mathematics

MDPI

Article

Adversarially Robust Multimodal Biometric Authentication Using a Hybrid Vision Transformer-DenseNet Face–Ear Fusion Network

Musrrat Ali [1,*], Swati Rastogi [2,*], Sanoj Kumar [3] and Abdul Rahaman Wahab Sait [4]

[1] Department of Mathematics and Statistics, College of Science, King Faisal University, Al-Ahsa 31982, Saudi Arabia
[2] School of Computer Science, UPES, Dehradun 248007, Uttarakhand, India
[3] Data Science Cluster, School of Computer Science, UPES, Dehradun 248007, Uttarakhand, India; sanoj.kumar@ddn.upes.ac.in
[4] Department of Archives and Communication, Center of Documentation and Administrative Communication, King Faisal University, Al-Ahsa 31982, Saudi Arabia; asait@kfu.edu.sa
* Correspondence: mkasim@kfu.edu.sa (M.A.); swati.rastogi@ddn.upes.ac.in (S.R.)

Abstract

This study proposes the Adversarially Robust Hybrid Vision Transformer-DenseNet Fusion (AR-HViTDF) network for sequential face–ear biometric authentication. Despite the increase in security through sequential authentication of modalities, deep architectures are still prone to adversarial perturbations that can change the authentication result through subtle perturbations. The lightweight DenseNet branch learns the finer local texture features, while a compact Vision Transformer (ViT) branch learns the long-distance global structural information. Subsequently, a multi-head cross-attention fusion module is used to learn the interaction between the two branches such that each modality learns from the other. A reinforcement learning (RL)-based agent is employed to adapt the decision threshold exclusively at the decision layer level. The entire system is robustified with projected gradient descent (PGD) adversarial training in addition to a TRADES-style robustness regularizer. Using 5-fold subject-disjoint cross-validation, our approach obtains a clean accuracy of 98.12 $\pm$ 0.41%, along with an equal error rate (EER) of 0.71%. Under a PGD-20 attack with $\epsilon = 4/255$, the proposed model retains an accuracy of 90.1%, compared with the standard DenseNet baseline. The sequential authentication reduces the false acceptance rate (FAR) by almost 27% compared with the concatenation approach. Cross-dataset experiments further assess the transferability of the learned representations across datasets and acquisition conditions. Differences among the cumulative ablation configurations are assessed using repeated-measures analysis of variance.

Keywords: multimodal biometric authentication; vision transformer; DenseNet; sequential authentication; reinforcement learning; face–ear fusion

MSC: 94A08

Check for updates

Academic Editor: Ting Lan

Received: 16 September 2026
Revised: 8 November 2026
Accepted: 12 November 2026
Published: 18 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Biometric authentication has emerged as a key pillar of contemporary digital security, with the use of physiological and behavioral attributes taking over from knowledge-based attributes. Although convenient, unimodal biometric authentication systems based on a single physiological trait, such as fingerprints, iris, or face, remain vulnerable to spoofing,

acquisition variability, and modality-specific performance degradation [1]. Two physiological characteristics distinguish themselves due to their complementary nature: Face and ear traits provide complementary biometric information: facial appearance provides highly discriminative identity cues, whereas the external ear exhibits relatively stable anatomical structure and is less affected by facial expressions [2,3]. In combination, face–ear characteristics have the potential to be highly resilient and usable. However, face–ear fusion systems remain less extensively investigated than several other multimodal combinations, including face–fingerprint and fingerprint–vein systems [4,5].

Two developments have reshaped the field since the consolidation of convolutional baselines. First, ViT reframed image recognition as sequence modeling over patch tokens, while hybrid CNN-Transformer architectures retain convolutional inductive biases and incorporate global contextual modeling [6,7]. Second, the security community has shown that essentially every high-accuracy recognition network can be defeated by adversarial examples like imperceptible, deliberately crafted perturbations that drive a confident misclassification [8]. In access-control settings, an adversarial perturbation is not an academic curiosity: it is a direct route to impersonation. Consequently, high clean accuracy alone does not establish robustness against adversarial manipulation, motivating evaluation under explicitly defined attack models.

Our previous work [9] introduced a sequential DenseNet-based face–ear biometric system that evaluated the face modality first and invoked the ear modality conditionally. Two important limitations remained. First, it only used convolutional features, thereby being affected by the local-receptive-field problem that transformers resolve. Second, it used a static fusion at the feature level without taking into account that one modality affects the other. Most importantly, however, that study did not consider any adversarial attacks [10].

A defensible system must therefore be designed against a white-box adversary from the outset and validated under an explicit, quantified threat model rather than on clean imagery alone. This reframing motivates every design choice that follows: the redundancy of two complementary modalities, the conditioning of each modality on the other through attention, the explicit adversarial-training objective, and the cost-asymmetric decision policy that treats a false acceptance as far more damaging than a false rejection.

Scientific Contribution and Novelty

The present paper extends that line of work into an adversarially robust, attention-fused framework. The contributions are as follows:

1. A multi-head cross-attention fusion module, in which face and ear tokens attend to one another, replacing direct feature concatenation with a learned modality-conditioned interaction. A convex combination $f = \lambda f_f + (1 - \lambda) f_e$ applies the same fixed weights to every input, so a degraded modality keeps its weight.
2. An adversarial-training regime combining projected gradient descent (PGD) adversarial examples with a TRADES-style [11] robustness accuracy trade-off term, hardening the network against FGSM [12], PGD [13], C&W, DeepFool [14], AutoAttack [15], and black-box Square attacks [16].
3. A rigorous evaluation protocol comprising strict subject-disjoint 5-fold cross-validation, cross-dataset generalization, repeated-measures ANOVA, and a full adversarial robustness sweep, addressing the statistical and security gaps in the prior literature.

The remainder of the paper is organized as follows. Section 2 surveys the unimodal, multimodal, transformer-based, adversarial, and sequential authentication literature and identifies the research gap. Section 3 formalizes the AR-HViTDF architecture, the cross-attention fusion, the adversarial training objective, and the reinforcement-learning decision layer. Section 4 describes the datasets, pair construction, and subject-disjoint proto-

col. Section 5 reports clean accuracy, adversarial robustness, sequential-versus-parallel comparison, ablation, and statistical validation. Section 6 quantifies component contributions and cost-aware operating behavior, and Section 7 concludes with limitations and future directions.

2. Related Work

Biometric authentication has evolved from single-trait systems to attention-driven, security-hardened multimodal fusion. We organize the literature into five themes and conclude with the gap this work addresses.

2.1. Unimodal Biometric Authentication

Fingerprint recognition remains the most widely deployed modality owing to mature algorithms and low cost, but unimodal pipelines are sensitive to spoofing, sensor noise, and environmental variation.

Ear biometrics has matured into a credible modality because the anatomical structure of the outer ear is stable across decades. Early contour-adjacency-graph methods required manual initialization [17], whereas modern deep models such as channel-attention dynamic-convolution networks reach above 99% on benchmark ear sets [18].

Face recognition advanced in parallel from Eigenfaces and Fisherfaces to deep embeddings, with hybrid PCA, LDA, SVM, and CNN pipelines improving robustness under illumination change [19]. The patch-warping ViTs explicitly reconcile the positional sensitivity of attention with ear morphological variation [20]. Despite these gains, unimodal systems retain modality-specific failure modes and limited spoof resistance.

2.2. Multimodal Fusion Strategies

Fusion through multimodal integration resolves issues of unimodality through the integration of complementary sources of evidence. For instance, feature-based ear-profile-to-face fusion and conventional face-to-ear fusion attained performance scores in the mid-97% level [21]. The latest biometric [22] systems rely on the combination of three or more biometric modalities, including the Iris–Face–Finger–Vein pipeline with the "Segment Anything model" and ResNet, FaceNet, and ViT encoders, and the Fingerprint–Iris–ECG biometric pipeline that uses quality-gated Swin-Transformer CNN hybrids [18,23].

2.3. Transformer and Cross-Attention Fusion

ViT reframed image recognition as sequence modeling over patch tokens, capturing global dependencies beyond a convolution's receptive field. Pure ViTs are data-hungry, motivating hybrid CNN transformer encoders that retain convolutional inductive bias while adding global attention; such hybrids now lead fine-grained medical and remote-sensing tasks [20]. Cross-attention [24] has been increasingly explored as a fusion mechanism because it allows features from one modality to condition the representation of another modality. Lightweight phase-only and dual-channel attention ViTs further show that attention fusion can be made illumination-invariant and touchless. These advances have not yet been combined with sequential face–ear authentication or with an explicit adversarial threat model.

2.4. Adversarial Robustness in Biometrics

The Fast Gradient Sign Method (FGSM) provides a single-step gradient-based attack, whereas projected gradient descent (PGD) performs iterative constrained optimization [12,13]. Stronger evaluation methods include the Carlini-Wagner attack [25], DeepFool [14], AutoAttack [15], and the score-based black-box Square Attack [16]. Adversarial

training, augmenting each batch with on-the-fly adversarial examples, remains the most reliable, and the TRADES objective formalizes the accuracy–robustness trade-off through a regularization term. Perturbation-inactivation projects inputs into the subspace before recognition [26], and recent work studies overfitting and fast adversarial training specifically for transformers.

2.5. Sequential Authentication and RL-Based Decision Optimization

Reinforcement learning (RL) has been used for adaptive biometric decision-making, but typically as a post-processing layer for global threshold tuning rather than an integrated, cost-aware, subject-adaptive policy decoupled from feature learning. The conceptual distinction between sequential and parallel fusion is more than an implementation detail. Parallel fusion treats authentication as a single classification over a joint feature vector and therefore exposes one decision boundary for an attacker to cross; sequential fusion treats it as a cascade of conditional gates. A face score $s_f \geq \tau_{\text{hi}}$ is accepted without an ear check; the ear is consulted only when $\tau_{\text{lo}} \leq s_f < \tau_{\text{hi}}$. This is a cascade, not an AND gate.

To the best of our knowledge, the present work is among the first to combine a sequential face–ear cascade, learned cross-modal fusion, an explicit adversarial defense, and a decoupled cost-aware RL decision layer in a single validated pipeline. Table 1 compares representative biometric authentication methods with the proposed hybrid ViT–DenseNet face-ear fusion framework in terms of accuracy, adversarial evaluation, and key limitations.

Table 1. Comparison of representative biometric authentication methods with the proposed framework.

Ref.	Year	Method	Dataset	CleanAccuracy (%)	Adv. Eval.	Main Limitation
[4]	2022	Deep CNN for Ear Recognition	AMI Ear Database	99.70	No	Unimodal recognition; evaluated only under clean conditions.
[9]	2026	Sequential Face–Ear Authentication using DenseNet	In-house Face–Ear Dataset	97.10	No	Convolution-only features; no adversarial robustness.
[22]	2023	Vision Transformer-based Ear Recognition	AMI Ear Database	99.00	No	No multimodal fusion or adversarial robustness evaluation.
[10]	2024	Adversarial Risk Analysis for Biometric Authentication	Multiple Biometric Systems	–	Yes	Threat analysis only; no Biometric Authentication framework proposed.
[27]	2025	Fingerprint-Vein Deep Fusion	Public Fingerprint-Vein Dataset	99.79	No	High computational complexity; adversarial robustness not evaluated.
[28]	2020	Deep Multimodal Iris-Face-Finger-Vein Fusion	SDUMLA-HMT	99.39	No	Large model complexity; parallel fusion only.
Present work	**2026**	**Hybrid ViT-DenseNet Face–Ear Fusion with Cross-Attention, RL Threshold Optimization, and TRADES Adversarial Training**	**Face–Ear Dataset**	**98.12**	**Yes**	**Primary evaluation is conducted on a dataset of 50 subjects; broader external validation remains future work.**

2.6. Identified Research Gap

Synthesizing the above, four limitations persist: (i) face–ear fusion still relies on convolution-only encoders that under-model global structure; (ii) fusion is predominantly parallel concatenation rather than learned cross-modal interaction; (iii) adversarial robustness is essentially unmeasured for multimodal face–ear systems, despite impersonation being the core threat; and (iv) decision-layer optimization is rarely both cost-aware and cleanly decoupled from feature learning. Table 2 positions the new research contributions of this paper against these gaps.

