

RAFA-BioAuth: Risk-Adaptive and Fairness-Aware Biometric Authentication for Secure Mobile Banking

Dendy K. Pramudito*

Department of Informatics Doctoral, Faculty of Information Technology, Universitas Nusa Mandiri, Jakarta, 12540, Indonesia

E-mail: 16240003@nusamandiri.ac.id

ORCID iD: <https://orcid.org/0000-0003-3079-8319>

*Corresponding author

Jufriadif Na'am

Department of Informatics Doctoral, Faculty of Information Technology, Universitas Nusa Mandiri, Jakarta, 12540, Indonesia

E-mail: jufriadifnaam@gmail.com

ORCID iD: <https://orcid.org/0000-0002-9586-7031>

Ferda Ernawan

Computing Program, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan Pahang, 26600, Malaysia

E-mail: ferda@umpsa.edu.my

ORCID iD: <https://orcid.org/0000-0002-6779-1594>

Received: April 6, 2026; Revised: May 5, 2026; Accepted: June 6, 2026; Published: August 8, 2026

Abstract: This research examines RAFA-BioAuth, a risk-adaptive, fairness-aware framework for mobile banking in the presence of presentation attacks and facial occlusions. The proposed solution combines aspects of identity similarity, passive presentation attack detection (PAD), asymmetric financial-risk estimation, and fairness regularization at the identity level. For evaluation purposes, all benchmark datasets were split into subject disjoint training, validation, and testing sets. The threshold values from the validation set were used. Bootstrap resampling was employed to estimate the variance. Monte Carlo simulations were performed to estimate the risk. The results showed that identity verification ($AUC = 0.548$) and PAD ($AUC \approx 0.55$) were poor. Comparing RAFA to the AND rule resulted in $FAR = 0.20$, $FRR = 0.31$, and expected risk = 340. On the other hand, the AND rule had lower FAR of 0.09, but increased FRR to 0.78. Finally, a conservative end-to-end approach produced an FRR of 0.686. Thus, our results indicate trade-offs rather than production-readiness, as we did not perform deployment, cross-device, or cross-dataset validations. We present the contributions of this research as being an interpretable integration and not a new algorithm.

Index Terms: Biometric Authentication, Mobile Banking, Presentation Attack Detection, Risk-Aware Fusion, Fairness-Aware Authentication, Financial Risk Optimization.

1. Introduction

The development of secure and convenient methods of authentication is one of the main drivers behind the rapid expansion of mobile banking. Biometric authentication technologies provide an attractive combination of user experience and security. Face recognition technology exemplifies this balance by enabling frictionless authentication without the need for additional hardware. Face recognition technology has therefore been widely adopted across various fields [1,2,3].

However, despite being widely adopted across many areas of use, face recognition technology still has a number of significant limitations. One major limitation of face recognition technology is its susceptibility to presentation attacks. Presentation attacks include printed photographs and digital video replays of facial images. If such attacks are not effectively mitigated, they may result in unauthorized access to individuals' accounts [4,5,6,7,8]. As a result, these forms of presentation attacks are particularly important to mitigate in the context of financial applications. This is

because false acceptance rates are directly correlated to potential financial loss [1,3].

A second limit is performance degradation due to facial occlusion. Masks represent a type of structured occlusion, removing or redistributing the identity features used by face-recognition models. Therefore, models trained mostly on unobstructed faces will have decreased separation between true and impostor comparisons when part of the face is obscured, thereby reducing their reliability in mobile banking [1,2,9].

Existing research covers presentation attacks, occlusion, mobile authentication and biometric fusion. However, each topic is usually examined with a different objective and protocol. Methods resistant to presentations are optimized for detecting presentations; methods aware of occlusions are optimized for robust recognition; mobile authentication studies usually focused on context and policies. In addition, the enhancements described in one area do not automatically imply lower total banking risk or greater fairness in user outcomes [1,3,4,18].

Research has also been conducted on detection of presentations through remote photoplethysmography, convolutional liveness classification, time-based analysis, physiological reactions, and multi-modal sensing [10,14,15,31,32,33]. The approaches exhibit great differences in sensor needs, attack coverage, and test protocols. This paper's unresolved question is not which of the presented PAD techniques is better, but how the identity and liveness evidence will interact with each other at a decision level, in a controlled protocol for a high-risk app.

Although many previous studies evaluate face verification and liveness detection separately, operational authentication decisions rely on the interaction of both [8,12,14,16,18]. Incorrect calibration can allow spoofer acceptance or cause rejecting valid users. An important end-to-end evaluation would therefore need to provide joint security, usability, cost, fairness, and uncertainty rather than just individual accuracy for each component [8,10,12,15,17].

In order to fill this gap, this study examines decision-level integration of face-identity similarity and passive PAD evidence for mobile-banking authentication. It does not develop a new face recognition backbone or PAD algorithm. The contribution of this study is the structure of combining well-established biometric evidence with asymmetric risk modeling, identity-level fairness regularization and end-to-end comparison within a common controlled protocol [8,10,12,15,17].

The identity and liveness threshold values were chosen using validation data. Subject-disjoint partitioning was used to prevent data leakage between training, validation and testing. Repeated resampling was used to assess internal sampling uncertainty. Although these controls enhance internal validity, they cannot replace cross-device, cross-dataset, longitudinal, or real mobile-banking validations.

Therefore RAFA-BioAuth should be considered as an interpretable decision framework that integrates identity similarity evidence (face), liveness evidence (liveness), estimated relative financial-risk (risk) and fairness-aware regularization. The biometric and statistical parts of the method are established methods and are not claimed to be individually novel algorithms.

This paper provides four verifiable contributions. Firstly, it represents authentication as a decision made jointly based upon identity similarities and liveness evidence. Secondly it compares verification only use of the methods listed above with AND rule use, weighted fused results and risk adaptive decision boundaries using the same measurements (scores), subject splits and validation procedures. Thirdly it examines false accept and false reject costs with variation due to identity rejection. Lastly it reported on security, usability, relative expected risk, fairness dispersion and internal sampling uncertainty. All conclusions were limited to the controlled bench mark protocol.

2. Methods

2.1. System Overview

The structure of the RAFA-BioAuth system is divided into four levels — identity verification, PAD, fairness regulation, and RFR — and these are based on previously described principles for biometric fusion and secure authentication systems [18,19] but instead of using an individualized threshold or single fusion-rule approach, this framework evaluates the overall success in authenticating a user as a cost-based and fairness-restricted decision-making process. The four layers in which this decision making occurs are: facial embedding generation; liveness determination; risk adaptive aggregation of evidence; multi-criteria assessment.

2.2. Datasets and Experimental Protocol

The experiments utilize public benchmark databases which have been used for both face recognition (FR) and presentation attack detection (PAD). These databases include variability in facial expressions, lighting, and head poses. They also contain images and videos of bonafide subjects, as well as replayed and printed photos of the subjects. This will allow us to perform comparative studies with a controlled protocol; however, these databases do not reflect how users interact with their mobile banking systems and thus we cannot assess the external validity of our results when deployed in a production environment.

We divide each database into disjoint subsets for testing, validating, and training. This ensures no overlap in subject identities across subsets. We compare the verification performance of the different methods by analyzing their similarity distributions for genuine and imposter pairs. We determine whether an image or video represents an attack

based on the PAD score. All thresholds, fusion weights, and regularization parameters are determined using only the validation set. Therefore, all evaluations occur on the test set only.

We use bootstrap resampling techniques to estimate the variance associated with internal sampling uncertainty. We also utilize Monte Carlo simulations to provide estimates of the distribution of relative expected loss assuming that both false accept and false reject rates incur equal losses. While these two methodologies help to alleviate dependency on a particular sample realization, neither methodology provides cross-dataset, cross-device, nor operational validation.

To ensure a fair comparison, each method utilizes the same subject disjoint partitions, identities, identity scores, PAD scores, and validation procedure as RAFA-BioAuth. Since each method has only been trained on the validation sets prior to testing, the comparison isolates differences in the decision rules employed rather than differences in the modern biometric processing backbone or sensor hardware.

Unfortunately, due to availability constraints, we did not have access to any mobile banking biometric log files, independent external datasets, or cross-device tests. Thus, our findings provide evidence of decision framework behavior in a controlled study. Subject-disjoint splitting, validation-only parameter optimization, resampling, and sensitivity analysis minimize selected internal validity risks; however, they cannot validate generalized behavior across different populations, devices, environments or forms of attacks.

2.3. Identity Feature Extraction

The first stage of the framework involves discriminative identity representation. A convolutional neural network backbone maps facial images into a compact embedding space. Specifically, the final classification layer is replaced with an identity-preserving projection layer, yielding embeddings:

$$f(x) \in \mathbb{R}^d \quad (1)$$

Similarity between two face images x_i and x_j is computed using cosine similarity:

$$S(x_i x_j) = \frac{f(x_i) f(x_j)}{\|f(x_i)\| \|f(x_j)\|} \quad (2)$$

The decision threshold τ_{id} is determined via Equal Error Rate (EER) analysis, defined as the operating point where:

$$FART(T_{id}) = FRR(T_{id}) \quad (3)$$

This embedding representation provides the quantitative identity evidence for the authentication decision.

2.4. Presentation Attack Detection

Although identity similarity measures user identity, it does not determine whether the sample originates from a live subject. Therefore, a PAD module is integrated to detect spoof attacks. The importance of liveness analysis in secure biometric authentication has been widely emphasized in both general biometric security surveys and multimodal anti-spoofing studies [19,20].

Given an input face image x , the PAD model produces a liveness probability:

$$L(x) = P(\text{live} | x) \quad (4)$$

A decision threshold τ_{live} is selected to minimize the Average Classification Error Rate (ACER):

$$ACER = \frac{APCER + BPCER}{2} \quad (5)$$

where APCER denotes attack presentation classification error rate and BPCER denotes bona fide presentation classification error rate. The PAD output constitutes a probabilistic measure that complements identity similarity in subsequent fusion.

2.5. Research Gaps and Position

Relevant literature can be categorized into four categories: Risk- or Context-Aware Access Control, Multimodal Biometric Fusion, Presentation Attack Detection, and Occlusion-Robust Verification. The above mentioned threads are optimized for different purposes and are normally tested with different data sets, sensors and attacks; direct numerical comparison is therefore not meaningful.

Studies focusing on Risk-Aware Mobile Authentication generally focus on the best way to escalate context or policy rather than integrating face similarity and passive liveness scores [43,44,45]. PAD research generally focuses on

discrimination against spoofs and multimodal or robustness oriented designs [28,31,32,33]; however, it generally does not translate individual errors into end-to-end banking loss and user rejection.

Occlusion-aware verification focuses on recognition robustness, and score-level fusion is based on empirical weight selection and aggregation metrics [18]. Therefore, the key is to evaluate evidence of identity, liveness, asymmetry costs, and identity-level variation upon rejection under the same process.

RAFA-BioAuth addresses this integration gap rather than a gap in algorithmic novelty. RAFA-BioAuth's methodological contribution is to show how security, usability, relative expected loss, and fairness disparity change together under the same decision model. Regardless of the quality of the overall system, it has limits determined by the weak identity and PAD components used in the test.

