors, (2026).

The images that are projected into the immune space are then sent to the SARGAN model along with their clean versions for supervised adversarial purification. An attention mechanism is built into the GAN framework to find both horizontal and vertical feature dependencies in facial images. The model also captures how feature maps interact with each other, which makes the contextual representation richer and the reconstruction more accurate. Residual blocks are added to the network because deep architectures can have the vanishing gradient problem. These residual connections create identity shortcuts that skip over middle layers and send feature information from the current layer to deeper layers.

This mechanism keeps the gradient flow steady during backpropagation and stops features from getting worse as the network gets deeper. The architecture also combines outputs from the pre-residual and post-residual stages to keep the structure intact and the inputs consistent during the reconstruction process. The algorithm also uses the VGG19 network to get high-level features from the generated (fake) images to improve perception. These traits help with reconstruction, making sure that the cleaned-up outputs keep identity-specific and semantic traits that are similar to those of the clean ground-truth images. Figure 5 illustrates the overall architecture and structural configuration of the SARGAN model.

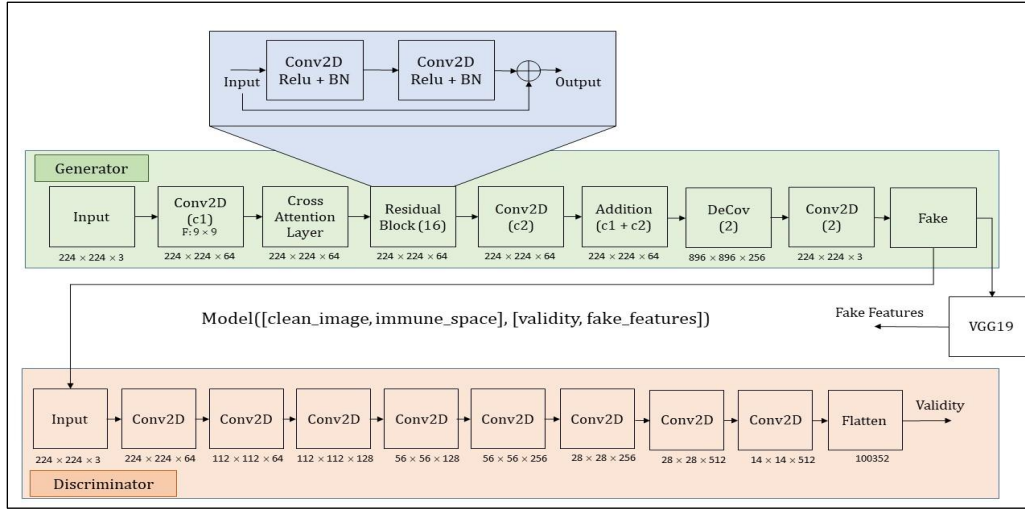


Figure 5: Visual representation of SARGAN model.

Source: Authors, (2026).

SARGAN puts together clear pictures from the immune-projected outputs made by the LCPCN model. The images that come out of this process have much less noise and very little remaining adversarial influence, which makes them clearer and more detailed. The reconstructed images using SARGAN are shown in following Figure 6



Figure 6: Original image, attacked image, immune space image and reconstructed image.

Source: Authors, (2026).

To increase the accuracy of face classification, the study suggests modifying MobileNetV1. The architecture incorporates two significant improvements. To achieve more efficient optimization, a hybrid loss function is first introduced. In particular, the pseudocode describes the combination of cross-triplet loss and categorical cross-entropy loss. This composite goal improves discriminative feature learning by simultaneously increasing intra-class compactness and inter-class separability. Second, an Osprey-based exploitation phase is added to the standard Fire Hawk Optimization (FHO) algorithm to improve weight optimization. By strengthening the optimizer's local search capability, this change allows for more accurate model parameter fine-tuning and better convergence behavior. Fire Hawk Optimization (FHO) is a metaheuristic algorithm based on how hawks hunt in the forest. The algorithm mimics how fire hawks hunt by starting and spreading fires to flush out prey and then catching them. A population of candidate solutions (X) is first made. Each solution is shown as a position vector that represents either a fire hawk or prey in the search space. These position vectors together define the starting point for the optimization process. A random initialization process is utilized to identify the initial positions of these vectors in the search space. Each d -dimensional vector X_i represent i -th weight vector to be optimized. Population is initialized with N such vectors. $x_i^j(0)$ is initialized as follow:

$$x_i^j(0) = x_{i,min}^j + rand(x_{i,max}^j - x_{i,min}^j), \quad \begin{cases} i = 1, 2, \dots, N \\ j = 1, 2, \dots, d \end{cases}$$

$x_{i,min}^j$ and $x_{i,max}^j$ are the minimum and maximum bounds of the j -th decision variable for the i -th solution candidate; and $rand$ is a uniformly distributed random number in the range of $[0,1]$. To find out where they are in the search space, all of the possible solutions are tested using the objective function that was set up for the optimization problem. Based on this assessment, candidates with better objective values are called Fire Hawks (FH), and the rest are called prey (PR). The chosen Fire Hawks then start the search process by spreading "fires" around the prey in the search space. This helps the search find promising areas and makes hunting (optimization) easier. The primary fire source is the global best solution found at each iteration. This high-level solution controls how search dynamics spread throughout the solution space, affecting both the exploration and exploitation phases of the algorithm. The position of l -th fire hawk is update as follow:

$$FH_l^{new} = FH_l + (r_1 \times GB - r_2 \times FH_{Near}), l = 1, 2, \dots, n$$

