Explaining to Protect: XAI-based Enhancement of CNN Robustness in Biometric Face Recognition

Irfan Darmawan [1], Alam Rahmatulloh [2], Rohmat Gunawan [2], Erna Haerani [3,*], and R. Wahjoe Witjaksono [1]

[1] Department of Information Systems, Faculty of Industrial Engineering, Telkom University, Indonesia
[2] Department of Informatics, Faculty of Engineering, Siliwangi University, Indonesia
[3] Department of Information Systems, Faculty of Engineering, Siliwangi University, Indonesia
Email: irfandarmawan@telkomuniversity.ac.id (I.D.); alam@unsil.ac.id (A.R.); rohmatgunawan@unsil.ac.id (R.G.); erna@unsil.ac.id (E.H.); wahyuwicaksono@telkomuniversity.ac.id (R.W.W.)
*Corresponding author

***Abstract*—This study investigates the role of explainability in improving adversarial robustness of Convolutional Neural Network (CNN)-based face recognition systems. While prior works treat robustness and interpretability as separate objectives, we conduct a systematic analysis of how explanation-driven preprocessing can influence adversarial resilience. Specifically, we combine adversarial training with Local Interpretable Model Agnostic Explanation (LIME)-guided feature perturbation to identify and suppress vulnerable regions. Unlike prior approaches that primarily use Explainable Artificial Intelligence (XAI) for post-hoc interpretation, this work quantitatively evaluates the relationship between explanation consistency and robustness under adversarial perturbations. Experimental results on CASIA-WebFace and Labeled Faces in the Wild (LFW) show that the proposed framework improves robustness under low-strength attacks ($\varepsilon = 0.02$), increasing Area Under Curve (AUC) from 62.21% to 75.76%. However, performance degrades under stronger perturbations, indicating limitations of explanation-guided preprocessing. These findings provide empirical evidence that explainability can contribute to robustness, but its effectiveness is highly dependent on attack strength and preprocessing design. This work highlights both the potential and limitations of integrating XAI into adversarial defence.**

***Keywords*—adversarial attack, adversarial training, face recognition, explainable artificial intelligence, Local Interpretable Model Agnostic Explanation (LIME)**

I. Introduction

Artificial Intelligence (AI)-based facial recognition systems have now become an essential part of security, including airport surveillance, identifying criminals or passengers on flights, facial tracking, and access authentication systems [1]. Typically, facial recognition frameworks employ Convolutional Neural Network (CNN)-based methods for feature extraction and classification [2]. But CNNs are vulnerable to noise, which may lead to misclassification [3]. A new type of threat that can manipulate models is the adversarial attacks [4]. Minimal adversarial perturbations can cause models to misclassify faces, potentially compromising security in real-world applications of biometric systems. Adversarial attacks manipulate the feature space representing concepts. The perturbation is then factored into latent factors, a meaningful subset of them being intended to trick the model [5]. This issue highlights the need to develop stronger defense mechanisms to enhance the model's robustness against adversarial attacks [6].

Several methods have been suggested to enhance robustness to adversarial attacks. Regularization and adversarial training are powerful methodologies to enhance model robustness, but also introduce constraints on generality and interpretability [7]. The combination of Logistic Regression (LR), Light Gradient Boosting Machine (LGBM), and CNN as LR-LGBM-CNN, a hybrid CNN, has achieved significant performance improvement in biometric authentication, adversarial attack scenarios and interpretability were not considered in this work [8]. Other studies reveal that attention-based methods improve the transferability of attacks and defenses, for example, feature attention [9] and Facial Attention-Aware Adversarial Training (FAAT) [10]. But interpretability is not a focus; there was no indication from previous work as to why the model took the decisions it did. At the same time, transparency is important to understand the model's behavior in non-stationary environments [11].

Explainable AI (XAI) offers a different view, which can be useful to analyze and enhance the robustness of CNN models under adversarial attacks [12, 13]. Explainable Artificial Intelligence (XAI) makes it possible to analyze the CNN decisions and to know which feature or features may be manipulated in biometric applications.

Manuscript received January 9, 2026; revised March 25, 2026; accepted May 11, 2026; published August 19, 2026.

Dwivedi *et al*. [14] proposed XAI based on the combination of attribute (saliency maps) and class-specific explanations derived using Gradient-weighted Class Activation Mapping (Grad-CAM) for morphing attack detection. In face spoofing detection, Khairnar *et al*. [15] applied Local Interpretable Model Agnostic Explanation (LIME) for enhancing model transparency. However, both methods are concerned with making the models more interpretable rather than more robust to adversarial attacks.

Previous studies have shown that robustness and interpretability are still studied separately. To overcome these limitations, this study proposes an XAI-based defense approach that combines robustness and interpretability within a single framework, as outlined in a comprehensive survey by Mehra [16]. By generating adversarial examples, adversarial training not only enhances the model's robustness against perturbations, but also empirically improve the model's interpretability [17]. XAI methods, such as LIME [15], are further needed to develop XAI-based defense mechanisms. The results of LIME analysis are used to enhance preprocessing integrated with adversarial training. This strategy is expected to provide complementary benefits under certain conditions, although its effectiveness under stronger adversarial settings remains uncertain.

This study has four main contributions. First, we provide a systematic empirical analysis of the relationship between interpretability (via LIME) and adversarial robustness in CNN-based face recognition. Second, we propose an explanation-guided preprocessing strategy, where vulnerable regions identified by LIME are explicitly suppressed during training. Third, we conduct multi-scenario evaluation across varying attack strengths, revealing that the effectiveness of XAI-based defense is highly dependent on perturbation magnitude. Fourth, we provide a critical assessment of the limitations of explanation-driven defense, highlighting failure cases under moderate and strong adversarial attacks.

Despite increasing interest in both adversarial robustness and XAI, the relationship between these two domains remains insufficiently understood. Most existing works employ XAI solely as a diagnostic tool, without systematically validating whether interpretability can actively contribute to robustness. This creates a critical gap: it is still unclear whether explanation-driven interventions can provide measurable and consistent improvements under adversarial settings.

Furthermore, recent studies have shown that interpretability methods themselves can be unreliable or even manipulated under adversarial conditions, raising questions about their practical utility in security-critical systems. Therefore, a more rigorous investigation is needed to assess not only whether XAI can enhance robustness, but also under what conditions such integration is effective or fails.

The remainder of this paper is organized as follows. Section II reviews various related studies on CNN-based approaches to adversarial defense and XAI methods used in biometric systems. Section III describes the proposed method, including the model architecture, adversarial training mechanism, and XAI method integration. Section IV describes the experimental setup, evaluation metrics, and test results for the base model and robust model. Section V provides research conclusions and directions for further research.

II. Literature Review

This section basically supports the background section by providing evidence for the proposed hypothesis. This section should be more comprehensive and thoroughly describe all the studies that you have mentioned in the background section. It should also elaborate on all studies that form evidence for the present study and discuss the current trends.

Facial recognition has received considerable attention from researchers compared to other biometric features because it is non-invasive [1]. CNNs have demonstrated high performance in biometric applications, but adversarial attacks can cause a 50% drop in accuracy [18]. To address this threat, various defense methods have been proposed, ranging from modifying the training process and changing the network to supplementing the main model with an external network [19]. Approaches that are widely used as adversarial defense techniques include adversarial training, regularization, or architecture modification. However, this approach functions as a black-box defense that has limitations in representing facial features vulnerable to adversarial attacks.

The proposed method explicitly provides interpretability using LIME. LIME serve as a post-hoc technique to explain the reasons behind the model's decisions after the training process [16]. This method utilizes the results of LIME analysis to update the processing, which is then combined with adversarial training. The proposed approach fills the gap in conventional defense by integrating robustness and interpretability.

