Journal of Science and Technology – Data Science and Artificial Intelligence
Volume 36, Issue 1, October 2026, 040–048

A Complete On-Device RGB–IR Face Presentation Attack Detection System with Independent ISO/IEC 30107-3 PAD Level 1 and Level 2 Evaluation by BixeLab

Huynh Duong Huu[1], Hung Pham Ngoc[1], Vinh La The[1], Thuan Nguyen Dinh[1], Hiep Hoang Van[1,*]

[1]*Hanoi University of Science and Technology, Ha Noi, Vietnam*
[*]*Corresponding author email: hiep.hoangvan@hust.edu.vn*

Abstract

Face presentation attack detection (PAD) systems intended for real-world access control must provide not only accurate spoof detection but also reliable on-device operation, multi-sensor integration, traceable decision making, and reproducible evaluation. This paper presents a complete on-device RGB–IR face PAD system implemented on a custom Android access-control device and evaluated independently under ISO/IEC 30107-3-based test protocols. The system employs temporally paired RGB and infrared observations and integrates three complementary sources of evidence: a primary RGB liveness score, an RGB–IR cross-spectral consistency score, and a secondary independent liveness score. Frame-level outputs are combined through a configurable transaction-level decision policy based on multi-frame aggregation, while intermediate scores, decisions, and transaction results are logged to support traceability and post-hoc analysis. The complete hardware–software system was independently evaluated by BixeLab, an ISO/IEC 17025-accredited biometric testing laboratory, at Level A and Level B. For each level, the evaluation involved 10 consenting subjects, 50 bona fide transactions, and 300 presentation attacks covering six presentation attack instrument species. No attack classification errors or attack non-responses were observed in either evaluation, corresponding to an observed APCER of 0/300 and APNRR of 0/300 at both levels. However, 10 of 50 bona fide presentations were incorrectly rejected at each level, corresponding to an observed BPCER of 20%. These results demonstrate the effectiveness of the proposed system-level integration against the evaluated attack species while also highlighting the security–usability trade-off introduced by conservative operating points. The study provides an independently validated example of a deployable and auditable RGB–IR PAD architecture for real-world biometric access-control systems.

Keywords: Android access control, face presentation attack detection, independent evaluation, ISO/IEC 30107-3, RGB–IR sensing.

1. Introduction

Face recognition has become increasingly common in biometric access control, mobile authentication, and unattended identity-verification systems because of its contactless operation and ease of use. However, face-based authentication remains vulnerable to presentation attacks in which an adversary presents manipulated or artificial biometric evidence, such as printed photographs, display-based replays, masks, or other fabricated artifacts, to the capture device. Face presentation attack detection (PAD), therefore, plays a critical role in preventing such attacks from being accepted as bona fide presentations. Although recent learning-based PAD methods have achieved strong performance on public benchmarks, deploying PAD in practical access-control systems introduces additional challenges beyond classification accuracy alone. A real-world system must operate reliably under variations in illumination, user pose, acquisition quality, and interaction behavior while satisfying constraints on processing latency, computational resources, and user experience. Consequently, effective PAD for operational biometric systems requires not only accurate liveness estimation but also a robust sensing and decision pipeline capable of producing stable and traceable authentication outcomes.

In deployed systems, PAD performance is determined not only by the accuracy of an individual classifier but also by the reliability of the complete acquisition and decision pipeline. Single-stream RGB solutions can be sensitive to illumination changes and visual patterns that may be reproduced by high-quality printed or display-based attacks, motivating the use of complementary sensing modalities such as infrared imaging. Multi-spectral sensing can provide additional evidence that is more difficult to reproduce consistently across spectral bands; however, integrating RGB and IR sensing into an operational system introduces its own challenges. Reliable deployment requires coordinated dual-camera acquisition, temporal pairing of observations, face-quality and geometric validation, cross-spectral consistency checking, transaction-level aggregation, and explicit handling of missing or unreliable sensor evidence. These components must operate together under real-time and resource constraints without introducing

p-ISSN 3093-3285
e-ISSN 3093-3315
https://doi.org/10.51316/jst.194.dsai.2026.36.1.5
Received: Aug 25, 2026; Revised: Oct 10, 2026;
Accepted: Oct 10, 2026; Online: Oct 15, 2026

unstable latency or inconsistent decisions. Therefore, the practical effectiveness of an RGB–IR PAD solution depends on hardware–software co-design and on the behavior of the complete transaction pipeline, rather than on the performance of any single PAD model in isolation.