GB is the global best (GB) solution in the search space considered as the main fire. FH_{Near} is one of the other Fire Hawks in the search space; r_1 and r_2 are uniformly distributed random numbers in the range of (0, 1) for determining the movements of Fire Hawks toward the main fire and the other Fire Hawks' territories. The new position of vector of the q -th prey (PR_q) surrounded by the l -th Fire Hawk FH_l is computed as,

$$PR_q^{new} = PR_q + (r_3 \times FH_l - r_4 \times SP_l), \begin{cases} l = 1, 2, \dots, n \\ q = 1, 2, \dots, r \end{cases}$$

SP_l is a safe place under the l -th Fire Hawk territory. r_3 and r_4 are uniformly distributed random numbers in the range of (0, 1) for determining the movements of prey toward the Fire Hawks and the safe place. Furthermore, prey exhibit adaptive movement behaviors within the search space. They may migrate toward the territories of other Fire Hawks, inadvertently approaching ambush zones, or attempt to escape toward relatively safer regions outside the dominant Fire Hawk's territory. These dynamic behaviors are incorporated into the position update mechanism of the algorithm. The movement and adaptive response of prey are mathematically modeled in the position updating process using the following equation.

$$PR_q^{new} = PR_q + (r_5 \times FH_{Alter} - r_6 \times SP), \begin{cases} l = 1, 2, \dots, n \\ q = 1, 2, \dots, r \end{cases}$$

Where PR_q^{new} is the new position vector of the q -th prey (PR_q) surrounded by the l -th fire hawk FH_l . FH_{Alter} is one of the other fire hawks in the search space; The algorithm might reach a local optimum too soon during the optimization process. To keep Fire Hawk Optimization (FHO) from getting stuck in these bad areas, a weight perturbation strategy based on how ospreys hunt is used. This mechanism simulates the osprey's dynamic strategy for diving and moving while catching fish, which improves the algorithm's ability to explore. The optimizer can get out of local minima and look for other promising areas of the solution space by adding controlled perturbations to the search process. The following equation is used to figure out the new position of the prey during this osprey-inspired exploitation phase.

$$PR^{op} = PR^{new} + \frac{lb + r \cdot (ub - lb)}{t}, t = 1, 2, \dots, T$$

Where PR^{new} is the position of prey (optimal weights) determined by FHO. Variable lb and ub represents lower and upper bound to avoid weights drifting too much on either side, T is number of iterations, and r is random number. Once the new position is computer using above formula, the decision is made weather to keep that position or not. If update position is within the specified bounds, then solution is accepted, otherwise it is clipped.

$$R^{op} = \begin{cases} PR^{op}, & lb \leq PR^{op} \leq ub \\ lb, & PR^{op} < lb \\ ub, & PR^{op} > ub \end{cases}$$

If the fitness value of new position is better than the old position, then accept the newly computed position (PR^{op}), otherwise reject it and continue with old position (PR^{New}).

$$PR = \begin{cases} PR^{op}, & F^{op} < F^{new} \\ PR^{new}, & Otherwise \end{cases}$$

Where, F^{op} is the fitness value of newly generated solution after applying osprey step. Whereas, F^{new} is the fitness value of the solution generated by FHO

IV. RESULTS AND DISCUSSIONS

IV.1 DATASET DESCRIPTION

For the purpose of the experimental evaluation, an internal dataset was created. A Nikon D3500 DSLR camera with a 24-megapixel APS-C CMOS sensor was used to take the facial photos. To guarantee uniform focal length and image quality across subjects, the standard 18–55mm f/3.5–5.6G VR kit lens was utilized during image capture. Table 2 provides a summary of the dataset. CLAHE is applied to preprocess the image. To incorporate the variability in data and to generalize the model, we applied augmentation, which creates 7 image out of each image. Then adversarial noise is added in each before passing it to OLD-ADM pipeline.

Table 2: Summary of dataset.

Number of subjects	Age Group	Male	Female	Modality	Resolution	Frame rate	Ethnicity
20	21-25	15	5	RGB	1080 x 1920	60fps	Filipino

Source: Authors, (2026).

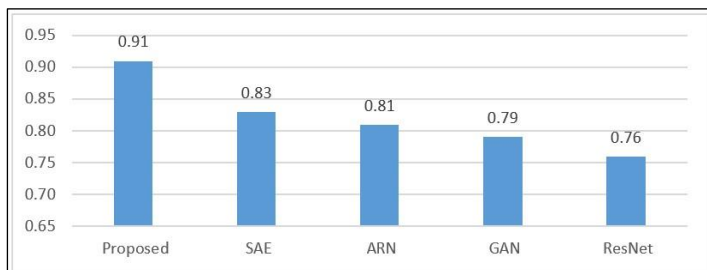
IV.2 RESULT ANALYSIS OF IMAGE RECONSTRUCTION

To evaluate the quality of reconstructed images generated by a GAN, quantitative image fidelity metrics are essential. The most widely used metrics are SSIM, PSNR, MSE, and MAE. MSE measures the average squared pixel-wise difference between the original image and reconstructed image. GAN outputs may look perceptually good but still have moderate MSE because MSE does not consider perceptual structure.