Table 1. Differentiation between RAFA-BioAuth and closely related authentication studies.

Study	Focus	Fusion	Risk / Fairness	PAD / Occlusion	Type	Difference from RAFA-BioAuth
Chen et al. (2024)	Risk-aware access	Policy or multi-stage	Risk-aware; no identity-level fairness	No joint face verification + passive PAD banking test	Access-control	Does not jointly combine identity score, liveness, financial loss, and fairness regularization.
Hintze et al. (2019)	Mobile risk authentication	Contextual multimodal fusion	Risk-aware; no explicit financial-loss objective	Not focused on passive face PAD	Continuous authentication	Addresses mobile risk, but not banking PAD + financial risk + identity fairness together.
Ayeswarya and Singh (2025)	Context-aware biometrics	Adaptive biometric fusion	Security-usability focus; limited cost modelling	Not centred on passive face liveness	Multi-biometric	Improves fusion but does not integrate passive PAD, banking loss, and fairness dispersion.
Mandalapu et al. (2021)	Audio-visual biometrics and PAD	Survey of multimodal recognition and attack detection	Discusses security; no RAFA-style optimization	Broad PAD and biometric coverage	Survey	Establishes the field context but does not test financial-risk/fairness fusion.
Tian et al. (2023)	Multimodal face PAD	Polarization-oriented multimodal PAD	No financial-risk or identity-fairness objective	Advanced PAD; not a joint banking decision test	PAD method	Provides a stronger PAD direction, but uses a different sensor/protocol and does not evaluate joint identity, banking cost, and fairness.
Proposed RAFA-BioAuth	Mobile-banking biometric authentication	Risk-adaptive identity + liveness fusion	Explicit financial risk and identity fairness	Joint verification + passive PAD under occlusion/spoofing	Interpretable decision framework	Integrates the four objectives in one controlled decision framework, but does not provide external validation or state-of-the-art component performance.

Table 1 presents selected studies based on their scope rather than their reported performance metrics. Due to the differences in each of the earlier works' respective datasets, sensors, type(s) of attack(s), preprocessing pathways, and assessment metrics, previous numerical results cannot be directly compared. This experiment will re-implement only four decision procedures using the exact same split and score dataset, and therefore this experiment will not compare the present biometric-based decision processes with the state-of-the-art for the underlying biometric component(s).

Therefore, the empirical claims of this study are limited to verification-only, AND-rule, weighted fusion, and RAFA-based decisions, all assessed using an identical controlled experimental procedure. Any strong PAD or face-verification backbone can significantly alter both the absolute false accept rates (FARs) and false reject rates (FRRs). To conduct an external comparison that accounts for all the potential variables, such as training and testing modern methods on the same data, devices, attacks, and evaluation criterion(s) would be required.

2.6. Conventional Fusion Strategies

Prior to introducing the proposed risk-aware authentication framework, several conventional fusion strategies are implemented as baseline models. These methods represent commonly used decision mechanisms in biometric authentication systems and provide reference points for evaluating the effectiveness of the proposed approach. Specifically, three representative strategies are considered: verification-only authentication, conjunctive fusion using an AND rule, and weighted linear score fusion. Each method combines identity similarity and liveness information in different ways, and therefore exhibits distinct trade-offs between security and usability, consistent with the broader biometric fusion literature [18].

2.6.1. Verification-Only (VO)

The verification-only strategy represents the most basic form of biometric authentication, in which the authentication decision relies exclusively on identity similarity between the probe image and the stored template. In this

approach, the system evaluates whether the similarity score exceeds a predefined identity threshold. The decision function is defined as:

$$D_{v_0}(x) = 1(S \geq \tau_{id}) \quad (6)$$

where S denotes the cosine similarity between the identity embeddings and τ_{id} represents the verification threshold. The indicator function returns a positive authentication decision when the similarity score exceeds the threshold. Although this strategy is computationally simple and widely used in conventional face verification systems, it neglects additional security signals such as liveness detection. Consequently, the model is vulnerable to presentation attacks and does not account for the asymmetric economic consequences of authentication errors in financial applications [19].

2.6.2. AND-Rule Fusion

To improve security against spoofing attacks, many biometric systems incorporate presentation attack detection (PAD) signals in addition to identity verification. A common strategy for combining these signals is conjunctive fusion, where authentication is granted only if both identity similarity and liveness probability exceed their respective thresholds. The decision rule is formulated as:

$$D_{AND}(x) = 1(S \geq \tau_{id}) \wedge 1(L \geq \tau_{live}) \quad (7)$$

where L denotes the liveness probability produced by the PAD model and τ_{live} is the corresponding liveness threshold.

The AND rule requires that both identity similarity and liveness be greater than their respective threshold values. A rectangular acceptance area as shown above, can provide a level of protection against false acceptances (when both signals are good) but will reject valid subjects anytime one of the two signals becomes corrupted due to some form of noise, occlusion, or variability in the capture process [18,20,21].

2.6.3. Weighted Linear Fusion (WF)

An alternative approach to multimodal authentication is weighted linear fusion, which combines identity similarity and liveness scores into a single continuous decision variable. This method allows the system to balance the contribution of each modality through adjustable weights. The fused score is defined as:

$$R_{(WF)} = w_1 S + w_2 L \quad (8)$$

where w_1 and w_2 represent empirically determined weights that control the relative importance of identity similarity and liveness detection. Authentication decisions are then determined using a single threshold applied to the fused score:

$$D_{(WF)}(x) = 1(R_{(WF)} \geq \tau) \quad (9)$$

Compared with the AND-rule strategy, weighted fusion produces a smoother decision boundary and allows gradual trade-offs between verification and liveness signals. However, the method remains heuristic in nature because the weights are typically selected empirically rather than derived from principled optimization. Moreover, the approach does not explicitly consider fairness constraints across identities or the asymmetric costs associated with authentication errors in financial systems. These characteristics are consistent with broader observations in the biometric fusion literature regarding the strengths and limitations of score-level integration methods [18].

2.7. Proposed Risk-Adaptive Fairness-Aware Fusion (RAFA-BioAuth)

To overcome these limitations, the proposed RAFA-BioAuth introduces fairness regularization within a risk-minimization framework. The composite risk score is defined as:

$$R_{(RAFA)} = w_1 S + w_2 L - \lambda \sigma_l \quad (10)$$

where σ_l denotes intra-identity instability and λ controls fairness regularization strength.

Authentication is determined by:

$$D_{(RAFA)}(x) = 1(R_{(RAFA)} \geq \tau_{(risk)}) \quad (11)$$

The threshold τ_{risk} is optimized by minimizing expected financial loss:

$$R = C_{(FA)} \cdot FAR + C_{(FR)} \cdot FRR \quad (12)$$

where CFA is the assumed cost of a false acceptance, and CFR is the assumed cost of a false rejection. The model changes authentication to an asymmetric decision-risk type issue. Therefore, the resultant risk values represent relative experimental units based on the cost assumptions; they do not measure actual bank loss [18,19,20].

2.8. Fairness Quantification

Fairness is evaluated at the identity level. For each identity i , the false rejection rate is computed as:

$$FRR_i = \frac{\text{False Rejections for identity } i}{\text{Total genuine attempts for } i} \quad (13)$$

To measure dispersion across identities, the coefficient of variation (CV) is defined:

$$CV = \frac{\sigma(FRR_B)}{\mu(FRR_B)} \quad (14)$$

Lower CV indicates more equitable authentication reliability. By incorporating σ_i into the fusion score, the proposed framework explicitly penalizes identities with high instability.

A very high level of FRR at the identity level is to be treated as a trigger to initiate a non-exclusionary fallback process instead of permanent refusal. Some candidate safeguard measures that could limit the damage caused by a false match include; a fixed number of attempts to re-capture (recaptures), quality guided re-enrolment, one or more additional live captures on top of what was captured during initial enrolment, and/or step-up authentication through a completely separate and independent communication channel. The thresholds selected must also remain below predetermined FAR limits.

2.9. Multi-Objective Optimization and Pareto Analysis

Because minimizing financial risk and minimizing fairness dispersion may conflict, the model supports multi-objective optimization. By varying λ , a Pareto frontier is generated in the objective space:

Operational Pareto tuning may also be performed as a fixed validation procedure: (1) establish an organization-specific upper bounds for FAR, FRR, identity-level CV, and customer-friction cost; (2) examine all combinations of a predeclared set of identity threshold levels, liveness threshold values, fusion parameter settings, and fairness regularization strength values against a validation buffer from recent time; (3) eliminate each combination which is in violation of an upper bound; (4) determine the subset of remaining feasible (non-violating) combinations such that no combination in this subset has higher values than another combination within the subset for at least one metric; (5) choose the feasible point with the minimum assumed expected loss value; and (6) validate the chosen feasible point on a separate "hold out" validation buffer before deploying it.

$$(R, CV) \quad (15)$$

Points in the Pareto frontier are only conditional optima for the specified objective(s), constraint(s), data, and cost assumption(s). The points do not constitute an optimal universal bank policy.

2.10. Calibration Analysis

To ensure probabilistic reliability of fused scores, logistic calibration is applied. Calibration performance is evaluated using the Brier score:

$$\text{Brier} = (1/N) \sum_{i=1}^N (p_i - y_i)^2 \quad (16)$$

where p_i denotes predicted probability and y_i denotes ground-truth label. Lower Brier scores indicate better alignment between predicted confidence and empirical correctness, strengthening risk-based decision validity.

The Controlled Adaptive-Threshold Deployment Protocol is a process made up of five steps: (1) Collecting a labeled validation buffer representative of approved devices, capture conditions, and existing usage; (2) Calculating false alarm rate (FAR), false rejection rate (FRR), relative expected loss, PAD uncertainty and identity level variance for each of the pre-declared candidates; (3) Rejection of those candidates that either violate organizational security or fairness criteria; (4) Selection of the point representing the lowest total loss from the remaining Pareto set, and verification of that point on a separate hold-out buffer; (5) Monitoring drift with explicit step-up and roll-back strategies. Thresholds should never be changed based upon verified production results or solely based upon acceptable transactional activity since this can cause feedback and could introduce security risks.

2.11. Statistical Validation

Bootstrap resampling was performed to calculate the variance due to internal sampling variability. In each resample a subset is drawn with replacement from the evaluation data. Then each subset computes FAR and the coefficient of variation (cv) as an indication of how much different identities are treated fairly. These results then form empirical distributions on which can be built the confidence intervals and visualizations of uncertainty. This analysis investigates how sampling variation occurs within the given benchmark data, i.e., this study does not examine the possibility of generalizing the results to new sets of data or new hardware.

To compute the confidence ellipse in FAR-cv space using the covariance matrix of the bootstrap estimates. If smaller ellipses result it indicates less resampling variability within the tested dataset and NOT better protection against new device and/or population types. McNemar's Test is used to make a comparison based on pairs between RAFA BioAuth and a baseline set of decision rules to see if there is a significant difference between what happens when those two systems disagree and where both systems agree.