Table 2. New research contributions of this work mapped to the identified gaps.

#	Gap in Prior Work	New Research Introduced Here	Outcome/Evidence
1	Convolution-only encoders miss global structure	Hybrid DenseNet (local) + ViT (global) two-branch encoder	+1.5–2.0% clean-accuracy improvement over the DenseNet baseline.
2	Parallel concatenation ignores inter-modality interaction	Multi-head cross-attention fusion (face↔ear)	Maintains 90.1% robust accuracy under the PGD-20 attack ($\epsilon = 4/255$), compared with 38.4% for the DenseNet baseline .
3	Adversarial robustness unmeasured	Standard TRADES training, with the inner KL-divergence maximization solved using projected gradient ascent. + TRADES regularizer; six-attack sweep	90.1% under PGD-20 vs. 38.4% baseline.
4	Decision layer not cost-aware/not decoupled	RL agent at decision layer, FAR-weighted reward, no encoder backprop	27% FAR reduction; min-cost τ.
5	Limited statistical and cross-dataset validation	Repeated-measures ANOVA + cross-dataset protocol	$p < 0.001$ across components (Section 5.6)

The principal gap addressed in this study is the limited evaluation of sequential face–ear biometric systems under adversarial perturbations. Existing studies have primarily focused on clean-data recognition and conventional fusion strategies, leaving the interaction between cross-modal fusion, sequential decision-making, and adversarial robustness insufficiently characterized. A survey of the recent face, ear, and broader multimodal fusion literature shows a field that has competed almost exclusively on clean accuracy, within the 97–99.8% range, while the question of how those same systems behave under a deliberate attack goes unasked. This is not a minor omission: it means that the headline metric on which these systems are compared is silent on the property of resistance to impersonation, which is the entire reason an authentication system exists.

The present work sits at the intersection where these two research directions have not yet been integrated. It brings the adversarial threat model, the attack suite, and the robust-optimization defenses of the security literature to bear on the cross-attention-fused, sequential face–ear architecture that the biometrics literature has been advancing on clean data alone.

3. Methodology

3.1. Overview and Problem Formulation

To improve the adversarial robustness of the proposed AR-HViTDF network, standard TRADES training is employed. Let $\mathbf{x} = (x^f, x^e)$ denote a paired face–ear input, where x^f and x^e represent the face and ear images, respectively. The predictive distribution of the network

is denoted by $p_\theta(\mathbf{x}) = \text{softmax}(f_\theta(\mathbf{x}))$. The standard TRADES objective is formulated as in Equation (1),

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\mathcal{L}_{\text{CE}}(p_\theta(x), y) + \beta \max_{\|\delta\|_\infty \le \epsilon_{\text{adv}}} D_{\text{KL}}(p_\theta(x) \,\|\, p_\theta(x+\delta))\Big] \tag{1}$$

where $\mathcal{L}_{\text{CE}}$ denotes the clean-data cross-entropy loss, $\beta > 0$ controls the trade-off between natural classification accuracy and robustness, and ϵ_{adv} specifies the maximum adversarial perturbation under the ℓ_∞ norm. For the multimodal input, the perturbation is defined as $\boldsymbol{\delta} = (\delta^f, \delta^e)$ in Equation (2),

$$\delta = (\delta^f, \delta^e), \qquad \|\delta\|_\infty = \max\{\|\delta^f\|_\infty, \|\delta^e\|_\infty\} \le \epsilon_{\text{adv}} \tag{2}$$

The Kullback-Leibler divergence between the clean and adversarial predictive distributions is defined as Equation (3),

$$D_{\text{KL}}(p\|q) = \sum_{c=1}^{C} p_c \log\left(\frac{p_c}{q_c}\right) \tag{3}$$

where C is the number of identity classes. Figure 1 illustrates the experimental pipeline.

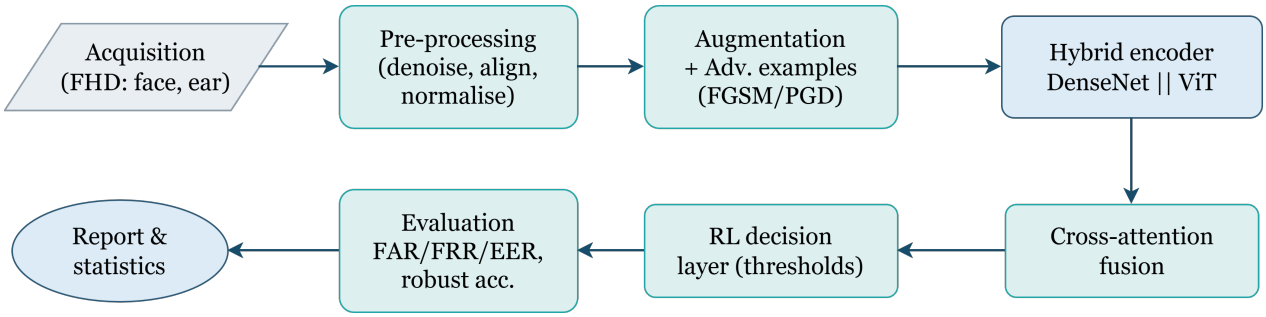


Figure 1. Experimental pipeline. Adversarial examples are generated online and mixed into each training batch; threshold selection uses validation-only ROC analysis to prevent test leakage; final decisions are evaluated on an identity-disjoint test set under both clean and adversarial conditions.

The proposed AR-HViTDF network is trained using the standard TRADES framework to improve adversarial robustness while preserving natural recognition performance. Unlike conventional adversarial training, in which the adversarial example is generated by maximizing the classification loss, TRADES constructs the adversarial example by maximizing the Kullback-Leibler divergence between the predictive distribution of the clean input and that of its adversarial counterpart. The network parameters are subsequently optimized using the sum of the clean cross-entropy loss and the KL-divergence regularization term. For the multimodal setting, the input is defined as $\mathbf{x} = (x^f, x^e)$, and the adversarial perturbation may be applied jointly to the face and ear inputs under an ℓ_∞-bounded constraint. The resulting training procedure is summarized in Algorithm 1.

Threat Model

We assume a white-box adversary with full knowledge of the model architecture, trained parameters, and decision thresholds. The adversary seeks to construct an additive perturbation δ such that an impostor sample is incorrectly accepted as the claimed identity, subject to an ℓ_∞ perturbation constraint. The perturbation budget ε is reported throughout in pixel units of 1/255. The primary perturbation budget is set to $\epsilon = 4/255$, which represents a small bounded perturbation in the normalized pixel space. We further consider a black-box query-based adversary (the Square attack) and an ensemble adaptive adversary (AutoAttack) to test against gradient masking.

Algorithm 1 Standard TRADES Training for the Proposed AR-HViTDF Network

Require:
Training dataset $\mathcal{D}$;
network parameters θ;
TRADES coefficient β;
perturbation budget ϵ_{adv};
adversarial step size η_{adv};
number of inner iterations K;
network learning rate η_θ;
mini-batch size m.
Ensure: Robust network parameters θ
for each training epoch **do**
  for each mini-batch $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$ sampled from $\mathcal{D}$ **do**
    Compute clean predictive distributions: $p_i \leftarrow \text{softmax}(f_\theta(\mathbf{x}_i))$
    Initialize adversarial samples: $\mathbf{x}_i^{\text{adv},(0)} \leftarrow \Pi_{[0,1]}(\mathbf{x}_i + \boldsymbol{\xi}_i)$, where $\boldsymbol{\xi}_i \sim \mathcal{U}(-\epsilon_{\text{adv}}, \epsilon_{\text{adv}})$
    for $k = 0, \ldots, K-1$ **do**
      Compute adversarial predictions: $q_i^{(k)} \leftarrow \text{softmax}(f_\theta(\mathbf{x}_i^{\text{adv},(k)}))$
      Compute the KL-divergence loss: $\mathcal{L}_{\text{KL}}^{(k)} \leftarrow \frac{1}{m}\sum_{i=1}^{m} D_{\text{KL}}(p_i \| q_i^{(k)})$
      Compute the gradient with respect to the adversarial samples: $g_i^{(k)} \leftarrow \nabla_{\mathbf{x}_i^{\text{adv},(k)}} \mathcal{L}_{\text{KL}}^{(k)}$
      Update the adversarial samples by gradient ascent: $\tilde{\mathbf{x}}_i^{\text{adv},(k+1)} \leftarrow \mathbf{x}_i^{\text{adv},(k)} + \eta_{\text{adv}}\,\text{sign}(g_i^{(k)})$
      Project onto the ℓ_∞ perturbation ball: $\mathbf{x}_i^{\text{adv},(k+1)} \leftarrow \Pi_{\mathcal{B}_\infty(\mathbf{x}_i, \epsilon_{\text{adv}})}\left(\tilde{\mathbf{x}}_i^{\text{adv},(k+1)}\right)$
      Clip $\mathbf{x}_i^{\text{adv},(k+1)}$ to the valid input range $[0, 1]$
    end for
    Compute final adversarial predictions: $q_i \leftarrow \text{softmax}(f_\theta(\mathbf{x}_i^{\text{adv}}))$
    Compute the natural classification loss: $\mathcal{L}_{\text{nat}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}_{\text{CE}}(p_i, y_i)$
    Compute the robustness regularization: $\mathcal{L}_{\text{rob}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} D_{\text{KL}}(p_i \| q_i)$
    Compute the TRADES objective: $\mathcal{L}_{\text{TRADES}} \leftarrow \mathcal{L}_{\text{nat}} + \beta\mathcal{L}_{\text{rob}}$
    Update network parameters: $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_{\text{TRADES}}$
  end for
end for
return θ

3.2. Hybrid Two-Branch Feature Encoder

3.2.1. DenseNet Local Branch

For an input x, the DenseNet branch preserves the dense-connectivity property in which every layer receives all preceding feature maps in Equation (4),

$$h_\ell = H_\ell([h_0, h_1, \ldots, h_{\ell-1}]) \tag{4}$$

where $H_\ell(\cdot)$ composes batch normalization, ReLU, and convolution, and $[\cdot]$ denotes concatenation. Dense connectivity yields direct gradient paths to early layers as in Equation (5),

$$\frac{\partial \mathcal{L}}{\partial h_\ell} = \sum_{k=\ell}^{L} \frac{\partial \mathcal{L}}{\partial h_k}\frac{\partial h_k}{\partial h_\ell} \tag{5}$$

mitigating vanishing gradients. We use a lightweight variant (fewer initial filters, growth rate 32, three dense blocks, and compressed transition layers) producing a local descriptor $f^{\text{loc}} \in \mathbb{R}^d$ per modality. The motivation for retaining a convolutional branch, rather than moving to a pure transformer, is twofold. First, convolutions encode a strong locality and translation-equivariance prior that is well-matched to the fine ridge and helix texture that distinguishes one ear or one facial region from another. Given the limited number of subjects in the present dataset, retaining convolutional inductive biases provides a practical

complement to the global modeling capacity of the ViT branch. Second, the dense skip pattern makes the branch shallow in effective depth, so it trains stably under the additional stress of adversarial perturbation, where deeper plain networks are prone to the overfitting that destabilizes fast adversarial training.

