EASE-PVNet: Robust Periocular Identity Verification Across Pre- and Post-Operative Facial Images

Ziyad Azzaz , Omar Mohamed , Esraa Khatab , Hany Said , Omar Shalash *

Posted Date: 9 May 2026

doi: 10.20944/preprints202605.0567.v1

Keywords: periocular biometrics; post-operative identity verification; siamese networks; autoencoder pretraining; ensemble learning; hard-negative mining; explainable AI

 doi:10.20944/preprints202605.0567.v1

Article

EASE-PVNet: Robust Periocular Identity Verification Across Pre- and Post-Operative Facial Images

Ziyad Azzaz [1], Omar Mohamed [1], Esraa Khatab [2], Hany Said [1] and Omar Shalash [3]*

[1] College of Artificial Intelligence, Arab Academy for Science, Technology & Maritime Transport
[2] School of Mathematics and Computer Science Herriot Watt University
[3] Artificial Intelligence Research Center (AIRC), College of Engineering and Information Technology, Ajman University, Ajman P.O. Box 346, United Arab Emirates
* Correspondence: o.shalash@ajman.ac.ae

Abstract

Identity verification across pre-operative and post-operative facial images remains a challenging task, particularly following eyelid surgery, where localized periocular changes can disrupt conventional face recognition systems. This research introduces a novel verification framework using an ensemble-based autoencoder-initialized siamese eye-region periocular verification network designed to remain resilient to surgically induced appearance variation. The proposed approach integrates anatomy-guided periocular normalization with a Siamese deep metric learning architecture initialized through unsupervised autoencoder pretraining, allowing the model to acquire periocular-specific representations prior to supervised learning. Robustness in this data-limited clinical setting is further enhanced through staged hard-negative mining, validation-weighted multi-seed ensemble learning, and bootstrap-based threshold calibration. Ensemble Grad-CAM is employed to provide visual explanations that support clinical interpretability. Experimental evaluation demonstrates strong and consistent performance, achieving recognition rates of 94.71% on training data, 96.77% on validation, and 96.08% on the test set, with an overall recognition rate of 95.24%. Compared to previously reported periocular verification methods which reported an overall recognition rate of only 91.8% under similar conditions. These results highlight the effectiveness and stability of the proposed framework for post-surgical periocular identity verification in clinical and forensic applications.

Keywords: periocular biometrics; post-operative identity verification; siamese networks; autoencoder pretraining; ensemble learning; hard-negative mining; explainable AI

1. Introduction

Artificial intelligence has increasingly shaped the evolution of facial analysis in security, clinical, and forensic applications, particularly in settings where appearance variability challenges conventional methods [1–3]. Traditional biometric systems, which rely on handcrafted features and fixed similarity metrics, tend to degrade rapidly when facial structure is altered by aging, trauma, or surgical intervention. By contrast, AI-driven approaches—most notably those based on deep learning—offer the flexibility to learn identity-relevant patterns directly from data, making them far better suited to handling complex and localized facial changes [4–6].

In medical and surgical contexts, AI has enabled more objective and reproducible analysis of facial imagery than was previously possible [7]. Deep neural networks can capture subtle anatomical cues while remaining tolerant to non-rigid deformations, illumination differences, and acquisition variability. This capability is particularly important in perioperative scenarios, where localized surgical modifications may significantly alter texture and geometry without fully obscuring underlying identity. From a practical standpoint, AI-based systems provide a level of consistency that is difficult to achieve through manual assessment or classical algorithmic pipelines [8]. Another important contribution of AI lies in addressing data limitations that are common in clinical environments. Large, fully annotated

 doi:10.20944/preprints202605.0567.v1

surgical datasets are rarely available due to privacy concerns and ethical constraints. Techniques such as unsupervised representation learning, encoder pretraining, and transfer learning allow models to leverage unlabeled data and prior knowledge, reducing reliance on extensive manual annotation. Ensemble strategies further enhance robustness by mitigating the effects of random initialization and optimization variability, which can otherwise have a disproportionate impact in data-scarce settings [9,10].

Beyond performance alone, the role of AI in facial analysis has expanded to emphasize transparency and trustworthiness. In applications such as post-surgical identity verification, incorrect decisions can carry meaningful clinical or legal consequences [11–13]. As a result, there is growing demand for systems that not only produce accurate predictions but also offer insight into the reasoning behind them. Explainable AI techniques, including gradient-based saliency visualization, make it possible to inspect where models focus their attention and to verify that decisions are driven by anatomically meaningful structures rather than spurious artifacts [14–16].

Taken together, these advances have shifted facial analysis away from rigid, feature-engineered systems toward adaptive, data-driven frameworks capable of operating reliably under real-world variability. In the context of post-surgical facial and periocular verification, AI has contributed the essential tools needed to balance discrimination, robustness, and interpretability. The approach proposed in this research builds directly on these developments by integrating unsupervised representation learning, metric-based verification, ensemble modeling, and explainable inference into a unified framework designed for surgically altered periocular imagery.

The main contributions of this research are fourfold. First,EASE-PVNet was proposed, a domain-specific periocular identity verification framework explicitly designed to address the challenging problem of matching pre-operative and post-operative eyelid surgery images, where conventional face recognition methods systematically fail. Second, an autoencoder-initialized Siamese verification architecture was introduced that leverages unsupervised periocular representation learning prior to metric supervision, improving robustness in data-limited clinical settings. Third, an enhanced verification stability through a combination of staged hard-negative mining was proposed, validation-weighted multi-seed ensemble learning, and bootstrap-based threshold calibration, resulting in consistent and well-balanced biometric performance across training, validation, and test splits. Finally, an integrated ensemble Grad-CAM explainability layer was added to provide anatomically meaningful visual evidence supporting each verification decision, improving transparency and clinical interpretability. Together, these contributions form a robust, interpretable, and application-ready solution for post-surgical periocular identity verification in clinical and forensic contexts.

1.1. Related Work

In 2022, Sabharwal et al. proposed a deep feed-forward neural network with a Hessian-trace-based weight update to recognize faces altered by plastic surgery. Their method reduces computational complexity by optimizing hidden layers without requiring GPU environments. On a surgical facial dataset, it achieved a recognition rate of 96.40% and an F-score of 1.00, outperforming traditional methods like PCA (29.70%) and SURF (49.60%), while reaching near-perfect results for Rhinoplasty (98.24%) and skin peeling (97.89%) [17].

In 2021, Hayasaka et al. developed an artificial intelligence model using a convolutional neural network (CNN) to classify tracheal intubation difficulty from facial images in 202 adult surgical patients. Sixteen different facial images were taken per patient and labeled as “easy” (Cormack-Lehane I-II) or “difficult” (Cormack-Lehane III-IV) based on actual intubation outcomes. The best-performing model used supine-side-closed mouth-base position images, achieving an accuracy of 80.5%, sensitivity of 81.8%, specificity of 83.3%, and an area under the curve (AUC) of 0.864 (95% CI: 0.731–0.969). Class activation heat maps revealed that the model focused on the neck region, particularly from the chin tip to the larynx, to identify easy intubations. This study benefits the field by being the first to apply deep learning for intubation difficulty classification, potentially providing an objective, rapid assessment tool to assist inexperienced medical staff in emergency situations [18].

 doi:10.20944/preprints202605.0567.v1

In 2025, Jerjes et al. prospectively evaluated the diagnostic accuracy of in vivo optical coherence tomography (OCT) for detecting, subtyping, and assessing margins of facial basal cell carcinomas (BCCs) in 136 patients with 220 suspicious lesions, using histopathology as the gold standard. OCT demonstrated excellent diagnostic performance, with a sensitivity of 96.8%, specificity of 98.2%, and accuracy of 97.5% (AUC = 0.97). Subtype sensitivity was highest for superficial (93.1%) and nodular BCC (92.1%), slightly lower for micronodular (89.3%) and infiltrative (90.0%) subtypes. OCT-derived tumour depth correlated strongly with histopathology (2.3 ± 0.9 mm vs. 2.2 ± 0.8 mm; p = 0.08). The study benefits the field by confirming OCT as a reliable, non-invasive preoperative tool for facial BCC management, potentially reducing unnecessary excisions and improving surgical planning, particularly in cosmetically sensitive areas [19].