Early face presentation attack detection research predominantly relied on visible-spectrum RGB imagery, using appearance, texture, temporal cues, or learned representations to distinguish bona fide faces from printed photographs and replay attacks [1–3]. Although these approaches have substantially advanced RGB-based PAD, increasingly realistic presentation attack instruments and variations in acquisition conditions have motivated the use of additional sensing modalities. Multi-modal benchmarks such as CASIA-SURF introduced aligned RGB, depth, and infrared observations for studying complementary spoofing cues [4], while the WMCA database extended multi-channel PAD to color, depth, near-infrared, and thermal sensing [5]. CASIA-SURF CeFA further investigated multi-modal PAD under cross-ethnicity and unseen-attack conditions [6]. More recent research has moved beyond conventional feature fusion toward flexible and generalizable multi-modal learning. Yu et al. introduced a flexible-modal benchmark covering different sensor configurations [7], and subsequently investigated Vision Transformer architectures, modality-aware representations, and self-supervised pre-training for RGB, depth, and IR PAD [8]. Lin et al. addressed modality unreliability and imbalance under domain shifts through uncertainty-aware fusion and gradient rebalancing [9], while Yang et al. further studied both intra-domain modality misalignment and inter-domain shifts in multi-modal FAS [10]. In parallel, cross-spectral biometric research has highlighted both the complementary information and the potential security challenges associated with visible and near-infrared sensing [11, 12]. Collectively, these studies demonstrate that complementary spectral and sensing modalities can strengthen PAD, while also showing that effective multi-modal operation depends on reliable cross-modal integration and appropriate handling of modality-specific uncertainty and domain variation.

Despite these advances, most existing face PAD studies remain primarily model- or benchmark-oriented, with evaluation typically performed on public datasets under predefined intra-dataset, cross-dataset, or cross-domain protocols. Such evaluations are essential for comparing learning algorithms, but they do not fully capture the challenges involved in deploying a complete PAD subsystem in an operational biometric device. In practice, system performance also depends on factors that are often abstracted away in benchmark studies, including multi-camera acquisition and temporal pairing, face-quality gating, handling of missing or unreliable sensor observations, fixed operating-point configuration, multi-frame transaction policies, and the generation of traceable intermediate and final decisions. Moreover, relatively few published studies provide sufficient implementation detail to reproduce these system-level behaviors together with results from independent, standards-based third-party evaluation. ISO/IEC 30107-3 establishes a framework for testing and reporting biometric presentation attack detection performance [13], yet a gap remains between PAD methods demonstrated on curated benchmarks and complete on-device systems whose acquisition, decision, and logging pipelines are evaluated under controlled, protocol-driven conditions. Addressing this gap is particularly important for access-control applications, where the practical value of PAD depends not only on attack detection accuracy but also on repeatability, auditability, and the security–usability trade-off of the deployed operating point.