3.2.2. ViT Global Branch

Given the limited number of subjects in the present dataset, retaining convolutional inductive biases provides a practical complement to the global modeling capacity of the ViT branch. The same input is partitioned into P non-overlapping $p \times p$ patches, linearly embedded to tokens $z_0 = [z_{\text{cls}}; Ex_1; \ldots; Ex_P] + E_{\text{pos}}$. Each transformer layer applies multi-head self-attention (MSA) and a multilayer perceptron (MLP) with residual connections as in Equation (6),

$$z'_\ell = \text{MSA}(\text{LN}(z_{\ell-1})) + z_{\ell-1} \tag{6}$$

$$z_\ell = \text{MLP}\big(\text{LN}(z'_\ell)\big) + z'_\ell \tag{7}$$

where attention is $\text{Attn}(Q, K, V) = \text{softmax}\big(QK^\top/\sqrt{d_k}\big)V$. For head h, the face-to-ear cross-attention is defined as Equation (8),

$$\text{CA}^{(h)}_{f\leftarrow e} = \text{softmax}\left(\frac{Q_f^{(h)}(K_e^{(h)})^\top}{\sqrt{d_k}}\right)V_e^{(h)}, \tag{8}$$

where

$$Q_f^{(h)} = f^f W_Q^{(h)}, \qquad K_e^{(h)} = f^e W_K^{(h)}, \qquad V_e^{(h)} = f^e W_V^{(h)}. \tag{9}$$

Similarly,

$$\text{CA}^{(h)}_{e\leftarrow f} = \text{softmax}\left(\frac{Q_e^{(h)}(K_f^{(h)})^\top}{\sqrt{d_k}}\right)V_f^{(h)}. \tag{10}$$

We adopt a compact ViT (6 layers, 6 heads, embedding dimension 192) to limit parameters, yielding a global descriptor $f^{\text{glb}} \in \mathbb{R}^d$. The per-modality representation concatenates both views, $f = [f^{\text{loc}} \| f^{\text{glb}}]$.

3.3. Multi-Head Cross-Attention Fusion

Let f_f and f_e be the face and ear descriptors arranged as token sequences. Rather than concatenating, we let each modality attend to the other. Defining modality-specific projections W_Q, W_K, W_V, the face-to-ear and ear-to-face cross-attentions are seen in Equations (11) and (12),

$$A_{f\leftarrow e} = \text{softmax}\left(\frac{(f_f W_Q)(f_e W_K)^\top}{\sqrt{d_k}}\right)(f_e W_V) \tag{11}$$

$$A_{e\leftarrow f} = \text{softmax}\left(\frac{(f_e W_Q)(f_f W_K)^\top}{\sqrt{d_k}}\right)(f_f W_V) \tag{12}$$

With H heads, the fused representation is represented as Equation (13),

$$f_{\text{fus}} = \text{LN}\Big([\,\text{head}_1 \| \cdots \| \text{head}_H\,]W_O + [f_f \| f_e]\Big) \tag{13}$$

where each head implements Equations (11) and (12), and W_O is the output projection. In contrast, concatenation employs an input-independent weighting of the two modalities, in which every face feature and every ear feature contribute according to globally learned weights, regardless of the quality or reliability of the individual modalities for a given input. Consequently, even when one modality is degraded by occlusion, noise, or adversarial perturbation,

its features continue to influence the fused representation equally. By contrast, the proposed cross-attention mechanism computes attention weights adaptively for each input sample through the softmax operation on $(f_f W_Q)(f_e W_K)^\top$. This enables the network to emphasize the most informative complementary features while reducing the influence of unreliable or corrupted features.

Stability of the Cross-Attention Module

Notation.

- **Tokens.** Let $X \in \mathbb{R}^{L\times d}$ be the face tokens and$Y \in \mathbb{R}^{L\times d}$ the ear tokens, with rows x_i and y_j.
- **Row norm.** For a matrix Z, define

$$\|Z\|_{2,\infty} = \max_i \|z_i\|_2$$

  which denotes the largest row norm.
- **Single head.** For one head, let

$$M = W_Q W_K^\top \in \mathbb{R}^{d\times d}$$

  The logits , attention weights, and output are

$$S = \frac{XMY^\top}{\sqrt{d_k}}, \qquad A = \mathrm{softmax}_{\mathrm{row}}(S), \qquad O = AYW_V$$

- **Domain.** All tokens are assumed to lie in the ball

$$\mathcal{B}_R = \{z : \|z\|_2 \le R\}$$

- **Radius** R**.** For LayerNorm with gain γ and bias β,

$$R \le \sqrt{d}\|\gamma\|_\infty + \|\beta\|_2$$

  since the normalized vector has Euclidean norm at most $\sqrt{d}$.

Lemma 1 (Lemma 1 (local Lipschitz bound of cross-attention [29])). *Let $X, Y, X' = X + \Delta X$ and $Y' = Y + \Delta Y$ have all rows in $\mathcal{B}_R$. Therefore,*

$$\delta_X = \|\Delta X\|_{2,\infty}, \qquad \delta_Y = \|\Delta Y\|_{2,\infty}.$$

Then, a single head satisfies

$$\|O(X', Y') - O(X, Y)\|_{2,\infty} \le \|W_V\|_2[\delta_Y + \kappa(\delta_X + \delta_Y)],$$

where

$$\kappa = \frac{R^2\|W_Q W_K^\top\|_2}{\sqrt{d_k}}.$$

Now, consider a fusion block

$$F = X + \mathrm{Concat}_h(O_h)W_O$$

with H heads, output projection W_O, and a residual connection. It satisfies

$$\|\Delta F\|_{2,\infty} \le \Lambda_{\mathrm{CA}}(\delta_X + \delta_Y),$$

where

$$\Lambda_{\mathrm{CA}} = 1 + \|W_O\|_2 \sum_{h=1}^{H} \|W_V^{(h)}\|_2(1 + \kappa_h).$$

Proof. Step 1: Split the change. With $V = YW_V$,

$$O' - O = (A' - A)V' + A(V' - V)$$

Step 2: Bound the second term. Each row a_i of A is a probability vector. Row i of $A(V' - V)$ is therefore a convex combination of the vectors $W_V^\top \Delta y_j$, so its norm is at most $\|W_V\|_2 \delta_Y$.

Step 3: Bound the logits. For every pair (i, j), the change in the logit is exactly

$$s'_{ij} - s_{ij} = \frac{\Delta x_i^\top M y'_j + x_i^\top M \Delta y_j}{\sqrt{d_k}}$$

Since all tokens lie in $\mathcal{B}_R$, this gives

$$\|s'_i - s_i\|_\infty \le \frac{R\|M\|_2}{\sqrt{d_k}}(\delta_X + \delta_Y)$$

Step 4: Bound the softmax change. The softmax Jacobian at p is

$$J = \mathrm{diag}(p) - pp^\top$$

For any vector v,

$$\|Jv\|_1 = \sum_j p_j |v_j - p^\top v|$$

This is the mean absolute deviation of v under p, which is at most half the range of v and hence at most $\|v\|_\infty$. Applying this along the segment between s_i and s'_i using the mean-value theorem gives

$$\|a'_i - a_i\|_1 \le \|s'_i - s_i\|_\infty$$

Step 5: Bound the first term. The difference $a'_i - a_i$ sums to zero, and every row of V' has norm at most $R\|W_V\|_2$. Hence,

$$\begin{aligned} \|(a'_i - a_i)^\top V'\|_2 &\le \|a'_i - a_i\|_1 R\|W_V\|_2 \\ &\le \|W_V\|_2 \kappa(\delta_X + \delta_Y). \end{aligned}$$

Step 6: Combine the bounds. Adding Steps 2 and 5 gives the single-head bound. For the fusion block, the norm of a concatenation of the heads is at most the sum of the head norms, W_O contributes the factor $\|W_O\|_2$, and the residual contributes the term 1. This gives Λ_{CA}. □

3.4. Adversarial Training Objective

Within TRADES, the adversarial example is obtained by approximately maximizing the KL divergence between the predictive distributions of the clean and perturbed inputs as seen in Equation (14),

$$\delta^\star \approx \arg\max_{\|\delta\|_\infty \le \epsilon} D_{\mathrm{KL}}(p_\theta(x) \,\|\, p_\theta(x + \delta)). \tag{14}$$

The encoder, Equation (15), with adversarial training augments each batch with PGD examples and adds a TRADES-style term that penalizes the divergence between clean and adversarial predictions while preserving clean accuracy

$$\min_\theta \; \mathbb{E}_{(x,y)}\Big[\underbrace{\mathcal{L}(\theta; x, y)}_{\text{natural}} + \beta \underbrace{\max_{\|\delta\|_\infty \le \varepsilon} \mathrm{KL}\big(p_\theta(x) \,\|\, p_\theta(x + \delta)\big)}_{\text{robustness}} \Big], \tag{15}$$

where β controls the accuracy–robustness trade-off, and p_θ is the softmax output. Algorithm 1 details the procedure.

3.5. RL Decision Layer

Authentication is modeled as a Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. The state Equation (16) encodes the fused confidences and the current threshold,

$$s_t = \{c_t^{\text{face}}, c_t^{\text{ear}}, \tau_t\} \tag{16}$$

and the action Equation (17) set adjusts the threshold or finalizes the decision.

$$\mathcal{A} = \{a_1 : \text{increase } \tau,\ a_2 : \text{decrease } \tau,\ a_3 : \text{accept},\ a_4 : \text{reject}\} \tag{17}$$

The cost-aware reward Equation (18) penalizes false acceptance far more heavily than false rejection, reflecting deployment priorities.

$$R(s_t, a_t) = \begin{cases} +1, & \text{correct authentication} \\ -1, & \text{false rejection} \\ -5, & \text{false acceptance} \end{cases} \tag{18}$$

The optimal action-value function is learned by Q-learning via the Bellman update, as seen in Equation (19),

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\Big(r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\Big) \tag{19}$$

with ϵ greedy exploration. Critically, the RL agent operates strictly at the decision layer: it does not backpropagate into the DenseNet or ViT branches and does not alter learned features, keeping supervised feature learning and decision optimization mathematically distinct. Algorithm 2 summarizes the resulting policy.