Monte Carlo simulations are also used to analyze the stability of the financial risk estimates under repeated stochastic sampling. Each time a Monte Carlo simulation is run, the authentication results are selected at random from the observed error distributions, and the expected financial risk is determined as follows:

$$R^{(k)} = C_{(FA)} \cdot FAR^{(k)} + C_{(FR)} \cdot FRR^{(k)} \quad (17)$$

The letters CFA and CFR represent the presumed costs associated with accepting something that is not supposed to be accepted and rejecting something that should have been accepted. The repeated application of the Monte Carlo procedure generates a distribution of relative expected risk. Therefore, using bootstrap confidence estimation, paired-hypothesis testing and Monte Carlo simulation generate complementary information regarding internal uncertainty, however none provide valid information about operational reliability or cost savings.

3. Results

3.1. Identity Verification Characteristics

Similarity-distributions for true (genuine) and false (impersonator) comparisons are good indicators of how well each of the methods represents the "identity" to be verified. True comparison-pairs show a greater average similarity value to one another. However, there is such a large amount of variability within these distributions that all decisions based on similarity alone will create a trade-off between FAR (false acceptance rate) and FRR (false rejection rate).

The simulated mask condition reduces the average similarity values of true-comparison pairs and increases the amount of variability among the two distributions. Therefore, the mask conditions produce a distributional shift and not just a small perturbation. Consequently, this results in reduced utility of similarity-based identification. An equal error rate (EER) operating point in Fig. 1(c), serves as a reference-threshold for the experimental analysis that follows; it does not represent an actual banking threshold.

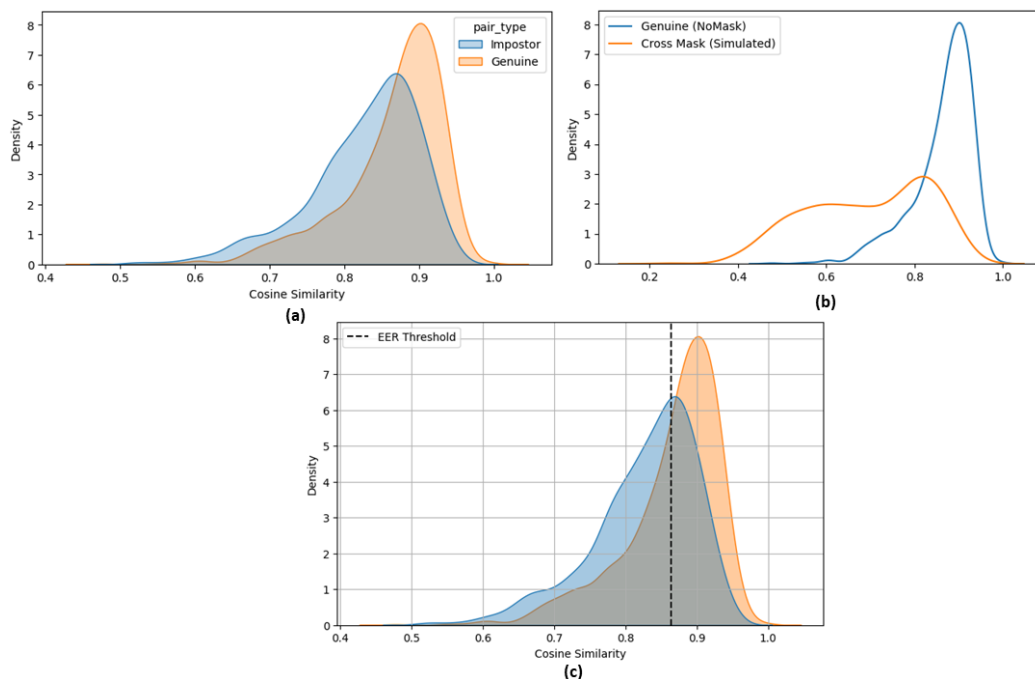


Fig. 1. Identity characteristics on (a) similarity score distribution, (b) effect of mask, and (c) score distribution with EER.

3.2. Verification Performance

If the threshold is increased to 0.85 then FAR decreased to 0.29 and FRR rose to 0.48. At a threshold of 0.9 FAR fell further to 0.18 but FRR rose further again to 0.63. Therefore, the level of false acceptances has fallen at the expense of much higher rates of legitimate user denial.

Finally, at the highest threshold value (0.95) FAR fell further still to 0.09 whilst FRR rose to 0.78. The implications of these results are clear; with such high values for FRR and low values for FAR no single threshold is suitable for use alone as an access control mechanism even though there may be a lower risk associated with fraudulent access.

Table 2 shows how the system performed when verified at four different levels (thresholds) and it highlights the relationship between the security and usability. When a threshold of 0.8 was used, this meant that an incorrect number of Impostors were being accepted (FAR=0.42), and that the correct number of Legitimate Users were being denied (FRR=0.31). It can be seen from these values that although there would have been some protection against fraud, it is likely that the majority of users will find this method of authentication unacceptable because it requires them to enter their password many times.

Table 2. Threshold Sensitivity.

Threshold	FAR	FRR
0.80	0.42	0.31
0.85	0.29	0.48
0.90	0.18	0.63
0.95	0.09	0.78

In general, the results reflect the inherent trade-off between security and usability associated with similarity-based verification systems. Thresholds set at low levels promote acceptance of legitimate users but make the system vulnerable to fraudulent access. In contrast, setting the thresholds at high levels improves the security of the system but may cause a significant portion of legitimate users to experience rejection.

3.3. Presentation Attack Detection Performance

A passive presentation attack detection evaluation occurred as a second signal. The AUC-ROC for PAD was about .55. As shown in Figure 2, the Bona Fide and Spoof Score Distributions show significant overlap. Therefore, while the PAD Component tested here provides limited independent discrimination, it does not provide sufficient evidence to support that a banking system would be resistant to spoofing on its own. Thus, when included in fusion, this component's presence should be viewed strictly as evidence of complementary evidence based upon the current standard, rather than an indication of a robust live test.

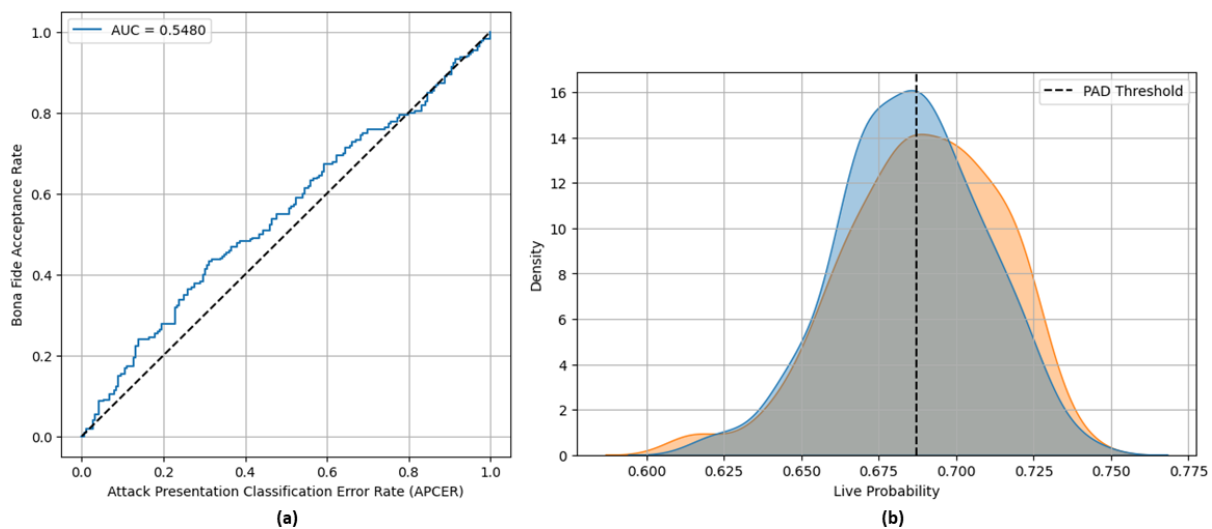


Fig. 2. Detection performance of (a) PAD ROC Curve and (b) PAD Liveness Score Distribution.

3.4. Joint Authentication Behavior

The joint similarity-liveness plots were developed as a method to analyze the interaction between the two signals. High-similarity (high confidence) and high-liveness areas on these plots represent authentication events that are most likely to be accepted. Low-confidence (low-similarity) regions represent areas where more authentication events will be

rejected. Unlike the single threshold which ignored one type of evidence, the dual-threshold approach utilizes both types of evidence. However, as stated above, a dual-threshold still contains a geometrical inflexibility.

Table 3. Authentication Behavior.

Method	Decision Rule	Boundary Type	Security Bias	Usability
Verification Only	$\text{similarity} \geq T$	vertical	low security	high usability
AND Rule	$\text{similarity} \geq T \text{ AND } \text{liveness} \geq T$	rectangular	very high security	low usability
Weighted Fusion	$w1S + w2L \geq T$	linear	balanced	moderate
RAFA (Proposed)	risk-aware decision	adaptive surface	balanced	balanced

These plots provide a visual explanation of the decision boundary testing and show how liveness may compensate for some of the lower confidence values from the weak component AUC under various conditions.

As seen in Table 3, the table provides an overview of the decision rules utilized in this study along with the operational trade-offs associated with each. The verification-only baseline establishes a purely vertical decision boundary that favors usability over liveness. The AND rule baseline creates a rectangular decision acceptance area by requiring both signals to exceed thresholds. While this approach has the potential to decrease false accepts, it also causes extreme legitimate user rejects if either of the two signals fall below the thresholds established.

3.5. End-to-End Authentication Performance

Table 4 provides the individual conservative system level configurations with their FRR = 0.686 and TAR = 0.314; therefore it rejects 68.6% of all legitimate access requests (i.e., only accepts 31.4% of legitimate requests). This type of configuration will produce considerable amounts of user frustration and can never be employed in a manner to deny users irreversibly from the system.

Table 4. System-Level Authentication Performance.

Metric	Value
True Acceptance Rate (TAR)	0.314
False Acceptance Rate (FAR)	0.200
False Rejection Rate (FRR)	0.686
Attack Success Rate	0.194
Total Authentication Attempts	600

The system-level FAR of .2 represents approximately one false positive for every 5 attacks attempted against the system, and the attack success rate at this system level is .194 according to the given definitions. Both numbers are significantly greater than what would be acceptable to use as part of production grade banking security. Therefore they define the parameters of the controlled environment which support more robust biometric backends, improved calibration, external verification, and additional step-up mechanisms beyond biometrics.

A failed biometric enrollment/verification process should result in the triggering of appropriate recovery processes commensurate with the failure. A typical sequence could include quality feedback, recapture, and if recapture continues to fail then reenrollment. If after reenrollment or recapture processes continue to fail then step up authentication using an independent channel to verify identity for higher risk transactional activity. In addition to developing and implementing these recovery mechanisms you must also test them to determine the successful rate of completion, residual FAR (False Acceptance Rate), added time required for each stage, and assess for accessibility prior to implementation.