To address this gap, we present a complete on-device RGB–IR face PAD system designed for deployment in biometric access-control applications and for reproducible third-party evaluation. The system is implemented on a custom Android device integrating RGB and infrared cameras, with RGB and IR observations temporally paired before PAD processing. Rather than relying on a single classifier, the transaction pipeline combines three complementary sources of evidence: a primary RGB liveness score (S1), an RGB–IR cross-spectral consistency score (S2), and a secondary independent RGB liveness score (S3). These frame-level signals are processed through face-quality and validity gates and aggregated over multiple frames using configurable transaction-level decision thresholds. The system also records frame-level scores, intermediate decisions, and final transaction outcomes to provide traceability and support post-hoc analysis. The complete hardware–software implementation was independently evaluated by BixeLab using test protocols implemented in accordance with ISO/IEC 30107-3 [13]. At each evaluation level, 10 subjects participated in 50 bona fide transactions and 300 presentation-attack trials spanning six PAI species. No attack classification errors or attack non-responses were observed in either Level A or Level B, corresponding to an observed APCER of 0/300 and APNRR of 0/300, while 10 of 50 bona fide presentations were incorrectly rejected at each level. These results provide evidence of strong attack resistance against the evaluated PAI species while also exposing the security–usability trade-off associated with the conservative operating points used during testing.

The main contributions of this work are summarized as follows. First, we present a complete on-device RGB–IR PAD system co-designed across sensing hardware and Android software, including dual-camera acquisition, temporal RGB–IR pairing, face-quality validation, multi-stage PAD processing, and transaction-level decision making for practical access-control deployment. Second, we develop an auditable multi-signal transaction pipeline that combines a primary RGB live-



Fig. 1. Prototype device integrating RGB and IR sensing

ness score, RGB–IR cross-spectral consistency, and a secondary independent liveness score through multi-frame aggregation, while preserving frame-level and transaction-level evidence for traceability and post-hoc analysis. Third, we report an independent third-party evaluation of the complete system by BixeLab under ISO/IEC 30107-3-based test protocols at both Level A and Level B, providing system-level evidence across multiple PAI species and explicitly characterizing the resulting security–usability trade-off at the tested operating points.

The remainder of this paper is organized as follows. Section 2 describes the hardware platform, RGB–IR acquisition pipeline, on-device software architecture, and logging mechanisms. Section 3 presents the per-frame PAD processing stages and transaction-level decision logic. Section 4 describes the independent BixeLab evaluation protocol, reports the PAD results, and analyzes the behavior of the internal decision signals and operating points. Finally, Section 5 concludes the paper and discusses the main limitations and directions for improving bona fide usability while preserving attack resistance.

2. System Design and On-Device Implementation

2.1. Hardware and Data Acquisition

The proposed face presentation attack detection system is implemented on a custom Android access-control device integrating an RGB camera, an infrared (IR) camera, and an IR illuminator, as illustrated in Fig. 1. All acquisition, preprocessing, PAD scoring, and transaction-level decision making are performed locally on the device without relying on cloud-based inference. The hardware and software components are organized as a complete PAD subsystem intended for continuous operation in practical access-control scenarios. At the system level, each authentication attempt is treated as a transaction, allowing the observations and intermediate processing results generated during an attempt to be associated with a single final PAD decision. An overview of the resulting on-device processing architecture and data flow is shown in Fig. 2.

The Android application acquires the RGB and IR streams concurrently through a USB host acquisition pipeline. The two cameras are handled independently, with separate capture handlers and frame callbacks, allowing the sensing streams to operate asynchronously. Rather than requiring hardware-level synchronization between the two cameras, the system performs temporal RGB–IR pairing in software. The RGB stream serves as the trigger for PAD processing, while incoming IR frames are sampled and cached. For each selected RGB frame, the processing pipeline uses the most recent available IR observation to form an RGB–IR pair. This latest-IR pairing strategy provides cross-spectral observations in close temporal proximity while avoiding the additional hardware and computational complexity associated with tightly synchronized dual-camera acquisition.

The acquisition and processing pipeline is designed to prevent uncontrolled frame accumulation during continuous operation. At most one RGB PAD processing task is active at a time; incoming RGB frames are not queued for processing while the previous frame is still being analyzed. The IR stream is similarly sampled at a controlled rate so that an up-to-date IR observation remains available without introducing unnecessary processing overhead. These mechanisms bound the amount of pending work and reduce the risk of latency growth caused by an accumulating frame queue. Consequently, the acquisition architecture prioritizes recent RGB–IR observations and predictable on-device processing behavior rather than attempting to process every captured frame.

