Al-Farouk journal of sciences
Volume (2), Issue (3) (JUNE, 2026)
ISSN: 3135-2359

مجلة الفاروق للعلوم
المجلد (2) العدد 3 يونيو 2026
ردمد: 3135-2359

Enhancing Face Recognition Performance: A Comparative Evaluation of ArcFace and FaceNet using Image Preprocessing and ROC–EER Metrics

[1st] Essra Sowan
Postgraduate office-College of Computer Technology Tripoli [CCTT], Tripoli, Libya
is2403003@cctt.edu.ly
[2nd] Amal Elbuaishi
Postgraduate office-College of Computer Technology Tripoli [CCTT], Tripoli, Libya
ab2403006@cctt.edu.ly
[3rd] Milad Areir
https://orcid.org/0000-0002-4053-6892
College of Computer Technology Tripoli [CCTT], Tripoli, Libya meladmohammed@gmail.com

تاريخ الاستلام: 2026/04/06 تاريخ المراجعة 2026 /05/08 تاريخ القبول:2026/05/19- تاريخ النشر: 2026 /06/07

Abstract— Deep learning has pushed face recognition to the point where standard benchmarks are no longer much of a challenge accuracy at near-human levels has become almost routine. What has received far less attention is a quieter but consequential question: how does image preprocessing affect the verification performance of models that have already been trained? This paper takes that question seriously, presenting a systematic comparison of two leading face recognition models FaceNet and ArcFace as they are subjected to a progressive preprocessing framework applied to the Labeled Faces in the Wild (LFW) dataset.The evaluation used 1,000 face pairs, split evenly between 500 matched and 500 mismatched, and ran each pair through five preprocessing stages. Stage 0 used raw images with no modification. By Stage 4, a full pipeline was in place, incorporating face detection, CLAHE histogram equalization, Gaussian blur, and pixel normalization. Performance at each stage was measured using the ROC curve, the Equal Error Rate (EER), classification accuracy, and error rate metrics chosen because they do not depend on selecting a single decision threshold and are well suited to binary verification tasks.At baseline, ArcFace was the stronger model across every metric, reaching 97.30% accuracy, an AUC of 0.9837, and an EER of 0.0340, against FaceNet's 95.60%, 0.9812, and 0.0480 respectively. What happened next ran counter to what most practitioners would expect. Rather than improving as preprocessing became more involved, both models declined consistently, across every stage, with no preprocessing step producing a measurable gain over the raw baseline. ArcFace felt the impact more sharply, losing 6.40% in accuracy from Stage 0 to Stage 4, while FaceNet dropped 3.20% over the same range. From Stage 2 onward, the ranking between the two models flipped: FaceNet posted lower EER values than ArcFace at every subsequent stage.The picture that emerges from these results is fairly clear. ArcFace is the better choice when imaging conditions can be controlled and images arrive at the model largely intact. FaceNet is more resilient when the pipeline involves external processing that the model has no control over. More broadly, the findings push back against the common assumption that image enhancement is a safe or neutral intervention applied without validation, it can quietly degrade the very systems it was meant to help.

Keywords—*face recognition; ArcFace; FaceNet; image preprocessing; ROC curve; equal*

error rate; LFW; deep learning; biometric verification.

I. INTRODUCTION

Facial recognition is one of the most widely used biometric identification applications in both security and commercial settings, thanks to its non-intrusive nature and applicability in diverse real-world environments [1]. Despite the remarkable results achieved by deep learning models in recent years, the process of facial recognition in real-world environments remains dependent on the quality of the processed images and the nature of the applied preprocessing pipeline [2].

In this context, a handful of CNN-based face recognition models have risen to particular prominence. Among them, FaceNet and ArcFace have attracted the most sustained attention from researchers and practitioners alike. FaceNet utilizes triplet loss to learn a compact representation of the face in a Euclidean space [3], while ArcFace employs additive angular margin loss, which enhances class separation in a spherical metric space [4]. Although both models have demonstrated strong performance in standardized benchmarks, to the best of our knowledge, comparative investigations of FaceNet and ArcFace under progressively applied preprocessing stages using threshold-independent ROC–EER evaluation remain relatively underexplored [5-7].

Based on this problem, the study uses the LFW dataset as an experimental framework for comparison. This database holds a well-established reference position in the facial recognition literature due to its inclusion of images taken in real-world conditions far removed from controlled environments [8]. LFW was selected because it provides unconstrained face images collected from real-world settings and remains one of the most widely adopted benchmarks for face verification, making it suitable for evaluating model robustness under non-controlled imaging conditions. The core research question investigates whether the preprocessing stages yield measurable performance variations in the performance of both models, and whether this effect differs between the two architectures [15].

The choice of the ROC and EER metrics for evaluation was not merely a conventional choice, but rather a response to the nature

of the problem itself; they enable performance evaluation independently of any predetermined decision threshold, which is a necessary condition to ensure the integrity of the comparison between the two models [7], [8]. Moreover, These metrics jointly characterize the trade-off between false acceptance and false rejection rates , giving the comparison a deeper analytical dimension than that provided by traditional accuracy measures.