2. Methodology

This section presents the periocular identity verification framework developed for the clinically demanding problem of matching subjects across pre-operative and post-operative eyelid surgery images. Surgical modification of the periorbital region introduces localized but substantial appearance changes that systematically undermine standard face recognition approaches, motivating a domain-specific solution. The proposed system is built around three foundational design decisions: first, the encoder is initialized through unsupervised autoencoder pretraining on the unlabeled periocular collection before any supervised signal is introduced, so that the shared feature extractor begins metric learning with periocular-aware representations rather than random weights; second, the pretrained encoder is transferred directly into a Siamese verification network and fine-tuned end-to-end under a contrastive objective; and third, five such networks are trained independently under different random seeds and their outputs are fused through a validation-weighted ensemble, reducing the sensitivity of the final prediction to any single optimization trajectory. These three decisions, together with anatomy-guided normalization, curriculum hard-negative mining, bootstrap-stabilized threshold estimation, and gradient-weighted explainability, constitute the complete pipeline described in the subsections below.

2.1. Overview of the Proposed Pipeline

The full system is depicted in Figure 1. Given an input pair comprising one pre-operative and one post-operative facial image, the pipeline proceeds through four sequentially coupled blocks. Block 1 performs anatomy-guided periocular region localization, normalization, and augmentation, yielding two standardized 192×96 ROI images. Block 2 passes both ROIs through five independently trained Siamese encoders with shared weights, each producing a 256-dimensional ℓ_2-normalized embedding pair and a cosine similarity score. Block 3 aggregates the five scores through validation-weighted fusion and applies a bootstrap-stabilized threshold to produce a binary decision. Block 4 generates gradient-weighted saliency maps via ensemble Grad-CAM to support qualitative interpretability of each decision.

 doi:10.20944/preprints202605.0567.v1

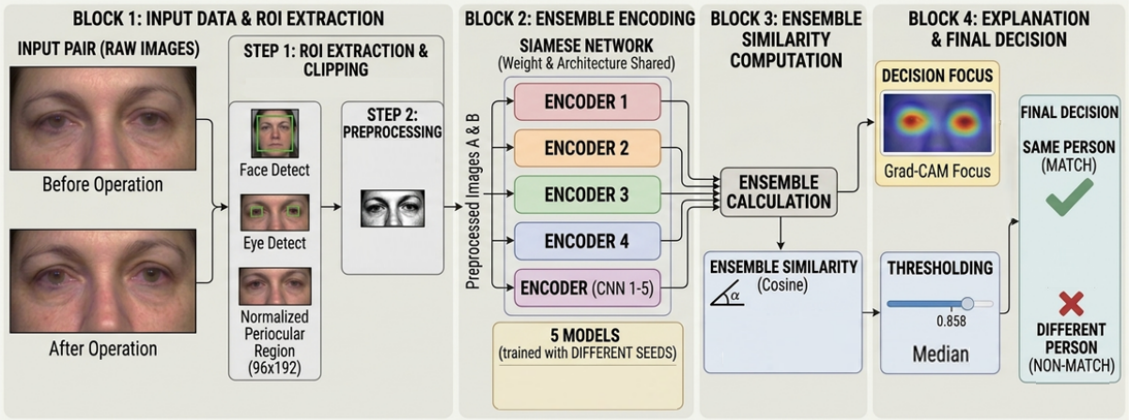

Figure 1. Overview of our proposed architecture EASE-PVNet four-block periocular verification pipeline.

2.2. Dataset Protocol and Verification Pair Generation

Images were organized using a subject-wise naming convention in which suffix *a* denotes the post-operative acquisition and suffix *b* the pre-operative acquisition. The complete corpus comprised 261 facial images from 131 subjects. Following preprocessing and quality control, 256 normalized periocular samples were retained; 5 samples were excluded due to detection failure, as described in Section 2.3. Genuine pairs were formed from the pre-operative and post-operative images of the same subject, while impostor pairs combined a pre-operative image from one subject with the post-operative image of another. The number of impostor pairs was matched exactly to the number of genuine pairs at each experimental seed to prevent class-imbalance bias during training. The resulting pair set was partitioned with stratified sampling into training, validation, and testing subsets of 170, 31, and 51 pairs, respectively, preserving the genuine-to-impostor ratio across all three splits.

2.3. Periocular ROI Extraction, Normalization, and Augmentation

Reproducible periocular localization is especially critical in the surgical context, where face-level appearance may differ substantially between sessions. Face detection was performed using a frontal Haar cascade applied to the full image. Eye candidates were searched within the detected face crop, and when two geometrically consistent eye centers were identified, the inter-eye axis was used to estimate in-plane rotation and align the image to canonical orientation. A periocular window centered on the inter-eye midpoint was then extracted using dimensions scaled proportionally to the inter-eye distance, ensuring consistent anatomical coverage across subjects with varying facial scales, and resampled to 192×96 pixels. When bilateral detection yielded insufficient confidence, a fallback crop of the upper face region was extracted from the bounding box. In both pathways, contrast-limited adaptive histogram equalization (CLAHE) was applied prior to resizing to enhance local periocular texture visibility while compensating for illumination differences between sessions. A sample was accepted only if it satisfied minimum sharpness and contrast criteria, operationalized as the variance of the Laplacian and the normalized grayscale standard deviation, respectively.

Data augmentation was applied to training pairs immediately after normalization and before any subsequent processing stage. The augmentation policy included stochastic brightness and contrast perturbation, additive Gaussian noise, Gaussian blur, small-angle in-plane rotation, and limited spatial translation. Transformation magnitudes were deliberately constrained to preserve periocular identity cues while exposing the model to the photometric and geometric variability characteristic of clinical acquisition conditions.

2.4. Autoencoder-Based Encoder Pretraining

Training a Siamese verification network from randomly initialized weights on a limited clinical dataset risks encoder underfitting, as the shared backbone must learn discriminative periocular representations from a small number of labeled pairs. To address this, the encoder was first trained

 doi:10.20944/preprints202605.0567.v1

through unsupervised reconstruction, as illustrated in Figure 2. Each normalized periocular image was compressed into a compact latent code and subsequently reconstructed by a symmetric decoder. By minimizing pixel-wise reconstruction error over the full unlabeled periocular collection, the encoder was exposed to periocular structure, local texture continuity, and illumination-robust spatial patterns before any pairwise supervision was introduced, providing a substantially more informative initialization than random weights.

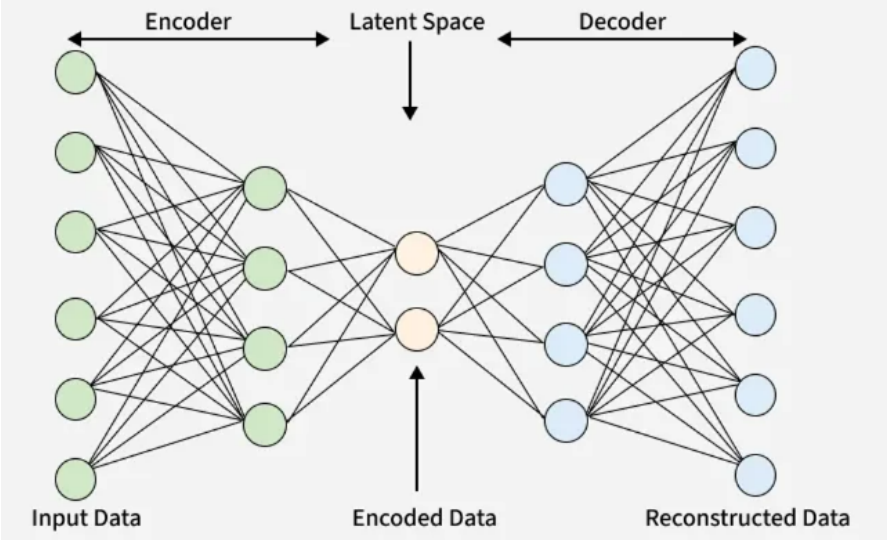

Figure 2. Convolutional autoencoder used for unsupervised encoder pretraining prior to Siamese transfer.