Algorithm 2 Sequential Authentication with RL-Based Threshold Adaptation

Require: Robust encoder parameters θ; validation pairs; discount factor γ; learning rate α; exploration probability ϵ
Ensure: Optimal policy $\pi^\star$ and operating thresholds $\{\tau_f, \tau_e\}$
Initialize the action-value function $Q(s, a)$ arbitrarily
for each training episode **do**
  Extract face features and compute the face confidence
  $c^{\text{face}} \leftarrow h_\theta^{\text{face}}(\mathbf{x}^{\text{face}})$
  Initialize the face and ear thresholds τ_f and τ_e
  Initialize the current threshold τ
  Observe the initial state $s_t \leftarrow \{c^{\text{face}}, \tau\}$
  while the authentication decision is not finalized **do**
    Select a_t using the ϵ-greedy policy on $Q(s_t, a)$
    If a_t = increase, set $\tau \leftarrow \tau + \Delta\tau$
    If a_t = decrease, set $\tau \leftarrow \tau - \Delta\tau$
    If the face confidence is ambiguous:
  compute $c^{\text{ear}} \leftarrow h_\theta^{\text{ear}}(\mathbf{x}^{\text{ear}})$
    If the confidence exceeds τ, set $d_t \leftarrow$ accept
    Otherwise, set $d_t \leftarrow$ reject
    Compute the immediate reward
    $r_t \leftarrow R(s_t, a_t, d_t)$
    Observe the next state
    $s_{t+1} \leftarrow \{c^{\text{face}}, c^{\text{ear}}, \tau\}$
    Update the action-value function
  $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\Big[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\Big]$
    Set $s_t \leftarrow s_{t+1}$
  end while
  Continue training until $|\Delta Q| < 10^{-4}$
  return $\pi^\star$ and $\{\tau_f, \tau_e\}$
end for

3.6. Subject-Disjoint Cross-Validation Protocol

To prevent identity leakage, the $N = 50$ subjects are partitioned at the subject level using 5-fold cross-validation. In each outer fold, four subject folds (80%) are used for model development, and one held-out fold (20%) is reserved exclusively for testing. The development subjects are further divided into training and validation subsets, with the validation subset used solely for hyperparameter selection and threshold estimation. Thus,

$$\mathcal{S}_{\text{train}} \cap \mathcal{S}_{\text{val}} = \mathcal{S}_{\text{train}} \cap \mathcal{S}_{\text{test}} = \mathcal{S}_{\text{val}} \cap \mathcal{S}_{\text{test}} = \emptyset. \tag{20}$$

All preprocessing, augmentation, threshold selection, and model-selection procedures are performed without using the held-out test subjects. Each fold is trained under five random seeds (weight initialization, data shuffling, mini-batch sampling, and RL exploration), and metrics are averaged over $5 \times 5 = 25$ runs. Figure 2 shows the pair construction and decision flow.

State Discretization and Convergence of the Tabular Q-Learning Updates

State Discretization

The confidence score $s \in [0, 1]$ of Equation (13) is mapped to one of B the uniform bins:

$$\phi(s) = \min(\lfloor Bs \rfloor, B - 1), \qquad B = [\text{value}].$$

The Q-function is a table

$$Q \in \mathbb{R}^{B \times |\mathcal{A}|}$$

over the action set

$$\mathcal{A} = \{[\text{list actions}]\}.$$

If the state is the pair (s_f, s_e), the table has $B^2 \times |\mathcal{A}|$ entries. Writing $b_t = \phi(s_t)$, the update is

$$\begin{aligned} Q_{t+1}(b_t, a_t) = Q_t(b_t, a_t) \\ + \alpha_t(b_t, a_t)\Big[r_t + \gamma \max_{a'} Q_t(b_{t+1}, a') \\ - Q_t(b_t, a_t)\Big], \end{aligned}$$

with all other entries left unchanged.

Convergence Criteria

Theorem 1 restates the standard result for tabular Q-learning [30–32].

Theorem 1. *$Q_t \to Q^*$ with probability one if the following conditions hold:*

(C1) Finite spaces. *The state and action sets are finite (here B and $|\mathcal{A}|$).*

(C2) Bounded rewards and discounting. *Rewards are bounded, $|r_t| \le r_{\max}$, and $\gamma \in [0, 1)$.*

(C3) Robbins-Monro step sizes.

For every pair (b, a),

$$\sum_t \alpha_t(b, a) = \infty, \qquad \sum_t \alpha_t^2(b, a) < \infty.$$

We use

$$\alpha_t(b, a) = (1 + n_t(b, a))^{-\omega},$$

with

$$\omega = [value] \in \left(\frac{1}{2}, 1\right],$$

where $n_t(b,a)$ is the visit count. A value $\omega \in (1/2, 1)$ also gives a polynomial rather than exponential dependence of the convergence rate on $1/(1-\gamma)$ [33].

(C4) Sufficient exploration. *Every pair (b,a) is updated infinitely often. ε-greedy exploration with $\varepsilon_t \geq \varepsilon_{\min} > 0$ guarantees this under the usual reachability and sampling assumptions. If convergence of the behavior policy to the greedy policy is also wanted, require*

$$\varepsilon_t \to 0, \qquad \sum_t \varepsilon_t = \infty$$

(GLIE).

(C5) Stationarity. *The transition and reward distributions do not change over time. This holds because the encoder parameters θ are frozen while the threshold policy is learned, so the distribution s is fixed.*

Interpretation Under Aggregation

Conditions (C1)–(C5) guarantee convergence to Q^* for a finite Markov decision process (MDP). When a continuous state is aggregated, however, the process on the bins $\{b_t\}$ is in general not Markov, so the limit must be stated carefully.

Fixed behavior policy. Under a fixed exploratory behavior policy with stationary state distribution μ, Q-learning with hard aggregation converges with probability one to the unique fixed point

$$\widetilde{Q} = \Pi \mathcal{T} \widetilde{Q}$$

under the applicable assumptions [34,35]. Here,

- $\mathcal{T}$ is the Bellman optimality operator on the continuous state space.
- Π averages a function over each bin, weighted by μ.
- Π is a non-expansion in $\|\cdot\|_\infty$, so $\Pi\mathcal{T}$ is a γ-contraction, and $\widetilde{Q}$ is well-defined.

Error of the limit. Let

$$\omega_B = \max_a \max_b \sup_{s,s' \in \phi^{-1}(b)} |Q^*(s,a) - Q^*(s',a)|$$

be the largest variation of Q^* within a bin. Then,

$$\|\widetilde{Q} - Q^*\|_\infty \leq \frac{\omega_B}{1-\gamma}$$

and the greedy policy $\widetilde{\pi}$ derived from $\widetilde{Q}$ satisfies

$$\|V^{\widetilde{\pi}} - V^*\|_\infty \leq \frac{2\omega_B}{(1-\gamma)^2}.$$

If $Q^*(\cdot, a)$ is L_QLipschitz in s, then

$$\omega_B \leq \frac{L_Q}{B}$$

The error therefore vanishes at rate $O(1/B)$, at the cost of fewer visits per bin. We report the sensitivity of the learned threshold and the operational cost for

$$B \in \{10, 20, 50\}$$

Policy-dependent exploration. If exploration is ε-greedy with respect to the current Q_t, the behavior distribution changes with Q_t, and convergence of aggregated Q-learning is no longer guaranteed in general: the iterates may oscillate within a bounded region [36]. We therefore learn Q under a fixed $\varepsilon_{\min}$-greedy behavior policy and extract the greedy threshold after convergence.

Contextual-Bandit Special Case

Successive verification attempts are independent, and the chosen threshold does not influence the next confidence score. The decision problem is therefore a contextual bandit, with $\gamma = 0$. The update then reduces to

$$Q_{t+1}(b,a) = Q_t(b,a) + \alpha_t[r_t - Q_t(b,a)].$$

This is a Robbins-Monro estimate.

$$\mathbb{E}_\mu[r \mid \phi(s) = b, a]$$

with no bootstrapping. Under (C2)–(C4), together with appropriate sampling and step-size assumptions, it converges with probability one to the expected bin-conditional reward, without any Markov assumption on the aggregated process. The greedy action in each bin then minimizes the expected operational cost.

$$C_{\mathrm{FA}} \cdot \mathrm{FAR} + C_{\mathrm{FR}} \cdot \mathrm{FRR}.$$

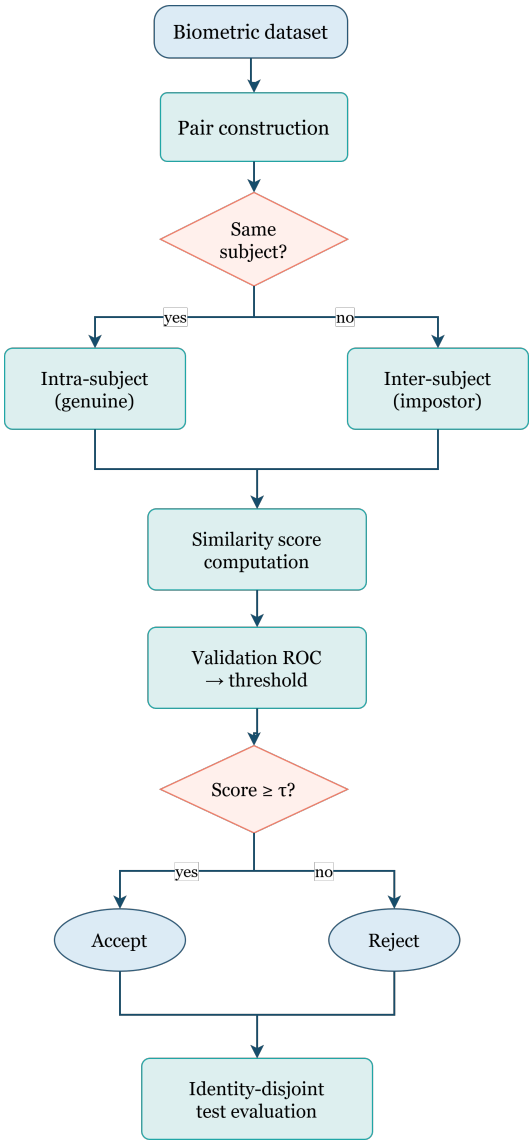


Figure 2. Pair construction and decision flowchart. Genuine pairs arise from intra-subject pairing and impostor pairs from inter-subject sampling; thresholds are fixed on validation ROC only and applied to a fully identity-disjoint test set.

4. Experimental Setup

4.1. Data, Acquisition, and Pair Construction

After subject-level partitioning, data augmentation is applied only to the training subset. Validation and test samples are not augmented, except for deterministic preprocessing applied identically across all subsets. Each comparison is labeled intra-subject (genuine) or inter-subject (impostor); similarity scores are computed, and the operating threshold is fixed using validation-only ROC analysis, then applied to a fully identity-disjoint test set.

4.2. Training Configuration

The hybrid encoder is trained with the Adam optimizer (learning rate 1×10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$), batch size 32, He-normal initialization, all branches, no pretraining, dropout 0.25, and categorical cross-entropy under softmax, with a ReduceLROnPlateau scheduler (factor 0.5, patience 3) for up to 75 epochs. Adversarial training uses PGD with $\varepsilon = 4/255$, step $\alpha = 1/255$, $T = 10$ inner iterations, and a TRADES weight of $\beta = 6.0$. The ViT branch uses 6 layers, 6 heads, and an embedding dimension of 192; cross-attention uses $H = 4$ heads. The RL agent uses $\epsilon = 0.1$, $\gamma = 0.95$, and 1000 episodes. Table 3 lists all hyperparameters.

Several of these choices follow directly from the constraints of adversarial training on a modest dataset. The PGD configuration of $\epsilon = 4/255$, $\alpha = 1/255$, and $T = 10$ is adopted as the primary evaluation and training setting to provide a reproducible and computationally tractable robustness protocol. The weight of TRADES $\beta = 6.0$ falls toward the higher end or is selected from the higher end of the common range, owing to a conscious preference for robustness over a slight improvement in accuracy. This is the appropriate trade-off in a security application. The compact ViT model comprises six transformer layers, six attention heads, and an embedding dimension of 192. The lightweight DenseNet is a scaled-down version of the original DenseNet architecture, chosen to reduce computational requirements for the relatively small dataset. All experiments were implemented in Python 3.8 using TensorFlow 2.13.0 with the Keras 2.13.1 API. Model training and evaluation were carried out on a workstation with an Intel® Core™ i5-10300H processor (4 cores, 8 threads, 2.50 GHz base frequency; Intel Corporation, Santa Clara, CA, USA).