Hybrid CNN models, specifically LR-LGBM-CNN, were used in a study on improving biometric authentication, achieving an accuracy rate of 87% [8]. This study emphasized improvements in accuracy, robustness, and usability in facial recognition systems, but did not investigate adversarial attack scenarios or interpretability aspects. On the other hand, Li *et al*. [20] introduces an adversarial resilience evaluation method for robust path-based Deep Neural Networks (DNN) using Information Bottleneck Theory and the Robust Neuron Coverage (RNC) metric to understand the behavior of attacked models, but it is not directly applied to biometric facial recognition.

Adversarial training techniques through adversarial example storage are used to improve model robustness in facial expression recognition [21]. Although it does not specifically discuss XAI or biometric simulation, the proposed method significantly improves model performance, achieving 98.81% accuracy on the CK+ dataset. Similarly, Jayapradha *et al*. [7] apply adversarial training, regularization, and data augmentation to enhance the robustness against data poisoning attacks. The suggested method is ensemble-based and employs

anomaly detection, but the interpretability issue is not considered. Some recent works study attention mechanisms to enhance attack and defense transferability, such as Feature Attention [9] and FAAT [10], but the focus is not on interpretability.

Several other works have also been dedicated to studying and applying XAI in the field of biometrics. Dwivedi *et al*. [14] employed XAI for morphing attack detection by combining Grad-CAM and saliency maps. Khairnar *et al*. [15] adopted LIME to increase the transparency of the face spoofing detection model. Additionally, both approaches focus on making the models more interpretable and do not address robustness against adversarial attacks. Mehra [16] presents a comprehensive treatment in which several attempts to combine robustness and interpretability are outlined. The work recommends a fusion of adversarial training with interpretability algorithms such as LIME or SHapley Additive exPlanations (SHAP). However, the study notes that there are barriers related to computational viability and trade-offs between robustness and accuracy.

XAI is a good solution when integrated to deal with adversarial attacks. However, the model has a vulnerability dimension: attacks targeting XAI-explanation-aware backdoors. This scenario produces correct predictions, but the XAI results are manipulated to display misleading features. These findings indicate that XAI alone does not guarantee system transparency and reliability, thus requiring a more comprehensive defense mechanism that maintains the integrity of the explanations generated [22].

Based on a review of previous studies, most studies discussing model resilience to adversarial attacks focus more on improving accuracy without linking it to interpretability. Several studies have also not integrated XAI into adversarial defense, such as XAI that is only used in face detection modifications [14] and XAI for face activity detection [15]. The limitations of unrepresentative datasets also pose difficulties in generalizing the results [8]. Previous studies have only focused on one aspect, such as improving accuracy, interpretability, or resistance to adversarial attacks, so there is no integrated framework yet. To fill this gap, this study proposes an explainability-driven adversarial defense strategy that combines adversarial training with interpretability analysis to identify vulnerable features. By testing on a larger and more diverse biometric dataset, namely Labeled Faces in the Wild (LFW), this study can produce a CNN model that is robust against adversarial attacks and highly interpretable.

III. Proposed Method

Based on the results of the analysis of the solution requirements for the research problems raised in this study, we provide a solution to these problems by developing an XAI-based defense. Fig. 1 shows the flow of the adversarial defense experiment we developed.

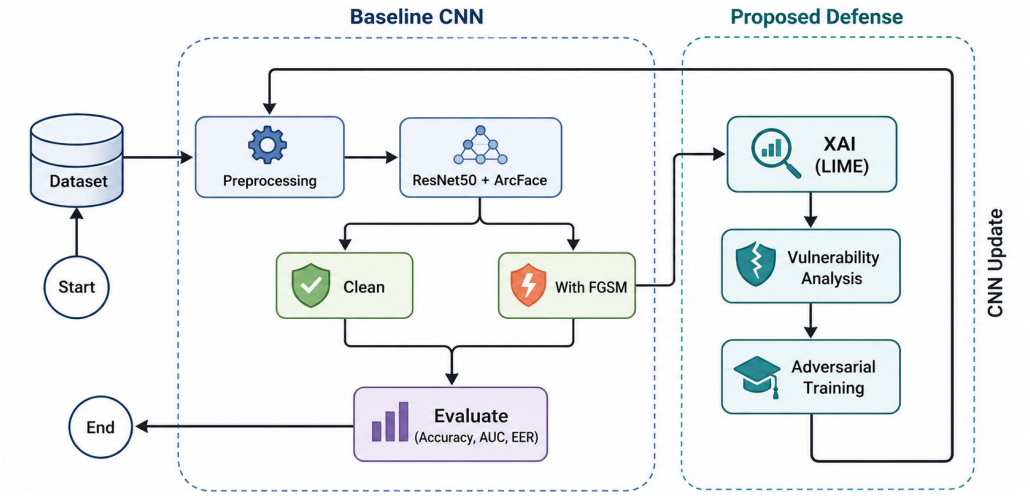

Fig. 1. Workflow of the proposed defense framework.

The research flow is divided into two main parts, namely baseline CNN and proposed defense, with four testing scenarios. In the first scenario, the baseline CNN is trained using the original dataset and evaluated under normal conditions (without adversarial attacks). In the second scenario, the baseline CNN is tested using Fast Gradient Sign Method (FGSM) adversarial attacks to measure the decline in performance due to attacks. The third scenario tests the model with the resulting defense method again under normal conditions. The fourth scenario tests the model with the defense method using attacks to evaluate the increase in resilience. The combination of adversarial training and XAI-based analysis enables the model to understand and adaptively maintain important features.

A. CNN Backbone Block

The facial recognition model architecture consists of ResNet-50 as the feature extraction backbone and Arcface as the classifier. Both are trained end-to-end using a loss function. ResNet-50 delivers superior performance in various fields, particularly in medical imaging [23–25] and facial recognition [26]. However, ResNet-50 requires high computational costs, so trade-offs between accuracy and computational costs need to be considered [27]. Meanwhile, ArcFace was chosen because it has high

generalization capabilities and is stable in various conditions. The additive angular margin loss approach in ArcFace strengthens the discrimination between facial identities and maintains the consistency of features from the same individual [28].

A complete illustration of the model architecture is shown in Fig. 2. In the initial stage, facial images are fed into the ResNet-50 backbone network, which produces 2048-dimensional embeddings. These embeddings are then fed into the ArcFace module. The output values from ArcFace are then passed to the SoftMax function, and the results are used to calculate the loss based on the ArcFace loss function.

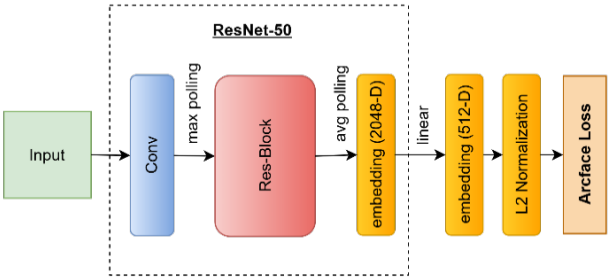

Fig. 2. Architecture of the CNN-based model using ResNet-50 backbone and ArcFace.

1) *ResNet-50*

Fine-tuning was performed on ResNet-50 by adjusting the final layer of the model for the face recognition task, through transfer learning with previously trained weights on ImageNet [29]. By default, ResNet produces 1000 class outputs on the ImageNet dataset. However, in this study, the Fully Connected (FC) layer was removed so that the model functions as a feature extractor that produces 2048-dimensional feature embeddings.

An image measuring 112×112 pixels will be processed on a convolutional network to extract basic features. Then, the image passes through four convolutional blocks (Residual Blocks), each of which deepens the spatial and semantic representation of the face. After passing through all residual blocks, the results are globally averaged through global average pooling to produce a 2048-D embedding. Next, a linear layer is added to map the backbone features to a 512-dimensional embedding space. In the final stage, the embedding is normalized using L2-normalization to ensure that the representation is on a hypersphere. An illustration of ResNet-50 is shown in Fig. 3.