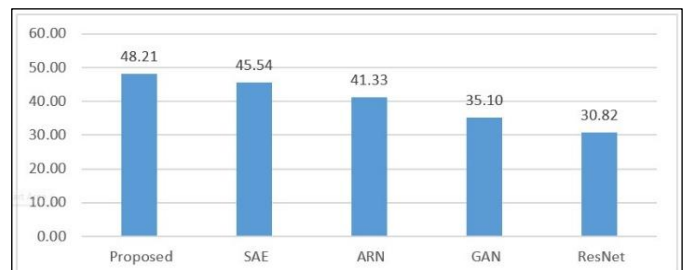
The Structural Similarity Index (SSIM) is the main performance metric used to evaluate the quality of reconstructed images following the application of the defense mechanism. SSIM is used to assess how well the model reconstructs facial images for accurate identification. SSIM offers a perceptually significant evaluation of reconstruction fidelity by simultaneously examining structural information, luminance, and pixel intensity correlations between the adversarially perturbed and original images. SSIM measures how well important facial features' structural integrity is maintained following reconstruction in the context of adversarial defense. The robustness of the defense mechanism can be methodically assessed by comparing SSIM values prior to and following the application of the suggested model.

Greater structural similarity between the reconstructed and ground-truth images is indicated by a higher SSIM value, indicating that the model effectively reduces adversarial perturbations while maintaining the discriminative facial features required for precise identification. A comparison of the SSIM values attained by the suggested method and current techniques is shown in Figure 7(a). Another important number used to test how well the proposed model works for face recognition is the Peak Signal-to-Noise Ratio (PSNR). It is very important for figuring out how well facial images are reconstructed after the process of mitigating adversarial perturbations. PSNR figures out how accurate the reconstructed image is by finding the logarithmic ratio between the highest pixel intensity and the mean squared reconstruction error. To test how strong the model is, the PSNR values of the reconstructed images are compared to those of the original, unperturbed images. Higher PSNR values mean less reconstruction error, which shows that the model can handle adversarial changes while still keeping important facial features that are needed for accurate recognition. So, a higher PSNR means that the system is better able to protect itself from attacks and restore the original image content more accurately. Figure 7(b) shows a comparison of the PSNR values between the proposed method and other methods. The graphical results clearly show that the proposed method gets better PSNR values than standard methods. Mean Squared Error (MSE) is a common way to measure how similar two images or video frames are. It finds the average of the squared differences between the pixel intensities of the ground-truth image and the reconstructed (or predicted) image.

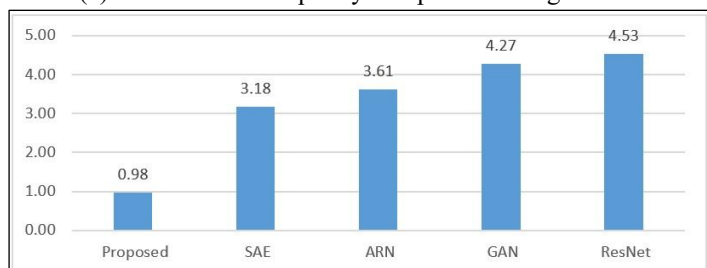
MSE is a basic measure of how well an image is reconstructed and how accurate a prediction is in image processing, machine learning, and statistical modeling. Lower MSE values mean that the reconstruction error is lower and the fidelity is higher. Figure 7(c) shows how the proposed framework and existing architectures compare in terms of MSE. The proposed model has the lowest MSE value because it uses the Stacked Attention Residual Generative Adversarial Network (SARGAN), which effectively hides adversarial perturbations while keeping facial features that help with discrimination. In contrast, traditional models like stacked autoencoder, attention-based ResNet, GAN, and ResNet have higher MSE values, which means they make more mistakes when trying to reconstruct something. Attack Success Rate (ASR) is an important measure of performance for testing how well the proposed method works in face recognition systems when there are threats. ASR measures how many adversarial attempts successfully trick the recognition model even when there is a defense mechanism in place. So, it shows how easy it is for an enemy to change the system. It is possible to systematically figure out how strong the proposed defense strategy is by looking at ASR values. A lower ASR means that the defense mechanism works well to stop adversarial changes and keep malicious inputs from messing up recognition accuracy. So, a low ASR shows that the face recognition system is more resilient and that the proposed framework can make security stronger against attacks. Figure 7(d) shows a comparison of the ASR values obtained by the proposed method and other methods, showing that the proposed framework is much more robust.



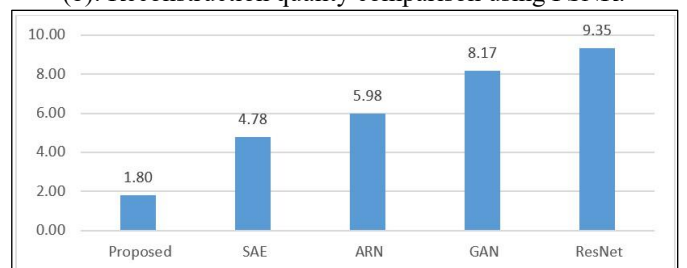
(a): Reconstruction quality comparison using SSIM.



(b): Reconstruction quality comparison using PSNR.



(c): Reconstruction quality comparison using MSE.



(d): Reconstruction quality comparison using ASR.

Figure 7: Comparison of quality of reconstructed images for various model with proposed approach.