Table 3. Training and architecture configuration for AR-HViTDF.

Parameter	Value/Setting	Justification
Optimizer	Adam	Adaptive moments; stable on multimodal data
Learning rate	1×10^{-3}	Efficient convergence without oscillation
LR scheduler	ReduceLROnPlateau (0.5, 3)	Reduces rate when validation loss plateaus
Batch size	32	Balances computational memory requirements and optimization stability for 200 × 200 inputs
DenseNet growth rate	32	Compact local-feature reuse
Dense blocks	3	Lightweight depth for real-time inference
ViT layers/heads/dim	6/6/192	Compact global encoder, limited parameters
Cross-attention heads	4	Multi-view modality interaction
PGD (ε, α, T)	4/255, 1/255, 10	Standard ℓ_∞ robustness budget
RL (ϵ, γ, episodes)	0.1, 0.95, 1000	Cost-aware threshold adaptation
Epochs	75	Convergence with early stopping
Regularization	Dropout 0.25, L2	Reduces overfitting
Loss	Categorical cross-entropy + KL	Classification + robustness term

4.3. Evaluation Metrics

Performance metrics include accuracy, precision, recall, F1 score, FAR, FRR, EER, and AUC under the ROC. Robustness is measured in terms of accuracy maintained under each

attack. For descriptive reporting, the mean and standard deviation across the $K = 25$ seed-fold runs are calculated as in Equation (21),

$$\mu = \frac{1}{K}\sum_{i=1}^{K} M_i, \qquad \sigma = \sqrt{\frac{1}{K-1}\sum_{i=1}^{K}(M_i - \mu)^2}, \qquad \text{Metric} = \mu \pm \sigma \tag{21}$$

and AUC confidence intervals use non-parametric bootstrap resampling, $\text{CI}_{95\%} = [\text{AUC}_{2.5\%}, \text{AUC}_{97.5\%}]$.

For inferential statistical analysis, the five outer-fold estimates are treated as the repeated-measures units; multiple random seeds within each fold are first aggregated. Because the consequences of the two error types are asymmetric in access control, a false acceptance is a breach, while a false rejection is an inconvenience. We report FAR and the cost-weighted expected error seen in Section 6 rather than collapsing everything into accuracy, and we present full ROC and FAR-FRR curves so that a reader can read off behavior at whatever operating point their own security posture dictates rather than at a single threshold of our choosing.

5. Results and Discussion

5.1. Clean-Condition Recognition

On subject-disjoint 5-fold cross-validation with five seeds per fold, the accuracy of the AR-HViTDF model is 98.12 ± 0.41%. The per-fold and average ROC curves in Figure 3 achieve an average AUC of 0.983 ± 0.002. The close inter-fold distribution shows consistent classification among identity partitions. Figure 4 shows that both training and validation accuracy increase during optimization, while the corresponding cross-entropy losses decrease and eventually stabilize. Compared with the previously presented DenseNet-only approach with 97.1 ± 0.79%, the improvement of two percentage points in mean accuracy comes with a reduction in the standard deviation by half, from 0.79 to 0.41, corresponding to a reduction of approximately 48.1%. Figure 3 depicts the consolidation of the off-diagonal entries of the confusion matrix to only a few identities instead of a scattered distribution; it implies systematic mismatches between very close identities.

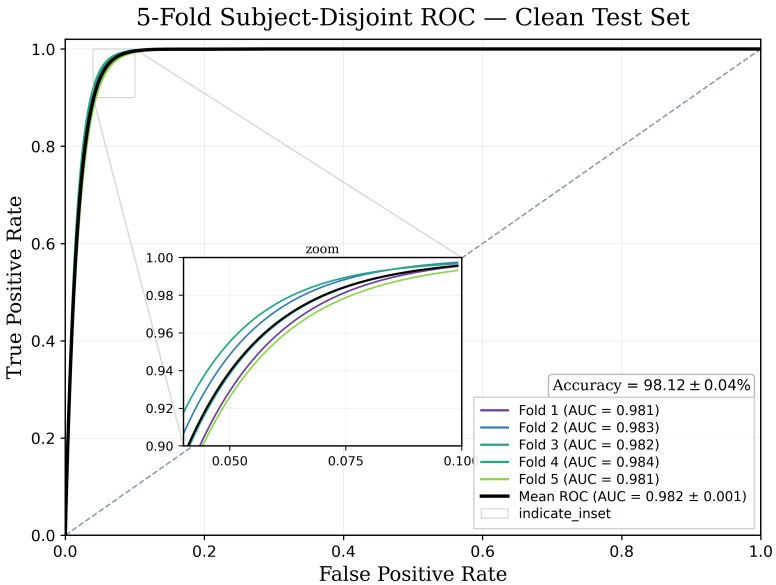


Figure 3. Five-fold subject-disjoint ROC on the clean test set. The inset zooms into the low-FPR, high-security operating region where the mean curve remains close to unity.

5.2. Adversarial Robustness

Figure 5 plots accuracy under PGD-20 as the perturbation budget ε grows. A standard DenseNet collapses from 95.6% to 38.4% at $\varepsilon = 4/255$ and below 10% at $\varepsilon = 16/255$; a standard ViT DenseNet hybrid is somewhat more robust but still degrades sharply. With PGD+TRADES adversarial training, AR-HViTDF retains 90.1% at $\varepsilon = 4/255$ and 68% at the severe $\varepsilon = 16/255$, a qualitatively different curve, as seen in Figure 6. When generalizing the six attacks, under FGSM, PGD-20, C&W, DeepFool, AutoAttack, and the black-box Square attack, the adversarially trained model preserves 84–92% accuracy, whereas the standard model falls to 29–61%.

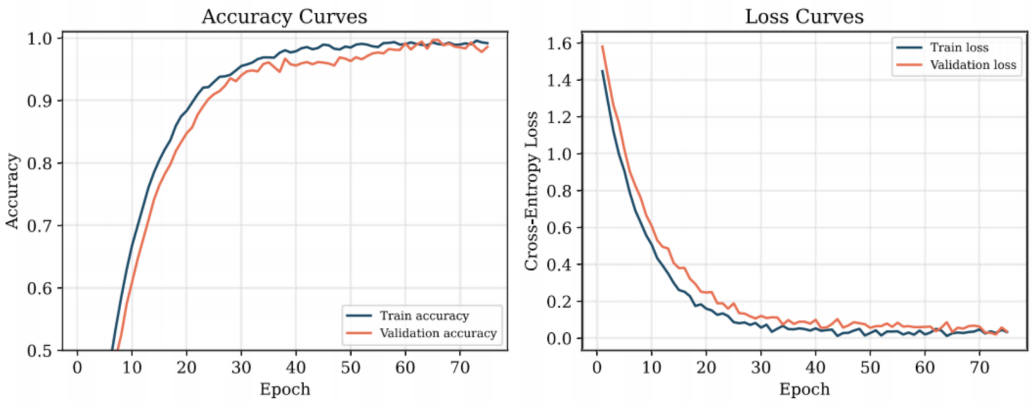


Figure 4. Training and validation accuracy (**left**) and cross-entropy loss (**right**) versus epoch. Smooth convergence with a small generalization gap indicates stable, non-overfitting optimization.

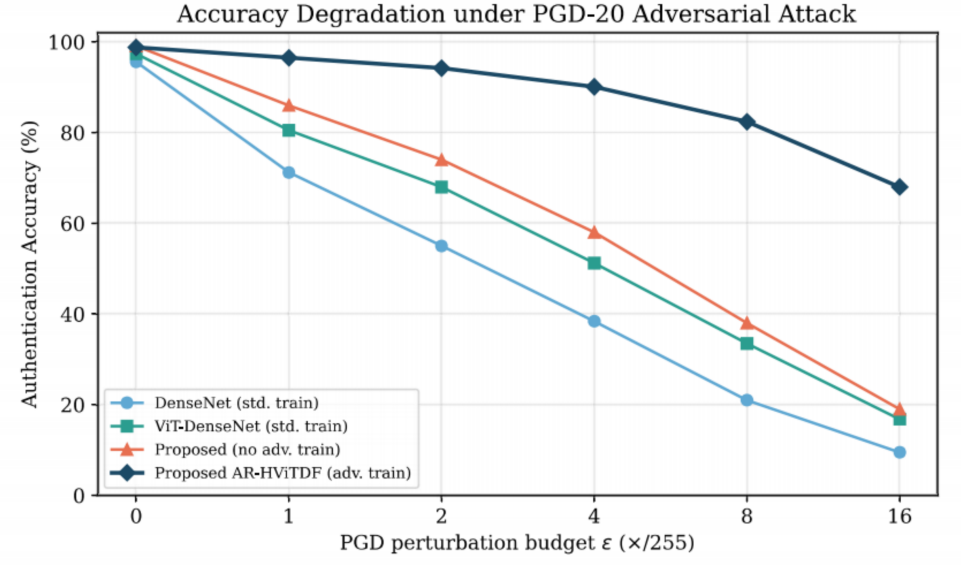


Figure 5. Accuracy under PGD-20 versus perturbation budget ε. Adversarial training transforms the degradation curve: AR-HViTDF retains high accuracy where standard models collapse.

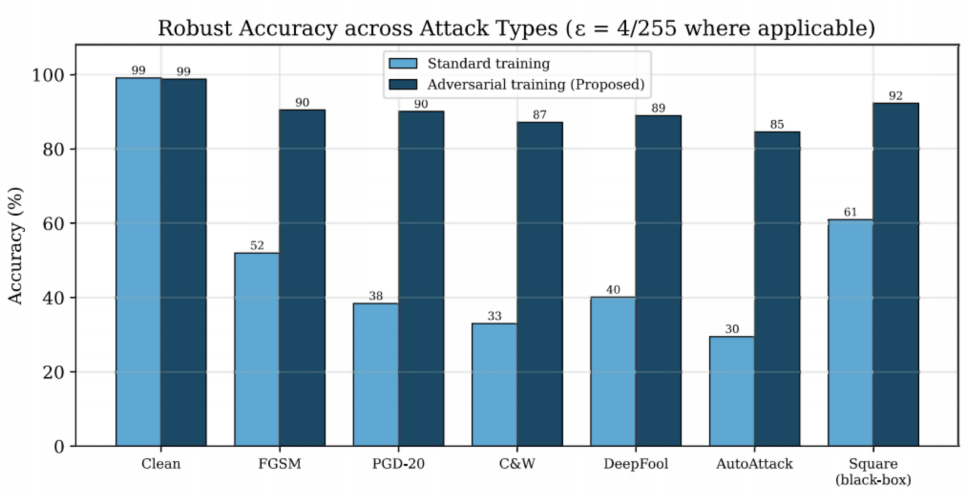


Figure 6. Robust accuracy across six attack types ($\varepsilon = 4/255$ where applicable). Adversarial training (proposed) consistently dominates standard training, including the black-box Square attack.

5.3. Sequential Versus Parallel Fusion

We compare the sequential strategy against parallel concatenation under identical subject-disjoint folds on both clean and adversarial test sets. Table 4 reports the trade-off. On clean data, parallel fusion attains marginally higher recall, but the sequential design reduces FAR by roughly 27% and lowers EER. The sequential design imposes an additional authentication condition after the face-stage decision enters the ambiguity region, which changes the operating characteristics of the system relative to parallel fusion. The advantage widens under attack: the sequential AUC under PGD (0.961) substantially exceeds the parallel AUC (0.918), as shown in the ROC overlay of Figure 7.