Most comparative studies in this space have evaluated models under standardized conditions and left it there, without asking how performance shifts as preprocessing becomes progressively more involved [5]. Furthermore, these studies have rarely relied on the EER as a central criterion for comparing the two models within a standardized experimental framework [10],[11]. Therefore, this paper aims to bridge this gap through three main contributions: (1) establishing a reproducible comparative evaluation framework for FaceNet and ArcFace using the LFW dataset; (2) investigating the progressive impact of preprocessing stages on the verification performance of each model individually; and (3) employing ROC curves and Equal Error Rate (EER) as threshold-independent quantitative metrics for fair and objective model comparison.

The rest of the paper is organized as follows : Section II reviews related work , section III describes the methodology while section IV introduces the results section V is the discussion and finally , section VI concludes the paper.

II. RELATED WORK

Deep learning has transformed face recognition. When convolutional neural networks arrived, they raised the bar in a way that nothing before them had managed delivering verification accuracy that held up even when images were noisy, poorly lit, or captured in the kind of unpredictable conditions that real deployment actually throws at a system. But progress on one front has a way of exposing problems on another, and the field is still working through a set of challenges that better architectures alone have not resolved — questions about how robust these models really are, how sensitive they remain to image quality, and how much the preprocessing applied before recognition shapes the results that follow. These questions are made harder to answer by the state of the existing literature. Studies comparing leading models such as FaceNet and ArcFace have produced results that are difficult to reconcile with one another, largely because they were conducted under different conditions — different datasets, different experimental setups, different evaluation protocols. What looks like a performance difference between two models may in some cases reflect little more than a difference in how the experiments were designed. With that in mind, this section reviews the work most directly relevant to three areas: how face recognition systems have developed over time, what comparative evaluations of FaceNet and ArcFace have found, and where the methodological limitations of existing studies lie.

A. Comparison of facial recognition accuracy between ArcFace, FaceNet, and FaceNet512 models

The comparison between FaceNet and ArcFace has attracted increasing research attention in recent years, although the methodologies employed and the conclusions drawn vary, reflecting primarily differences in experimental environments

rather than in the models themselves. In a 2023 study by Firmansyah et al. using the DeepFace framework on an Indonesian dataset, FaceNet512 outperformed both models with an accuracy of 97.4%, while ArcFace recorded the lowest rate at 87.8% [12].

B. Comparative analysis of FaceNet and ArcFace in access control application

Juwanda et al. took a different approach in their 2024 study, testing both models in real-world security control environments rather than clean benchmark conditions. Their findings pointed in FaceNet's favor it produced lower false acceptance rates when lighting conditions changed or subjects were positioned at varying distances, which made it a stronger candidate for access control systems where wrongly letting someone in carries real consequences [13]. That said, the study has a notable gap: it did not systematically examine how preprocessing affects these results, since the experiments were run on raw images without any standardized preparation pipeline. That omission makes it harder to know how much of the performance difference was down to the models themselves and how much might shift with proper image normalization in place.

C. Comparative analysis of FaceNet, ArcFace, and OpenFace models over five years

A 2024 study that tracked FaceNet, ArcFace, and OpenFace across five years of benchmarks (2018–2023) painted a more nuanced picture [14]. FaceNet proved more resilient when lighting conditions were poor, while ArcFace tended to deliver more consistent results when the subjects varied in ethnicity or age. The takeaway was not that one model is better it is that each one has a comfort zone, and performance depends heavily on who is in the image and how it was captured.

D. Comparative evaluation of ArcFace and FaceNet using the AHP framework and the EER

metric

Perhaps the closest work to ours is that of Ronquillo-Figueroa et al. (2026), who compared FaceNet and ArcFace across three datasets including LFW, and were among the few to incorporate EER as a core evaluation metric [15]. Their findings leaned toward ArcFace in terms of classification accuracy and F1, while FaceNet held its ground in speed and verification stability.

E. The effect of facial image quality on recognition accuracy in uncontrolled environments

In a line of inquiry directly relevant to the concerns of this study, Zhang (2023) took on a question that the face recognition literature has often been too quick to dismiss: does image quality still pose a genuine challenge for deep learning models when they operate in uncontrolled environments [9]? To find out, she divided facial images into three groups according to quality level and ran parallel human verification experiments on the same data, allowing for a direct comparison between machine and human performance. The findings were sobering. Image quality remains a meaningful problem even for state-of-the-art deep learning models, and the widespread claim that these models have surpassed human-level performance turns out to rest on assumptions that do not always hold in practice the optimism, it seems, has been somewhat premature.

For the purposes of this study, that finding carries particular weight. It reinforces the case for treating preprocessing not as a background technicality but as an independent variable that deserves serious, controlled investigation one that directly shapes the quality of the images a model receives before it ever attempts recognition. If quality matters as much as Zhang's work suggests, then understanding exactly how preprocessing influences that quality is not a peripheral question.

F. Critical Literature Review and Research Gaps

Taken together, the five studies reviewed here share a common limitation that quietly undermines the reliability of their conclusions: none of them treated preprocessing as a controlled, independent variable within a standardized experimental protocol. Firmansyah et al. [12] and Juwanda et al. [13] confined their experiments to specific data environments, which raises questions about how far their findings can be generalized. Iype and Sebastian [14] covered a longer time horizon but without consistent control over experimental conditions across that period. Ronquillo et al. [15] came closest to the methodological approach taken here using LFW and EER as evaluation tools yet even they stopped short of systematically isolating preprocessing as a variable in its own right. Zhang [9] perhaps made the most pointed contribution by demonstrating that image quality continues to challenge deep learning models in ways the field has been slow to acknowledge, which only strengthens the argument for studying preprocessing effects with greater rigor.