Source: Authors, (2026).

IV.3 ABLATION STUDY

IV.3.1 Results with No Defense

First, using clean, attack-free images, a thorough assessment was carried out on both the suggested and baseline models. The suggested framework outperformed well-known architectures like MobileNetV1, VGG19, CNN, and LSTM with an accuracy of 99%. It is crucial to remember that LSTM is not the best option for static image classification tasks because it is naturally built for sequential data modeling. Although MobileNetV1 is lightweight and computationally efficient, the current study improves its performance by combining it with the Enhanced Fire Hawk Optimization (EFHO) algorithm for better parameter tuning. In almost every evaluation metric, the suggested Cross-Triplet MobileNetV1 classifier performs better overall. Figure 8 shows a comparison of the suggested model with four current classifiers.

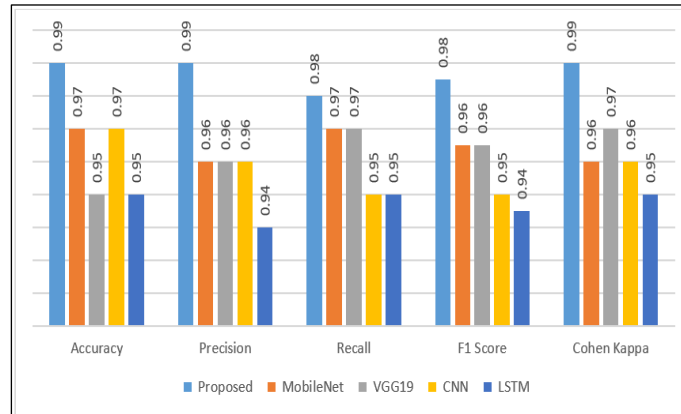


Figure 8: Comparison of proposed model (CTM) and existing models on clean images.
Source: Authors, (2026).

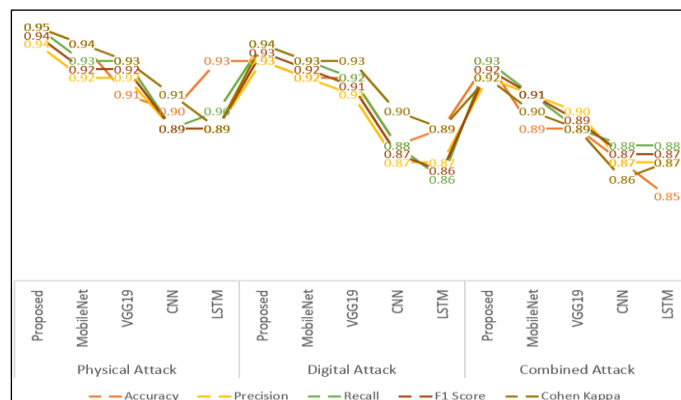


Figure 9: Comparison of Defense with LCPCN only.
Source: Authors, (2026).

V.3.2 Defense with LCPCN only

A thorough performance analysis of various classifiers tested on images processed by the Learnable Convolutional Principal Component Network (LCPCN) is provided in this subsection. By projecting input images into a learned Eigen space (immune space), LCPCN functions as an encoder–decoder architecture that reduces adversarial noise. The network is specifically trained to maintain discriminative identity features while suppressing adversarial perturbations. The suggested and baseline classifiers under each of the three types of adversarial attacks are graphically compared in Figure 9. The experimental results confirm the robustness and improved generalization capability of the proposed classifier in adversarial settings by showing that it consistently outperforms the current models across almost all evaluation metrics. Figure 10 shows a comparison of how well the standard Cross-Triplet MobileNetV1 (CTM) model works against physical, digital, and combined attacks. We compare images that were directly attacked with images that were processed through the LCPCN framework. Before classifying, LCPCN uses a denoising auto encoder to get rid of adversarial noise. PCA is an unsupervised, non-trainable linear projection technique, while LCPCN is a learnable architecture that is improved through backpropagation. Through supervised training, LCPCN learns a strong immune space that can handle changes well. This learned projection keeps important local spatial features while also capturing global variance structures in the data, making it more resistant to manipulation by bad actors.

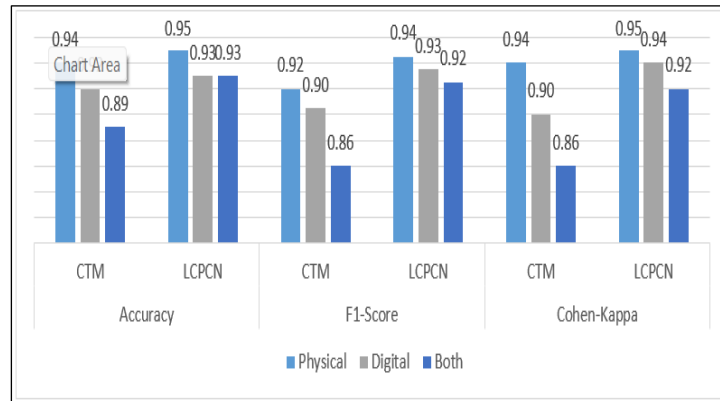


Figure 10: Results comparison of CTM and LCPCN.
Source: Authors, (2026).