The encoder comprised four convolutional blocks with channel progression $3 \to 32 \to 64 \to 128 \to 256$, each consisting of a 3×3 convolution with stride 2, batch normalization, and ReLU activation. The decoder mirrored this structure with transposed convolutions and a sigmoid output layer. Our proposed architecture for autoencoder was trained for 30 epochs using Adam [20] with a learning rate of 10^{-3} and mean squared error as the reconstruction objective. The checkpoint achieving the lowest validation reconstruction loss was retained for weight transfer. The autoencoder was used exclusively for encoder initialization and performed no generative role during verification inference.

2.5. Siamese Verification Architecture

Following pretraining, the encoder weights were transferred to our proposed architecture for Siamese verification network shown in Figure 3. The two branches share identical weights and process each periocular ROI independently, producing spatial feature maps that are flattened and passed through a three-layer fully connected verification head.

The verification head projects features through dimensions of 1024, 512, and 256, with batch normalization and ReLU following the first two layers. Dropout is applied at rates of 0.35 and 0.25 after the first and second layers, respectively; the higher rate at the first layer, which handles the highest-dimensional representation, applies stronger regularization where overfitting risk is greatest, while the reduced rate at the second layer preserves the discriminative structure progressively built up by the network. The head output is ℓ_2-normalized to yield a unit embedding in $\mathbb{R}^{256}$; the 256-dimensional space balances representational capacity against the risk of sparse coverage when training data is limited.

Given two normalized embeddings $\mathbf{e}_1, \mathbf{e}_2 \in \mathbb{R}^{256}$, the verification score is the cosine similarity:

$$s(\mathbf{e}_1, \mathbf{e}_2) = \frac{\mathbf{e}_1^\top \mathbf{e}_2}{\|\mathbf{e}_1\|_2 \, \|\mathbf{e}_2\|_2}. \tag{1}$$

 doi:10.20944/preprints202605.0567.v1

Network parameters were optimized under the contrastive loss [21]:

$$\mathcal{L} = y\, d^2 + (1 - y)\, \big[\max(0,\, m - d)\big]^2, \tag{2}$$

where $y \in \{0, 1\}$ is the pair label, $d = \|\mathbf{e}_1 - \mathbf{e}_2\|_2$ is the Euclidean distance between normalized embeddings, and m is the margin. The margin was set to $m = 1.1$: for unit-normalized embeddings the maximum achievable distance is 2, so a margin of 1.1 places the repulsion boundary comfortably above the mid-range of the distance space, keeping the repulsive term actively informative for confusable impostor pairs without approaching the theoretical maximum at which the penalty would become trivially satisfied.

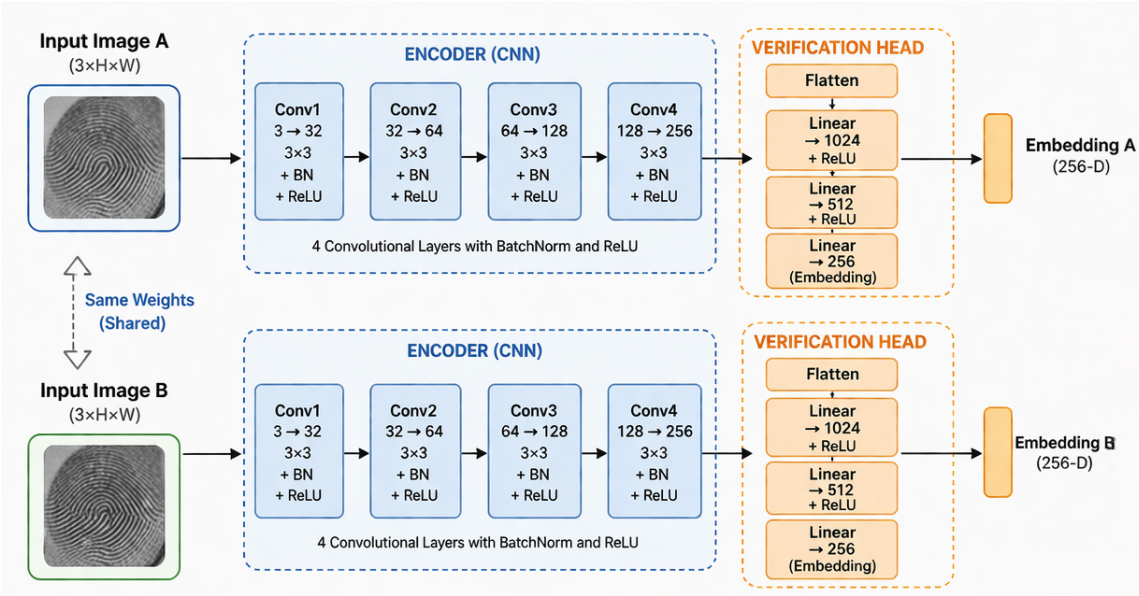

Figure 3. Our proposed architecture Siamese verification network with shared convolutional encoder and fully connected verification head producing 256-D normalized embeddings.

Optimization used AdamW [22] with separate learning rates for the head (2×10^{-4}) and the encoder (scaled by 0.2, yielding 4×10^{-5}). The lower encoder rate reflects the finer gradient adjustments appropriate for pretrained weights, preventing premature overwriting of the reconstruction-learned representation. Weight decay of 10^{-4} and a cosine annealing schedule were applied throughout, with gradient clipping at a maximum ℓ_2-norm of 5.0 to prevent instability during the early fine-tuning phase. The encoder was frozen for the first four epochs and unfrozen from epoch 5 onward, allowing the verification head to establish a coherent embedding geometry before joint end-to-end optimization began. Training ran for 55 epochs in total.

2.6. Staged Hard-Negative Mining

During standard contrastive training, randomly sampled impostor pairs are often easily separable and provide weak gradient signal once the model has learned coarse discrimination. To drive the embedding boundary toward fine-grained periocular discrimination, a curriculum-based hard-negative mining schedule was introduced.

The mechanism operates as follows. At each epoch, all negative training pairs are rescored using the current model state. Pairs that receive a high cosine similarity score—those whose embeddings are closest to the positive cluster in the shared space, and are therefore the most likely to be misclassified—are designated as hard negatives. These confusable pairs are then oversampled in proportion to their rank, mixed with randomly drawn negatives, and used to construct the training batch for the next optimization round. The effect is a form of iterative importance reweighting, as illustrated in Figure 4:

 doi:10.20944/preprints202605.0567.v1

pairs that the current model fails to discriminate acquire higher effective weight, directing gradient updates precisely where the decision boundary is weakest.