This paper takes that gap as its starting point. Rather than treating preprocessing as an afterthought or a fixed background condition, it builds a comparative evaluation of FaceNet and ArcFace on the LFW dataset around a unified experimental framework in which preprocessing is adjusted progressively and systematically across conditions. ROC curves and the Equal Error Rate (EER) serve as the primary tools for comparison metrics that the earlier studies either did not use [12, 13, 14] or applied only in part [9]. The goal is to produce the kind of controlled, reproducible evidence that the existing literature has so far left on the table.

TABLE I. Previous Studies Comparison.

study	Study /Approach description	strengths	Weakness
Zhang [9] 2023	Investigating the impact of face image quality on deep learning model performance using LFW, with human verification comparison	- Demonstrates image quality impact - Uses LFW benchmark Human vs. machine comparison	- Does not directly compare FaceNet and ArcFace- No ROC/EER metrics
Firmansyah et al. [12] 2023	Comparing FaceNet, ArcFace, and FaceNet512 using DeepFace framework on an Indonesian dataset	-Simultaneous comparison of three models - Unified DeepFace framework	- Dataset not generalizable to LFW No ROC/EER metrics Preprocessing not specified
Juwanda et al. [13] 2024	Comparing FaceNet and ArcFace in real- world access control environments, measuring FAR across varying distances and lighting	Comparing FaceNet and ArcFace in real-world access control environments, measuring FAR across varying distances and lighting	- Does not Include LFW No ROC/EER Metrics No standardized preprocessing.
Iype & Sebastian [14] 2024	Five-year comparative analysis (2018–2023) of FaceNet, ArcFace, and OpenFace using accuracy and speed metrics	Long-term longitudinal perspective Three -model comparison	- Inconsistent experimental variables No ROC/EER metrics - No unified preprocessing
Ronquillo-Figueroa et al. [15] 2026	Comparing ArcFace and FaceNet on LFW, CASIA, and Celeb-DF using AHP framework with Accuracy, F1, and EER metrics	- Includes LFW and EER - Systematic AHP framework - Multi-dataset	Preprocessing not treated as an independent variable - No

III. METHODOLOGY

The central aim of this study is to understand how FaceNet and ArcFace respond as the preprocessing applied to face images becomes progressively more involved. Rather than

treating preprocessing as a fixed background condition, the experiments were designed to vary it systematically so that its effect on each model's verification performance could be observed directly and compared. The methodology that supports this is built around four components: the dataset selected for evaluation, the preprocessing pipeline through which images were passed, the configuration under which each model was run, and the metrics used to assess and compare the results.

A. Dataset

The Labeled Faces in the Wild (LFW) dataset [8] was chosen as the evaluation benchmark for this study. LFW brings together 13,233 face images spanning 5,749 individuals, all collected from the web rather than controlled capture sessions. This characteristic is particularly important because the images reflect the kind of variation that real-world systems actually encounter shifting lighting conditions, different head poses, a wide range of facial expressions, and frequent cases of partial occlusion. It is precisely this diversity that makes LFW a meaningful testing ground for comparing how well each model holds up outside the clean conditions of a laboratory.

For this experiment, a balanced subset of 1,000 face pairs was generated using a custom random sampling procedure with a fixed random seed (42) to ensure experimental reproducibility. Genuine pairs were constructed by generating all possible image combinations for identities containing two or more samples, followed by random shuffling and selection of 500 matching pairs. Impostor pairs were built by randomly drawing images from two different identities and repeating that process until 500 non-matching pairs had been assembled. This sampling strategy ensured balanced class representation while maintaining computational consistency across all preprocessing stages:

TABLE II. LFW Pair Configuration**.**

Pair Type	Count	Description
Matched	500	Two images of the same person under varying lighting, expression, and viewpoint
Mismatched	500	Random pairs of two different individuals selected from the dataset

B. Preprocessing Pipeline

- One of the more distinctive contributions this study makes is taking image preprocessing seriously as an independent experimental variable rather than a background condition to be assumed and ignored. A progressive pipeline was designed in which techniques are applied incrementally across five stages (Stage 0 through Stage 4), enabling the isolated measurement of each technique's contribution to model performance.

TABLE III. Progressive Preprocessing Stages.

Stage	Applied Techniques	Purpose
Stage 0	No preprocessing raw images	Baseline reference performance
Stage 1	Face Detection + Resize (160×160)	Consistent face localization and geometric normalization
Stage 2	Stage 1 + Histogram Equalization (CLAHE)	Enhance contrast and reduce illumination variation
Stage 3	Stage 2 + Gaussian Noise Reduction	Remove high-frequency noise artifacts
Stage 4	Stage 3 + Pixel Normalization [0, 1]	Standardize pixel intensity across images

A unified input resolution of 160×160 was adopted for both models to ensure consistency across all experimental conditions and to eliminate input-size variability as a confounding factor during comparative evaluation.

C. Model Configuration

Both models were deployed via the DeepFace framework using pre-trained weights, with no fine-tuning applied on LFW to ensure a fair and reproducible comparison. FaceNet [3] generates 128-dimensional face embeddings trained with Triplet Loss, while ArcFace [4] produces 512-dimensional embeddings trained with Additive Angular Margin Loss on MS-Celeb-1M. Verification decisions were made by computing the cosine similarity between embedding pairs.