IV.3.3 Defense with SARGAN only

As explained in the conceptual framework, the images that LCPCN sends into the immune space are then sent to the Stacked Attention Residual Generative Adversarial Network (SARGAN) for more adversarial purification. SARGAN adds an attention mechanism to the GAN architecture so that it can effectively capture both vertical and horizontal feature dependencies in different parts of the face. The attention module lets the network focus on important structural parts while getting rid of extra noise artifacts. The GAN is trained with pairs of clean and adversarially perturbed images. This lets the generator learn how to turn noisy inputs into their clean reconstructions. Because of this, the network makes purified images with much less adversarial noise while keeping important identity-related features. SARGAN uses multiple stacked attention layers to focus on the most informative parts of the facial image instead of treating the whole image the same way. This selective attention mechanism boosts important spatial features and hides unimportant ones, which makes feature representations more distinct and strong. The vanishing gradient problem is a common problem with deep architectures that can make it hard for backpropagation to work well in deeper layers. To fix this problem, SARGAN adds residual connections that make it easier for gradients to flow steadily between layers. These skip connections not only make training more stable, but they also help keep important identity-related information safe when reconstructing an image. Figure 11 shows a graph that compares the proposed classifier to other models when they are attacked physically, digitally, or in both ways. The findings indicate that the proposed framework consistently surpasses baseline classifiers on nearly all evaluation metrics.

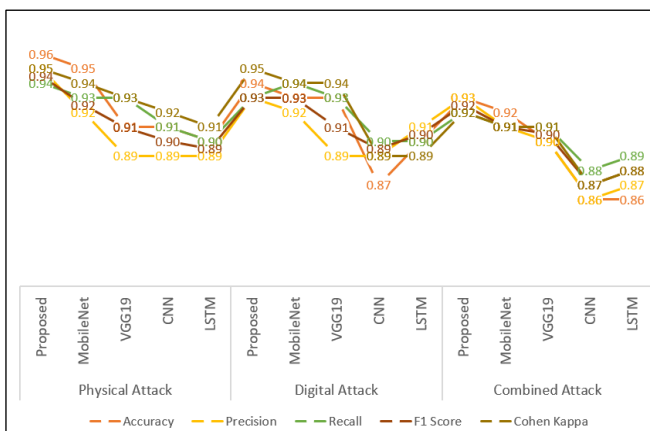


Figure 11: Graphical comparison of various models for SARGAN defense.
Source: Authors, (2026).

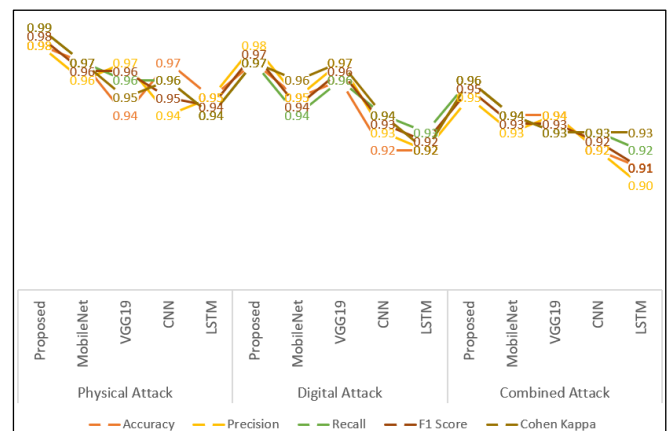


Figure 12: Graphical comparison of various models for ODL-ADM defense (right).
Source: Authors, (2026).

IV.3.4 Defense with Full ODL-ADM pipeline

The proposed framework's end-to-end pipeline is called ODL-ADM. The Learnable Convolutional Principal Component Network (LCPCN) processes images that have been attacked by an adversary. It does this by projecting the inputs into a learned immune space, which reduces noise. The immune-projected images are then used to teach the Stacked Attention Residual Generative Adversarial Network (SARGAN) how to make cleaner versions of the messed-up inputs. Finally, the Cross-Triplet MobileNetV1 (CTM) classifier gets the better images. At this point, the baseline MobileNetV1 architecture is improved by adding a hybrid loss function and the Enhanced Fire Hawk Optimization (EFHO) algorithm for better weight tuning. This makes the classification performance and robustness better. Figure 12 highlights the performance of complete ODL-ADM pipeline i.e. images processed through LCPCN and SARGAN both. Figure 13 shows a comparison of three configurations' key evaluation metrics: accuracy, F1-score, and Cohen's kappa. These configurations are images processed using LCPCN only, SARGAN only, and the full ODL-ADM pipeline. The results show that digital attacks cause more damage than physical attacks.

For example, when there are digital attacks, LCPCN's accuracy drops from 95% to 93%, and SARGAN's accuracy drops from 96% to 95%. The full ODL-ADM pipeline, on the other hand, shows better resilience by keeping accuracy almost the same in both physical and digital attack situations.