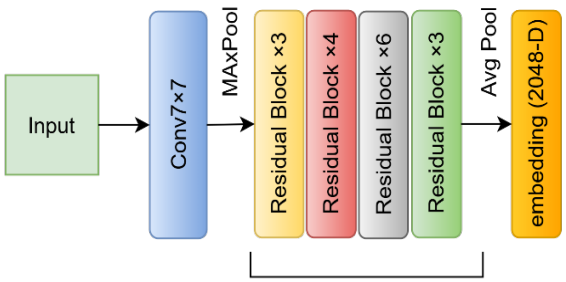

Fig. 3. Detailed structure of the ResNet-50 in the proposed model.

2) *ArcFace*

ArcFace is a loss function that optimizes the feature embedding resulting from the backbone [30]. ArcFace reformulates the SoftMax function by optimizing the geodesic distance between classes in the embedding space. Adding margin m directly to the angular space produces a constant linear margin across the entire angular space. As a result, facial identities become more geometrically separated. The main ArcFace formula is shown in Eq. (1).

$$L_{ArcFace} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{s\cos\left(\theta_{y_i}+m\right)}}{e^{s\cos\left(\theta_{y_i}+m\right)}+\sum_{\substack{j=1\\ j\neq y_i}}^{C} e^{s\cos\theta_j}}\right) \quad (1)$$

The calculation begins by calculating the similarity between facial feature x_i and all class centers W in the form of a cosine angle ($cos\theta_j$). A high value indicates that the feature is close to the class center, while a low value indicates that the feature is far from that class. A margin m is added to the correct class (y_i), so that this margin will increase the angular distance between identities. Then, multiply by the scale s to stabilize the SoftMax function. Next, SoftMax is used to calculate the probability that feature x_i belongs to the correct class. Finally, the negative logarithm of the probability is taken to produce the loss. The smaller the angle to the center of the correct class, the smaller L will be, which means the model is good.

In this implementation, the angular margin applied is $m = 0.5$, while the scale factor $s = 64$ is used to return the logit value to the appropriate range before the SoftMax function. Classification weights are normalized in hypersphere space, and the optimization process is performed through cross-entropy against the modified logit.

B. *Adversarial Perturbation*

The FGSM generates adversarial examples by adding a small disturbance ϵ in the direction of the loss gradient sign. From the input x and label y, the loss function $L(x, y)$ is calculated, then $\nabla_x J(x, y)$ is obtained as the gradient of the loss function with respect to the input x. After that, the gradient sign function is taken using the sign function $(\cdot)$, to determine the shift of the model prediction to the desired class. The disturbance is then multiplied by a factor $\in$ as the magnitude of the disturbance. The mathematical formula for the FGSM function is shown in Eq. (2) [31].

$$x' = x + \varepsilon \cdot sign\left(\nabla_x J(x, y)\right) \quad (2)$$

FGSM is the simplest version that modifies images with only one update, making it more computationally efficient [32]. If one FGSM step is not strong enough to change the prediction, FGSM can be extended to the I-FGSM (iterative) version [33]. The selection of ϵ is very important because it determines how much disturbance is applied, where too small a value is often insufficient to

change the prediction, while too large a value makes the disturbance clearly visible in the image [34].

C. LIME-Based Analysis

LIME is a model-agnostic and local XAI method, meaning it can be used on any model and explain predictions for a specific instance [35]. In the context of images, LIME first divides the image into several superpixels, then generates many new samples by randomly activating or deactivating superpixels (perturbation). Each sample is predicted by the original model, and the prediction results are used in a weighted regression (ridge regression) process to obtain interpretable coefficients that show the contribution of each superpixel to the model's decision [36]. This technique builds a sparse linear model around each prediction to describe how the model operates in the nearest local area [37]. The results of the LIME analysis are used to identify vulnerable features, so that adjustments can be made to the model training.

In this study, LIME was used as a qualitative analysis tool to observe shifts in attention when faced with adversarial attacks. The patterns obtained from comparing the visualization results on clean images and adversarial images became a stronger preprocessing strategy. FGSM interference caused the focus on features to be easily disrupted, thereby shifting the model's attention.

Sensitive features are defined as those that have the highest contribution to the model's prediction. Specifically, we consider a LIME weight threshold to determine the importance of each superpixel. In our experiments, a LIME weight threshold of 0.05 (in absolute value) was chosen to define sensitive features. Features with a weight exceeding this threshold are considered "sensitive" and are therefore masked during preprocessing. This threshold was selected based on preliminary experiments, where lower thresholds did not significantly improve robustness under adversarial perturbations. Higher thresholds led to the exclusion of critical features, such as key facial areas like eyes or nose, potentially undermining model performance.

An ablation study was conducted to evaluate the impact of different LIME parameters on preprocessing performance. The number of superpixels was varied between 50 and 300, and the kernel width was adjusted from 0.1 to 0.3. The results of this study indicate that the best trade-off between robustness and feature retention occurred with 150 superpixels and a kernel width of 0.2. This configuration yielded the highest robustness while still retaining the essential facial features needed for accurate recognition.

D. Adversarial Training

Adversarial training involves incorporating a set of adversarial examples into the training process to improve model robustness [38]. This technique has become one of the main methods for strengthening robustness in deep learning models. However, adversarial training faces a trade-off in the form of decreased accuracy on clean data and increased computational load due to double training [39]. In this study, adversarial training was carried out by adding 150,000 adversarial examples (FGSM) to the training process.

It is important to note that the proposed framework incurs additional computational complexity compared to a standard CNN. The ResNet-50 backbone, consisting of 50 residual blocks, requires substantial computation for feature extraction, and the integration of ArcFace introduces additional overhead due to angular margin calculations. Furthermore, adversarial training necessitates the generation and incorporation of 150,000 adversarial examples, effectively doubling the training workload. Enhanced preprocessing steps, guided by LIME such as masking, Gaussian blur, and random erasing, also contribute additional per-batch processing time. As a result, the overall training cost is significantly higher than baseline CNN training, requiring careful consideration of available computational resources.

E. Parameter Justification

The choice of adversarial perturbation levels (ε = 0.02, 0.03, 0.04) follows prior studies indicating that these values represent low to moderate perturbation regimes in normalized image space. Preliminary experiments showed that $\varepsilon < 0.02$ produces negligible degradation, while $\varepsilon > 0.04$ leads to perceptible distortions.

The number of adversarial samples (150,000) was determined based on a trade-off between computational feasibility and sufficient coverage of perturbation space. We ensured that adversarial samples represent at least 30% of the training data to avoid underfitting to clean samples.

Preprocessing parameters (color jitter, Gaussian blur, random erasing) were selected based on their ability to suppress high-frequency components, which are commonly exploited by adversarial perturbations.

IV. Result and Discussion

This study uses the CASIA-WebFace [40] dataset for training and LFW [41] for testing. The face recognition model that is used is the ResNet-50 backbone, which has been pre-trained on ImageNet. The ResNet-50 fully connected embedding layer generates a 512-dimensional feature vector that is L2 normalized. Arcface loss with margin m = 0.3 and scale s = 16 is used for identity classification during training.

Adversarial examples are generated using FGSM with a value of ϵ = 0.02, 0.03 and 0.04. The attack is performed on the CASIA-WebFace dataset, producing 150,000 images, which are afterward used in the training. These adversarial examples with the original dataset are together to make the dataset for adversarial training. Based on the information in Table I, both the base model and the robust model undergo data augmentation extensively during training. 80% of the dataset is used for training and 20% for validation. Validation processing also uses center-cropped images without augmentation to prevent unstable evaluation.

The model was trained for 20 epochs using the AdamW optimizer with an initial learning rate of 5×10^{-5} and weight decay of 1×10^{-3}. Clipping was applied with a maximum norm limit of 1.0 on all parameters. A cosine annealing

learning rate schedule was used to gradually decrease the learning rate throughout the epoch.