Similarity thresholds were systematically swept across the full cosine similarity range to generate ROC curves and estimate the Equal Error Rate (EER) for each preprocessing stage.

D. Evaluation Metrics

The following metrics were used to evaluate model performance [10, 11]:
True Positive (TP): Matched pairs correctly identified as the same person.
True Negative (TN): Mismatched pairs correctly identified as different people.
False Positive (FP): Mismatched pairs incorrectly identified as the same person (False Acceptance).
False Negative (FN): Matched pairs incorrectly identified as different people (False Rejection).

$$Accuracy = \frac{TP + TN}{\text{Total Pairs}} \times 100\% \quad (2)$$

$$Error\ Rate = \frac{FP + FN}{\text{Total Pairs}} \times 100\% \quad (3)$$

$$FAR = \frac{FP}{FP + TN} \times 100\% \quad (4)$$

$$\text{FRR} = \frac{FN}{TP+FN} \times 100\% \quad (5)$$

$$\text{TAR} = \frac{TP}{TP+FN} \times 100\% \quad (6)$$

The ROC curve works by plotting the True Accept Rate against the False Accept Rate at every possible decision threshold, which means the picture it gives of a model's verification performance does not hinge on choosing any particular operating point it captures the full range of trade-offs in a single curve.
The Equal Error Rate takes a different angle. Rather than summarizing the whole curve, it zeroes in on one specific operating point the threshold at which the False Accept Rate and the False Rejection Rate happen to be equal, meaning the system is letting in exactly as many unauthorized users as it is turning away legitimate ones. That balance point is a useful reference.

E. Experimental Setup

Both models were always evaluated on exactly the same image pairs at each preprocessing stage, removing any variation that could come from differences in the input data. The evaluation used 1,000 face pairs in total 500 matched pairs and 500 mismatched pairs applied consistently across all experiments. Preprocessing operations were handled through OpenCV and scikit-image. The full experimental environment ran on Python
3.10 with TensorFlow 2.x and the DeepFace library. Finally, results from each preprocessing stage were recorded independently rather than averaged together, which makes it possible to

trace how performance shifts as preprocessing becomes progressively more involved.

IV. RESULTS

This section presents the experimental results obtained from evaluating FaceNet and ArcFace across five progressive preprocessing stages on 1,000 LFW face pairs (500 matched, 500 mismatched) Performance is reported in terms of EER, AUC, Accuracy and Error Rate. Table IV summarizes the complete results, followed by a stage-by-stage analysis.

A. Experimental Configuration

Table IV details the hardware specifications, software framework, and evaluation parameters used to ensure a unified and reproducible experimental environment for both models.

TABLE IV. Experimental Configuration and Parameters.

Category	Parameter / Specification
Hardware	**Intel Core i7, 32 GB RAM**
Operating System	**Windows 11**
Software Framework	**Python 3.10, TensorFlow 2.x, DeepFace**
Preprocessing Libraries	**OpenCV, scikit-image, scikit-learn**
Face Detector	**MTCNN (via DeepFace backend)**
Input Image Size	**160 × 160 px (Unified for both models)**
Distance Metric	**Cosine Similarity**
Dataset	**LFW — 1,000 pairs (500 matched + 500 mismatched)**
Evaluation Metrics	**AUC, EER, Accuracy, Error Rate**
Preprocessing Stages	**5 stages (Stage 0: Baseline → Stage 4: Full Pipeline)**
Verification Threshold	**Derived from ROC analysis (EER operating point)**

B. Performance Results

Table V brings together the full set of performance metrics for both models across all five preprocessing stages in one place.

TABLE V. Performance Comparison Across Preprocessing Stages.

Model	stage	Preprocessing	AUC	EER	Accuracy	Error Rate (%)
FaceNet	0	Baseline (Raw)	0.9812	0.0480	95.60	4.40
	1	+ Detection & Resize	0.9812	0.0480	95.60	4.40
	2	+ CLAHE	0.9608	0.0850	93.40	6.60
	3	+ Gaussian Blur	0.9528	0.1020	92.40	7.60
	4	+ Normaliza tion	0.9529	0.1020	92.40	7.60
ArcFace	0	Baseline (Raw)	0.9837	0.0340	97.30	2.70
	1	+ Detection & Resize	0.9837	0.0340	97.30	2.70
	2	+ CLAHE	0.9365	0.1040	92.50	7.50
	3	+ Gaussian Blur	0.9263	0.1110	90.90	9.10

C. Baseline Performance (Stages 0 and 1)

At the baseline stage (Stage 0 — raw images), ArcFace achieved an AUC of 0.9837 and an EER of 0.0340, corresponding to an accuracy of 97.30% and an error rate of 2.70%. FaceNet came in at an AUC of 0.9812 and an EER of 0.0480, with an accuracy of 95.60% and an error rate of 4.40%. ArcFace thus outperformed FaceNet on raw LFW images by 1.70 percentage points in accuracy and 0.0140 in EER. Stage 1 (Detection + Resize) produced results identical to Stage 0 for both models across all metrics, as illustrated in Fig. 2. This outcome indicates that the DeepFace framework applies face detection and resizing internally prior to embedding extraction, rendering the external Stage 1 preprocessing redundant under this configuration.