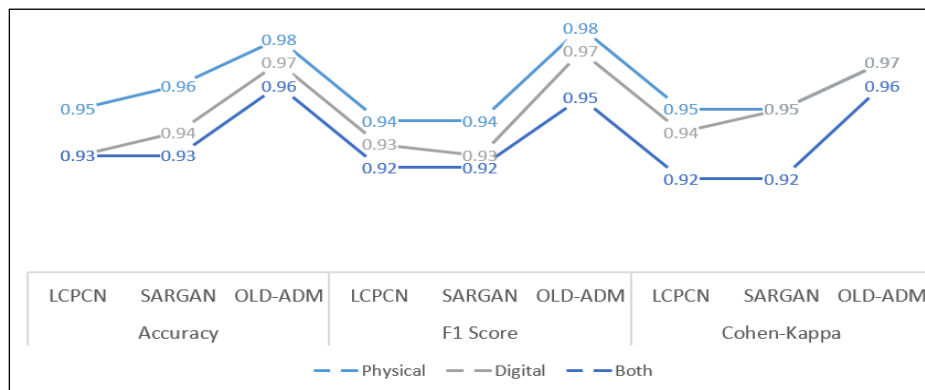


Figure 13. Comparison of LCPCN, SARGAN and ODL-ADM models.

Source: Authors, (2026).

The simultaneous occurrence of both physical and digital attacks significantly exacerbates the overall attack severity, a conclusion supported by the experimental analysis. When both LCPCN and SARGAN are attacked at the same time, they both get 93% accuracy. The full ODL-ADM pipeline, on the other hand, has an accuracy of 96%, which is 3% better than LCPCN alone and 3% better than SARGAN alone. The integrated, multi-stage defense strategy is clearly better based on these results. There is a consistent trend in performance across other evaluation metrics as well. For example, LCPCN gets an F1-score of 0.91 under combined attacks, SARGAN gets 0.92, and the new ODL-ADM pipeline raises this score to 0.95, which means that the balance between precision and recall has gotten better. The Cohen's kappa coefficient shows a similar pattern: both LCPCN and SARGAN get a score of 0.92, but ODL-ADM does much better with a score of 0.96. This shows that there is stronger agreement than chance and that classification is more reliable.

IV.3.5 Comparison across types of images

Three types of images were used in the experiments: original (clean), adversarially attacked, and reconstructed images. Figure 14 shows the results of five different evaluation metrics side by side. The addition of adversarial perturbations significantly reduces recognition performance, as seen. But after reconstruction, the model's performance is still very similar to that of the original clean images. This shows how strong and recoverable the proposed algorithm is. The graph shows that the proposed model gets higher metric values when there are no attacks. Even though performance goes down when adversarial perturbations are added, the reconstructed images greatly improve classification effectiveness, proving the resilience of the framework.

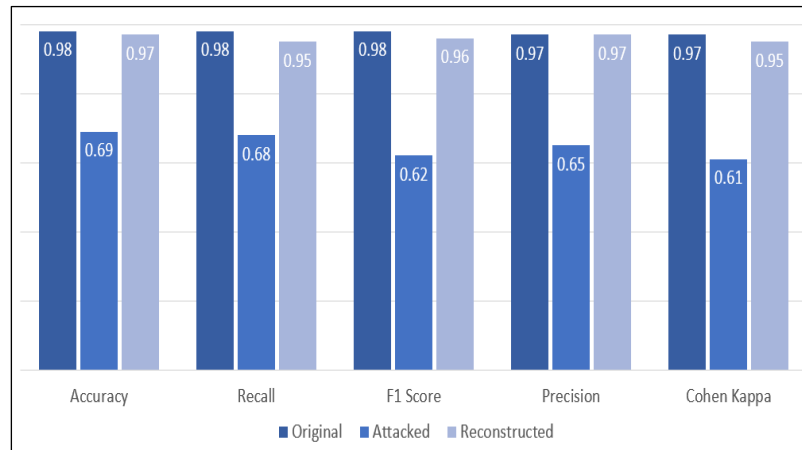


Figure 14: Analysis of image reconstruction with and without attack.

Source: Authors, (2026).

V. CONCLUSIONS

The research presents an innovative framework for adversarial mitigation and resilient face verification. We propose a Learnable Convolutional Principal Component Network (LCPCN) that can project images that have been changed in a way that makes them look different into a learned immune space. This reduces the effect of noise and keeps important features. After that, a Stacked Attention Residual Generative Adversarial Network (SARGAN) is trained with both immune-projected and clean images to make clean versions of adversarial inputs. This step makes the images even better by getting rid of any remaining noise while keeping the unique features of each person. For final face verification, the baseline MobileNetV1 architecture is improved by adding the Enhanced Fire Hawk Optimization (EFHO) algorithm, which makes tuning parameters and classifying better.

Experimental results show that the proposed integrated model is better than current methods on a number of evaluation metrics. The four experimental scenarios presented in Section 4.5 unequivocally illustrate the enhanced efficacy of the proposed ODL-ADM framework. The model gets 98%, 98%, and 96% correct when attacked physically, digitally, and in combination. This is much better than both LCPCN and SARGAN. ODL-ADM is 6% more accurate than LCPCN and 3% more accurate than SARGAN. In terms of F1-score, ODL-ADM is 4% better than LCPCN and 3% better than SARGAN. This means that it is better at balancing precision and recall when things go wrong. Also, the proposed framework has a Cohen's kappa score that is 4% higher than both LCPCN and SARGAN. This means that the classification is more reliable and there is more agreement than chance. Overall, these results show that the proposed method is strong and works well against many types of adversarial attacks. Results shows that the proposed model is not only good at recognizing the face images but also very strong in reconstructing images from the attacked images. In future work, the suggested framework may be enhanced to create adaptive defense mechanisms proficient in managing a variety of adversarial attacks, rather than being tailored to particular attack types. Adding other security measures, like encryption methods or secure multi-party communication protocols, could make face recognition systems even more reliable when they are used in the real world. Also, making strong models that need fewer training samples could save a lot of time and money on training while still giving good generalization performance. For example, Jinal et al. [43] came up with a deep learning-based way to lessen the effects of facial masks on recognition tasks. A comparable conceptual framework may be modified to mitigate adversarial perturbations and improve the robustness of face recognition systems. Additionally, the thorough survey conducted by Payal et al. [44], which examines the challenges and progress in face recognition, offers significant insights and can establish a robust basis for determining future research trajectories in this field.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Sheilla Ann Pacheco and Mahesh Goyani.