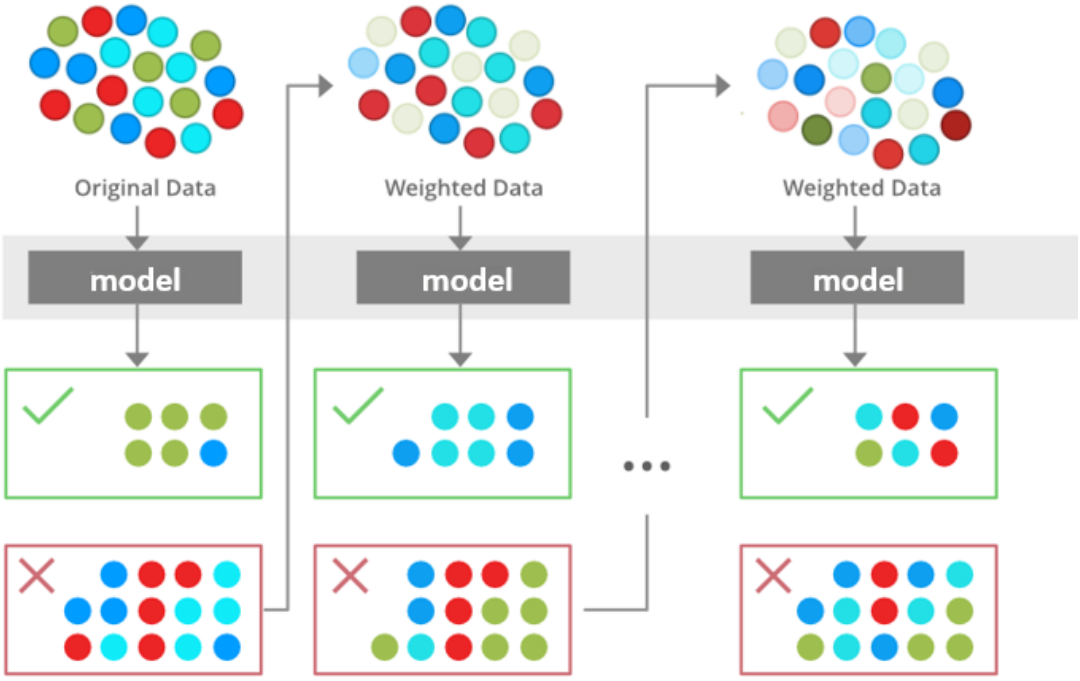

Figure 4. Iterative hard-negative reweighting: misclassified impostor pairs receive increasing sampling weight across successive training rounds.

To avoid the optimization instability caused by exclusive hard-negative training from initialization, the hard-negative fraction was increased gradually through three stages: 0.15 during the early phase, 0.35 in the middle phase, and 0.50 in the late phase. This graduated schedule ensures the model first develops broad discrimination before being progressively focused on the most confusable pairs, transitioning from coarse identity separation to fine-grained boundary refinement.

2.7. Weighted Multi-Seed Score-Level Ensemble

A single training run is subject to variability from stochastic weight initialization, data ordering, and local optimization trajectories. In the data-limited clinical setting considered here, this variability can produce non-trivial performance fluctuations that do not reflect the underlying model capability. To mitigate this, five independent Siamese networks were trained using the seed set $\mathcal{S} = \{42, 123, 777, 2026, 3407\}$, each following the identical preprocessing, architecture, and training protocol. All five models were evaluated on the common reference split defined by the first-seed checkpoint, ensuring that every member is scored on identical partitions.

For a given input pair, the k-th model produces a cosine similarity score $s_k \in [-1, 1]$. The ensemble score is computed as:

$$S_{\text{ens}} = \frac{\sum_{k=1}^{K} w_k \, s_k}{\sum_{k=1}^{K} w_k + \varepsilon}, \tag{3}$$

where $K = 5$, $\varepsilon = 10^{-8}$ ensures numerical stability, and each weight w_k is proportional to the validation recognition rate of the k-th member. Weighting by validation performance assigns greater influence

 doi:10.20944/preprints202605.0567.v1

to members that are more consistent on held-out data, without requiring model-level selection or additional calibration.

The rationale for selecting five ensemble members is evident from Table 1 and Figure 5. As cardinality increases from one to three models, all metrics improve substantially—recognition rate rises from 64.7% to 94.1% on the test set and EER drops from 11.54% to 7.69%—reflecting the high sensitivity of a single model to its initialization. From three to five models, gains are incremental and the five-model configuration consistently achieves the highest values across all reported metrics, confirming that the performance plateau has been reached. Increasing cardinality further would add computational overhead without meaningful performance benefit; five models therefore represent the operating point that balances stability against efficiency.

Table 1. Verification performance as a function of ensemble cardinality.

Models	Train RR	Val RR	Test RR	Overall RR	Test AUC	Test F1	Test EER
1	88.8%	87.1%	64.7%	83.7%	0.9262	72.7%	11.54%
2	87.0%	90.3%	66.7%	83.3%	0.9415	73.9%	11.54%
3	92.3%	93.6%	94.1%	92.9%	0.9615	94.1%	7.69%
4	94.1%	93.6%	94.1%	94.1%	0.9631	94.1%	3.85%
5	94.7%	96.8%	96.1%	95.2%	0.9692	96.0%	3.85%

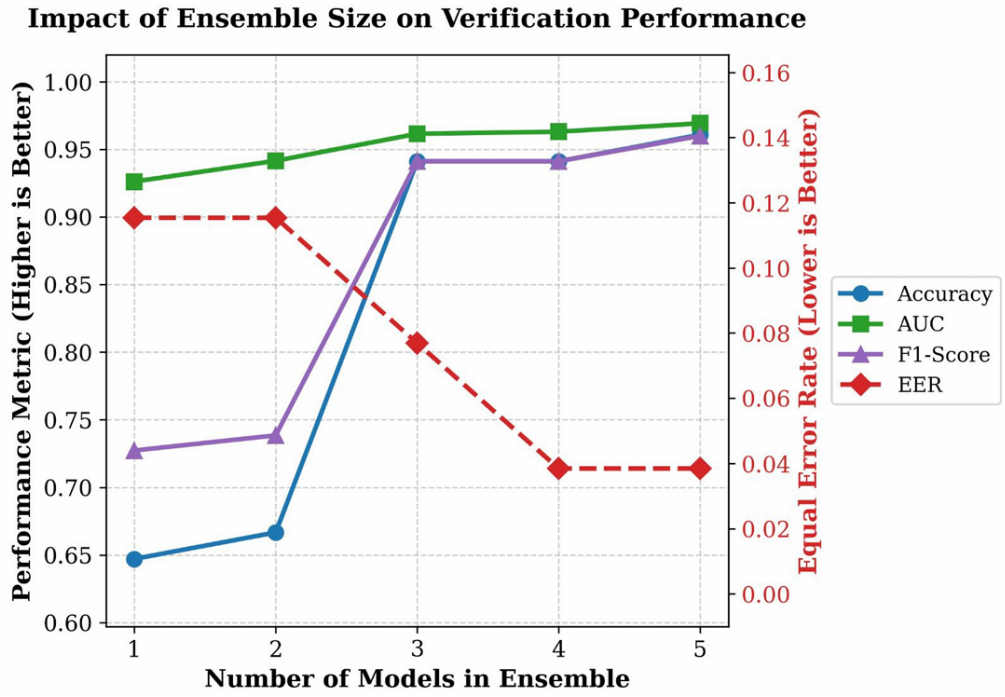

Figure 5. Impact of ensemble cardinality on accuracy, AUC, F1-score, and EER across model counts.

2.8. Threshold Estimation and Decision Rule

The operating threshold was estimated directly from validation-set scores of the weighted ensemble through a principled data-driven procedure, rather than being fixed heuristically. A dense search was conducted over the full range of observed validation scores, and each candidate threshold t was evaluated using:

$$\mathcal{J}(t) = 0.7\,\mathrm{ACC}(t) + 0.3\,F_1(t) - 0.15\,\mathrm{HTER}(t), \tag{4}$$

where $\mathrm{HTER}(t) = (\mathrm{FAR}(t) + \mathrm{FRR}(t))/2$. The dominant weight on accuracy (0.7) reflects the primary verification objective; the F_1 term (0.3) preserves sensitivity to class-balanced recognition on the small validation set, where a threshold optimized purely on accuracy could silently favor one class; and the

 doi:10.20944/preprints202605.0567.v1

negative HTER term (-0.15) applies a mild but explicit penalty against operating points with strongly asymmetric false acceptance and false rejection rates, which would be clinically undesirable.

To reduce sensitivity to the specific composition of the single validation partition, the search was repeated over 200 non-parametric bootstrap resamples drawn with replacement from the validation scores. The threshold maximizing $\mathcal{J}(t)$ was recorded for each resample, and the final operating threshold t^* was defined as the median of the resulting empirical distribution. The median was chosen over the mean because it is robust to the occasional outlier bootstrap sample that may arise from the small validation set. The binary verification decision on the test set was then:

$$\hat{y} = \begin{cases} 1, & S_{\text{ens}} \geq t^*, \\ 0, & S_{\text{ens}} < t^*, \end{cases} \tag{5}$$

where $\hat{y} = 1$ denotes a same-identity match and $\hat{y} = 0$ a different-identity non-match. The threshold t^* was fixed entirely from validation data and applied without modification to the test set, ensuring a clean separation between model selection and final evaluation.