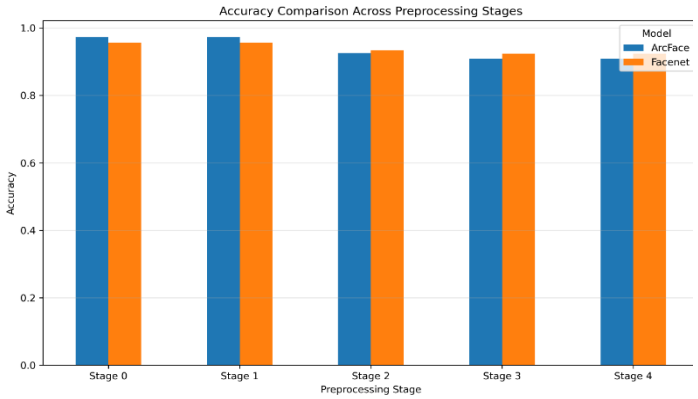

Fig. 2. Accuracy comparison of FaceNet and ArcFace across five preprocessing stages.

D. Effect of CLAHE Histogram Equalization (Stage 2)

The application of CLAHE in Stage 2 resulted in a consistent performance degradation for both models. FaceNet's AUC decreased from 0.9812 to 0.9608 (ΔAUC = −0.0204) and EER increased from 0.0480 to 0.0850 (ΔEER = +0.0370), while accuracy dropped by 2.20 percentage points to 93.40%. ArcFace experienced a more pronounced decline, with AUC falling from 0.9837 to 0.9365 (ΔAUC = −0.0472) and EER rising from 0.0340 to 0.1040 (ΔEER = +0.0700), with accuracy dropping by 4.80 percentage points to 92.50%. As illustrated in Fig. 3, ArcFace's EER after Stage 2 (0.1040) exceeded that of FaceNet (0.0850), representing a reversal of the performance ranking observed at baseline. This finding suggests that both models — and particularly ArcFace — are sensitive to contrast enhancement operations that alter the original pixel intensity distribution learned during pre-training on large-scale datasets.

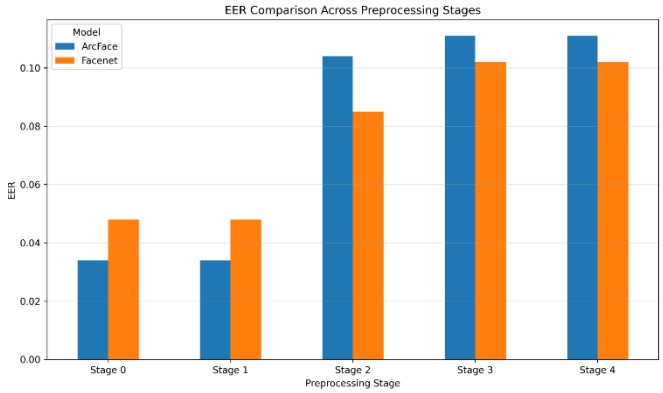

Fig. 3. EER comparison of FaceNet and ArcFace across five preprocessing stages.

E. Effect of Gaussian Blur and Normalization

FaceNet's AUC declined to 0.9528 and EER rose to 0.1020, while ArcFace recorded AUC=0.9263 and EER=0.1110. Stage 4 (pixel normalization) produced no measurable change from Stage 3 for either model across all four metrics, indicating that normalization alone does not influence the cosine similarity computation once blurring has been applied. As shown in Fig. 4, the error rate curves plateau between Stages 3 and 4 for both models, confirming the negligible contribution of this step under the current experimental configuration. Across Stages 2–4, ArcFace consistently recorded lower AUC and higher EER than FaceNet, indicating that

its sensitivity to preprocessing-induced distortions is greater than that of FaceNet.

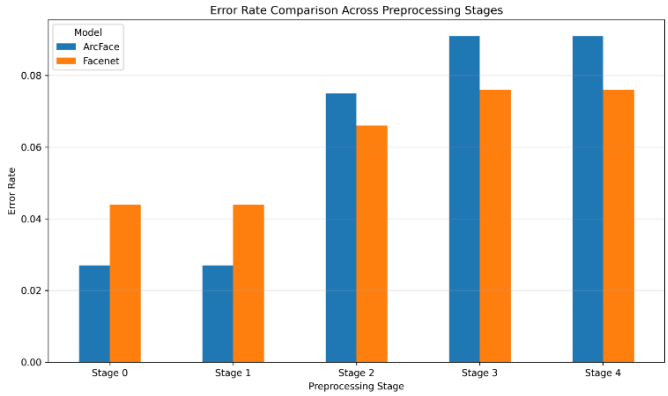

Fig. 4. Error rate comparison of FaceNet and ArcFace across five preprocessing stages.

F. ROC Curve Analysis

Fig. 5 presents the ROC curves for both models across all five preprocessing stages. At Stages 0 and 1, the curves of both model's cluster near the upper-left corner of the ROC space, confirming strong discriminative ability under minimal preprocessing conditions. Notably, the baseline ROC curves for ArcFace (AUC=0.9837) and FaceNet (AUC=0.9812) are closely aligned, yet ArcFace's curves descend more steeply as

preprocessing intensifies — a visual confirmation of its greater sensitivity to image manipulation relative to FaceNet.