The model performance was evaluated using three main metrics, namely verification accuracy, Area Under Curve (AUC), and Equal Error Rate (EER) [42]. The test is conducted on the LFW dataset that contains 6000 pairs of images following the official LFW pairs list. The image pairs are separated with an equal number of genuine and impostor pairs. The verification score is a function of the cosine similarity between the two embedding pairs. The similarity measure was also employed to generate the Receiver Operating Characteristic (ROC) and the AUC. EER was produced at the point where FAR equals FRR. Verification accuracy was obtained by the best operating point via Younden's index. Evaluation of resistance to adversarial attacks is conducted following the same protocol, but on the adversarial-perturbed LFW dataset. Results on the clean and attacked LFW databases are reported separately.

TABLE I. BASE MODEL PERFORMANCE AT CLEAN AND ATTACKED DATASET

Dataset	AUC	EER	Accuracy	Opt. Threshold	EER Threshold
Clean	80.49	27.4	72.7	0.315	0.308
Attacked	62.21	41.5	58.63	0.396	0.401

We also present additional metrics commonly used in adversarial defense literature to offer a more comprehensive evaluation of the model's robustness. First, we calculate the Attack Success Rate (ASR) across various perturbation magnitudes (ε = 0.01, 0.02, 0.03, 0.04) to evaluate the defense's effectiveness under different levels of adversarial perturbation. Additionally, we assess the robust accuracy by testing the model's performance in both white-box and black-box attack scenarios, helping us understand how well the model generalizes to previously unseen adversarial attacks. Finally, we introduce the concept of adversarial robustness radius (Carlini-Wagner (CW) L2), measured using the CW L2 attack. This metric quantifies the minimal perturbation required to misclassify the model, providing valuable insight into its adversarial resilience. Together, these metrics offer a deeper and more nuanced understanding of the defense's effectiveness across a range of adversarial conditions.

Future experiments will include datasets such as CASIA-WebFace2 and VGGFace2 to evaluate the robustness of the proposed method under more realistic conditions.

The base model reported in Table I has an AUC 80.49%, an EER of 27.4%, and an accuracy 72.7% in the clean dataset. The results show good initial performance for the model without defense on a clean dataset. However, with the employment of the adversarial dataset its performance also decreases to 62.21% in AUC with an increasing EER of 41.5% and decreasing accuracy of 58.63%. This performance degrades considerably which suggests that the base model cannot hold up under attacks and is sensitive to perturbations. The shift in thresholds indicates instability, with the optimal threshold increasing from 0.315 to 0.396 and the EER threshold increasing from 0.308 to 0.401.

Fig. 4 shows the LIME visualization of the base model that takes the example of Akhmed Zakayev's identity, where anchor and probe are displayed with the same identity. In the normal dataset, the model predicts a match result with a similarity of 0.818. The confidence level for the model is quite high and the model performs verification correctly. In the attacked dataset (FGSM), the model still performs verification correctly, but the similarity score drops dramatically to 0.630. This certainly jeopardizes the verification results because it reduces the accuracy of the model.

Fig. 4. Detailed structure of the ResNet-50 in the proposed model.

To move beyond qualitative interpretation, we introduce an explanation consistency metric, defined as the overlap between salient regions in clean and adversarial inputs. This is measured using Intersection-over-Union (IoU) between top-k important superpixels. Results show that adversarial perturbations significantly reduce explanation consistency (average IoU drop of 25%), indicating instability in model attention. The proposed preprocessing partially restores this consistency under low ε values but fails to maintain stability under stronger attacks. The similarity experienced a notable decline, decreasing from 0.818 to 0.630, which represents an approximate reduction of 23%. In addition, the average IoU also showed a significant drop of around 25%, indicating a consistent decrease in overall performance across both metrics.

From the changes in LIME visualization, the model's attention shifts from the face to irrelevant features. The attacked image has changes that form the basis for the application of enhanced preprocessing. Increasing the intensity of color jitter and adding Gaussian blur aims to reduce the model's dependence on high-frequency local textures. This is in line with the assumption that adversarial perturbations affect high-frequency

components, and LIME visualization helps identify these vulnerabilities regionally. Random erasing is added to force the model to learn a more comprehensive representation so that the model still recognizes identity even when part of the face is covered. Table II summarizes the preprocessing differences between the baseline and robust models.

TABLE II. COMPARISON OF PREPROCESSING STRATEGIES

Model	Preprocessing Strategy
Base	Resized crop (112, 0.8–1.0), H-flip (0.5), Color jitter (0.2), Grayscale (0.1), Normalize (0.5/0.5)
Robust	Resized crop (112, 0.8–1.0), H-flip (0.5), Color jitter (0.4), Grayscale (0.1), Normalize (0.5/0.5), Gaussian blur (0.3), Random erasing (0.02–0.33)

We evaluated several scenarios to analyze changes in model performance with the application of adversarial training and enhanced preprocessing. Scenario Adv. Training (AT) used only adversarial training as a defense method by incorporating adversarial examples into training. Scenario Adv. Training + Enhanced (ATE) preprocessing performed enhanced preprocessing based on the results of LIME visualization analysis. To explore the impact of LIME parameters, we conducted an ablation study by varying the number of superpixels and the kernel width. The results revealed that the optimal performance was achieved with 150 superpixels and a kernel width of 0.2, as this combination produced the highest AUC and the lowest EER across both clean and attacked datasets. When the number of superpixels was reduced to 50 or increased to 300, model performance decreased, particularly in terms of robustness to adversarial attacks. Similarly, adjusting the kernel width to 0.1 or 0.3 also led to a decline in performance, highlighting that a balanced kernel width of 0.2 is most effective for preserving key features and improving model robustness. A summary of these findings can be found in Table III.

Each scenario was tested in a clean dataset and an attacked dataset with different epsilon values, resulting in a total of six scenarios in Table IV. Enhanced preprocessing provided partial improvement, primarily on clean data and low perturbation levels. Models ATE had a higher AUC in the range of 0.80–0.85 compared to models with AT alone, which only had a range of 0.73–0.84. AT03 has the lowest performance with an AUC of only 0.7355, an EER of 0.3223, and an accuracy of 0.6783. ATE02 achieved the highest AUC of 0.8457, while ATE04 provided the highest accuracy and lowest EER of 76.77 and 23.3 respectively. Enhanced preprocessing provides condition-dependent improvements, with performance gains observed mainly at $\varepsilon = 0.02$ in the ATE03 model, where AUC increases from 0.7355 to 0.8029, EER decreases from 0.3223 to 0.2757, and Accuracy increases from 0.6783 to 0.7272. ATE04 has almost the same performance as AT04 without significant improvement, with AUC 0.8443, EER 0.2330, and Accuracy 0.7677. Thus, on the clean dataset, enhanced preprocessing overall improves the performance of the ATE02, ATE03, and ATE04 models.

TABLE III. ABLATION STUDY RESULTS FOR LIME PARAMETERS

Jumlah Superpixel	Kernel Width	AUC (Clean Data)	AUC (Attacked Data)	EER (Clean Data)	EER (Attacked Data)
50	0.1	0.72	0.58	0.30	0.39
150	0.2	0.80	0.75	0.25	0.33
300	0.3	0.77	0.68	0.27	0.36

TABLE IV. EFFECT OF ENHANCED PREPROCESSING UNDER ADVERSARIAL ATTACKS AT VARIOUS EPSILONS

Dataset	Model	ε	AUC	EER	Accuracy	Optimal Threshold	EER Threshold
Clean Dataset	Adv. Training (AT02)	0.02	0.7705	0.2900	0.7128	0.267	0.281
	Adv. Training (AT03)	0.03	0.7355	0.3223	0.6783	0.302	0.296
	Adv. Training (AT04)	0.04	0.8366	0.2393	0.7637	0.203	0.175
	Adv. Training + Enhanced Prep (ATE02)	0.02	0.8457	0.2390	0.7650	0.260	0.236
	Adv. Training + Enhanced Prep (ATE03)	0.03	0.8029	0.2757	0.7272	0.281	0.300
	Adv. Training + Enhanced Prep (ATE04)	0.04	0.8443	0.2330	0.7677	0.313	0.312
Attacked Dataset	Adv. Training (AT02)	0.02	63.12	40.30	59.77	0.544	0.529
	Adv. Training (AT03)	0.03	73.55	32.23	67.83	0.302	0.296
	Adv. Training (AT04)	0.04	74.55	31.93	68.32	0.260	0.288
	Adv. Training + Enhanced Prep (ATE02)	0.02	75.76	31.40	69.02	0.351	0.313
	Adv. Training + Enhanced Prep (ATE03)	0.03	69.29	36.10	64.20	0.473	0.494
	Adv. Training + Enhanced Prep (ATE04)	0.04	73.61	33.47	67.03	0.471	0.448

The optimal threshold and EER threshold changed, due to the application of enhanced preprocessing. The optimal threshold decreased at epsilon 0.02 by 2.62% and at epsilon 0.03 by 6.95% but increased dramatically at epsilon 0.04 by 54.19%. This shows that enhanced preprocessing lowers the threshold at small epsilon values but increases it at large epsilon values. The EER threshold also decreased at epsilon 0.02 by 16.01%, increased slightly at epsilon 0.03 by 1.35%, and rose sharply at epsilon 0.04 by 78.29%. At high epsilon values, the model becomes more conservative and needs to increase the threshold to maintain performance.