3.6. Robustness Analysis

To assess robustness in relation to controlled perturbations at the representation level, two experiments were conducted. The first used additive feature noise and showed that the Equal Error Rate (EER) for all three modalities was affected by an increase in the size of the perturbation. Similarly, simulated feature drift had the effect of degrading the accuracy of the system (as shown in Figure 3). While these findings demonstrate that the system has some susceptibility to perturbation in its representations, they do not simulate many forms of realistic disruption due to variability in light levels, camera resolution, motion blur, user behavior, variations in the hardware devices used for capture, or adaptive spoofing attacks. Therefore, any claims made about deploying such technology into operational environments would need to be substantiated with experimental evaluations both of how well each type of environment affects the system's performance, and how well it performs when evaluated across multiple types of device.

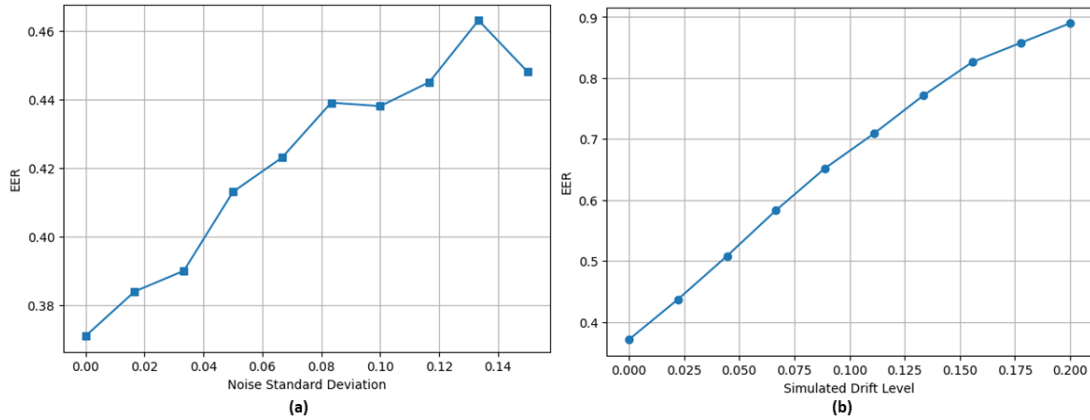


Fig. 3. Robustness analysis for (a) Feature Noise and (b) Feature Drift.

3.7. Fairness Analysis across Identities

The mean FRR at the identity level is .298, and the SD for this value is .344. In addition, the CoV for this value is 1.15. As can be seen from these statistics there is considerable variance among rejection rates for different identities; however, these statistics alone will provide no information about the duration of authentication attempts or demographic characteristics.

At the identity level, the highest FRR is 0.92 while the lowest is .01. Because aggregate FRR may mask individual users who experience rejection in nearly every single attempt at authentication, both identity-level tracking and non-exclusionary fallbacks should be considered during the design phase. However, the current experiment does not measure if the safeguards provided in this paper affect either the max FRR or CV associated with those designs.

Table 5. Identity-Level Authentication.

Metric	Value
Mean FRR	0.298
Std FRR	0.344
Coefficient of Variation	1.150
Max Identity FRR	0.920
Min Identity FRR	0.010

3.8. Fusion Strategy Comparison

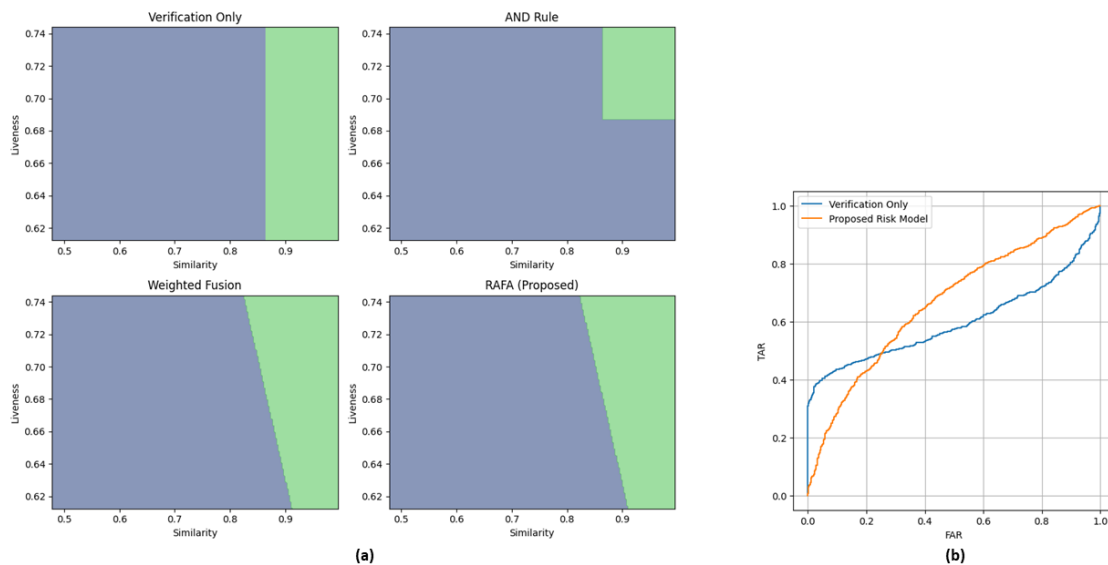


Fig. 4. Fusion strategy of (a) Boundary comparison and (b) ROC comparison.

For all four decision-making strategies (verification-only VO, AND rule-based fusion AF, Weighted Fusion WF, Risk Adaptive Fusion Algorithm RAFA) described in this paper, decisions regarding acceptance/rejection were based upon identical identity and liveness scores. The geometric arrangement of their decision boundaries illustrated in figure

4a provides insight as to why strict rules lead to very high FRR rates while relaxed rules lead to very high FAR rates. Figure 4b illustrates comparative ROC behavior amongst these four decision-making strategies within this controlled study.

3.9. Security–Usability Tradeoff

Table 6 compares false acceptance rates (FAR), False Rejection Rates (FRR), relative expected risk, and Fairness CV. This table clearly indicates how the And Rule has the lowest false acceptances (FAR = 0.09) and the highest false rejections (FRR = 0.78); these results show that there is a strong security bias to its operation. VO and WF each have a higher false acceptance rate than VO. Each of the three methods also has a relatively high false rejection rate. The three methods occupy intermediate positions on both FAR-FRR axes. The findings in this table provide decision-makers with a basis to choose among the same decision rules based on identical scores and splits; however, it provides no evidence that would support a claim of superiority by any of these methods to all other biometric systems.

Table 6. Security Evaluation.

Model	FAR	FRR	Expected Risk	Fairness CV
VO	0.31	0.41	374	0.59
AND	0.09	0.78	212	0.21
WF	0.34	0.37	380	0.57
RAFA	0.20	0.31	340	0.61

A cost-optimal operating point for the RAFA method was chosen so that the decision-maker could implement the method under specified conditions, given the stated costs and decision-making objectives. The choice of the optimal operating point did not consider minimizing the value of the fairness CV. As such, the fairness CV of 0.61 associated with the RAFA method is larger than those of VO (CV=0.59), AND (CV=0.21), and WF (CV=0.57). In summary, the inclusion of a fairness CV as part of the formulation provided no additional improvement in terms of fairness at the chosen operating point. The results indicate that fairness can be incorporated into the decision-analysis process as another tunable objective and should be reported along side the selected security and cost constraints.

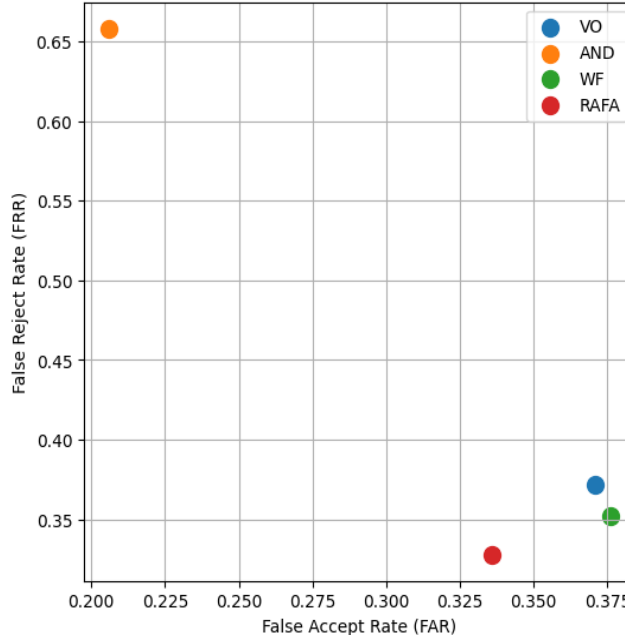


Fig. 5. FAR-FRR Comparison.

Figure 5 displays the test operating points in FAR/FRR space. A trade-off exists where the And Rule sacrificed usability for lower FAR. RAFA occupied an intermediate position. The optimal operating point will depend upon institutional documentation of fraud costs, customer-friction costs, security limits, and fairness constraints.

In order to isolate the effects of the decision-rules from the remainder of modern commercial face-recognitions and Passive Authentication Devices (PADs), the authors chose to include a baseline comparison. However, this baseline comparison represents neither a standard nor a benchmark for existing face-recognitions or PADs. Modern systems utilize different backends, sensors, datasets, pre-processing algorithms, and attack protocols than were used during testing, and therefore will generate vastly different absolute FAR and FRR values. Thus claims regarding the

effectiveness of the four rules on absolute FAR and FRR values are limited to their performance on the same experimental data.

This work represents an incremental and application-oriented contribution to prior work: it includes fraud exposure, legitimate user rejection, relative expected risk, and identity level variability in one decision-analysis. Furthermore, it does not represent a step-change in biometric recognition accuracy.

3.10. Financial Risk Evaluation

Table 7 provides an estimate of the relative expected risk using the assumed costs for false acceptance and false rejection. These results are expressed in simulation units. Under this assumption VO produces FAR = 0.31, FRR = 0.41, and an estimated risk of 374.

Table 7. Financial Risk Evaluation.

Model	FAR	FRR	Expected Risk
VO	0.31	0.41	374
AND	0.09	0.78	212
WF	0.34	0.37	380
RAFA	0.20	0.31	340

AND uses the least amount of resources with respect to FAR = 0.09 and estimated risk = 212; however, it also had the largest FRR = 0.78. WF produced FAR = 0.34, FRR = 0.37, and estimated risk = 380. RAFA produced FAR = 0.20, FRR = 0.31, and estimated risk = 340. Therefore, while RAFA reduced both FAR as compared to VO and WF, and FRR as compared to AND, it did not produce the smallest estimated risk given the above mentioned cost estimates. The ordering may vary based on a change in institutional cost assumptions or hard constraints.

3.11. Multi-Objective Optimization

In addition to assessing the performance of the strategies listed above (Table 8), we also compare their structural capabilities. Both VO and AND rely on fixed decision rules. WF allows for continuous score fusion via weighted aggregation. RAFA allows for continuous decision boundaries, fairness regularization, explicit cost sensitivity, and Pareto-based policy tuning. While these capabilities allow for greater degrees of configuration, none of them reflect deployment readiness.

Table 8. Multi-Objective Optimization.