2.2. On-Device Software Architecture and Logging

The on-device software is organized around a transaction-oriented processing architecture in which each authentication attempt is assigned a unique transaction identifier and is processed as an independent sequence of RGB–IR observations. The pipeline follows a validity-first design: lightweight acquisition and quality checks are performed before computationally more expensive PAD stages, and processing does not continue when required sensing evidence is unavailable or invalid. In particular, the IR observation is treated as mandatory cross-spectral evidence. If a valid IR face observation cannot be obtained for the current attempt, the system terminates the attempt rather than producing a PAD decision from RGB evidence alone. This policy prevents incomplete multi-sensor observations from being silently accepted by downstream decision stages and supports conservative operation in unattended access-control scenarios.

The behavior of the transaction pipeline is controlled by an explicit set of configurable parameters stored on the device. These parameters include the number

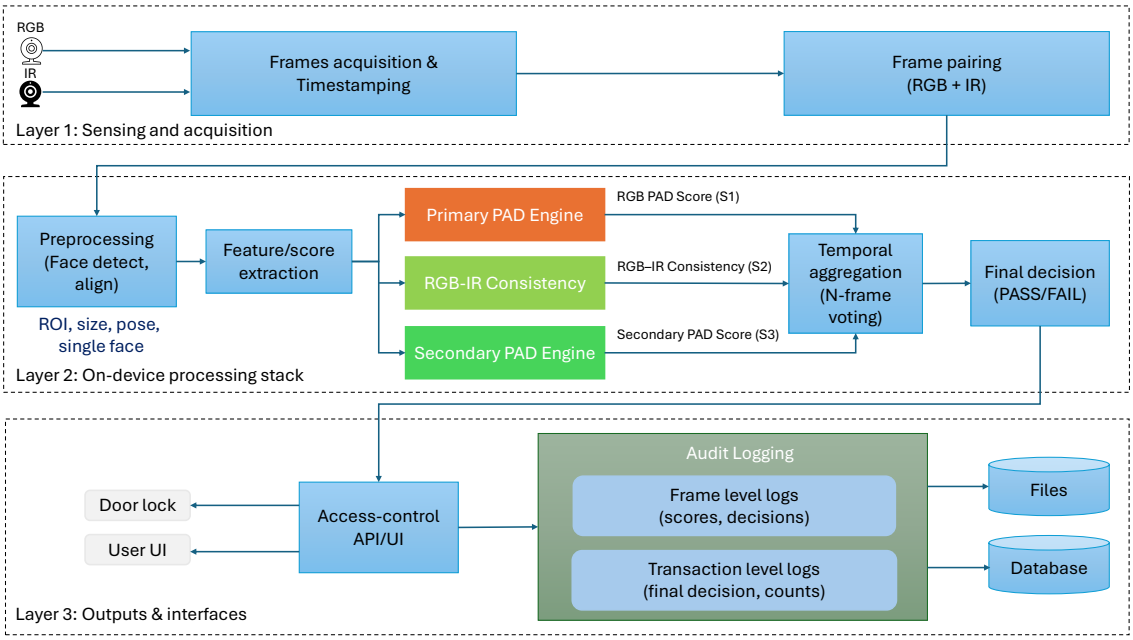


Fig. 2. On-device processing stack and data flow

of frames accumulated for a transaction, face-quality and geometric constraints, and the thresholds used by the internal PAD signals and transaction-level aggregation logic. During evaluation, the selected configuration remains fixed within each test condition so that the operating point does not adapt to individual presentations or subjects. This separation between configurable system parameters and runtime observations is important for reproducibility and for protocol-driven evaluation, since the exact operating point can be recorded and subsequently associated with the resulting PAD metrics. The parameter values used in the independent Level A and Level B evaluations are reported in Section 4.