2.9. Gradient-Weighted Explainability

Clinical deployment of identity verification systems requires not only high accuracy but also the capacity to expose the spatial evidence underlying each decision, enabling domain experts to audit model behavior and identify failure modes. To serve this role, the pipeline incorporates an ensemble Grad-CAM [23] module that produces a pair of spatially registered saliency overlays—one per input image—highlighting the periocular regions that most influenced the similarity score. Per-model maps are fused using the same validation-derived weights w_k adopted in score fusion, so that models contributing more to the verification decision also contribute more to the explanation. A representative output is shown in Figure 6, illustrating the decision heatmaps and the top discriminative points identified for both images in a matched pair.

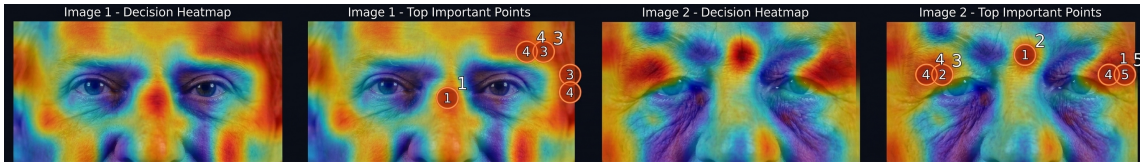

Figure 6. Ensemble Grad-CAM output for a matched pair: decision heatmaps and top discriminative points for Image 2 (pre-operative) and Image 1 (post-operative).

This module is strictly a qualitative analysis component. It does not participate in parameter optimization, threshold selection, or ensemble-member weighting, and its output is not used to modify any upstream decision. Its sole function is to provide visual evidence that model attention is concentrated on anatomically meaningful periocular structures—eyelid contours, canthal regions, and periorbital texture—rather than on background content or ROI boundary artifacts, thereby supporting clinical confidence in the system's decision basis.

2.10. Evaluation Protocol

System performance will be assessed across three dimensions. Verification accuracy is evaluated using recognition rate, accuracy, precision, recall, macro F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Security robustness is characterized by false acceptance rate (FAR), false rejection rate (FRR), equal error rate (EER), and half total error rate (HTER). System stability and efficiency are assessed through cross-seed variability of accuracy and AUC across the five ensemble members, score-distribution separation between genuine and impostor pairs, total test-set inference time, and mean per-pair latency. All verification metrics are computed on the training, validation, and test splits separately, with the threshold fixed at t^* as estimated from validation data.

 doi:10.20944/preprints202605.0567.v1

3. Results

3.1. Core Classification and Biometric Security Performance

Table 2 summarize classification and biometric security metrics across training, validation, and test, while Figure 8 show the same metrics across all splits.

Table 2. Performance and Security Metrics Across Data Splits

Category	Metric	Training	Validation	Testing
Performance	Accuracy (ACC)	0.9471	0.9677	0.9608
	Precision	0.9872	0.9412	0.9600
	Recall (Sensitivity)	0.9059	1.0000	0.9600
	F1-Score	0.9448	0.9697	0.9600
	AUC-ROC	0.9871	0.9833	0.9692
Security / Error	EER	0.0353	0.0667	0.0385
	FAR	0.0118	0.0667	0.0385
	FRR	0.0941	0.0000	0.0400
	HTER	0.0529	0.0333	0.0392

On test set, ensample achieves an accuracy 96.08%, with precision, recall, and F1-score all at 96.00%, indicating a strong balanced system with no meaningful asymmetry between false acceptance and false rejection. The AUC-ROC is 96.92% show a strong rank between genuine (pre-surgery) and imposter (after-surgery) pairs it illustrated by ROC curve in Figure 7 , The three ROC Curves training, validation, and test rise up sharply in top-left corner. From the biometric security, the system achieves an EER of 3.85%, a FAR of 3.85%, an FRR of 4.00%, and an HTER of 3.92% on the test set. The semi-similarity between FAR and FRR demonstrates how well the bootstrap threshold calibration in balancing the two error types. The HTER of 3.92% is highly significant in the post-surgical verification contexts, involving both unauthorized access and false rejection of legitimate patients carries real-world consequences that are roughly equal in costly

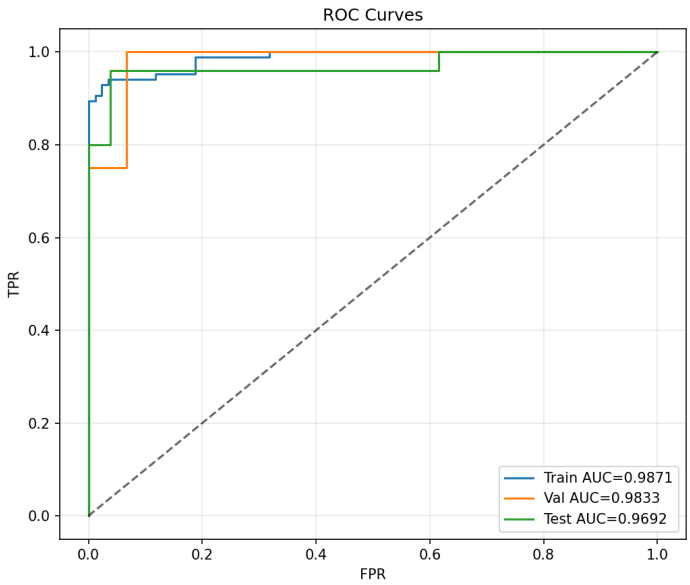

Figure 7. Receiver Operating Characteristic (ROC) curve for training, validation, test.

 doi:10.20944/preprints202605.0567.v1

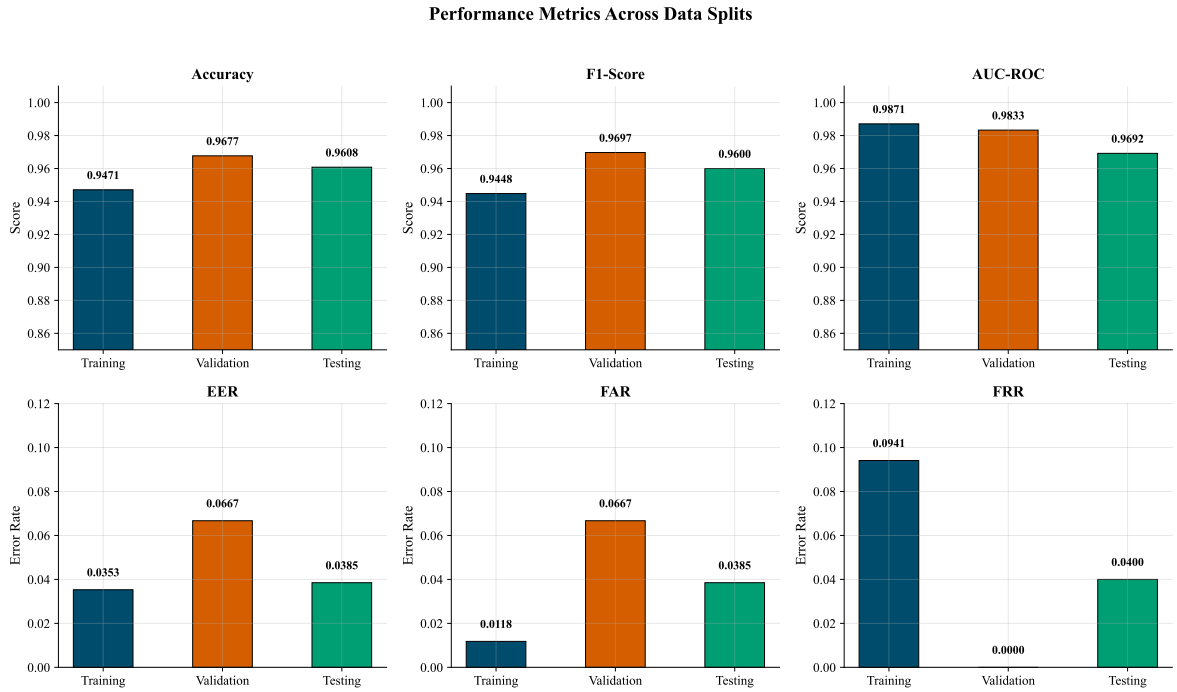

Figure 8. Summary of all performance and security metrics across training, validation, and testing splits. Upper row: accuracy, F1-score, and AUC-ROC. Lower row: EER, FAR, and FRR.