Model	Cost-Sensitive Optimization	Fairness Regularization	Continuous Decision Boundary	Pareto Tunable
Verification-Only (VO)	No	No	No	No
AND Rule	No	No	No	No
Weighted Fusion (WF)	Partial	No	Partial	No
RAFA (Proposed)	Yes	Yes	Yes	Yes

A Pareto Frontier represents those configurations at which improvement in one objective results in loss in another. For operational usage of RAFA, defined security and fairness constraints will need to be specified, all infeasible candidates will need to be removed from consideration, the feasible candidate with the lowest loss will need to be selected, and validation using a test (hold-out) buffer will be required. The experiment did not find an optimal amount of regularization for a wide variety of applications. It is recommended that each time the data distributions, hardware, attacks, or institutional costs vary, the process should be re-run.

3.12. Statistical Uncertainty Analysis

Bootstrapping resampling was used to calculate the uncertainty associated with both FAR and Fairness CV. Additionally, 95% confidence ellipses were calculated to represent the joint variability due to bootstrapped resampling. Smaller ellipses represent greater stability within the evaluated samples. Bootstrapping resampling can reduce dependence on a particular sample realization, however it does not provide evidence regarding how well the method will perform when applied to different datasets, hardware platforms, environmental conditions, population types, or attack methodologies.

4. Discussion

4.1. Security Limitations of Similarity-Based Biometric Verification

The performance of the identity-verification module was assessed via ROC and DET curves. Given the low AUC

of approximately 0.548, there is minimal separation between real users and attackers using this system under the specified test conditions. The system lacks sufficient reliability to serve as an independent means of identity verification as an independent means of verifying identity. Therefore, we will evaluate better representations before any other operational use of this technology [22].

The DET curve also demonstrates that the reduction in False Acceptance Rate (FAR), at each level of increasing threshold, corresponds with an increase in False Rejection Rate (FRR). This behavior is typical of systems whose comparison score distribution has a substantial amount of overlap [22] and supports the notion that adjusting the threshold cannot fix both types of error for the current representation [23].

In addition to showing the threshold trade-off of FAR vs. FRR, the plot shown in Fig. 6(c) provides visual confirmation of how moving the decision threshold impacts the False Acceptance Rate (FAR) and the False Rejection Rate (FRR): moving the threshold up decreases FAR while increasing FRR; and moving the threshold down increases FAR while decreasing FRR. Thus, the choice of which threshold to select should be made on the basis of specific user-based security/usability considerations and not solely upon whether or not EER occurs [23].

These results suggest that some form of supplementary information and/or enhanced features will be required to provide meaningful authentication/evaluation of risk. However, because fusion is not able to automatically correct poor performing modules, the overall performance of any fused system is ultimately dependent on the quality of the individual discrimination and calibration of both PAD and Identity Scores [18,23].

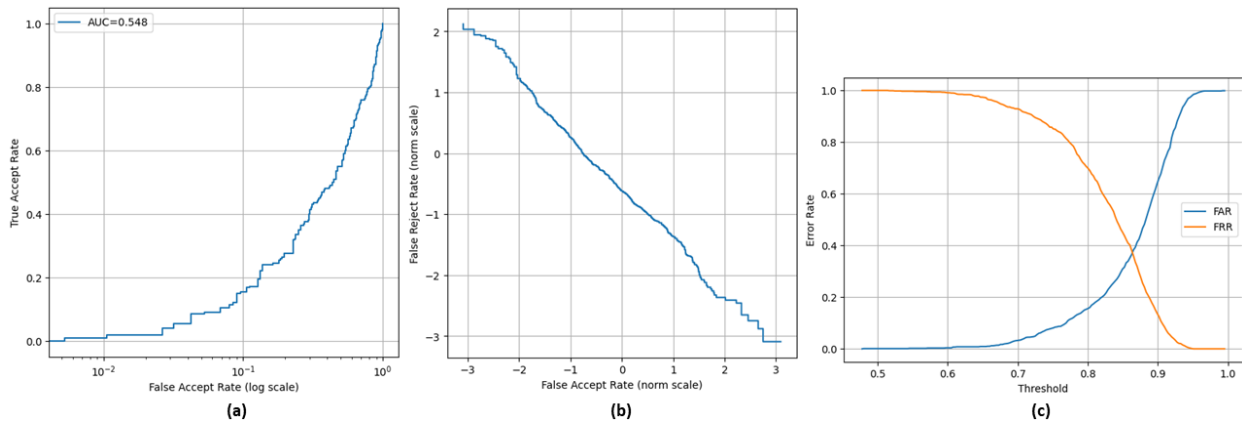


Fig. 6. Identity Verification Performance Analysis: (a) Verification ROC Curve, (b) DET Curve for Identity Verification, and (c) Threshold Sensitivity Analysis.

4.2. Complementary Role of Presentation Attack Detection

Presentation Attack Detection (PAD) determines if an input has been entered by an individual or through some form of manipulation, such as being a print-out, replay, etc. [24,25,26] In this research, we have found our PAD ROC-AUC to be approximately .55, which is only marginally better than chance. However, capture illumination, camera resolution, presentation quality, and type of attack can each significantly impact PAD performance. [25,26,27,28,29,30,31] Therefore, based solely on these findings, the PAD component utilized within this research is insufficient as a single security mechanism.

The PAD results provide insight into how the liveness score from this portion of the experiment will contribute to the overall fusion result. These results cannot serve as a basis for comparing the results of this particular PAD method to other more sophisticated methods as no contemporary models implemented using the same dataset, sensors, attacks, pre-processing techniques, and evaluation methodology. Therefore, a more thorough study of comparative PAD methods must employ current state-of-the-art models as baselines within comparable experimental frameworks.

4.3. End-to-End Authentication Behavior and Operational Impact

Ultimately, both PAD and Identity Verification function jointly to form an end-to-end authentication system. Fig. 7(a) provides a confusion matrix showing that while the system successfully rejected most of the attacks presented to it during testing, it also allowed a sizeable amount of attacks to pass through while also denying a significant number of legitimate attempts. Thus, this represents both a security issue and an accessibility issue and may not represent a viable production operating point.

As shown in Fig. 7(b), the aggregate measures further supports this finding; FAR = 0.200 remains unacceptable at a level required for a banking security application, and FRR = 0.686 would produce excessive login friction. Overall, the results clearly indicate the limitations associated with the biometric mechanisms employed in this research and clearly show that a more robust biometric backbone (in terms of identity recognition), along with additional PAD mechanisms and backup recovery mechanisms are necessary prior to any deployment [35, 37].

Enterprise banks should formally identify their assumptions regarding fraudulent losses and costs associated with customer friction, establish maximum allowable False Acceptance Rate (FAR) and False Rejection Rate (FRR), and formalize any constraints related to identity-based fairness. Failed biometric attempts should not automatically lead to denial of service; instead they should invoke a limited recovery process which includes providing users with quality feedback, allowing users to recapture/attempt again, permitting enrollment/re-enrollment as needed, and independently verifying user identity through additional authentication processes. As such, this paper recommends such a recovery strategy; however, we did not evaluate its operational feasibility.

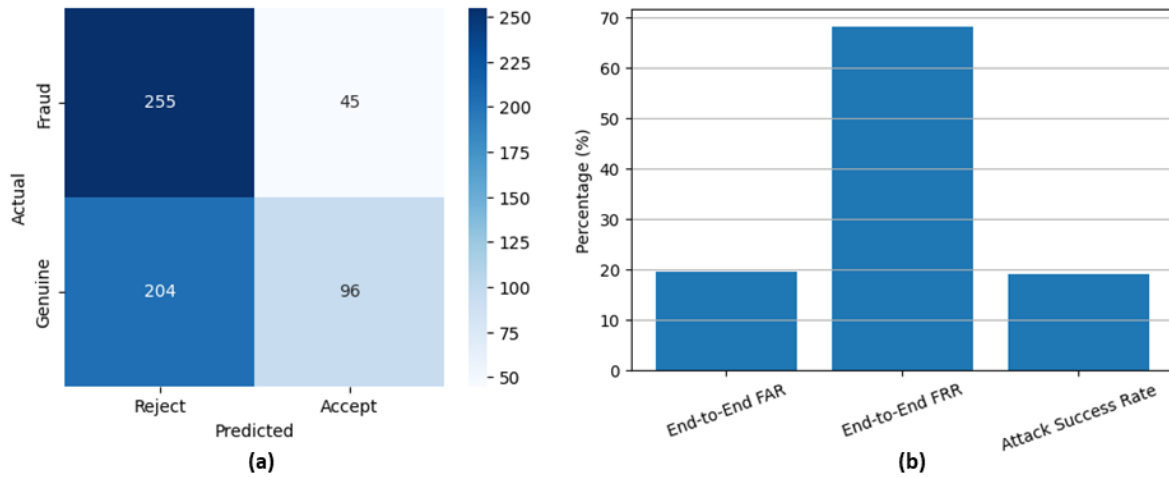


Fig. 7. Full Banking Authentication Evaluation: (a) End-to-End Authentication Confusion Matrix and (b) Overall Banking Authentication Performance.

4.4. Decision Boundary Geometry and Interpretability

The three tested methods form different areas of acceptability within similarity-liveness space (as described in Table 3 and Figure 8), which provides an intuitive understanding of how each decision rule balances security and usability.

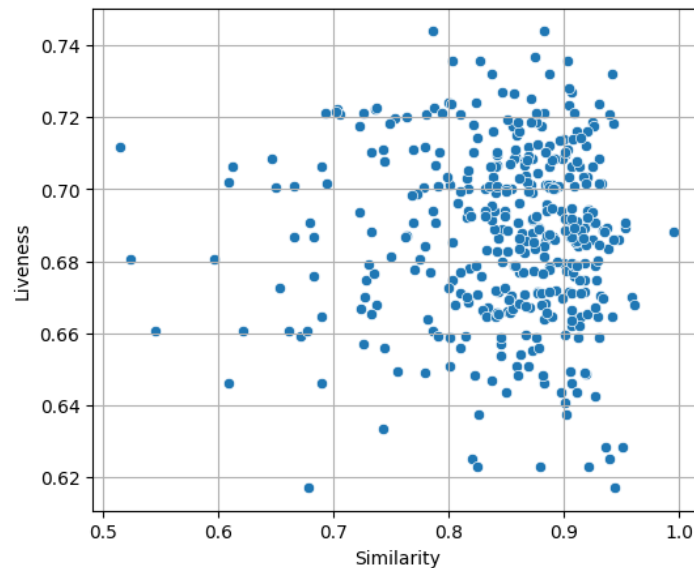


Fig. 8. Joint Distribution of Similarity and Liveness.

The verification-only method utilizes a single similarity threshold, ignoring all evidence of liveness. In contrast, the AND-rule accepts an input if it exceeds two pre-specified thresholds in terms of similarity and liveness. Although this approach reduces the size of the acceptable area relative to other approaches, it rejects valid users when either one of the scores has been decreased. The geometric boundaries of the AND-rule are simple to interpret yet rigid.

The weighted fusion strategy combines similarity and liveness information into a single score via a weighted linear combination. This results in a decision boundary that forms a linear decision boundary within the similarity-liveness space. This allows for partial trade-offs between these modalities. Thus, while this model offers greater flexibility than

the AND rule, the chosen weights remain arbitrary until some explicit optimization criterion can be used to determine them [18],[38].