Auditability is a central design requirement of the implementation. For every processed frame, the application records the transaction identifier, the frame index or timestamp, the internal PAD scores and their corresponding per-frame decisions, and relevant processing outcomes. At the transaction level, the system records summary quantities derived from the accumulated frame window, including valid-frame counts, spoof counts, and the final transaction result. This creates a traceable hierarchy from the final decision back to the individual observations and intermediate PAD signals that contributed to it. Such traceability is particularly important in a multi-stage PAD system because it enables post-hoc inspection of rejected transactions and analysis of the behavior of individual decision components.

The same logging mechanism also supports independent evaluation. Device-generated frame-level and transaction-level records can be exported and aligned with externally recorded ground-truth information using transaction identifiers and timestamps. The resulting records provide the information required to reconstruct the system outcome and analyze the contribution of the internal PAD signals without modifying the on-device decision process during testing. In the BixeLab evaluation described in Section 4, these audit logs were used as part of the protocol-driven analysis workflow for computing and examining PAD performance at the tested operating points.

3. Presentation Attack Detection Pipeline and Decision Logic

This section describes the PAD pipeline and transaction-level decision logic executed within the proposed on-device RGB–IR stack illustrated in Fig. 2. The method operates on paired RGB–IR observations and outputs a final PASS/FAIL decision after accumulating a fixed number of frames. Each authentication attempt is assigned a timestamp-based TestId to associate frame-level evidence with the final transaction outcome for auditability and evaluation. For each selected RGB frame, the system pairs it with the most recent IR frame and executes a gated pipeline consisting of face detection, quality control, per-frame scoring, and multi-frame aggregation.

3.1. Presentation Attack Detection Pipeline and Per-Frame Scoring

To ensure stable operation during real-world user interactions, the system applies quality gates before liveness scoring. Face detection is performed on the RGB stream within a predefined checking region; frames are rejected if multiple faces are present or if geometric constraints such as face scale and head-pose limits are violated. To prevent inconsistent observations from contaminating a transaction, temporal identity consistency is enforced by comparing the current RGB face template with the previous template; if similarity drops

below a threshold, the transaction state is reset and accumulation restarts. On the IR channel, a valid face detection within the designated IR checking region is mandatory; if no valid IR face is available, the attempt terminates immediately as a failure to avoid decisions based solely on RGB evidence.

For each RGB–IR pair that passes the gates, the method computes three complementary per-frame signals. We employ two independent commercial PAD engines on-device, one as the primary RGB liveness scorer (KBY) and another as a secondary liveness scorer (Neurotechnology), and fuse them with an RGB–IR consistency check. We define the following scores:

- S1: RGB PAD score (primary liveness). The RGB face crop is brightness-normalized (CLAHE-based equalization) and evaluated by a primary commercial liveness estimator. The resulting score is compared to a configurable threshold to produce a per-frame REAL/SPOOF outcome for S1.
- S2: RGB–IR consistency score (cross-spectral agreement). Face templates are extracted independently from RGB and IR observations and compared using a similarity function. If similarity is below a configurable threshold, the frame is marked SPOOF for S2, capturing cross-spectral inconsistency that is difficult to reproduce reliably by many presentation attack instruments (PAIs).
- S3: Secondary PAD score (independent liveness). A second liveness signal is computed from the RGB crop using liveness attributes provided by an independent commercial estimator (different from the primary one). When available, the score is used to assign a REAL/SPOOF outcome; otherwise, it is treated as UNKNOWN for that frame. All per-frame scores and decisions are recorded to support traceability and later analysis.