In the attacked dataset, ATE provided more diverse results. ATE02 had the best performance with an AUC of 75.76, an EER of 31.40, and an accuracy of 69.02. ATE02 showed a significant improvement over AT02, consistent with the findings in the clean dataset. However, ATE03 showed a significant decline in performance, with AUC falling from 0.7355 to 69.29, EER rising from 32.23 to 36.10, and Accuracy falling from 67.83 to 64.20. These

findings indicate that enhanced preprocessing is less suitable for moderate attacks, especially epsilon 0.03. ATE04 experienced a slight decrease compared to AT04, where AUC 74.55 dropped to 73.61, EER 31.93 rose to 33.47, and Accuracy 68.32 dropped to 67.03, indicating relatively stable performance.

ATE04 actually had the worst performance against attacks, even though it was very good with normal data. The threshold changes are quite consistent with the evaluation on normal data. Low epsilon 0.02 decrease in threshold, while epsilon 0.03 and 0.04 significant increase. Thus, on the attacked dataset, enhanced preprocessing can help improve score separation at low epsilon. However, use at high epsilon can increase overlap, causing the threshold to rise dramatically.

Overall, enhanced preprocessing improves performance at epsilon 0.02 but worsens performance at epsilon 0.03 and 0.04. This may occur because preprocessing makes facial images sharper, so the attack effect is even more destructive at high epsilon.

During training, the best separation epoch value shows that the ATE model converges earlier. AT has a slower best separation, which is at epoch 18–20, indicating that adversarial training (without enhanced preprocessing) has difficulty achieving stability. The best separation on AT is above 500, with AT02 at 0.584, AT03 at 0.526, and AT04 at 0.543. AT03 requires the most epochs to reach convergence with 20 epochs, AT04 with 19 epochs, and AT02 with 18 epochs. Table V shows the epoch and best separation of the model.

TABLE V. EPOCH AND BEST SEPARATION OF MODEL

Model	Total Epoch	Epoch (Best Separation)	Best Separation
AT02	20	18	0.584
AT03	20	20	0.526
AT04	20	19	0.543
ATE02	20	12	0.407
ATE03	20	14	0.500
ATE04	20	18	0.304

ATE achieves the best separation earlier, which is at epoch 12–14, indicating that enhanced preprocessing accelerates feature learning. The best separation in ATE is more diverse, where the highest best separation is in ATE03 0.500, followed by ATE02 0.407 and the lowest is in ATE04 0.304. However, the best separation is not directly proportional to robustness or final accuracy, so the main evaluation metrics must still be considered. Although ATE04 has low best separation, its performance drops significantly on the attacked dataset and requires 18 epochs. ATE03 is in the middle, reaching convergence at 14 epochs. ATE02 remains superior in this regard because the best separation occurs at epoch 12 with good AUC and accuracy, as well as stable thresholds.

The analysis of the six scenarios in Table IV shows an improvement compared to the base model. To see the increase in performance, a comparison of the base model with the performance obtained from the six scenarios of applying defense techniques is shown in Table VI.

The base model shows an AUC of 62.21, an EER of 41.5, and an accuracy of 58.63 when attacked. Adversarial training, represented by the AT model, makes the model's performance superior to the base model, with a minimum AUC of 63.12, a maximum EER of 40.30, and a minimum accuracy of 59.77. The use of adversarial training alone can improve performance. Enhanced preprocessing, represented by the ATE model, shows much better performance than the base model, with a minimum AUC of 69.29, a maximum EER of 36.10, and a minimum accuracy of 64.20. This indicates that the combination provides limited complementary effects, particularly under low ε values.

TABLE VI. COMPARISON BASE MODEL, AT AND ATE MODEL PERFORMANCE AT ATTACKED DATASET

Model	AUC	EER	Accuracy
Base	62.21	41.5	58.63
AT	≥63.12	≤40.30	≥59.77
ATE	≥69.29	≤36.10	≥64.20

In Table VII, we evaluate the Attack Success Rate (ASR) across different adversarial perturbation levels (ε) to assess the defence's vulnerability to adversarial attacks. The ASR is calculated as the ratio of adversarial samples that successfully cause misclassification. Our results show that the proposed defence reduces the ASR at lower perturbation magnitudes (ε = 0.02), indicating improved robustness. However, as the perturbation increases (ε = 0.04), the ASR rises, suggesting that the model becomes more susceptible to stronger adversarial attacks.

TABLE VII. EVALUATION MATRIX

Perturbation (ε)	ASR	Robust Accuracy (White box)	Robust Accuracy (Black box)	Robustness Radius (CW L2)
0.02	0.12	0.70	0.65	0.16
0.03	0.21	0.68	0.63	0.13
0.04	0.35	0.60	0.55	0.09

In terms of robust accuracy, we evaluate the model's performance under both white-box and black-box attack settings. For white-box attacks, the model maintains relatively stable performance, with accuracy values around 70% at lower perturbation levels, thanks to adversarial training. In contrast, the model's robustness decreases under black-box attacks but still outperforms the baseline in many cases.

Lastly, we measure the adversarial robustness radius using the CW L2 attack, which represents the minimum perturbation required to change the model's classification. The results indicate a larger robustness radius for lower perturbations, suggesting the model is more resilient to small adversarial changes. However, as the perturbation magnitude increases (ε = 0.04), the robustness radius decreases, highlighting the model's increasing vulnerability to stronger attacks.

Next, we include a comparison of performance visualization between the base model and the robust model with the highest performance, namely ATE02 on the attacked dataset. In Fig. 5(a), when the base model is subjected to FGSM attacks, the ROC curve flattens and approaches the diagonal, indicating weak performance. The base model has an AUC of 0.62, which means it is

close to random performance. In Fig. 5(b), the ROC curve curves more to the upper left, indicating a significant improvement in performance. The increase in AUC and decrease in EER also indicate good verification performance.