3.2. Confusion Matrix

Figure 9 presents the confusion matrix for the five-model ensemble on the test set. for 51 test pairs, the system correctly classifies 49, with one false acceptance and one false rejection. which describe a strong model can differentiate the genuine and imposter persons.

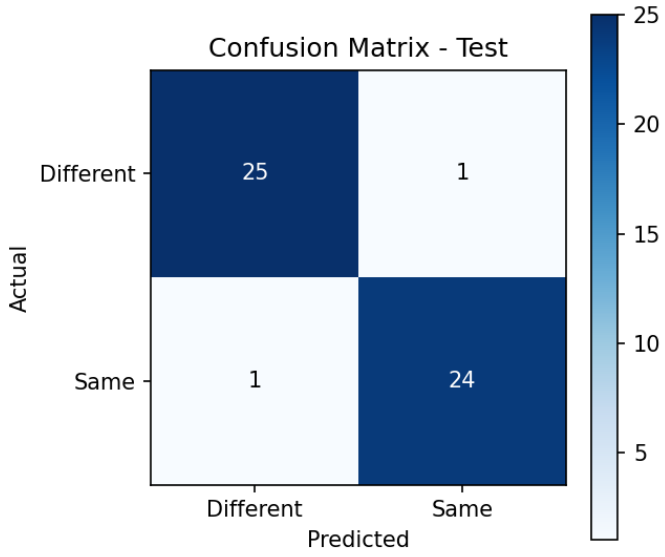

Figure 9. Confusion matrix on the test set.

3.3. Detection Error Tradeoff and Precision–Recall curve

Figure 10 presents the Detection Error Tradeoff (DET) curve for the test set, which is plotted on a logarithmic scale to provide enhanced resolution in the operationally critical low-error regions. the DET plot clearly illustrates the precise balance between the False Alarm Rate (FAR) and False Reject Rate (FRR).The Equal Error Rate (EER) operating point of 3.85% is explicitly marked, Also in

 doi:10.20944/preprints202605.0567.v1

Figure 11, Precision–Recall curve (PR) illustrates the balanced between precision and recall on the test set, PR curve shows that the model holds onto a very high precision even as we push for more matches, which means we aren't seeing a spike in false alarms just to get better coverage. Reaching an Average Precision of 0.9776 is a strong result, proving that the system stays dependable despite the visual challenges typical of post-surgical recovery.

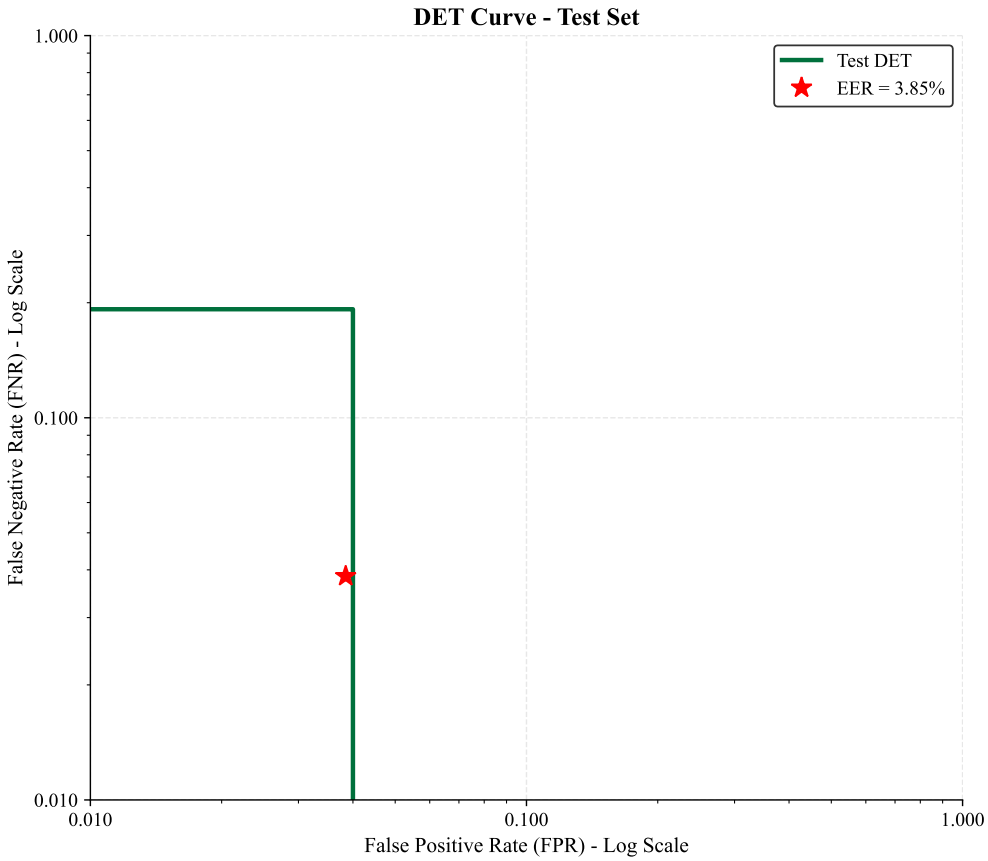

Figure 10. Detection Error Tradeoff (DET) curve on the test set. The star marker indicates the Equal Error Rate (EER ≈ 3.85%).

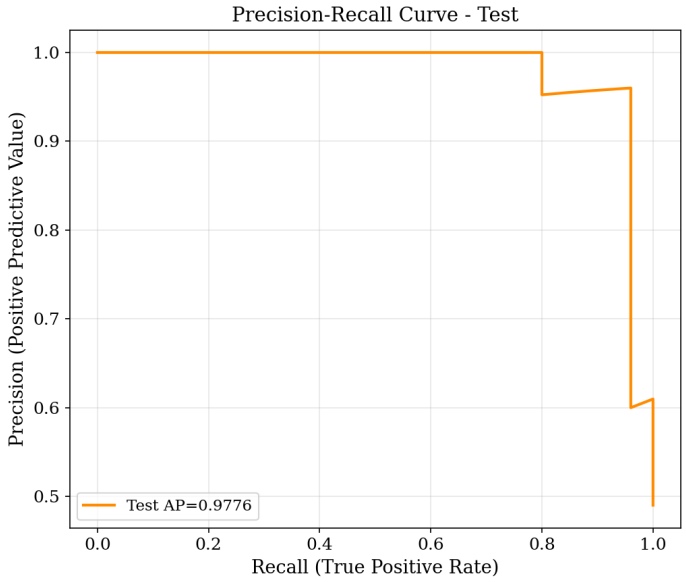

Figure 11. Precision–Recall (PR) curve on the test set.

 doi:10.20944/preprints202605.0567.v1

3.4. Ensemble Size Ablation

Table 3 and Figure 12 represent the recognition rate,Accuracy, AUC, F1-score, EER, and inference time as the ensemble is grown incrementally from one to five models.

Table 3. Test performance with evaluation time in different ensemble sizes.

Models	Test RR (%)	Test AUC	Time (s)
1	64.71	0.9262	328.09
2	66.67	0.9415	319.22
3	94.12	0.9615	299.10
4	94.12	0.9631	407.69
5	96.08	0.9692	404.23

As shown Performance is improves consistently with larger ensembles, with 5 models achieving the highest recognition rate (96.08%) and AUC (0.9692).