Methodology: Sheilla Ann Pacheco, Mahesh Goyani and Jayzel P. Bangoy and Born Christian A. Isip.

Investigation: Sheilla Ann Pacheco and Mahesh Goyani.

Discussion of results: Sheilla Ann Pacheco, Mahesh Goyani, Jayzel P. Bangoy and Born Christian A. Isip.

Writing – Original Draft: Sheilla Ann Pacheco, Mahesh Goyani, Jayzel P. Bangoy.

Writing – Review and Editing: Sheilla Ann Pacheco, Mahesh Goyani, Jayzel P. Bangoy and Born Christian A. Isip.

Resources: Mahesh Goyani, Jayzel P. Bangoy.

Supervision: Mahesh Goyani and Born Christian A. Isip.

Approval of the final text: Sheilla Ann Pacheco.

VII. REFERENCES

- [1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Computer Vision – ECCV 2020*, Cham: Springer International Publishing, 2020, pp. 213–229. doi: 10.1007/978-3-030-58452-8_13.
- [2] A. Gulati et al., "Conformer: Convolution-augmented Transformer for Speech Recognition," May 16, 2020, arXiv: arXiv:2005.08100. doi: 10.48550/arXiv.2005.08100.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, arXiv: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.
- [4] C. Badue et al., "Self-driving cars: A survey," *Expert Systems with Applications*, vol. 165, p. 113816, Mar. 2021, doi: 10.1016/j.eswa.2020.113816.
- [5] A. Polydoros and N. Lazaros, "Survey of Model-Based Reinforcement Learning: Applications on Robotics," vol. 86, no. 2, pp. 153–173.
- [6] S. Bhattacharya et al., "Deep learning and medical image processing for coronavirus (COVID-19) pandemic: A survey," *Sustainable Cities and Society*, vol. 65, p. 102589, Feb. 2021, doi: 10.1016/j.scs.2020.102589.
- [7] D. Wang, W. Yao, T. Jiang, G. Tang, and X. Chen, "A Survey on Physical Adversarial Attack in Computer Vision," Sep. 18, 2023, arXiv: arXiv:2209.14262. doi: 10.48550/arXiv.2209.14262.
- [8] C. Szegedy et al., "Intriguing properties of neural networks," Feb. 19, 2014, arXiv: arXiv:1312.6199. doi: 10.48550/arXiv.1312.6199.
- [9] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," Mar. 22, 2017, arXiv: arXiv:1608.04644. doi: 10.48550/arXiv.1608.04644.
- [10] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," Feb. 11, 2017, arXiv: arXiv:1607.02533. doi: 10.48550/arXiv.1607.02533.
- [11] Y. Dong et al., "Boosting Adversarial Attacks with Momentum," Mar. 22, 2018, arXiv: arXiv:1710.06081. doi: 10.48550/arXiv.1710.06081.
- [12] C. Li, H. Wang, J. Zhang, W. Yao, and T. Jiang, "An Approximated Gradient Sign Method Using Differential Evolution for Black-Box Adversarial Attack," *IEEE Transactions on Evolutionary Computation*, vol. 26, no. 5, pp. 976–990, Oct. 2022, doi: 10.1109/TEVC.2022.3151373.
- [13] R. Waelen, "The struggle for recognition in the age of facial recognition technology," *AI and Ethics*, vol. 3, no. 1, pp. 215–222, 2023.
- [14] M. Rusia and S. Dushyant, "A comprehensive survey on techniques to handle face identity threats: challenges and opportunities," *Multimedia Tools and Applications*, vol. 82, no. 2, pp. 1669–1748, 2023.
- [15] S. Ahmed, M. Frikha, T. D. H. Hussein, and J. Rahebi, "[Retracted] Optimum Feature Selection with Particle Swarm Optimization to Face Recognition System Using Gabor Wavelet Transform and Deep Learning," *BioMed Research International*, vol. 2021, no. 1, p. 6621540, 2021, doi: 10.1155/2021/6621540.
- [16] A. B. Shetty, Bhoomika, Deeksha, J. Rebeiro, and Ramyashree, "Facial recognition using Haar cascade and LBP classifiers," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 330–335, Nov. 2021, doi: 10.1016/j.gltp.2021.08.044.