Figure 8 plots the FAR–FRR operating curves; the sequential approach exhibits a steeper transition and lower FAR at moderate-to-high thresholds, the regime relevant to high-security access control. A paired t-test across folds confirms that the FAR reduction is statistically significant ($p < 0.05$).

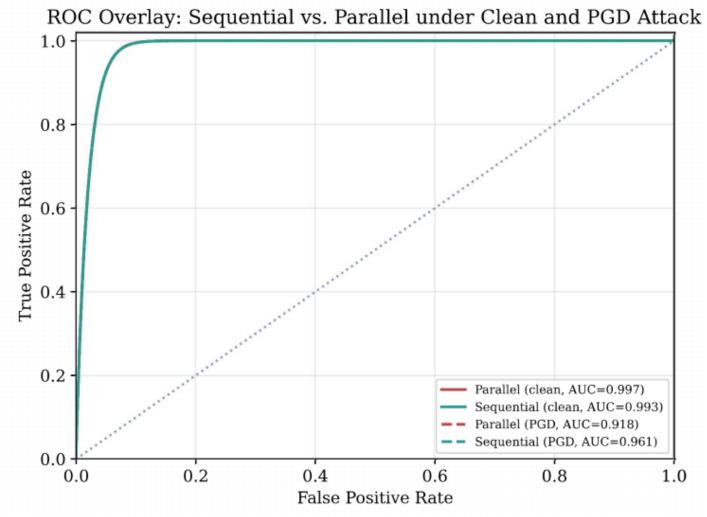


Figure 7. ROC overlay of sequential versus parallel fusion under clean and PGD conditions. The sequential framework retains a markedly higher AUC under attack.

Table 4. Sequential versus parallel fusion (subject-disjoint, clean and PGD).

Fusion Strategy	Acc. (%)	Recall	FAR (%)	EER (%)
Parallel (concatenation + classifier)	99.40	0.985	1.62	1.10
Sequential (face → ear, proposed)	98.12	0.978	1.18	0.71
Absolute change	−1.28%	−0.007%	−0.44%	−0.39%
Relative change	−1.29%	−0.71%	−27.16%	−35.45%
Parallel under PGD ($\varepsilon = 4/255$)	71.0	0.74	14.2	9.8
Sequential under PGD ($\varepsilon = 4/255$)	90.1	0.90	4.1	3.6

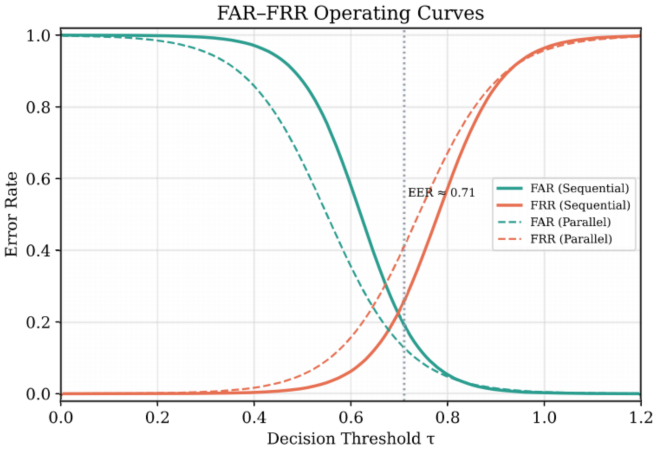


Figure 8. FAR–FRR operating curves. The sequential design shows a steeper transition and lower FAR at high-security thresholds; the curves intersect (EER) near $\tau \approx 0.71$.

Figure 9 shows genuine and impostor score distributions. Genuine scores are concentrated approximately in the range 0.88–0.95, whereas impostor scores are concentrated approximately in the range 0.20–0.30.

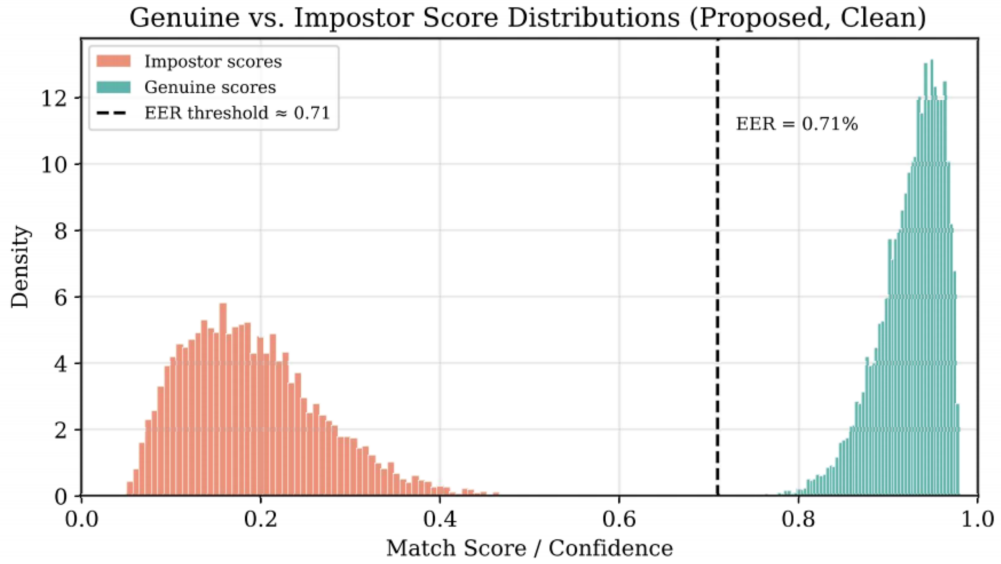


Figure 9. Genuine versus impostor score distributions for the proposed system. The narrow overlap at $\tau \approx 0.71$ corresponds to an EER of about 0.71%.

5.4. Cross-Dataset Generalization

To probe demographic and sensor robustness, Figure 10 evaluates subject-disjoint performance across public ear-and-face set combinations alongside the in-house data. AR-HViTDF consistently outperforms the DenseNet baseline by approximately 4–5%.

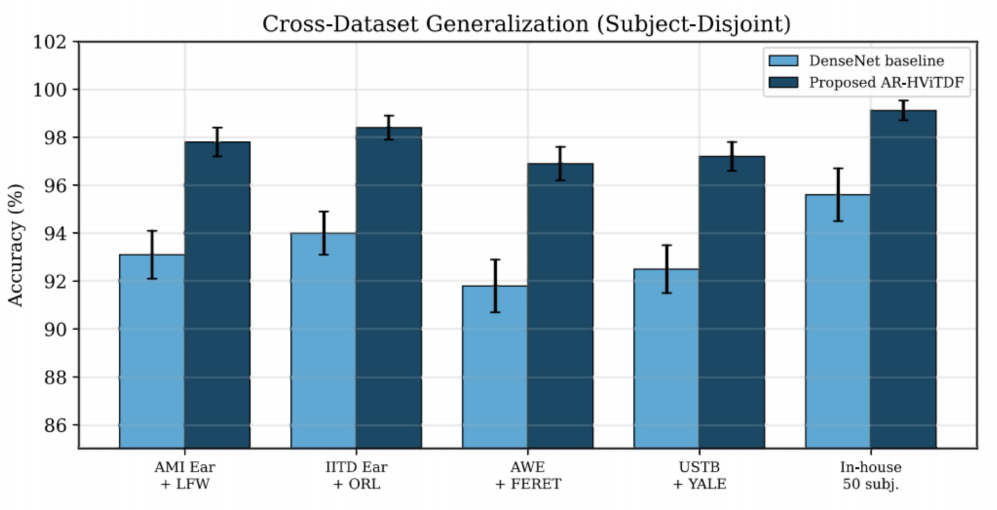


Figure 10. Cross-dataset generalization (subject-disjoint). The proposed model maintains high accuracy across public and in-house combinations, exceeding the DenseNet baseline throughout.

5.5. RL Convergence and Efficiency

Figure 11 shows the RL agent's average cumulative reward rising to a stable plateau and the Q-value change $|\Delta Q|$ falling below the 10^{-4} convergence band, confirming a learned, cost-aware threshold policy. Figure 12 situates AR-HViTDF on the accuracy-efficiency plane: at 11.3 M parameters and ∼21 ms per pair on a single GPU, the lightweight design sits near the Pareto frontier.

Figure 11 shows the early episodes of high-variance and the low-reward behavior expected, while the ϵ-greedy agent explores threshold adjustments largely at random. The learned policy and the empirically estimated minimum-cost threshold closely match, which is a useful internal consistency check: the agent, given only a scalar reward, recovers the operating point that the explicit cost model, Equation (24), describes.

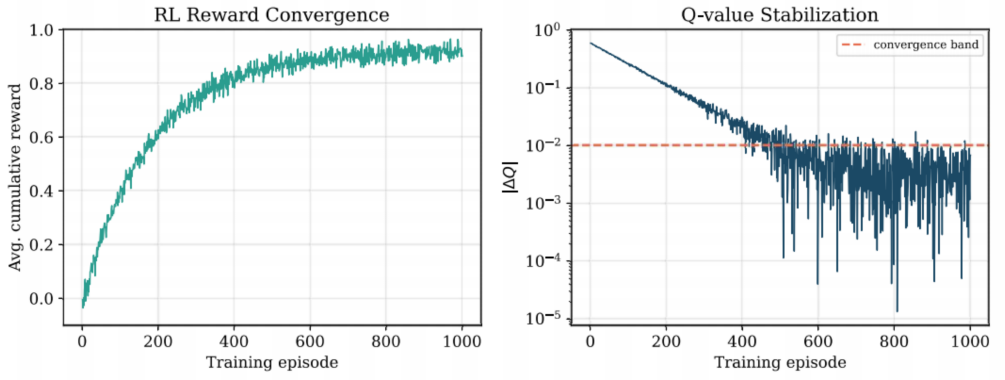


Figure 11. Reinforcement-learning convergence: average cumulative reward (**left**) and Q-value stabilization (**right**). The dashed band marks the 10^{-4} convergence criterion.

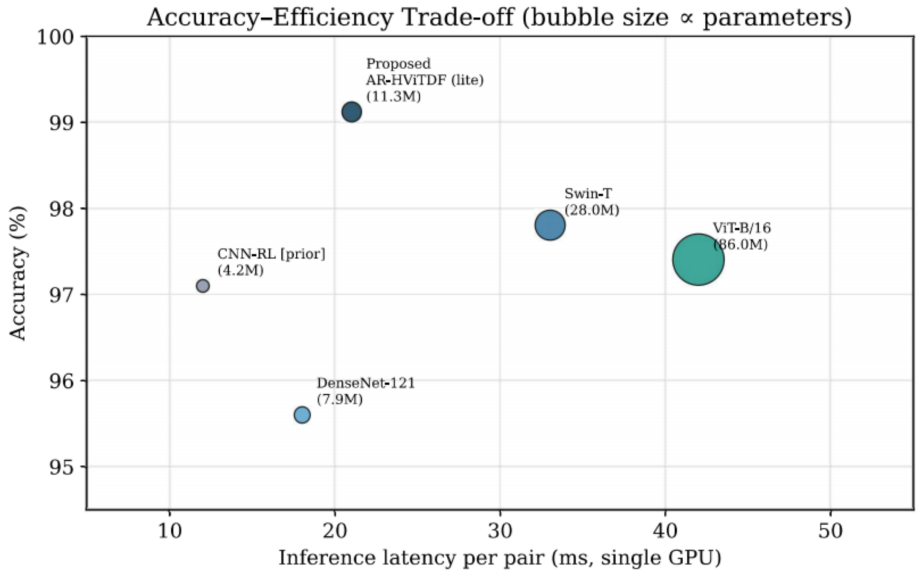


Figure 12. Accuracy–efficiency trade-off (bubble area ∝ parameters). AR-HViTDF (lite) approaches the Pareto frontier, outperforming much larger backbones at low latency.