A similar strategy is employed in RAFA, where the resulting decision surface is formed based upon stated risk and fairness criteria, rather than a predetermined rectangle or line. This allows for visualizations of individual decisions and their respective error-prone areas. However, it must be noted that this type of representation does not imply that the RAFA decision boundary is inherently superior to others; the decision boundary's value relies heavily on the quality and calibration of the scores, costs associated with misclassification, regularization, and the specific validation procedure.

4.5. Risk-Aware Authentication Formulation

EER and associated statistical standards are commonly employed as metrics for selecting appropriate biometric thresholds [39] – but these standards have no institutional consequence built in. Thus, RAFA uses an assumed expected loss function to evaluate its threshold selections.

$$R = C_{FA} \cdot FAR + C_{FR} \cdot FRR \quad (18)$$

where C_{FA} represents the cost associated with unauthorized access and C_{FR} represents the cost associated with rejecting legitimate users.

The ratio of the costs of False Acceptance (a) vs. False Rejection (b) will need to be defined by the institution. As shown in Fig. 9, changing this cost ratio can move the selected operating point through the iso-risk contours. The reported expected risk values represent simulated results based on stated conditions; they are neither empirically derived bank losses nor proof of a realized economic impact [40].

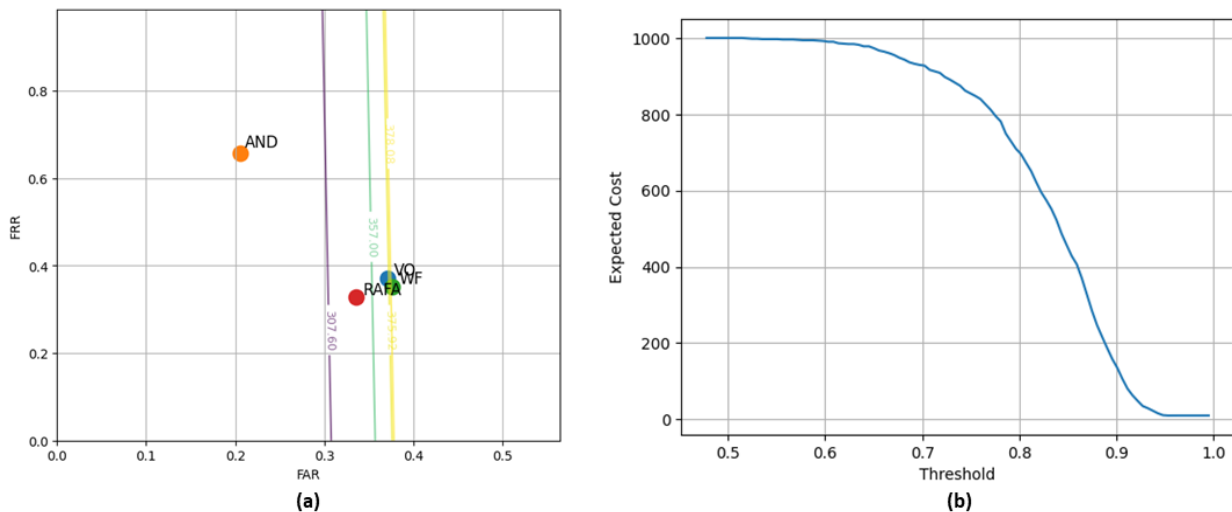


Fig. 9. Risk-Based Decision Analysis: (a) Iso-Risk Analysis and (b) Bayesian Risk Optimization.

4.6. Fairness and Identity-Level Variability

The aggregate error rate does not indicate how reliably the system authenticates each user. The per-identity False Rejection Rate (FRR) in Figure 10 clearly indicates significant variability. Many identities have FRRs which are less than one percent, while others experienced nearly a 100% FRR. Therefore, this measure assesses the dispersion of outcomes for each individual identity and should not be misinterpreted as an assessment of demographic fairness as protected attributes were not analyzed.

Variability of outcomes at the individual level may result from variations in capture quality, pose, illumination, occlusion, representation quality, and disparate data coverage. However, the current experiment did not control for these variables and thus should not be used to make conclusions about cause or demographics based solely upon the FRR distribution.

The Lorenz Curve presented in Figure 11 deviates from the equality line, indicating that False Reject Rates occur disproportionately among a subset of the sampled identities. The curve is a descriptive statistic measuring the concentration of outcomes and does not describe why such disparity exists or whether it occurs within a protected class [41].

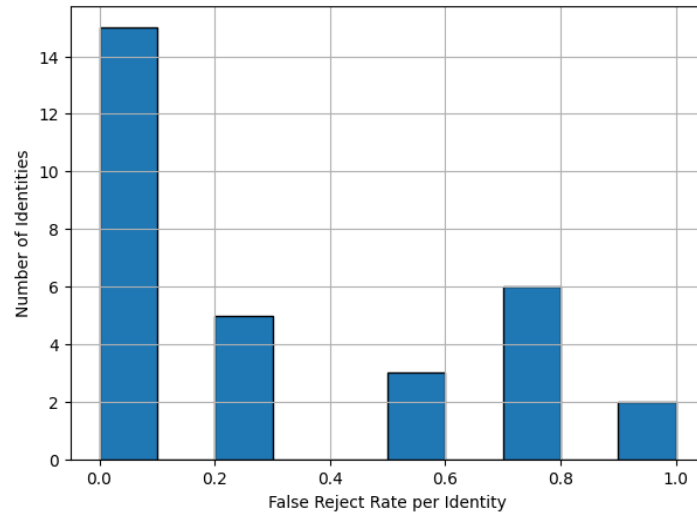


Fig. 10. Fairness Across Sampled Identities.

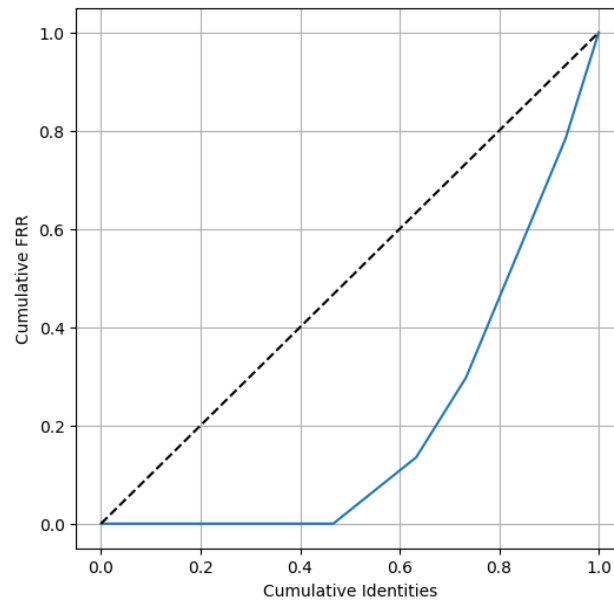


Fig. 11. Lorenz Curve of Identity FRR Inequality.

Persistent identity level rejection in a financial service may create disparate access to services, require multiple interactions with customer support staff and lead to abandonment. Therefore, monitoring will need to include mean FRR, maximum identity-level FRR, dispersion and performance of the fallback option.

Individuals experiencing persistent high levels of FRR should prompt diagnostic reviews and non-exclusionary fallback processes. Potential candidate interventions include Capture Quality Assessments, Controlled Recaptures, Re-enrollment Under Improved Conditions, Live Samples To Verify Additional Information Regarding Enrollment Channel, Bounded Threshold Adjustments While Maintaining Constant FAR Constraints. This study does not evaluate prospectively the effectiveness of these interventions on CV or maximum individual-level FRR.

Therefore, fairness evaluations should consider both aggregate FAR/FRR measures along with individual level measures and protected group analysis where possible. Current findings justify continued examination into monitoring and intervention strategies; however, they do not demonstrate that RAFA improves fairness at its current operational setting [40, 42]

4.7. Multi-Objective Optimization and Fairness Regularization

Fairness regularization represents a quantitative method of making an explicit compromise between relative expected risk and identity level variation. This is visualized in Fig. 12(a), where each combination exists at a unique location. The analysis is also consistent with prior work in the area of context aware and risk based authentication that views the process of choosing policies as a multi-criteria decision problem [43,44,45].

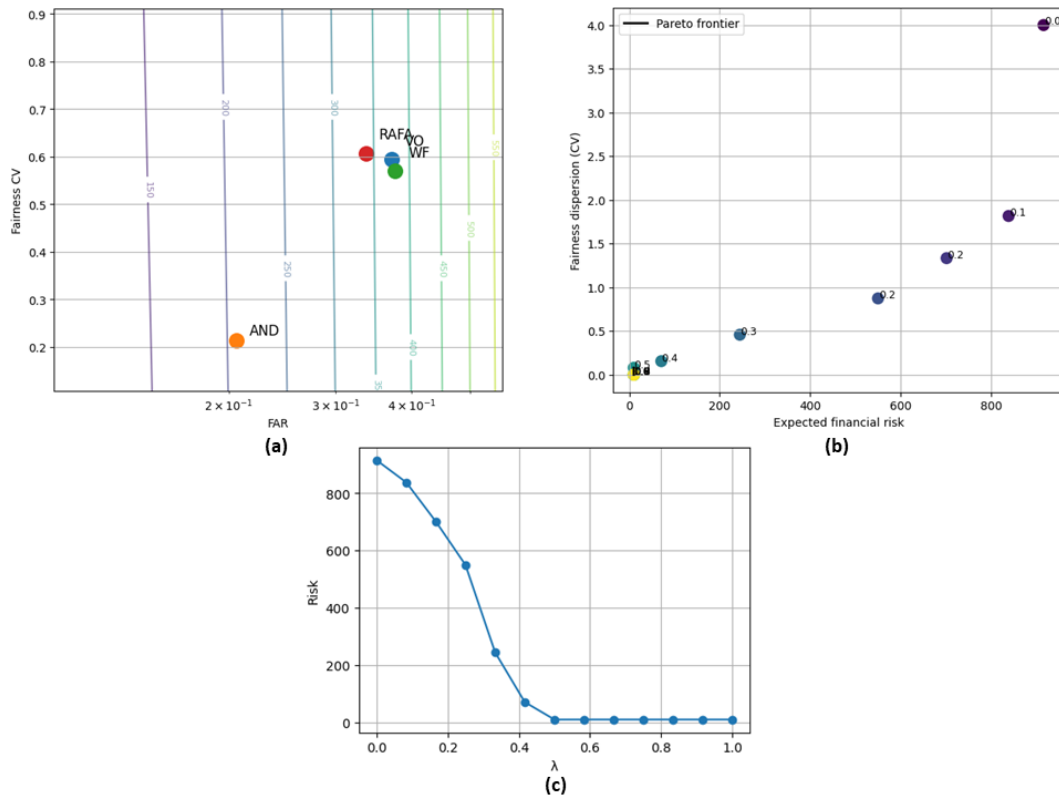


Fig. 12. Risk-Fairness Trade-off Analysis: (a) Risk-Fairness Surface, (b) Pareto Frontier, and (c) RAFA Risk Sensitivity.