- [17] S. Allagwail, O. S. Gedik, and J. Rahebi, "Face Recognition with Symmetrical Face Training Samples Based on Local Binary Patterns and the Gabor Filter," *Symmetry*, vol. 11, no. 2, Jan. 2019, doi: 10.3390/sym11020157.
- [18] S. Ejaz, I. Rabiul, S. Md, and S. Ananya, "Implementation of Principal Component Analysis on Masked and Non-masked Face Recognition," presented at the 1st international conference on advances in science, engineering and robotics technology, IEEE, 2019, pp. 1–5.
- [19] M. Hassan, H. Haval, E. Adel, and M. Ramadhan, "Face Recognition Based on Gabor Feature Extraction Followed by FastICA and LDA," *Computers, Materials and Continua*, vol. 68, no. 2, pp. 1637–1659, Mar. 2021, doi: 10.32604/cmc.2021.016467.
- [20] A. Ouyang, Y. Liu, S. Pei, X. Peng, M. He, and Q. Wang, "A hybrid improved kernel LDA and PNN algorithm for efficient face recognition," *Neurocomputing*, vol. 393, pp. 214–222, Jun. 2020, doi: 10.1016/j.neucom.2019.01.117.
- [21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv.org*, Dec. 2014.
- [22] A. Rozsa, M. Günther, E. M. Rudd, and T. E. Boulton, "Are facial attributes adversarially robust?," in 2016 23rd International Conference on Pattern Recognition (ICPR), Dec. 2016, pp. 3121–3127. doi: 10.1109/ICPR.2016.7900114.
- [23] V. Mirjalili and A. Ross, "Soft biometric privacy: Retaining biometric utility of face images while perturbing gender," in 2017 IEEE International Joint Conference on Biometrics (IJCB), Oct. 2017, pp. 564–573. doi: 10.1109/BTAS.2017.8272743.
- [24] S. Chhabra, R. Singh, M. Vatsa, and G. Gupta, "Anonymizing k-Facial Attributes via Adversarial Perturbations," Sep. 28, 2018, *arXiv*: arXiv:1805.09380. doi: 10.48550/arXiv.1805.09380.
- [25] X. Ying, R. Kiran, R. Raghavendra, and B. Christoph, "Adversarial Attacks on Face Recognition Systems," *Advances in Computer Vision and Pattern Recognition*, pp. 139–161.
- [26] Q. Wang et al., "Detecting Adversarial Faces Using Only Real Face Self-Perturbations," May 04, 2023, *arXiv*: arXiv:2304.11359. doi: 10.48550/arXiv.2304.11359.
- [27] G. Wang, G. B. Giannakis, and J. Chen, "Learning ReLU Networks on Linearly Separable Data: Algorithm, Optimality, and Generalization," *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2357–2370, May 2019, doi: 10.1109/TSP.2019.2904921.
- [28] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing Robust Adversarial Examples," Jun. 07, 2018, *arXiv*: arXiv:1707.07397. doi: 10.48550/arXiv.1707.07397.
- [29] A. Thamizharasi and S. Jayasudha, "Illumination pre-processing method for face recognition using 2D DWT and CLAHE," vol. 7, no. 4, pp. 380–390, 2017.
- [30] A. Rehmer and A. Kroll, "On the vanishing and exploding gradient problem in Gated Recurrent Units," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 1243–1248, Jan. 2020, doi: 10.1016/j.ifacol.2020.12.1342.
- [31] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, in CCS '16. New York, NY, USA, Oct. 2016, pp. 1528–1540. doi: 10.1145/2976749.2978392.
- [32] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022, doi: 10.1109/TPAMI.2021.3087709.
- [33] Z. Zhou, D. Tang, X. Wang, W. Han, X. Liu, and K. Zhang, "Invisible Mask: Practical Attacks on Face Recognition with Infrared," Mar. 13, 2018, *arXiv*: arXiv:1803.04683. doi: 10.48550/arXiv.1803.04683.
- [34] D.-L. Nguyen, S. S. Arora, Y. Wu, and H. Yang, "Adversarial Light Projection Attacks on Face Recognition Systems: A Feasibility Study," Apr. 17, 2020, *arXiv*: arXiv:2003.11145. doi: 10.48550/arXiv.2003.11145.
- [35] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach," May 28, 2018, *arXiv*: arXiv:1703.08448. doi: 10.48550/arXiv.1703.08448.
- [36] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger, "Simple Black-box Adversarial Attacks," Aug. 15, 2019, *arXiv*: arXiv:1905.07121. doi: 10.48550/arXiv.1905.07121.
- [37] G. Sun, H. Hu, Y. Su, Q. Liu, and X. Lu, "ApaNet: adversarial perturbations alleviation network for face verification," *Multimed Tools Appl*, vol. 82, no. 5, pp. 7443–7461, Feb. 2023, doi: 10.1007/s11042-022-13641-1.
- [38] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," Nov. 24, 2015, *arXiv*: arXiv:1511.07528. doi: 10.48550/arXiv.1511.07528.
- [39] J. Su, D. V. Vargas, and S. Kouichi, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Computat.*, vol. 23, no. 5, pp. 828–841, Oct. 2019, doi: 10.1109/TEVC.2019.2890858.
- [40] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks," Jul. 04, 2016, *arXiv*: arXiv:1511.04599. doi: 10.48550/arXiv.1511.04599.
- [41] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models," *arXiv.org*, Aug. 2017, doi: 10.1145/3128572.3140448.
- [42] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," *arXiv.org*.
- [43] J. Bhagat and M. Goyani, "Masked Face Recognition by Reconstructing Face Using Image Inpainting," *India: Recent Trends in Computational Intelligence and Its Application*, 2022, pp. 26–34.
- [44] P. Parekh and M. Goyani, "A comprehensive study on face recognition: methods and challenges," *The Imaging Science Journal*, vol. 68, no. 2, pp. 114–127, 2019.