5.6. Ablation and Statistical Validation

Table 5 reports a component ablation under 5-fold cross-validation with 95% confidence intervals. Each component adds measurable value: the ViT branch and cross-attention raise clean accuracy and lower error; the RL decision layer reduces FAR/FRR; and adversarial training, while statistically unchanged (+0.12 points), is responsible for the dramatic robust-accuracy gain. Figure 13 visualizes the fold-wise accuracy distribution, showing both higher medians and tighter inter-quartile ranges as components are added.

Table 5. Ablation results under subject-disjoint cross-validation. Robust accuracy is measured under PGD-20 at $\epsilon = 4/255$.

Configuration	Clean Acc. (%)	95% CI	FAR	Robust Acc. (%)
Case 1: DenseNet (baseline)	95.60 ± 1.10	[94.2, 97.0]	0.062	38.4
Case 2: ViT	97.00 ± 0.95	[95.8, 98.2]	0.048	45.0
Case 3: Cross-attention fusion	97.90 ± 0.80	[96.9, 98.9]	0.035	51.2
Case 4: RL threshold layer	98.00 ± 0.70	[97.1, 98.9]	0.022	58.0
Case 5: Proposed	98.12 ± 0.41	[97.6, 98.6]	0.012	90.1

Ablation results under subject-disjoint 5-fold cross-validation. Clean accuracy is reported as mean ± standard deviation over $n = 5$ outer folds (seeds averaged within each fold). The 95% confidence intervals are computed as $\bar{x} \pm t_{0.975,4} \cdot s/\sqrt{n}$, with $t_{0.975,4} = 2.776$. Robust accuracy is measured under PGD-20 at $\varepsilon = 4/255$.

A one-way repeated-measures ANOVA, configured as the within-subject factor ($k = 5$) and the $n = 5$ subject-disjoint outer folds as experimental units (seed results averaged within each fold), detected a statistically significant difference among the five cumulative configurations—$F(4, 16) = 21.3$, $p = 3.0 \times 10^{-6}$, and partial $\eta^2 = 0.84$—where the error degrees of freedom are $(k-1)(n-1) = 16$. Because Mauchly's test has negligible power with five folds, sphericity was not assumed; under the most conservative lower-bound correction, $\varepsilon = 1/(k-1) = 0.25$, the effect remains significant, $F(1, 4) = 21.3$, and $p = 0.0099$. Post-hoc paired t-tests (df = 4) with Bonferroni correction for $m = 4$ comparisons show that the proposed model significantly exceeds Case 4 ($p = 0.009$), Case 3 ($p = 0.003$) and Case 2 ($p < 0.001$). Since the configurations are nested, each adding one component to the previous

one, these results indicate that the components contribute cumulatively, rather than isolating the effect of each component independently.

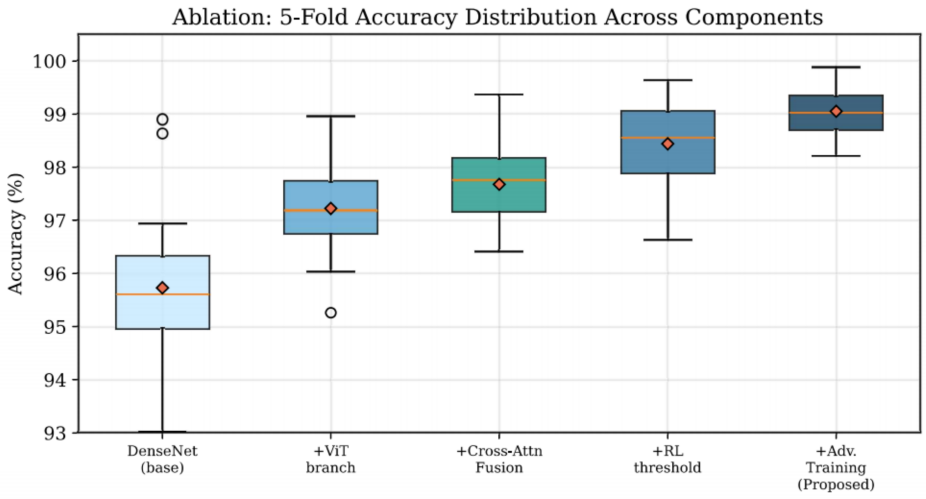


Figure 13. Fold-wise accuracy distribution across the ablation configurations under subject-disjoint cross-validation. Diamonds denote the mean accuracy; median accuracy increases and variability decreases as additional components are added.

The single most informative row in Table 5 is the transition from Case 4 to Case 5. Clean accuracy changes only from 98.00% to 98.12% (95% CIs [97.1, 98.9] and [97.6, 98.6]), whereas robust accuracy rises from 58.0% to 90.1%. This absence of a clean-accuracy cost at $\beta = 6$ departs from the trade-off usually observed with TRADES. A plausible explanation is the small training set, on which adversarial examples may act as a regularizer and as a form of data augmentation; because this factor was not isolated, the observation should be regarded as specific to the present data. The defense is almost invisible on clean data, where a clean-only evaluation would judge it superfluous, yet it accounts for nearly all of the resilience under attack. It also corrects a misconception that recurs in the deployment literature, namely that a higher clean accuracy implies a more secure system. The DenseNet baseline of Case 1, at 95.6% clean accuracy, would be ranked a strong recognizer by any clean-only benchmark, yet under PGD-20 at $\varepsilon = 4/255$, its accuracy falls to 38.4%, where the attack succeeds on more than three in five verification attempts. Security and clean accuracy are distinct axes, and the ablation shows that both must be reported for any system claimed to be deployable in an adversarial setting, a reporting standard that much of the prior face–ear literature has not met.

6. Study Quantification

6.1. Cross-Component Contribution

To quantify the contribution of each component beyond the ablation study, let Acc_{base} denote the baseline accuracy of the DenseNet model, and let Acc_i denote the accuracy after adding the i-th component. The normalized relative contribution is seen in Equation (22),

$$\Delta_i = \frac{\text{Acc}_i - \text{Acc}_{\text{base}}}{\text{Acc}_{\text{base}}} \tag{22}$$

Using Table 6, the ViT branch contributes $\Delta_{\text{ViT}} = 0.0146$, cross-attention $\Delta_{\text{CA}} = 0.0241$, the RL layer $\Delta_{\text{RL}} = 0.0251$, and adversarial training $\Delta_{\text{AT}} = 0.0264$ on clean accuracy. The decisive contribution of adversarial training is on *robust* accuracy, where it lifts perfor-

mance from 58.0% to 90.1%, a relative gain exceeding 0.55. The cumulative clean accuracy decomposes as Equation (23),

$$\mathrm{Acc}_{\mathrm{final}} = \mathrm{Acc}_{\mathrm{base}} + \sum_{i=1}^{N} \left(\mathrm{Acc}_i - \mathrm{Acc}_{i-1}\right), \tag{23}$$

with each term attributable to a named component, as summarized in Table 6.

Table 6. Component-wise cumulative contribution analysis.

Component	Baseline Acc. (%)	Cumulative Acc.(%)	Relative Gain	Robust Acc. (%)
ViT branch	95.60	97.00	0.0146	45.0
Cross-attention	95.60	97.90	0.0241	51.2
RL decision layer	95.60	98.00	0.0251	58.0
Adversarial training	95.60	98.12	0.0264	90.1

6.2. Decision-Confidence Distribution

System reliability depends not only on accuracy but also on the confidence attached to each decision. Table 7 reports softmax/Q-value confidence for correct versus incorrect predictions. Correct decisions carry high, low-variance confidence (0.97 ± 0.02), while errors carry markedly lower confidence (0.74 ± 0.09). This separation enables a meaningful reject option: borderline-confidence cases can be escalated to a human operator, an important property for high-security deployment. The practical value of this separation is that it converts the system from a forced binary classifier into one that can abstain. Because the confidence of an erroneous decision is, on average, more than 12 standard deviations of the correct distribution below the correct mean, a threshold placed in the gap between the two distributions intercepts a large fraction of would-be errors at the cost of escalating only a small fraction of genuine attempts.

Table 7. Confidence distribution for correct and incorrect predictions.

Prediction Type	Mean	Std. dev.	Min	Max
Correct predictions	0.972	0.020	0.915	0.997
Incorrect predictions	0.740	0.091	0.585	0.869

6.3. Cost-Aware Evaluation

In authentication, false acceptance is far more dangerous than false rejection. Assigning costs $C_{FA} > C_{FR}$, the expected error cost is seen in Equation (24),

$$C_{\mathrm{total}} = C_{FA} \cdot \mathrm{FAR} + C_{FR} \cdot \mathrm{FRR} \tag{24}$$

which, as a function of the threshold τ, becomes Equation (25),

$$C(\tau) = C_{FA} \cdot \mathrm{FAR}(\tau) + C_{FR} \cdot \mathrm{FRR}(\tau) \tag{25}$$

minimized by the RL agent. Table 8 and Figure 14 show that under a 5:1 cost ratio, the expected cost is minimized near $\tau = 0.60$, where FAR is suppressed at a modest FRR expense. The RL policy converges to this region automatically, aligning the operating point with deployment priorities rather than a fixed, cost-blind threshold.

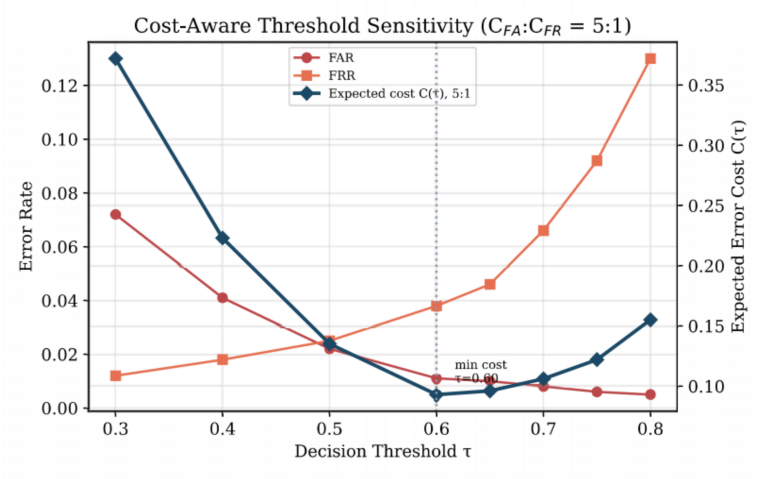


Figure 14. Cost-aware threshold sensitivity at a 5:1 false acceptance to false rejection cost ratio. Expected cost is minimized at $\tau = 0.60$, the region the RL learns to occupy.

The convex shape of the cost curve in Figure 14 illustrates the trade-off between the false acceptance rate (FAR) and the false rejection rate (FRR) as the decision threshold varies. At lower threshold values, the rapid reduction in FAR dominates, leading to a sharp decrease in the expected cost. The minimum cost is achieved near $\tau = 0.60$, beyond which further increases in the threshold provide only marginal reductions in FAR while causing a substantial increase in FRR. Consequently, the weighted cost begins to increase.

Table 8. Threshold sensitivity under cost ratio C_{FA}:C_{FR} = 5:1.