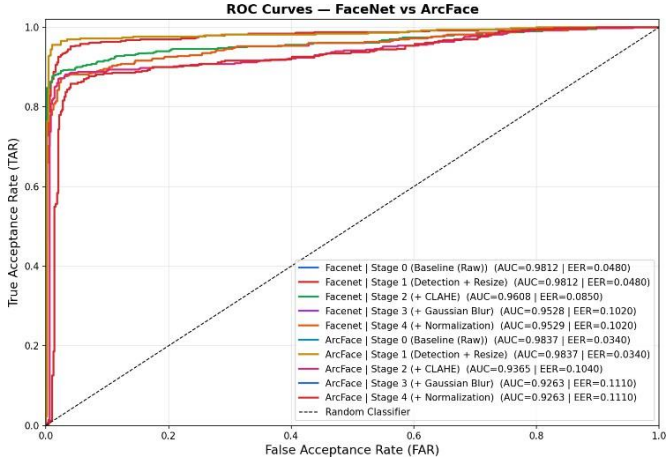

Fig. 5. ROC curves for FaceNet and ArcFace across all preprocessing stages on LFW (1,000 pairs).

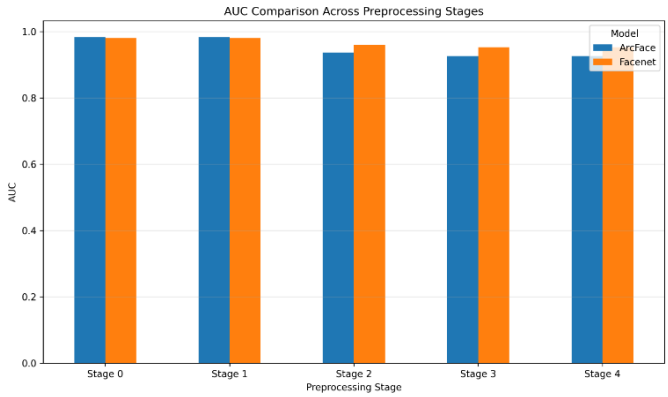

Fig. 6. AUC comparison of FaceNet and ArcFace across five preprocessing stages

TABLE VI. Performance Delta: Stage 0 vs Stage 4.

Model	ΔAUC	ΔEER	Accuracy	Best Stage
FaceNet	−0.0283	+0.0540	−3.20%	Stage 0 / 1
ArcFace	−0.0574	+0.0770	−6.40%	Stage 0 / 1

Table V and Fig. 2 collectively confirm that both models achieve peak performance on raw or minimally processed images. ArcFace, despite its superior baseline performance (AUC=0.9837, EER=0.0340), exhibited high sensitivity to preprocessing operations, with a total accuracy drop of 6.40% from Stage 0 to Stage 4 approximately twice the degradation observed for FaceNet (3.20%).

V. DISCUSSION

This section works through the key findings in relation to the questions this study set out to answer, places them alongside the , broader body of literature reviewed in Section II, and reflects honestly on what the results imply and where the study runs up against its own limitations.

A. Baseline Superiority of ArcFace and Its Theoretical Basis

The superior baseline performance of ArcFace (AUC=0.9837, EER=0.0340, Accuracy=97.30%) over FaceNet (AUC=0.9812, EER=0.0480, Accuracy=95.60%) on raw LFW images is consistent with theoretical expectations grounded in the architectural differences between the two models. ArcFace's Additive Angular Margin Loss enforces a larger inter-class angular margin on the hypersphere, which promotes tighter intra-class compactness and greater inter-class separability in the embedding space [4]. This geometric advantage directly translates into more reliable verification decisions at the cosine similarity level, particularly for face pairs exhibiting natural variation in pose and illumination conditions that characterize the LFW dataset [8]. This sits comfortably alongside what Ronquillo-Figueroa et al. found in their own work [15], who similarly observed ArcFace's superiority in classification accuracy and F1 score on LFW under controlled evaluation conditions, and corroborate the broader trend identified in the comprehensive review by Amirgaliyev et al. [6], putting ArcFace in strong company when measured against the best-reported results on LFW

by ROC-AUC.

B. Preprocessing-Induced Performance Degradation: An Unexpected Finding

One of the most notable findings of this study is that progressive image preprocessing consistently degraded the performance of both models rather than improving it. This runs counter to the assumption that tends to go unquestioned in the field that preprocessing helps by bringing input images into a more consistent, normalized state A plausible explanation lies in the pre-training data distribution of both models: FaceNet was trained on large-scale datasets such as VGGFace2, and ArcFace on MS-Celeb-1M both comprising images that include natural illumination variation without heavy contrast manipulation. when CLAHE is introduced at stage 2, it shifts the pixel distribution away from the statistical properties the networks were trained on and that shift is enough to disturb the internal feature representations both models had learned to rely on. This interpretation is consistent with the findings of Zhang [9], who demonstrated that image quality alterations even those intended as enhancements can adversely affect deep learning-based face recognition in unconstrained environments.

C. Differential Sensitivity: ArcFace vs. FaceNet

A critical observation is that ArcFace exhibited substantially greater sensitivity to preprocessing-induced distortions than FaceNet, with a total accuracy decline of 6.40% compared to 3.20% for FaceNet across Stages 0 to 4. Furthermore, the performance ranking reversed after Stage 2: while ArcFace outperformed FaceNet at baseline, FaceNet recorded lower EER values from Stage 2 onward (EER=0.0850 vs. 0.1040 at Stage 2; EER=0.1020 vs. 0.1110 at Stages 3 and 4). This differential sensitivity may be partially attributed to the higher dimensionality of ArcFace embeddings (512 dimensions vs. 128 for FaceNet), which may encode finer-grained texture and illumination features that are more susceptible to distortion when pixel statistics are altered. FaceNet's lower-dimensional embedding space, by contrast, may capture more abstract and invariant facial representations that are inherently more robust to surface-level image manipulations. This finding extends the observations of Juwanda et al. [13], who reported FaceNet's more stable performance in variable lighting conditions, by providing a quantitative account of this robustness within a controlled preprocessing framework.