Algorithm 1 Transaction-level PAD decision

Require: Per-frame decisions for S_1, S_2, S_3 over N frames; thresholds T_1, T_2, T_3
Ensure: PASS/FAIL

$c_1 \leftarrow 0,\ c_2 \leftarrow 0,\ c_3 \leftarrow 0,\ v_3 \leftarrow 0$
for $i = 1$ **to** N **do**
  if $S_1(i) = SPOOF$ **then**
    $c_1 \leftarrow c_1 + 1$
  end if
  if $S_2(i) = SPOOF$ **then**
    $c_2 \leftarrow c_2 + 1$
  end if
  if $S_3(i)$ is valid **then**
    $v_3 \leftarrow v_3 + 1$
    if $S_3(i) = SPOOF$ **then**
      $c_3 \leftarrow c_3 + 1$
    end if
  end if
end for
$r_1 \leftarrow c_1/N,\ r_2 \leftarrow c_2/N$
if $r_1 > T_1$ **or** $r_2 > T_2$ **then**
  return FAIL
end if
if $v_3 > 0$ **then**
  $r_3 \leftarrow c_3/v_3$
  if $r_3 > T_3$ **then**
    return FAIL
  end if
end if
return PASS

3.2. Transaction-Level Aggregation and Final Decision

The final PAD decision is computed at the transaction level by aggregating evidence over a fixed window of NumFrame frames. The system maintains spoof counters and computes spoof ratios for the three signals. A transaction is rejected if either the S1 spoof ratio or the S2 spoof ratio exceeds its configured threshold, yielding a conservative policy that requires both RGB liveness and RGB–IR consistency to pass. If S1 and S2 pass, S3 is evaluated when sufficient valid S3 frames exist; the transaction is rejected if the S3 spoof ratio exceeds its threshold, while UNKNOWN frames are excluded from the S3 ratio computation. Finally, the transaction-level PASS/FAIL outcome and summary counts are written to the result log together with the corresponding TestId. Algorithm 1 presents the pseudocode for this transaction-level decision rule.

4. Experimental Results: Independent BixeLab Evaluation

4.1. BixeLab Evaluation Setup and Test Set

The proposed system was independently evaluated as a PAD subsystem by BixeLab under ISO/IEC 30107-3 at both PAD Level 1 (Level A) and PAD Level 2 (Level B) in a controlled laboratory environment in Australia. The evaluation specifically targets the presentation attack detection (PAD) subsystem rather than biometric matching accuracy, with the objective of testing resilience against selected attacks executed on the client-provided hardware. Level 1 (Level A) attacks cover high-quality 2D print and cutout variants. The Level 1 report lists six presentation attack instrument (PAI) species, each with 50 trials: passport photo on matte, selfie on matte, half-face (bottom) mask, 2D paper without cutout, 2D paper with eyes and mouth cutout, and 2D paper with eyes, nose and mouth cutout. Level 2 (Level B) attacks are more sophisticated and use fabrication-enhanced and cross-spectral PAIs. The Level 2 report lists six PAI species (50 trials each): balaclava (face printed on wearable fabric), selfie mounted on mobile cover, selfie mounted on iPad cover, selfie printed with IR ink, half-face (top) mask with IR ink, and deepfake video on iPad.

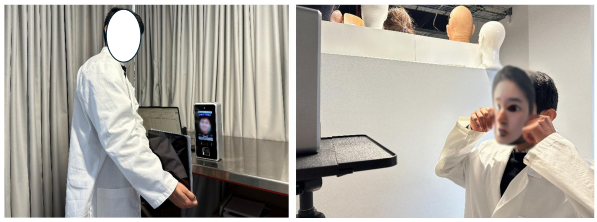


Fig. 3. Example laboratory test setup used during the independent BixeLab ISO/IEC 30107-3 evaluation (faces anonymized)

BixeLab's executed plan documents the test setup and logging workflow, including setting the device time zone for traceability, applying the threshold configuration of the item under test (IUT), conducting transactions through the on-device UI, and exporting device logs over ADB for analysis. Fig. 3 shows an example of the laboratory setup (faces anonymized). This setup supports repeatable execution of bona fide and presentation attack transactions under protocol-driven constraints.

For each evaluation level, BixeLab recruited 10 consenting test subjects and conducted 50 bona fide transactions and 300 presentation attack trials (six PAI species × 50 attacks each). The reports also specify operational "levels of effort" and decision-policy constraints, including a maximum of 45 seconds per attempt and 3 attempts per transaction.