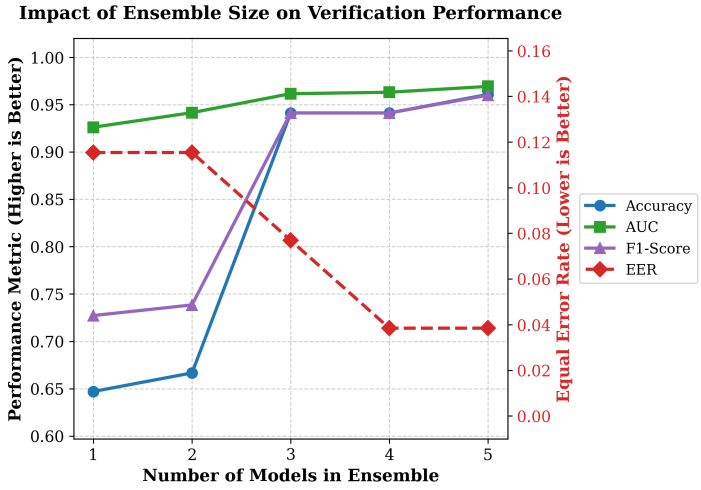

Figure 12. Performance scaling with ensemble size.

3.5. Cross Seed Robustness

Table 4 and Figure 13 report the test performance of each model. Mean test accuracy across five seeds is 0.9216 ± 0.0277, and mean test AUC is 0.9471 ± 0.0231. The average inference speed was 56.11 ms per pair, and approximately 17.82 FPS.

Table 4. System diagnostics, cross-seed stability, and final recognition rates (five-model ensemble).

Category	Metric	Value
System Diagnostics	Total Test Inference Time	2.86 s (51 pairs)
	Average Speed Per Pair	56.11 ms
	Selected Threshold (τ)	0.8583
Cross-Seed Stability	Test Accuracy (Mean $\pm$ Std)	0.9216 ± 0.0277
	Test AUC (Mean $\pm$ Std)	0.9471 ± 0.0231
Final Recognition Rate	Train RR	94.71%
	Val RR	96.77%
	Test RR	96.08%
	Overall RR	95.24%

 doi:10.20944/preprints202605.0567.v1

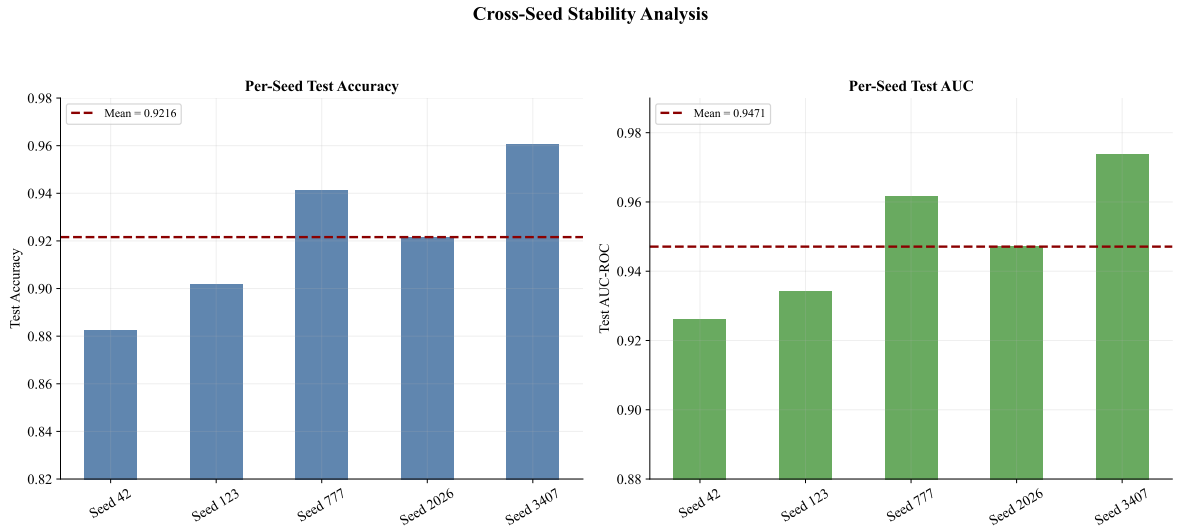

Figure 13. Cross-seed stability analysis showing per-seed Test Accuracy (left) and Test AUC (right). Dashed lines denote the mean values, while shaded areas represent the ± 1 standard deviation (SD) interval.

3.6. Inference Efficiency

The five model ensemble process 51 pairs in total 2.8s, while achieving an average latency of 56.11 milliseconds per pair. These images measured on standard cpu hardware, each pair latency is acceptable range in clinic and verification workflows due to it's real time without any GPU acceleration.

4. Comparison with Prior Work

Table 5 compares between our proposed EASE-PVNet (Ensemble based Autoencoder initialized Siamese Eye-region Periocular Verification Network) and perior work of [17] which is compared in same Surgery (Blepharoplasty) and same Dataset (HDA Plastic Surgery [24]), This figure show visually summary of comparison Figure 14

Table 5. Recognition rate (RR) comparison between the proposed EASE-PVNet and the [17] in Blepharoplasty (eyelid surgery)

Method	Train RR	Val RR	Test RR	Overall RR
Prior Work [17]	92.5%	90.5%	89.6%	91.8%
Proposed (Ours)	**94.71%**	**96.77%**	**96.08%**	**95.24%**
Improvement	+2.21pp	+6.27pp	+6.48pp	+3.44pp

 doi:10.20944/preprints202605.0567.v1

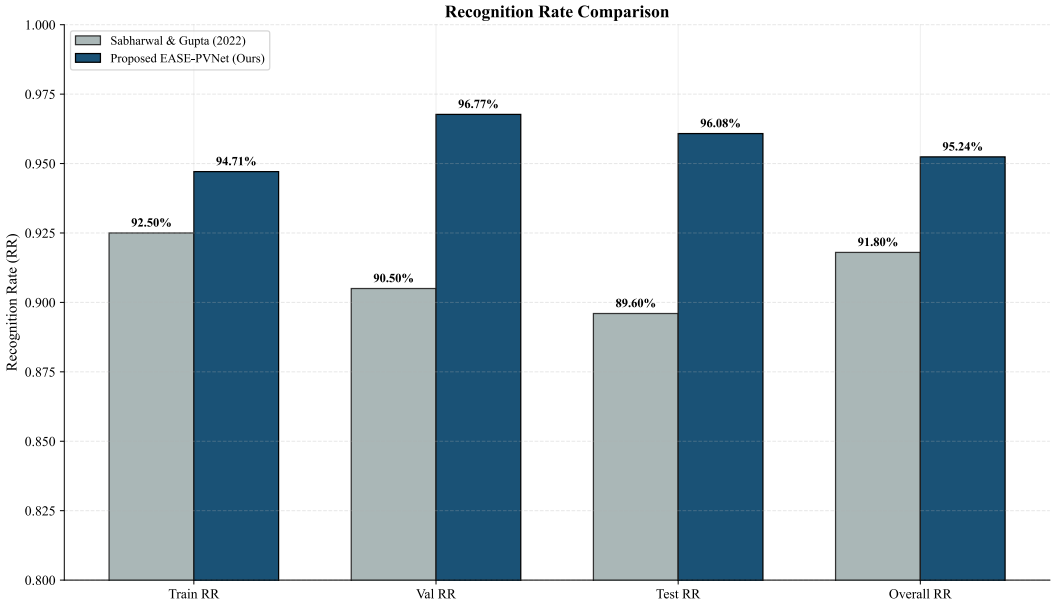

Figure 14. Comparative analysis: Proposed EASE-PVNet vs. Prior Work [17].

5. Discussion

The results demonstrate that periocular-based verification remains a reliable biometric modality even under the challenging conditions introduced by eyelid surgery. By focusing on the periocular region rather than the full face, the proposed framework effectively mitigates the impact of localized surgical alterations that commonly disrupt conventional facial recognition systems. The consistent performance observed across training, validation, and test sets indicates that the adopted combination of unsupervised pretraining and ensemble learning contributes meaningfully to representation stability in this data-limited clinical setting.