Fig. 12(b) represents a Pareto front of all combinations that have no improvement possible in one criterion, without causing degradation in the other. Fig. 12(c) illustrates the effect of varying the strength of the fairness regularizer. In order to create a deployment methodology, you could employ a predetermined grid for the regularizer, apply strict constraints to both the FAR and FRR, constrain the CV as well, determine the minimal loss feasible point, test it against a hold-out buffer and maintain roll-back strategies. While this paper conceptualized such a methodology, it did not produce a production controller that will continue to optimize itself dynamically [43,44].

4.8. Statistical Robustness and Uncertainty Analysis

Finite benchmarks are based on a finite number of data samples which are subject to sampling variance. Bootstrap confidence ellipses can express how FAR and Fairness-CV vary together when bootstrapping over different samples. A small ellipse indicates a precise estimate of the range of FAR and Fairness-CV using just the benchmark samples. This does not imply that the estimates will generalize with all other populations, devices, or spoofing methods.

Mobile banking applications need to have some level of tolerance for variability in motion, camera conditions, illumination, and user behavior, as well as evolving attacks [43,45]. Feature noise and feature drift experiments were only partially useful for testing and validating our proposed method, and should not be used as substitutes for actual device-to-device, time-based, and real-time mobile banking trials.

Monte Carlo simulation allows for quantifying the amount of variation in calculating the assumed-risk, ; however, we did not perform cross-dataset validations, cross-device experiments, or real-world trials on a mobile banking application. Therefore, confidence intervals and ellipses must not be taken as proof of real-world generalization.

This manuscript currently does not include all the information needed for independent replication of this study (such as: what the dataset was called; who gave permission for use; how many images were produced by each subject; how many subjects were included in each image; how the images were divided into training and test sets; how many times the Monte Carlo or Bootstraps were repeated; How big the parameters were; What version of software was used; Which Random Seeds were selected).

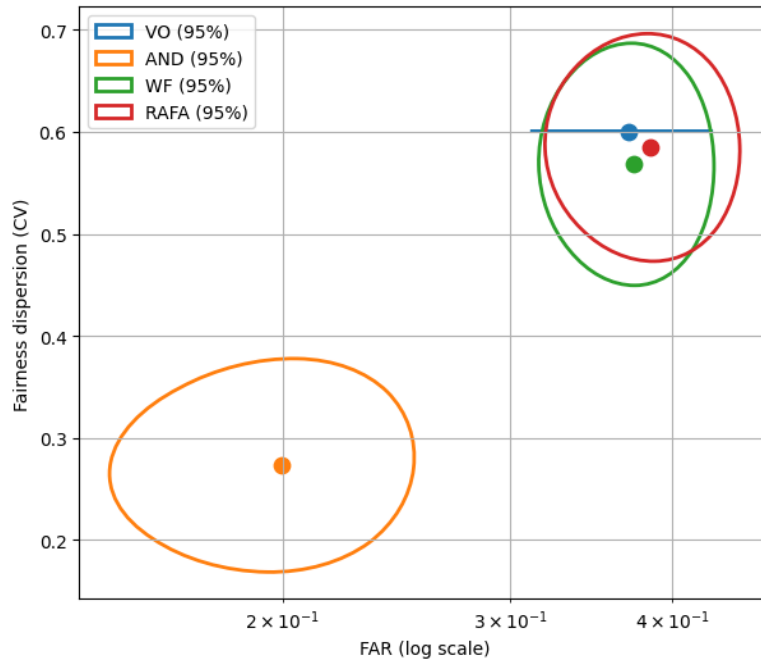


Fig. 13. Confidence Ellipse Comparison.

4.9. Implications for Secure Banking Authentication Systems

A number of key design considerations can be derived from the controlled results of this study. Design implication 1: weak identity or PAD scores cannot be used as independent proof of authenticity. Instead, such scores should be viewed as additional (complementary) evidence. Design implication 2: threshold selection should combine statistical performance with explicit security constraints, usability constraints and cost constraints. Design implication 3: individual-level rejection must be monitored due to aggregate average metrics that can conceal severe individual failure. Design implication 4: biometric failure should trigger proportionate recovery paths rather than automatic permanent denial. This is a list of design implications, not validated deployment-specific outcomes [18,43,44,45,46,47].

4.10. Limitations and Deployment Scope

There are five deployment limitations to this study. Limitation 1: The study is based on public benchmark data and simulated cost assumptions rather than real banking transactions, and lacks cross-dataset or cross-device test for independence. Limitation 2: the identity and PAD components have AUC values near .50 which constrains the entire system. Limitation 3: modern face-verification and PAD baselines were not re-implemented in the same protocol as this study; therefore, no state-of-the-art superiority claim can be made. Limitation 4: robustness testing was limited to simulated feature noise and drift only and did not include any realistic lighting, camera quality, motion, behavioral changes, or adaptive attacks. Limitation 5: the proposed fallback and adaptive-threshold safeguards were not prospectively evaluated; therefore, latency, throughput, scalability, longitudinal calibration, and mitigation effectiveness remain unknown.

Future work will need to provide complete dataset and split metadata, evaluate stronger backbones for face verification and passive liveness detection (PAD), conduct matched comparisons to state-of-the-art systems, conduct cross-dataset and cross-device validation tests, test realistic and dynamic attacks, measure runtime and scalability on target devices, and monitor identity level and demographic fairness over time.

5. Conclusion

RAFA-BioAuth combines features of both identity similarity, passive liveness evidence, asymmetric relative-risk estimation and identity-level fairness regularization into an interpretable decision framework. At the controlled benchmark protocol, both component classifiers are very poorly performing, thus all four tested decision rules expose different trade offs between security and usability. While rafa occupies an intermediate position in terms of false acceptance rate (far)-false rejection rate (FRR), neither expected risk nor fairness CV are minimized at the selected operating point. The identity level analysis further demonstrates that aggregate metrics can obscure severe rejection rates for certain users. Due to its use of benchmark data, simulated costs, incomplete reproducibility metadata and lack of external and cross-device validation tests, this study does not demonstrate production-readiness. The contribution of this study is a transparent framework for future controlled evaluations of systems based on this methodology and/or these algorithms; not a production ready banking authentication system.

All the Declarations and Statements

Author Contributions Statement

Dendy K. Pramudito – Conceptualization, Methodology, Supervision, and Writing - Review and Editing: Proposed the RAFA-BioAuth research idea, designed the overall risk-adaptive and fairness-aware authentication framework, guided the study direction, and reviewed the manuscript critically.

Jufriadif Na'am – Data Curation, Software, Validation, and Visualization: Managed the face verification and presentation attack detection datasets, implemented the experimental pipeline and fusion models, conducted validation and calibration experiments, and prepared the figures and tables.

Ferda Ernawan – Formal Analysis, Investigation, Statistical Analysis, and Writing - Original Draft: Performed the end-to-end performance, fairness, and financial risk analyses, carried out bootstrap and Monte Carlo-based statistical evaluation, reviewed the related literature, and drafted the manuscript.

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

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement

The processed data, experimental configuration, and additional results supporting the findings of this study are available from the corresponding author upon reasonable request.

Ethical Declarations

Ethics approval: none, because this study used publicly available benchmark datasets and did not involve direct human-subject data collection by the authors. Consent to participate: none, for the same reason. Consent for publication: all authors approved the manuscript for submission and publication. Data availability: the processed data, experimental configuration, and additional results are available from the corresponding author upon reasonable request, subject to the licenses of the original public datasets. Materials availability: none beyond the public datasets and reported experimental configuration. Code availability: the implementation can be made available from the corresponding author upon reasonable request. Competing interests: none. Funding: none.

Acknowledgments

The authors would like to acknowledge the providers and maintainers of the publicly available face recognition and presentation attack detection datasets used in this study. The authors also appreciate the valuable academic and technical discussions that helped improve the quality of this research.

Declaration of Generative AI in Scholarly Writing

The authors used AI-assisted tools only for language editing, formatting support, and consistency checking during revision. All scientific content, experimental interpretation, and final responsibility for the manuscript remain with the authors.

Abbreviations

The following abbreviations are used in this manuscript:

RAFA - BioAuth - Risk-Adaptive Fairness-Aware Biometric Authentication
PAD - Presentation Attack Detection
AUC - Area Under the Curve
ROC - Receiver Operating Characteristic
DET - Detection Error Tradeoff
TAR - True Acceptance Rate
FAR - False Acceptance Rate
FRR - False Rejection Rate
EER - Equal Error Rate
CV - Coefficient of Variation
VO - Verification-Only