D. Redundancy of Stage 1 and Stage 4

Two preprocessing stages yielded no measurable performance change relative to their predecessors. Stage 1 (face detection and resizing) produced results identical to Stage 0 for both models, confirming that the DeepFace framework applies MTCNN-based detection and resizing internally prior to embedding extraction — making external Stage 1 operations redundant in this configuration. Similarly, Stage 4 (pixel normalization) produced no change from Stage 3, suggesting that once Gaussian blurring has been applied, subsequent normalization does not influence the cosine similarity computation used for verification. These findings point to something that evaluation pipelines for pretrained face recognition models often fail to account for: the preprocessing that happens at the framework level, quietly and automatically, before the model ever sees the image. When those implicit steps are not acknowledged and controlled for, they can distort the picture either masking the effect that an experimental preprocessing stage was meant to produce or introducing confounds that make it harder to know what is actually driving the results.

E. Comparison with Prior Work

The baseline accuracy of ArcFace (97.30%) reported in this study is broadly consistent with values reported in prior comparative evaluations on LFW, including Firmansyah et al. [12], who

reported ArcFace accuracy of 87.8% on a domain-specific dataset — a lower figure attributable to dataset characteristics rather than model capability — and Ronquillo-Figueroa et al. [15], whose AHP-based evaluation placed ArcFace ahead of FaceNet on accuracy and F1 across multiple datasets. The baseline FaceNet accuracy (95.60%) is likewise aligned with published benchmarks, though lower than the 99.63% originally reported by Schroff et al. [3] — a discrepancy expected given the difference in the number of test pairs and the absence of fine-tuning in our evaluation. The five-year comparative analysis by Iype and Sebastian [14] similarly reported FaceNet's relative resilience under varying photometric conditions, a pattern corroborated by our preprocessing sensitivity results.

F. Practical Implications

First, for systems where image quality can be controlled such as access control environments with fixed cameras and lighting ArcFace may represent the preferred choice, given its superior baseline performance. Second, for systems deployed in unconstrained environments where images are subject to external processing pipelines (e.g., video surveillance, forensic identification), FaceNet may be the more prudent choice due to its greater robustness to preprocessing-induced distortions. Third, the results also carry a warning that is worth stating plainly: applying standard image enhancement techniques to

face images before feeding them into a deep learning verification system is not something that should be done on autopilot. The intuition that better-looking images will produce better results does not always hold, and in some cases these techniques actively work against the model rather than helping it. Anyone designing such a system would do well to test the effect of their preprocessing pipeline on the specific model they are using, rather than assuming that what works in general will work here. There are no safe defaults in this regard only choices that have been validated and choices that have not.

G. Limitations and Future Work

No study of this kind is without its constraints, and it is worth being straightforward about where this one falls short. The evaluation was built around a subset of 1,000 LFW pairs. That sample size was adequate for the comparative purposes the study had in mind, but it does not capture the full statistical variability that comes with running experiments across the complete LFW benchmark of 6,000 pairs — and that gap is worth keeping in mind when interpreting the results. Second, both models were used in their pre-trained state without fine-tuning on LFW, which may limit the absolute accuracy figures reported relative to fully optimized systems. Third, the preprocessing pipeline examined in this study represents one of many possible configurations; other techniques such as super-resolution, facial landmark-based alignment, or domain-adaptive normalization may yield different outcomes and warrant investigation in future work. Finally, the evaluation was limited to two models; extending the framework to include other state-of-the-art architectures such as AdaFace or MagFace would provide a more comprehensive characterization of preprocessing sensitivity across the current landscape of face recognition models.

VI. CONCLUSION

This paper presented a systematic comparative evaluation of two state-of-the-art face recognition models FaceNet and ArcFace under a progressive image preprocessing framework applied to the LFW benchmark dataset. The study addressed a gap identified in the existing literature: the absence of a controlled experimental protocol that treats image preprocessing as an independent variable and employs threshold-independent metrics, specifically ROC curves and the Equal Error Rate (EER), as the primary basis for comparison.

The experimental results yielded three principal conclusions. First, ArcFace demonstrated superior performance over FaceNet under baseline conditions (raw images), achieving an accuracy of 97.30%, an AUC of 0.9837, and an EER of 0.0340 compared to FaceNet's 95.60%, 0.9812, and 0.0480, respectively. This advantage may be attributed to the geometric properties of its Additive Angular Margin Loss, which enforces tighter intra-class compactness and greater inter-class separability on the unit hypersphere.

Second, and most significantly, progressive preprocessing consistently degraded the performance of both models across all stages, with no preprocessing technique yielding measurable improvement over the baseline. This finding pushes back against an assumption that often goes unquestioned in the field that enhancing an image will naturally make it easier for a model to

work with. What the results actually show is that pre-trained models can be remarkably particular about their inputs shift the images far enough from what the model encountered during training, and performance can degrade in ways that are neither obvious nor easy to predict.

Push the images too far from that distribution, even in ways that seem like improvements, and performance can quietly deteriorate rather than improve.