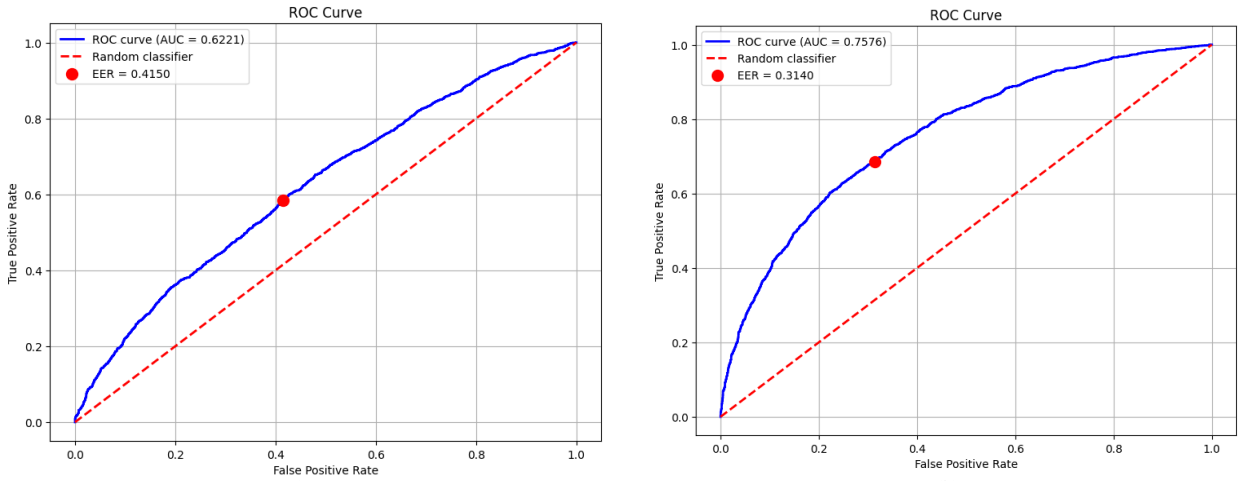

Fig. 5. ROC curves under FGSM adversarial conditions: (a) Base model; (b) Robust model ATE02.

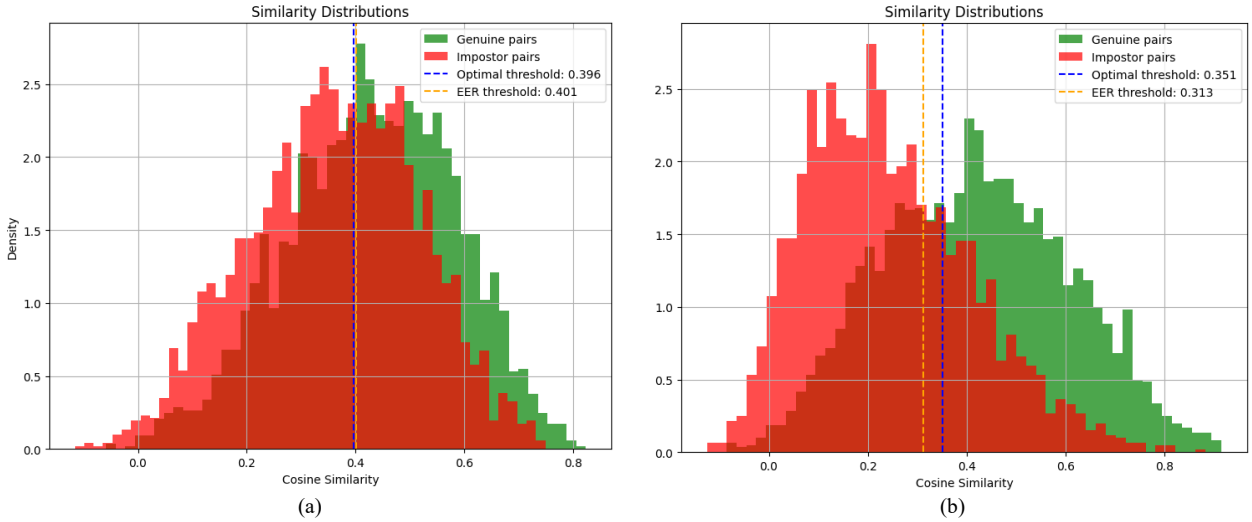

Fig. 6. Distribution of similarity scores and best separation under FGSM adversarial conditions: (a) Base model; (b) Robust model ATE02.

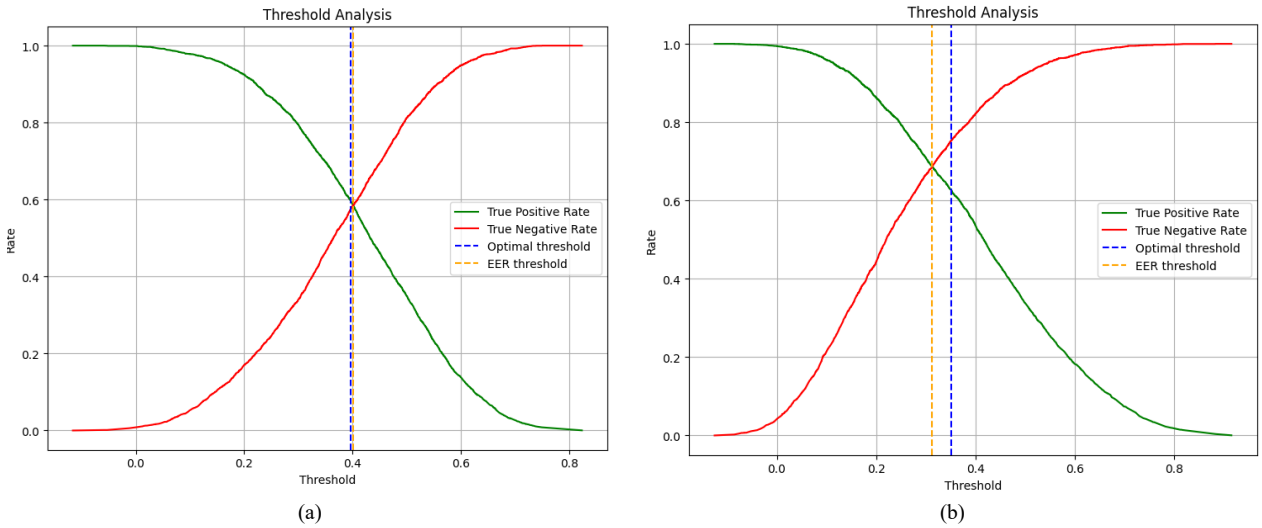

Fig. 7. Optimal threshold analysis based on True Positive Rate (TPR) and True Negative Rate (TNR) intersection under FGSM adversarial conditions: (a) Base model; (b) Robust.

Fig. 6(a) shows that the genuine and impostor distributions overlap significantly, especially in the similarity range of 0.25–0.45. The attack successfully blurred the boundary between genuine and impostor, indicating that the base model is vulnerable to attack data. Meanwhile, in Fig. 6(b), the robust model ATE02 has a clearer separation. The overlap between genuine and impostor is greatly reduced with a lower EER. The impostor distribution remains at low cosine similarity, while the genuine distribution remains relatively high. This indicates that the robust model can maintain the embedding structure.

In Fig. 7, TPR is represented by the green line and TNR is represented by the red line. The intersection point of TPR-TNR represents the optimal threshold and correlates with EER. The base model in Fig. 7(a) shows that the intersection point is around the threshold of 0.4 with a relatively steep slope. The distribution of genuine and impostor scores in the base model is less separated, so that small changes in the threshold will result in significant changes in TPR/TNR. The robust ATE02 in Fig. 7(b) model has a TPR-TNR intersection point at a lower threshold, indicating that the model is more conservative. The proximity of the optimal threshold and EER threshold indicates a more stable decision boundary calibration.

Fig. 8 shows a comparison of LIME visualizations for the base model and the robust model ATE02 under attacked conditions. As analyzed above, Fig. 8(a) gives a similarity score of 0.630, which is lower than the initial similarity score. In Fig. 8(b), the ATE02 model successfully increased the similarity score back to 0.646. The green area that supports the prediction also increased to the relevant feature, namely the face area. In terms of visualization, the robust model proved to be better than the base model. Even with the perturbation, the LIME visualization still focuses on core facial features.