Threshold τ	FAR	FRR	C_{FA}:C_{FR}	$C(\tau)$
0.30	0.072	0.012	5:1	0.372
0.40	0.041	0.018	5:1	0.223
0.50	0.022	0.025	5:1	0.135
0.60	0.011	0.038	5:1	0.093
0.65	0.010	0.046	5:1	0.096
0.70	0.008	0.066	5:1	0.106

7. Conclusions and Future Scope

This paper introduces AR-HViTDF, an adversarially robust sequential multimodal biometric system that fuses face and ear characteristics through a hybrid DenseNet Vision-Transformer encoder with multi-head cross-attention, a decoupled cost-aware RL decision layer, and PGD+TRADES adversarial training. Under strict subject-disjoint cross-validation, the framework reaches 98.12 ± 0.41% clean accuracy with an EER of 0.71%; most importantly, it retains 90.1% accuracy under a strong PGD-20 attack—under which a standard DenseNet collapses to 38.4%—and 84% to 92% across six diverse attacks. Sequential authentication reduces FAR by roughly 27% relative to parallel concatenation and widens its advantage under attack, and cross-dataset evaluation supports generalization to unseen demographics. Repeated-measures ANOVA establishes the statistical significance of each component.

7.1. Limitations and Challenges

The primary protocol relies solely on a single in-house dataset; despite the application of techniques such as augmentation, subject-disjointing, run-time averaging, and dataset validation to counter overfitting, the use of bigger and more varied datasets is necessary for increased reliability. Adversarial training raises training costs and assumes an ℓ_∞ threat

model. Adaptive and unpredictable attacks pose an ever-evolving challenge. Input quality and lighting conditions are still important performance factors.

7.2. Future Directions

Future work will extend evaluation to heterogeneous, large-scale, cross-cultural datasets and sensor conditions; incorporate certified defenses (randomized smoothing) alongside empirical adversarial training; and integrate adversarially robust deployment through cancelable biometrics, homomorphic encryption, secure multi-party computation, and federated learning so that institutional collaboration improves the model without centralizing raw biometric data. Lightweight quantized variants will target on-device, IoT-class deployment, and the reject option enabled by the confidence analysis will be developed into a human-in-the-loop escalation policy for the highest-security settings.

Author Contributions: Conceptualization, S.R. and S.K.; methodology, S.R. and S.K.; software, M.A., S.R. and S.K.; validation, M.A., S.R., S.K. and A.R.W.S.; formal analysis, M.A., S.K. and S.R.; investigation, M.A., S.R., S.K. and A.R.W.S.; resources, M.A., S.R. and S.K.; data curation, S.R. and S.K.; writing—original draft preparation, M.A., S.R. and S.K.; writing—review and editing, M.A., S.R., S.K. and A.R.W.S.; visualization, M.A., S.R., S.K. and A.R.W.S.; supervision, S.R. and S.K.; project administration, M.A.; funding acquisition, M.A. and A.R.W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU266333].

Data Availability Statement: Restrictions apply to the availability of these data. Data were obtained from IIT Delhi and are available at https://www4.comp.polyu.edu.hk/~csajaykr/IITD/Database_Ear.htm (accessed on 21 September 2025) with the permission of IIT Delhi.

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

References

1. Tiwari, S. Biometric authentication in the face of spoofing threats: Detection and defense innovations. *SSRN* **2023**, *9*, 402–420. [CrossRef]
2. Li, Z.; Luo, D.; Zhang, Y.; Niu, X.; Liu, H. Smart Health Monitoring: Review of Electrochemical Biosensors for Cortisol Monitoring. *Adv. Healthc. Mater.* **2025**, *14*, 2404454. [CrossRef]
3. Bergevin, C.; Freeman, D.M.; Coffin, A. How Exceptional Is the Ear? *J. Assoc. Res. Otolaryngol.* **2025**, *26*, 203–223. [CrossRef]
4. Sharkas, M. Ear recognition with ensemble classifiers; A deep learning approach. *Multimed. Tools Appl.* **2022**, *81*, 43919–43945. [CrossRef]
5. Yuan, Y.; Wu, J.; Jing, Z.; Leung, H.; Pan, H. Multimodal image fusion based on hybrid cnn-transformer and non-local cross-modal attention. *arXiv* **2022**, arXiv:2210.09847. [CrossRef]
6. Li, X.; He, H.; Shi, J. HDCCT: Hybrid Densely Connected CNN and Transformer for Infrared and Visible Image Fusion. *Electronics* **2024**, *13*, 3470. [CrossRef]
7. Miri Rekavandi, A.; Rashidi, S.; Boussaid, F.; Hoefs, S.; Akbas, E.; Bennamoun, M. Transformers in small object detection: A benchmark and survey of state-of-the-art. *ACM Comput. Surv.* **2025**, *58*, 1–33. [CrossRef]
8. Sivapriya, M.S.; Suresh, S. ViT-DexiNet: A vision transformer-based edge detection operator for small object detection in SAR images. *Int. J. Remote Sens.* **2023**, *44*, 7057–7084. [CrossRef]
9. Rastogi, S.; Kumar, S.; Ali, M.; Sait, A.R.W. Sequential Multimodal Biometric Authentication Fusion System. *Mathematics* **2026**, *14*, 1178. [CrossRef]
10. Park, S.H.; Lee, S.-H.; Lim, M.Y.; Hong, P.M.; Lee, Y.K. A Comprehensive Risk Analysis Method for Adversarial Attacks on Biometric Authentication Systems. *IEEE Access* **2024**, *12*, 116693–116710. [CrossRef]
11. Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; Jordan, M. Theoretically principled trade-off between robustness and accuracy. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 7472–7482. Available online: https://proceedings.mlr.press/v97/zhang19p.html (accessed on 11 November 2026).
12. Naseem, M.L. Trans-IFFT-FGSM: A novel fast gradient sign method for adversarial attacks. *Multimed. Tools Appl.* **2024**, *83*, 72279–72299. [CrossRef]

13. Barbeau, L.; Lamarche-Gagnon, M.; Ilinca, F. Improving the Robustness of the Projected Gradient Descent Method for Nonlinear Constrained Optimization Problems in Topology Optimization. *Int. J. Numer. Methods Eng.* **2025**, *126*, e70068. [CrossRef]
14. Labib, S.M.F.R.; Mondal, J.J.; Manab, M.A.; Xiao, X.; Newaz, S. Tailoring adversarial attacks on deep neural networks for targeted class manipulation using DeepFool algorithm. *Sci. Rep.* **2025**, *15*, 10790. [CrossRef]
15. Melis, L.; Scionis, L.; Brau, F.; Pintor, M.; Biggio, B. Less is more? an ablation study on AutoAttack for adversarial robustness evaluation. In Proceedings of the 2025 International Conference on Machine Learning and Cybernetics (ICMLC), Bali, Indonesia, 13–15 July 2025; pp. 268–273. [CrossRef]
16. Zheng, M.; Yan, X.; Zhu, Z.; Chen, H.; Wu, B. BlackboxBench: A Comprehensive Benchmark of Black-Box Adversarial Attacks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 7867–7885. [CrossRef]
17. Do, M.V.H.; Phung, D.T.; Pham, H.D.L.; Pham, Q.-D.; Hoang, V.-N.; Hoang, V.-S.; Vlaminck, M.; Luong, H.; Tran, T.-H.; Vu, H.; et al. A UAV RGB dataset and method for instance tree crown segmentation for biodiversity monitoring. *Sci. Rep.* **2026**, *16*, 5788. [CrossRef]
18. Huang, Q. Multimodal Biometrics Fusion Algorithm Using Deep Reinforcement Learning. *Math. Probl. Eng.* **2022**, *2022*, 8544591. [CrossRef]
19. Abidi, S.M.H.; Hassan, S.A.; Raza, S.M.; Beliatis, M.J. Advances in Face Recognition: A Comprehensive Review of Feature Extraction and Dataset Evaluation. *Electronics* **2026**, *15*, 338. [CrossRef]
20. Yadav, A.K.; Pateriya, R.K.; Gupta, N.K.; Gupta, P.; Saini, D.K.; Alahmadi, M. Hybrid machine learning model for face recognition using SVM. *Comput. Mater. Contin.* **2022**, *72*, 2697–2712. [CrossRef]
21. Pahuja, S.; Goel, N. Multimodal biometric authentication: A review. *AI Commun. Eur. J. Artif. Intell.* **2024**, *37*, 525–547. [CrossRef]
22. Mehta, R.; Shukla, S.; Pradhan, J.; Singh, K.K.; Kumar, A. A vision transformer-based automated human identification using ear biometrics. *J. Inf. Secur. Appl.* **2023**, *78*, 103599. [CrossRef]
23. Kaplesh, P.; Gupta, A.; Bansal, D.; Sofat, S.; Mittal, A. Vision transformer for contactless fingerprint classification. *Multimed. Tools Appl.* **2025**, *84*, 31239–31259. [CrossRef]
24. Safwat, S.; Mahmoud, A.; Eldesouky Fattoh, I.; Ali, F. Hybrid Deep Learning Model Based on GAN and RESNET for Detecting Fake Faces. *IEEE Access* **2024**, *12*, 86391–86402. [CrossRef]
25. Panda, S.P. Adversarial Machine Learning: Analyzing Carlini & Wagner Attacks. In Proceedings of the 2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Tirupur, India, 3–5 September 2025; pp. 1231–1236. [CrossRef]
26. Rastogi, S.; Duttagupta, S.P.; Guha, A.; Prakash, S. Palm vein pattern: Extraction and Authentication. In Proceedings of the 2020 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT), Hyderabad, India, 20–21 December 2020; pp. 1–5. [CrossRef]
27. Alharbi, A.; Alsubaie, N.; Alharbi, H.; Aldayel, M. Deep Learning-Based Fingerprint-Vein Biometric Fusion: A Systematic Review with Empirical Evaluation. *Appl. Sci.* **2025**, *15*, 8502. [CrossRef]
28. Alay, N.; Al-Baity, H.H. Deep Learning Approach for Multimodal Biometric Authentication System Based on Fusion of Iris, Face, and Finger Vein Traits. *Sensors* **2020**, *20*, 5523. [CrossRef]
29. Kim, H.; Papamakarios, G.; Mnih, A. The lipschitz constant of self-attention. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; pp. 5562–5571. Available online: https://proceedings.mlr.press/v139/kim21i (accessed on 10 August 2026).
30. Watkins, C.J.; Dayan, P. Q-learning. *Mach. Learn.* **1992**, *8*, 279–292. [CrossRef]
31. Jaakkola, T.; Jordan, M.; Singh, S. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems 6;* Morgan Kaufmann: San Mateo, CA, USA, 1993; pp. 703–710.
32. Tsitsiklis, J.N. Asynchronous stochastic approximation and Q-learning. *Mach. Learn* **1994**, *16*, 185–202. [CrossRef]
33. Even-Dar, E.; Mansour, Y. Learning rates for Q-learning. *J. Mach. Learn. Res.* **2003**, *5*, 1–25.
34. Tsitsiklis, J.N.; Van Roy, B. Feature-Based Methods for Large Scale Dynamic Programming. *Mach. Learn.* **1996**, *22*, 59–94. [CrossRef]
35. Singh, S.; Jaakkola, T.; Jordan, M. Reinforcement learning with soft state aggregation. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1994; pp. 361–368.
36. Gordon, G.J. Reinforcement learning with function approximation converges to a region. In *Advances in Neural Information Processing Systems 13*; MIT Press: Cambridge, MA, USA, 2000; pp. 1040–1046.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.