Third, ArcFace exhibited substantially greater sensitivity to preprocessing-induced distortions than FaceNet, with a total accuracy decline of 6.40% compared to 3.20% across Stages 0 to 4. From Stage 2 onward, the performance ranking reversed: FaceNet recorded lower EER values than ArcFace in all subsequent stages, suggesting that FaceNet's lower-dimensional embedding space encodes more preprocessing-invariant facial representations. These findings indicate that the choice between the two models should be guided not solely by baseline accuracy, but by the specific preprocessing conditions of the target deployment environment.

From a practical standpoint, the findings point toward different deployment contexts for each model. ArcFace is the stronger choice when images can be kept clean and consistent controlled environments where the integrity of the input is preserved from capture through to recognition. FaceNet, by contrast, shows more resilience when images have been through external processing before they reach the model.

There is also a cautionary note worth taking seriously for anyone designing these systems: standard image enhancement techniques histogram equalization, Gaussian filtering, and similar operations should not be applied to face images as a matter of routine. The assumption that cleaner or smoother images will naturally produce better recognition results does not hold across the board. Before introducing any such step into a pipeline, its effect on the specific model being used should be validated empirically. What helps one model may quietly hurt another.

Future work should extend this evaluation to the full LFW 6,000-pair protocol and to additional architectures such as AdaFace and MagFace, explore the effect of alternative preprocessing strategies including super-resolution and domain-adaptive normalization, and investigate whether fine-tuning on domain-specific data can recover performance lost through preprocessing.

REFERENCES

[1] S. Minaee et al., "Biometrics recognition using deep learning: A survey," *Artificial Intelligence Review*, 2023.vol. 56, pp. 8647–8695, doi: 10.1007/s10462-022-10237-x

[2] A. J. Shepley, "Deep learning for face recognition: A critical analysis," arXiv preprint, arXiv:1907.12739, 2019.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proc. IEEE Conf. Computer Vision and Pattern

Recognition (CVPR), Boston, MA, USA, Jun. 2015, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, Jun. 2019, pp. 4690–4699, doi: 10.1109/CVPR.2019.00482.

[5] X. Wang, J. Peng, S. Zhang, B. Chen, Y. Wang, and Y. Guo, "A survey of face recognition," arXiv preprint, arXiv:2212.13038, Dec. 2022, doi: 10.48550/arXiv.2212.13038.

[6] 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, p. 9390, Aug. 2025, doi: 10.3390/app15179390.

[7] D. Tolegenova and A. Moldagulova, "Comparative analysis of face biometric identification methods: From traditional algorithms to modern deep learning models," Computing & Engineering, vol. 3, no. 2, pp. 34–40, Jun. 2025, doi: 10.51301/ce.2025.i2.06. [Online]. Available: https://ce.journal.satbayev.university/index.php/journal/article/view/133 5

[8] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2007. [Online]. Available: http://vis-www.cs.umass.edu/lfw/

[9] BenHusein, A. H., Shakrum, F. M., & Elgdiri, E. M. (2026). A Comprehensive Performance Evaluation of a Wi-Fi-Based Wireless Sensor Network Using Raspberry Pi and ESP8266 Under Multi-Rate Traffic Conditions. *Al-Farooq Journal of Sciences*, *2*(3), 816-818. https://doi.org/10.65405/m5hagm61

[10] N. Zhang, "A study on the impact of face image quality on face recognition in the wild," arXiv preprint, arXiv:2307.02679, Jul. 2023, doi: 10.48550/arXiv.2307.02679.

[11] S. Kilany and A. Mahfouz, "A comprehensive survey of deep face verification systems adversarial attacks and defense strategies," Scientific Reports, vol. 15, art. no. 30861, Aug. 2025, doi: 10.1038/s41598-025-15753-8.

[12] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Computer Vision and Image Understanding*, vol. 189, p. 102805, 2019.

[13] A. M. Firmansyah, T. F. Kusumasari, and E. N. Alam, "Comparison of face recognition accuracy of ArcFace, FaceNet and FaceNet512 models on DeepFace framework," in *Proc. Int. Conf. Computer Science, Information Technology and Engineering (ICCoSITE)*, 2023, pp. 535–539, doi: 10.1109/ICCoSITE57641.2023.10127799.

[14] R. C. Juwanda, F. A. Alunjati, U. Elviani, and F. Hidayat, "Comparative analysis of FaceNet and ArcFace in minimizing false positives for enhanced access control security," in *Proc. Int. Conf. ICT for Smart Society (ICISS)*, 2024, doi: 10.1109/iciss62896.2024.10750931.

[15] J. K. Iype and S. Sebastian, "Comparative analysis of state-of-the-art face recognition models: FaceNet, ArcFace, and OpenFace using image classification metrics," in S. Aurelia et al. (Eds.), Computational Sciences and Sustainable Technologies, Communications in Computer and Information Science, vol. 1973, pp. 222–234, Springer, Cham, 2024, doi: 10.1007/978-3-031-50993-3_18.

[16] C. Ronquillo-Figueroa, M.-A. Quiroz-Martinez, and M.-D. Gomez-Rios, "Comparative evaluation of ArcFace and FaceNet models for real-world facial recognition using the analytic hierarchy process," in *Lecture Notes in Networks and Systems*, Cham, Switzerland: Springer, 2026, doi: 10.1007/978-3-031-87065-1_8.