An important practical observation is the close balance achieved between false acceptance and false rejection rates, which is particularly relevant in post-surgical verification scenarios where both security and patient convenience are critical. Moreover, the Grad-CAM visualizations suggest that the model relies on anatomically meaningful periocular features rather than spurious cues, supporting the interpretability and trustworthiness of the system. While the dataset size reflects realistic clinical constraints, future validation on larger and more diverse cohorts will be essential to further assess generalizability across surgical types and post-operative timelines.

6. Conclusions

This research proposed EASE-PVNet, a robust periocular identity verification framework designed to address the challenges posed by pre-operative and post-operative eyelid surgery, where localized anatomical changes undermine conventional face recognition systems. The proposed approach combines unsupervised autoencoder-based encoder pretraining with a Siamese metric learning architecture, enabling the model to acquire periocular-specific representations prior to supervised verification, which is particularly effective in data-limited clinical settings.

System robustness and stability are further enhanced through staged hard-negative mining, validation-weighted multi-seed ensemble learning, and bootstrap-based threshold calibration. Together, these strategies reduce sensitivity to random initialization and decision-boundary instability while maintaining balanced biometric error characteristics. Experimental evaluation demonstrates strong and consistent performance, achieving an overall recognition rate of 95.24 percent, with test-set accuracy of 96.08 percent, an AUC-ROC of 96.92 percent, and a low equal error rate of 3.85 percent. The close alignment between false acceptance and false rejection rates confirms the effectiveness of the proposed threshold calibration strategy.

 doi:10.20944/preprints202605.0567.v1

In addition to quantitative performance, the integration of ensemble Grad-CAM provides visual explanations that confirm model attention is focused on anatomically meaningful periocular regions, enhancing transparency and supporting clinical interpretability. Overall, EASE-PVNet offers an accurate, stable, and interpretable solution for post-surgical periocular identity verification, with clear potential for clinical and forensic deployment. Future work will focus on validating the framework on larger and more diverse surgical datasets and extending it to longitudinal and cross-spectral periocular scenarios.

Author Contributions: Z.A. (Ziyad Azzaz) contributed to conceptualization, methodology, software development, formal analysis, investigation, data curation, visualization, and writing of the original draft.

O.M. (Omar Mohamed) contributed to software development, formal analysis, data curation, visualization, and writing of the original draft.

E.K. (Esraa Khatab) contributed to validation, investigation, and critical review and editing of the manuscript.

H.S. (Hany Said) contributed to methodology, validation, and critical review and editing of the manuscript.

O.S. (Omar Shalash) contributed to conceptualization, supervision, project administration, funding acquisition, and manuscript review and editing.

All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The dataset used for this research is available online

Conflicts of Interest: The authors declare no conflicts of interest.

Acknowledgments: The researchers acknowledge Ajman University for its support in this research.

Abbreviations

The following abbreviations are used in this manuscript:

ACC	Accuracy
AUC	Area Under the Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BCC	Basal Cell Carcinoma
CNN	Convolutional Neural Network
CLAHE	Contrast-Limited Adaptive Histogram Equalization
CRediT	Contributor Roles Taxonomy
DET	Detection Error Tradeoff
EER	Equal Error Rate
EASE-PVNet	Ensemble Autoencoder-Initialized Siamese Periocular Verification Network
FAR	False Acceptance Rate
FRR	False Rejection Rate
HTER	Half Total Error Rate
OCT	Optical Coherence Tomography
PR	Precision–Recall
ROI	Region of Interest
RR	Recognition Rate
SGD	Stochastic Gradient Descent
XAI	Explainable Artificial Intelligence

References

1. Rasheed, S. Lightweight Deep Learning Models for Face Mask Detection in Real-Time Edge Environments: A Review and Future Research Directions. *Machine Learning and Knowledge Extraction* **2026**, *8*, 102.
2. Fawzy, H.; Elbrawy, A.; Amr, M.; Eltanekhy, O.; Khatab, E.; Shalash, O. A systematic review: Computer vision algorithms in drone surveillance. *J. Robot. Integr* **2025**, *2*, 1–10.

 doi:10.20944/preprints202605.0567.v1

3. Alonso-Fernandez, F.; Bigun, J.; Fierrez, J.; Damer, N.; Proenca, H.; Ross, A. Periocular biometrics: A modality for unconstrained scenarios. *IEEE Computer* **2024**, *55*, 54–63. https://doi.org/10.1109/MC.2021.3126382.
4. Mattioli, M.; Cabitza, F. Not in my face: Challenges and ethical considerations in automatic face emotion recognition technology. *Machine Learning and Knowledge Extraction* **2024**, *6*, 2201–2231.
5. Talreja, V.; Nasrabadi, N.M.; Valenti, M.C. Attribute-Based Deep Periocular Recognition: Leveraging Soft Biometrics. In Proceedings of the Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2022, pp. 3540–3549. https://doi.org/10.1109/WACV51458.2022.00360.
6. Alonso-Fernandez, F.; Bigun, J. A survey on periocular biometrics research. *Pattern Recognition Letters* **2016**, *82*, 92–105.
7. Aggarwal, G.; Biswas, S.; Flynn, P.J.; Bowyer, K.W. A Sparse Representation Approach to Face Matching Across Plastic Surgery. In Proceedings of the Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV). IEEE, 2012, pp. 113–119. https://doi.org/10.1109/WACV.2012.6163008.
8. Holzinger, A.; Longo, L.; Cangelosi, A.; Ser, J.D. Research Frontiers in Machine Learning & Knowledge Extraction. *Machine Learning and Knowledge Extraction* **2025**, *8*, 6.
9. Pavel, M.S.; Moldovanu, S.; Aiordachioaie, D. On classification of the human emotions from facial thermal images: a case study based on machine learning. *Machine Learning and Knowledge Extraction* **2025**, *7*, 27.
10. Borza, D.L.; Yaghoubi, E.; Frintrop, S.; Proenca, H. Adaptive Spatial Transformation Networks for Periocular Recognition. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23052456.
11. Zhang, L.; Arandjelović, O. Review of automatic microexpression recognition in the past decade. *Machine Learning and Knowledge Extraction* **2021**, *3*, 414–434.
12. Dietterich, T.G. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*; Springer, 2000; pp. 1–15. https://doi.org/10.1007/3-540-45014-9_1.
13. Weng, W.H. Learning Representations for Limited and Heterogeneous Medical Data. Ph.d. dissertation, Massachusetts Institute of Technology, 2022.
14. Zhang, H.; Ogasawara, K. Grad-CAM-Based Explainable Artificial Intelligence Related to Medical Applications. *Bioengineering* **2023**, *10*. https://doi.org/10.3390/bioengineering10091070.
15. Suara, S.; Jha, A.; Sinha, P.; Sekh, A.A. Is Grad-CAM Explainable in Medical Images? *arXiv* **2023**, [arXiv:cs.CV/2307.10506].
16. Zhuang, M.; Wang, H. Unsupervised Representation Learning of Medical Images for Downstream Segmentation. *OpenReview* **2025**. Preprint.
17. Sabharwal, T.; Gupta, R. Deep facial recognition after medical alterations. *Multimedia Tools and Applications* **2022**, *81*, 25675–25706.
18. Hayasaka, T.; Kawano, K.; Kurihara, K.; Suzuki, H.; Nakane, M.; Kawamae, K. Creation of an artificial intelligence model for intubation difficulty classification by deep learning (convolutional neural network) using face images: an observational study. *Journal of Intensive Care* **2021**, *9*, 38. https://doi.org/10.1186/s40560-021-00551-x.
19. Jerjes, W.; Hamdoon, Z.; Rashed, D.; Hopper, C. In Vivo Optical Coherence Tomography for the Detection, Subtyping, and Margin Assessment of Facial Basal Cell Carcinoma: A Comparative Study with Histopathology. *Journal of Clinical Medicine* **2025**, *14*, 949. https://doi.org/10.3390/jcm14030949.
20. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)* **2015**.
21. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality Reduction by Learning an Invariant Mapping. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006, pp. 1735–1742.
22. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *International Conference on Learning Representations (ICLR)* **2019**.
23. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.
24. Rahmani, A.; Ghedasati, S. HDA Plastic Surgery Face Database, 2024. Accessed: 2025-11-24.