WF - Weighted Fusion
CFA - Cost of False Accept
CFR - Cost of False Reject

References

- [1] E. Birgisdóttir, "Exploring the security of mobile face recognition: Attacks, defenses, and future directions," *Appl. Sci.*, vol. 15, no. 24, p. 13232, 2025, doi: 10.3390/app152413232.
- [2] M. Taşkıran, N. Kahraman, and Ç. Erdem, "Face recognition: Past, present and future (a review)," *Digit. Signal Process.* 106, vol. 102809, 2020, doi: 10.1016/j.dsp.2020.102809.
- [3] N. Waykole, "Secure facial recognition systems: A machine learning review of spoofing detection via parameter quality metrics," *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 31s, pp. 493–501, 2025, doi: 10.52783/jisem.v10i31s.5103.
- [4] A. Adedapo, O. Odejobi, and T. Taiwo, "Countermeasures against bias and spoofing in modern facial recognition systems," *World J. Adv. Res. Rev.*, vol. 25, no. 1, pp. 1914–1930, 2025, doi: 10.30574/wjarr.2025.25.1.0011.
- [5] A. Aziz, "Effects of visible and near infrared polarized lights on spoofing face detection," *J. Comput. Sci.*, vol. 15, no. 2, pp. 288–301, 2019, doi: 10.3844/jcssp.2019.288.301.
- [6] B. Chettri, T. Kinnunen, and E. Benetos, "Deep generative variational autoencoding for replay spoof detection in automatic speaker verification," *Comput. Speech Lang.* 63, vol. 101092, 2020, doi: 10.1016/j.csl.2020.101092.
- [7] J. Solanki, "Robust framework for antispoofing face recognition mechanism for mitigating image attacks," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 6, no. 3, pp. 1436–1439, 2018, doi: 10.22214/ijraset.2018.3222.
- [8] Z. Wang, S. Wang, W. Yu, B. Gao, C. Li, and T. Wang, "Accurate real-time live face detection using snapshot spectral imaging method," *Sensors*, vol. 25, no. 3, p. 952, 2025, doi: 10.3390/s25030952.
- [9] A. Guleria, K. Krishan, V. Sharma, and T. Kanchan, "Impact of prolonged wearing of face masks—Medical and forensic implications," *J. Infect. Dev. Ctries.*, vol. 16, no. 10, pp. 1578–1587, 2022, doi: 10.3855/jidc.16618.
- [10] M. Gündoğar and Ç. Erdem, "Presentation attack detection for face recognition using remote photoplethysmography and cascaded fusion," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 29, no. 7, pp. 3240–3258, 2021, doi: 10.3906/elk-2010-93.
- [11] N. Kumar, S. Sharma, P. Sharma, P. Mishra, B. Sharma, and A. Sharma, "Digital banking under siege: Trends and challenges in emerging security threats," *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 3, 2025, doi: 10.52783/jisem.v10i3.4760.
- [12] D. Muhtasim, M. Pavel, and S. Tan, "A patch-based CNN built on the VGG-16 architecture for real-time facial liveness detection," *Sustainability*, vol. 14, no. 16, p. 10024, 2022, doi: 10.3390/su141610024.
- [13] M. Parvadhi, S. Kishore, N. S. L. K., and P. S., "Enhanced face spoofing detection through geometric temporal dynamic analysis," *Int. J. Tech. Res. Sci.*, vol. 9, no. Spl, pp. 27–35, 2024, doi: 10.30780/specialissue-iset-2024/012.
- [14] P. Prasad et al., "Robust facial biometric authentication system using pupillary light reflex for liveness detection of facial images," *Comput. Model. Eng. Sci.*, vol. 139, no. 1, pp. 725–739, 2024, doi: 10.32604/cmescs.2023.030640.
- [15] S. Santhoshini, "Deep Guard: Face spoofing detection using Swin transformer and rPPG signal," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 13, no. 10, pp. 923–928, 2025, doi: 10.22214/ijraset.2025.74700.
- [16] K. Jha, A. Jain, and S. Srivastava, "Feature-level fusion of face and speech based multimodal biometric attendance system with liveness detection," *AIP Adv.*, vol. 14, no. 11, 2024, doi: 10.1063/5.0234430.
- [17] S. Kumar, Y. Asish, and S. Ganapathy, "A new hybrid CNN–LSTM model with non-softmax functions for face spoof detection," *Soft Comput.*, vol. 26, no. 19, pp. 10151–10162, 2022, doi: 10.1007/s00500-022-07418-9.
- [18] M. Singh, R. Singh, and A. Ross, "A comprehensive overview of biometric fusion," *Inf. Fusion*, vol. 52, pp. 187–205, 2019, doi: 10.1016/j.inffus.2018.12.003.
- [19] R. Zhang and Z. Yan, "A survey on biometric authentication: Toward secure and privacy-preserving identification," *IEEE Access*, vol. 7, pp. 5994–6009, 2019, doi: 10.1109/ACCESS.2018.2889996.
- [20] A. K. Hanumanthaiah and M. B. Eraiah, "Challenge responsive multi modal biometric authentication resilient to presentation attacks," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 2, pp. 494–507, 2022, doi: 10.22266/ijies2022.0430.44.
- [21] K. Okerefor, O. Osuagwu, and C. Onime, "Biometric anti-spoofing technique using randomized 3D multi-modal traits," *Int. J. Simul. Syst. Sci. Technol.*, vol. 19, no. 5, pp. 5.1–5.8, 2018, doi: 10.5013/ijssst.a.19.05.05.
- [22] H. Heidari and A. Chalechale, "A new biometric identity recognition system based on a combination of superior features in finger knuckle print images," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 28, no. 1, pp. 238–252, 2020, doi: 10.3906/elk-1906-12.
- [23] D. R. Kale and S. Kulkarni, "Integrating fusion levels for biometric authentication system," *IOSR J. Electron. Commun. Eng.*, vol. 12, no. 1, pp. 65–72, 2017, doi: 10.9790/2834-1201016572.
- [24] J. Devi, S. Parveen, N. Naem, and N. Abbas, "Facial verification along with spoof attacks," *Int. J. Adv. Res.*, vol. 5, no. 5, pp. 2264–2268, 2017, doi: 10.21474/IJAR01/5291.
- [25] M. N. Favorskaya, "Face presentation attack detection: Research opportunities and perspectives," *Intell. Decis. Technol.*, vol. 17, no. 1, pp. 159–193, 2023, doi: 10.3233/IDT-220197.
- [26] Z. Ming, M. Visani, M. Luqman, and J.-C. Burie, "A survey on anti-spoofing methods for facial recognition with RGB cameras of generic consumer devices," *J. Imaging*, vol. 6, no. 12, p. 139, 2020, doi: 10.3390/jimaging6120139.
- [27] J. Bok, K. Suh, and E. Lee, "Verifying the effectiveness of new face spoofing DB with capture angle and distance," *Electronics*, vol. 9, no. 4, p. 661, 2020, doi: 10.3390/electronics9040661.
- [28] A. Hassani, J. Diedrich, and H. Malik, "Improving monocular facial presentation-attack-detection robustness with synthetic noise augmentations," *Sensors*, vol. 23, no. 21, p. 8914, 2023, doi: 10.3390/s23218914.
- [29] D. Karmakar, P. Mukherjee, and M. Datta, "Spoofed facial presentation attack detection by multivariate gradient descriptor in micro-expression region," *Pattern Recognit. Image Anal.*, vol. 31, no. 2, pp. 285–294, 2021, doi: 10.1134/S1054661821020097.

- [30] H. Vinutha and G. Thippeswamy, "Antispoofing in face biometrics: A comprehensive study on software-based techniques," *Comput. Sci. Inf. Technol.*, vol. 4, no. 1, pp. 1–13, 2023, doi: 10.11591/csit.v4i1.pp1-13.
- [31] Y. Tian, Y. Huang, K. Zhang, Y. Liu, Z. Sun, and F. Chen, "Polarized image translation from nonpolarized cameras for multimodal face anti-spoofing," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 5651–5664, 2023, doi: 10.1109/TIFS.2023.3310348.
- [32] K. Kavita, G. Walia, and R. Rohilla, "A contemporary survey of multimodal presentation attack detection techniques: Challenges and opportunities," *SN Comput. Sci.*, vol. 2, no. 1, 2021, doi: 10.1007/s42979-020-00425-3.
- [33] H. Mandalapu *et al.*, "Audio-visual biometric recognition and presentation attack detection: A comprehensive survey," *IEEE Access*, vol. 9, pp. 37431–37455, 2021, doi: 10.1109/ACCESS.2021.3063031.
- [34] G. Ali, M. Dida, and A. Sam, "Two-factor authentication scheme for mobile money: A review of threat models and countermeasures," *Futur. Internet*, vol. 12, no. 10, p. 160, 2020, doi: 10.3390/fi12100160.
- [35] S. Tatineni, "Customer authentication in mobile banking-MLOps practices and AI-driven biometric authentication systems," *J. Econ. Manag. Res.*, pp. 1–5, 2022, doi: 10.47363/jesmr/2022(3)201.
- [36] M. Waliullah, M. George, M. Hasan, M. Alam, M. Munira, and N. Siddiqui, "Assessing the influence of cybersecurity threats and risks on the adoption and growth of digital banking: A systematic literature review," *ajates*, vol. 1, no. 01, pp. 226–257, 2025, doi: 10.63125/fh49gz18.
- [37] S. Alotaibi, A. Alruban, M. Alotaibi, A. Alshumrani, and A. Altamimi, "Smartphone users transparent verification approach for mobile applications," *Int. J. Intell. Comput. Res.*, vol. 10, no. 2, pp. 977–987, 2019, doi: 10.20533/ijicr.2042.4655.2019.0119.
- [38] N. Ammour, Y. Bazi, and N. Alajlan, "Multimodal approach for enhancing biometric authentication," *J. Imaging*, vol. 9, no. 9, p. 168, 2023, doi: 10.3390/jimaging9090168.
- [39] C. X. Tan *et al.*, "A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction," *Multimed. Tools Appl.*, vol. 80, no. 21–23, pp. 32725–32762, 2021, doi: 10.1007/s11042-021-11235-x.
- [40] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informatics*, vol. 17, no. 1, pp. 168–192, 2020, doi: 10.1016/j.aci.2018.08.003.
- [41] D. Koutsoyiannis, "Trade-off between entropy and Gini index in income distribution," *Entropy*, vol. 28, no. 1, p. 35, 2025, doi: 10.3390/e28010035.
- [42] V. Conti, L. Rundo, C. Militello, V. Salerno, S. Vitabile, and S. M. Siniscalchi, "A multimodal retina-iris biometric system using the Levenshtein distance for spatial feature comparison," *IET Biometrics*, vol. 10, no. 1, pp. 44–64, 2020, doi: 10.1049/bme2.12001.
- [43] S. Ayeswarya and K. J. Singh, "Enhancing security and usability with context aware multi-biometric fusion for continuous user authentication," *Sci. Rep.*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-14833-z.
- [44] J. Chen, U. Hengartner, and H. Khan, "MRAAC: A multi-stage risk-aware adaptive authentication and access control framework for Android," *ACM Trans. Priv. Secur.*, vol. 27, no. 2, pp. 1–30, 2024, doi: 10.1145/3648372.
- [45] D. Hintze *et al.*, "CORMORANT: Ubiquitous risk-aware multi-modal biometric authentication across mobile devices," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–23, 2019, doi: 10.1145/3351243.
- [46] A. Lonkar, S. Dharmadhikari, N. Dharurkar, K. Patil, and R. A. Phadke, "Tackling digital payment frauds: A study of consumer preparedness in India," *J. Financ. Crime*, vol. 32, no. 2, pp. 257–278, 2024, doi: 10.1108/JFC-01-2024-0029.
- [47] I. Riasat, M. Shah, and M. S. Gonul, "Strengthening cybersecurity resilience: An investigation of customers' adoption of emerging security tools in mobile banking apps," *Computers*, vol. 14, no. 4, p. 129, 2025, doi: 10.3390/computers14040129.

Authors' Profiles



Dendy K. Pramudito is a professional working in Information Technology. He is also a permanent lecturer at Pelita Bangsa University, Cikarang, West Java, and has served as a lecturer at several other universities in the Greater Jakarta area. He earned a Dr. (Doctorate) degree in Business Management in 2021 at Bina Nusantara University, Jakarta. Currently he is pursuing another Dr. (Doctorate) degree focusing on Informatics at Nusa Mandiri University, Jakarta. The author is also an active researcher and implementer as well.



Jufriadif Na'am is a professor and lecturer at Nusa Mandiri University specializing in Computer Vision and Image Processing. He earned his Professorship in 2019 and his Dr. (doctorate) degree on Information Technology in 2017. Previously, he was a lecturer at Putra Indonesia University (YPTK) Padang. His research focuses on computer vision, image processing and artificial neural networks. He also has many publications, actively writes in scientific journals and books, including on information systems and blockchain.



Ferda Ernawan is an academic and researcher in the field of computer science who currently serves as an Associate Professor at the Faculty of Computer Science at Al-Sultan Abdullah University, Pahang, Malaysia. He earned his Doctorate (Ph.D.) in Computer Technology from Universiti Teknikal Malaysia Melaka (UTeM) in 2014. He is widely known for his contributions in digital security research, image processing and expert systems. Currently, in addition to UMPsA, he is also listed as having a profile in SINTA (Science and Technology Index) Indonesia which is affiliated with Nusa Mandiri University. He also actively collaborates with other institutions such as FMIPA UNNES in research programs and guest lecturers.

How to cite this paper

Dendy K. Pramudito, Jufriadif Na'am, Ferda Ernawan, "RAFA-BioAuth: Risk-Adaptive and Fairness-Aware Biometric Authentication for Secure Mobile Banking", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.18, No.4, pp. 50-71, 2026. DOI:10.5815/ijieeb.2026.04.04