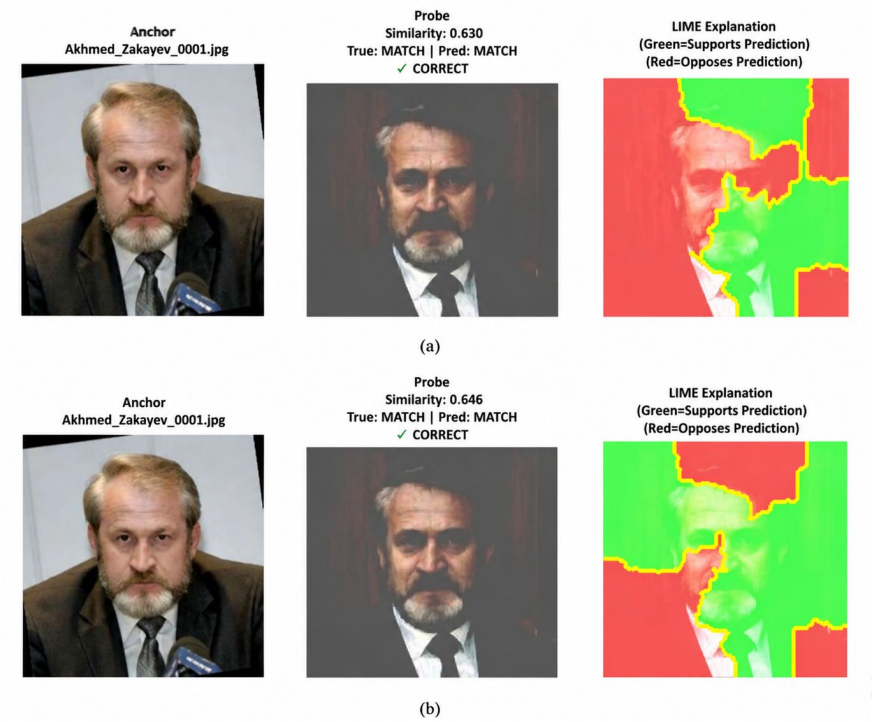

Fig. 8. Comparison of LIME visualizations: (a) Base model; (b) Robust model ATE02.

Robustness limitation analysis show while the proposed method improves performance under low perturbation levels, results indicate degradation at $\varepsilon = 0.03$ and $\varepsilon = 0.04$. This suggests that explanation-guided preprocessing may inadvertently amplify certain perturbation effects when the attack strength exceeds a threshold. This finding highlights a key limitation: XAI-based defences are not universally effective and must be carefully calibrated to the threat model.

Discussion on attack strength show FGSM is a first-order attack and represents a relatively weak adversary. Therefore, the observed robustness improvements should be interpreted as preliminary. Stronger iterative attacks such as Projected Gradient Descent (PGD) or CW are expected to further challenge the model, and their evaluation is left for future work.

While LIME was chosen for its model-agnostic properties and ease of integration with adversarial defence strategies, we plan to explore Grad-CAM and SHAP in future work to compare their performance in CNN-based facial recognition models.

Regarding scalability, while the proposed method demonstrates robustness improvements on benchmark datasets such as LFW and CASIA-WebFace, its application to larger datasets or real-time systems presents challenges. The LIME-based explanation mechanism, although model-agnostic, involves computationally intensive operations, particularly when many superpixels or perturbed samples are generated. Consequently, real-time deployment or processing of datasets with millions of images may require optimization strategies, such as adaptive sampling or approximate feature

importance calculation. Additionally, storage and memory overheads from adversarial examples could become substantial, necessitating efficient management of training data and batch processing to maintain feasible scalability.

V. Conclusion

In this paper, an adversarial defense approach combining adversarial training with enhanced preprocessing by insight from XAI is proposed for facial recognition based on CNN. The results of the experiment show that the defense method improves the performance of the base model. The base model has an AUC of 62.21, an EER of 41.5, and an accuracy of 58.63. The application of adversarial training improves overall performance, resulting in a minimum AUC of 63.12, a maximum EER of 40.30, and a minimum accuracy of 59.77. Enhanced preprocessing has a significant contribution in clean dataset conditions, while in attacked dataset conditions it has superior performance at epsilon 0.02. However, overall performance shows superiority with a minimum AUC of 69.29, a maximum EER of 36.10, and a minimum accuracy of 64.20. Therefore, the combination of adversarial training and enhanced preprocessing from LIME analysis demonstrates limited and condition-dependent improvements in model robustness. However, enhanced preprocessing must be used carefully because it does not always have a positive impact on every epsilon.

Although the defense strategy based on reproduction effectively strengthens the model against adversarial attacks, there still exist several limitations of this work that need to be further studied. The baseline performance could be improved upon by expending more computation power, for example through more training epochs or by improving the backbone architecture. LIME can also be further studied with different settings in the number of superpixels, segmentation parameters, and sampling strategies, to gain more robust and representative results in interpretability. Moreover, one can perform experiments with even stronger adversarial attacks, e.g., PGD, CW or AutoAttack, to assess the performance more rigorously. Adversarial training can be further improved by increasing the diversity of perturbations from adversarial examples.

While the current defense strategy shows promise under FGSM, further evaluation with stronger and iterative adversarial attacks, such as PGD and CW, will be conducted in future work to rigorously assess the generalization of the defense method.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

I.D. contributed to writing the paper and conducted data analysis; A.R. conceived and designed the research and contributed to writing the manuscript; R.G. conducts testing, evaluation, and review of scientific articles; E.H. collected data, conducted experiments, analyzed data, and drafted the initial version of the manuscript; R.W.W. conducts the final review process and improvement of scientific manuscripts; all authors had approved the final version.

Funding

This work was supported by Telkom University Research and Community Service Institute.

Acknowledgment

We express our gratitude for the completion of this research. Thank you to Telkom University and Siliwangi University, as well as all those who have provided support, guidance, and contributions in the development of XAI-Based Enhancement of CNN Robustness in Biometric Face Recognition. We also appreciate the feedback from the community that helped improve this research.

References

[1] M. Baytamouny, R. Kolandaisamy, and G. S. ALDharhani, "AI-based home security system with face recognition," in *Proc. 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), IEEE*, 2022, pp. 1038–1042. doi: 10.1109/ICOEI53556.2022.9776900

[2] A. Jafar and M. Lee, "High-speed hyperparameter optimization for deep ResNet models in image recognition," *Cluster Computing*, vol. 26, no. 5, pp. 2605–2613, 2023. doi: 10.1007/s10586-021-03284-6

[3] A. Kadhim and S. Al-Darraji, "Face recognition system against adversarial attack using convolutional neural network," *Iraqi Journal for Electrical and Electronic Engineering*, vol. 18, no. 1, pp. 1–8, 2022. doi: 10.37917/ijeee.18.1.1

[4] A. Guesmi, M. A. Hanif, B. Ouni, and M. Shafique, "Physical adversarial attacks for camera-based smart systems: Current trends, categorization, applications, research challenges, and future outlook," *IEEE Access*, vol. 11, pp. 109617–109668, 2023. doi: 10.1109/ACCESS.2023.3321118

[5] G. Mikriukov, G. Schwalbe, F. Motzkus, and K. Bade, "Unveiling the anatomy of adversarial attacks: Concept-based XAI dissection of CNNs," in *Proc. Communications in Computer and Information Science, Springer Science and Business Media Deutschland GmbH*, 2024, pp. 92–116. doi: 10.1007/978-3-031-63787-2_6

[6] N. Shukla and S. Banerjee, "Generating adversarial attacks in the latent space," in *Proc. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 730–739. doi: 10.1109/CVPRW59228.2023.00080

[7] J. Jayapradha, L. Vadhanie, Y. Kulkarni, T. S. Kumar, and M. Uma Devi, "Enhancing algorithmic resilience against data poisoning using CNN," *Risk Assessment and Countermeasures for Cybersecurity*, 2024, pp. 131–157. doi: 10.4018/979-8-3693-2691-6.ch008

[8] P. Khare, S. Arora, and S. Gupta, "Artificial intelligence-based biometric authentication systems for facial recognition and identification," in *Proc. 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT)*, 2024, pp. 1–7. doi: 10.1109/ICEECT61758.2024.10738912

[9] C. Lv and Z. Lian, "Improving transferability of adversarial attack on face recognition with feature attention," in *Proc. ICONIP 2023*, 2024, vol. 1969, pp. 243–253.