For performance analysis, BixeLab used both transaction-level and frame-level outputs exported by the IUT. The Level 2 report enumerates the fields expected in the transaction log (e.g., TestId, frame counts and spoof counts per internal component, and the final liveness result) and the per-frame log (the frame image, internal scores, and frame decisions). For PAD subsystem evaluation, the reports identify the mandatory ISO/IEC 30107-3 metrics used in the analysis: APCER and APNRR for attacks, and BPCER and BPNRR for bona fide presentations.

Consistent with the decision rule described in Section 3, BixeLab documents the IUT operating point (threshold values) used during testing (e.g., the number of frames and the score thresholds). The workflow of the IUT is illustrated in Fig. 4.

4.2. Evaluation Outcomes

To report PAD performance, we use the ISO/IEC 30107-3 metrics that separately capture errors and non-responses for attack and bona fide presentations. APCER (Attack Presentation Classification Error Rate) measures the proportion of attack presentations that are incorrectly classified as bona fide (false accepts). APNRR (Attack Presentation Non-Response Rate) measures the proportion of attack presentations for which the system does not return a decision. Similarly, BPCER (Bona Fide Presentation Classification Error Rate) measures the proportion of bona fide presentations that are incorrectly classified

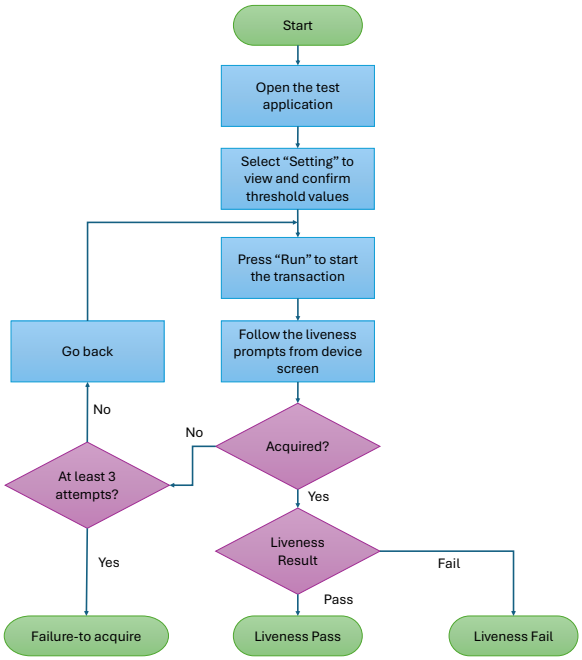


Fig. 4. Brief overview of the IUT workflow

as attacks (false rejects), and BPNRR (Bona Fide Presentation Non-Response Rate) measures the proportion of bona fide presentations that yield no decision. These metrics jointly reflect the security–usability trade-off of the PAD subsystem under a fixed operating point.

4.2.1. Overall results

Table 1 summarizes the overall PAD metrics. For Level A, BixeLab reports 10 bona fide classification errors out of 50 and 0 bona fide non-responses, while all 300 Level 1 attacks were rejected with 0 attack errors and 0 attack non-responses. For Level B, BixeLab reports APCER = 0% and APNRR = 0% across 300 Level 2 attacks, with 10 of 50 bona fide presentations incorrectly rejected.

4.2.2. Results by presentation attack detection species

Tables 2 and 3 report attack performance by PAI species. For Level A, BixeLab provides a detailed per-species summary with 0 errors observed and 0 non-responses across all six 2D print/cutout categories. For Level B, BixeLab reports the same outcome for all six more sophisticated attack types, including wearable/fabrication and cross-spectral IR-ink attacks as well as a deepfake video presentation. Overall, the proposed on-device RGB–IR system and its transaction-level PAD pipeline demonstrate acceptable PAD performance at both Level 1 and Level 2 within the scope and under the configurations and conditions of the protocol-driven ISO/IEC 30107-3 evaluations.