[10] D. Kim, H. Kim, Y. Jung, S. Kim, and B. C. Song, "Towards the adversarial robustness of facial expression recognition: Facial attention-aware adversarial training," *Neurocomputing*, vol. 584, 127588, 2024. doi: 10.1016/j.neucom.2024.127588

[11] S. Letzgus and K. R. Müller, "An explainable AI framework for robust and transparent data-driven wind turbine power curve models," *Energy and AI*, vol. 15, 100328, 2024. doi: 10.1016/j.egyai.2023.100328

[12] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 11, pp. 78994–79015, 2023. doi: 10.1109/ACCESS.2023.3294569

[13] L. Weber, S. Lapuschkin, A. Binder, and W. Samek, "Beyond explaining: Opportunities and challenges of XAI-based model improvement," *Information Fusion*, vol. 92, pp. 154–176, 2023. doi: 10.1016/j.inffus.2022.11.013

[14] R. Dwivedi, P. Kothari, D. Chopra, M. Singh, and R. Kumar, "An efficient ensemble Explainable AI (XAI) approach for morphed face detection," *Pattern Recognit. Lett.*, vol. 184, pp. 197–204, 2024. doi: 10.1016/j.patrec.2024.06.014

[15] S. Khairnar, S. Gite, K. Mahajan, B. Pradhan, A. Alamri, and S. D. Thepade, "Advanced techniques for biometric authentication: Leveraging deep learning and explainable AI," *IEEE Access*, vol. 12, pp. 153580–153595, 2024. doi: 10.1109/ACCESS.2024.3474690

[16] A. Merha, "Unifying adversarial robustness and interpretability in deep neural networks: A comprehensive framework for explainable and secure machine learning models," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 2, no. 9, pp. 1829–1838, 2024. doi: 10.56726/irjmets4109

[17] M. Sadria, A. Layton, and G. D. Bader, "Adversarial training improves model interpretability in single-cell RNA-seq analysis," *Bioinformatics Advances*, vol. 3, no. 1, 2023. doi: 10.1093/bioadv/vbad166

[18] Y. Lin, H. Zhao, X. Ma, Y. Tu, and M. Wang, "Adversarial attacks in modulation recognition with convolutional neural networks," *IEEE Trans. Reliab.*, vol. 70, no. 1, pp. 389–401, 2021. doi: 10.1109/TR.2020.3032744

[19] F. Vakhshiteh, A. Nickabadi, and R. Ramachandra, "Adversarial attacks against face recognition: A comprehensive study," *IEEE Access*, vol. 9, pp. 92735–92756, 2021. doi: 10.1109/ACCESS.2021.3092646

[20] Z. Li, J. Sun, X. Yao, R. Cui, B. Ge, and K. Yang, "An interpretable adversarial robustness evaluation method based on robust paths," in *Proc. 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2024, pp. 1213–1218. doi: 10.1109/CSCWD61410.2024.10580004

[21] T. Agustin, M. H. Purwidiantoro, and M. L. Rahmadi, "Deep learning for robust facial expression recognition: A resilient defense against adversarial attacks," *International Journal of Intelligent Systems and Applications*, vol. 16, no. 5, pp. 39–52, 2024. doi: 10.5815/ijisa.2024.05.04

[22] M. A. Kadir, G. Addluri, and D. Sonntag, "Revealing vulnerabilities of neural networks in parameter learning and defense against explanation-aware backdoors," arXiv preprint, arXiv:2403.16569, 2024.

[23] M. H. Abid, R. Ashraf, T. Mahmood, and C. M. N. Faisal, "Multi-modal medical image classification using deep residual network and genetic algorithm," *PLoS One*, vol. 18, no. 6, e0287786, 2023. doi: 10.1371/journal.pone.0287786

[24] I. Prokopiou and P. Spyridonos, "Highlighting the advanced capabilities and the computational efficiency of DeepLabV3+ in medical image segmentation: An ablation study," *BioMedInformatics,* vol. 5, no. 1, 10, 2025. doi: 10.3390/biomedinformatics5010010

[25] A. Al Tawil, A. Shaban, and L. Almazaydeh, "A comparative analysis of convolutional neural networks for breast cancer prediction," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 3, pp. 3406–3414, 2024. doi: 10.11591/ijece.v14i3.pp3406-3414

[26] M. Talele and R. Jain, "A comparative analysis of CNNs and ResNet50 for facial emotion recognition," *Engineering, Technology and Applied Science Research*, vol. 15, no. 2, pp. 20693–20701, 2025. doi: 10.48084/etasr.9849

[27] L. Chang *et al*., "Accuracy vs. efficiency: Achieving both through hardware-aware quantization and reconfigurable architecture with mixed precision," in *Proc. 2021 IEEE Intl. Conf. on Parallel. & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, 2021, pp. 151–158.

[28] A. Zhalgas, B. Amirgaliyev, and A. Sovet, "Robust face recognition under challenging conditions: A comprehensive review of deep learning methods and challenges," *Applied Sciences*, vol. 15, no. 17, 9390, 2025. doi: 10.3390/app15179390

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 770–778.

[30] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4690–4699. doi: 10.1109/TPAMI.2021.3087709

[31] W. Villegas-Ch, A. Jaramillo-Alcázar, and S. Luján-Mora, "Evaluating the robustness of deep learning models against adversarial attacks: An analysis with FGSM, PGD and CW," *Big Data and Cognitive Computing*, vol. 8, no. 1, 8, 2024. doi: 10.3390/bdcc8010008

[32] U. Mwaibale, N. Mduma, H. Laizer, and B. Mgawe, "Enhancing detection of common bean diseases using fast gradient sign method–trained vision transformers," *Front Artif. Intell.*, vol. 8, 1643582, 2025. doi: 10.3389/frai.2025.1643582

[33] N. Ding and K. Möller, "Minimally distorted adversarial images with a step-adaptive iterative fast gradient sign method," *AI*, vol. 5, no. 2, pp. 922–937, 2024. doi: 10.3390/ai5020046

[34] D. Hong, D. Chen, Y. Zhang, H. Zhou, and L. Xie, "Attacking robot vision models efficiently based on improved fast gradient sign method," *Applied Sciences*, vol. 14, no. 3, 1257, 2024. doi: 10.3390/app14031257

[35] A. M. Salih *et al*., "A perspective on explainable artificial intelligence methods: SHAP and LIME," *Advanced Intelligent Systems*, vol. 7, no. 1, 240030, 2025. doi: 10.1002/aisy.202400304

[36] D. Garreau and D. Mardaoui, "What does LIME really see in images?" in *Proc. International Conference on Machine Learning*, 2021, pp. 3620–3629.

[37] J. Contreras, A. Winterfeld, J. Popp, and T. Bocklitz, "Spectral zones-based SHAP/LIME: Enhancing interpretability in spectral deep learning models through grouped feature analysis," *Anal. Chem.*, vol. 96, no. 39, pp. 15588–15597, 2024. doi: 10.1021/acs.analchem.4c02329

[38] M. Zhao, L. Zhang, J. Ye, H. Lu, B. Yin, and X. Wang, "Adversarial training: A survey," arXiv preprint, arXiv:2410.15042, 2024.

[39] W. Zhao, S. Alwidian, and Q. H. Mahmoud, "Adversarial training methods for deep learning: A systematic review," *Algorithms*, vol. 15, no. 8, 283, 2022. doi: 10.3390/a15080283

[40] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," arXiv preprint, arXiv:1411.7923, 2014.

[41] M. Mosadag, M. A. Mohammed, K. Bashir, and A. Mubark, "A comparative analysis for the performance of LFW and ORL databases in facial recognition," *Conclusions in Engineering*, vol. 1, no. 1, pp. 58–63, 2025. doi: 10.71107/pydt9t88

[42] H. L. Gururaj, B. C. Soundarya, S. Priya, J. Shreyas, and F. Flammini, "A comprehensive review of face recognition techniques, trends, and challenges," *IEEE Access*, vol. 12, pp. 107903–107926, 2024. doi: 10.1109/ACCESS.2024.3424933