4.2.3. Additional analysis

BixeLab also conducted a score-level analysis for Level A using labeled transaction logs and ROC curves. The analysis reports that the two internal liveness scores

Table 1. Overall ISO/IEC 30107-3 PAD results reported by BixeLab

Evaluation	APCER	APNRR	BPCER	BPNRR
PAD Level 1 (A)	0/300	0/300	10/50	0/50
PAD Level 2 (B)	0/300	0/300	10/50	–

Table 2. PAD Level 1 (A) results by PAI species (50 trials per species)

PAI species (Level A)	#Tests	APCER	APNRR
Passport photo on matte	50	0/50	0/50
Selfie on matte	50	0/50	0/50
Half-face (bottom) mask	50	0/50	0/50
2D paper without cutout	50	0/50	0/50
2D paper with eyes and mouth cutout	50	0/50	0/50
2D paper with eyes, nose and mouth cutout	50	0/50	0/50

provide strong discriminative power, while the RGB–IR consistency score provides limited separability at the Level A operating point (ScoreRGBIR threshold of 0.1). Specifically, BixeLab reports AUC values of 0.97 for the primary liveness score and 0.92 for the secondary liveness score, compared with 0.54 for the RGB–IR consistency score (Table 4). Fig. 5 shows substantial overlap in ScoreRGBIR values between bona fide and Level 1 attack presentations, particularly in the 0.7 to 0.95 range. This overlap indicates that ScoreRGBIR alone has limited discriminative utility.

BixeLab also analyzed the RGB–IR consistency signal for Level B and reported that all three internal liveness scores provide strong discriminative power with AUC values of 0.96, 0.92, and 0.93 for the primary liveness score, the secondary liveness score, and the RGB–IR consistency score, respectively (Table 5); the RGB–IR consistency AUC is substantially higher than its Level A value. The Level B evaluation also used a stricter S2 operating threshold of 0.7. As shown in Fig. 6, the score distributions exhibit greater separation between bona fide and Level B attack presentations, indicating that the RGB–IR consistency signal is substantially more discriminative for the more sophisticated PAIs evaluated.

4.2.4. Discussion

BixeLab notes that for Level A, while 315 Level 1 attack trials were present in exported logs, only 300 configuration-verified trials were included in the final analysis (with superseded early trials retained for traceability).

Regarding bona fide usability, the system rejected all attack presentations at both levels but also incorrectly rejected 10/50 bona fide transactions under the fixed operating point used during the independent evaluation. This outcome is consistent with a conservative, security-first configuration in which strict quality gates and multi-signal, multi-frame aggregation favor rejection when any required evidence is missing or inconsistent. Potential contributors to these bona fide rejections include illumination-dependent variations in the RGB liveness score, deviations in face pose, scale, or position that trigger the face-quality gates, incomplete or unreliable IR observations, RGB–IR inconsistency, and borderline frame-level decisions amplified by the multi-frame aggregation policy. However, the independent evaluation data do not provide a unique root-cause annotation for each rejected bona fide transaction. Therefore, the relative contribution of these factors cannot be quantified from the present experiment. In practical deployments, efforts to reduce BPCER while maintaining low APCER can focus on controlled operating-point calibration and policy refinement. Examples include tuning decision thresholds on a held-out bona fide set captured under representative lighting conditions and with representative user behavior, introducing a calibrated "retry" policy that separates quality failures from spoof decisions, and applying confidence-aware aggregation so that a small number of borderline frames does not dominate the transaction outcome. Because the system logs per-frame scores and intermediate decisions, these adjustments can be performed in a traceable manner and validated using the same ISO/IEC 30107-3 metrics reported by BixeLab, enabling an explicit security–usability trade-off analysis at deployment time.

A further limitation of the present evaluation is the relatively small bona fide participant pool. Although the independent evaluation involved 10 subjects, 50 bona fide transactions, and 300 attack presentations at each PAD evaluation level, the limited number of participants restricts the generalizability of the observed bona fide rejection rate to a broader user population. Accordingly, the reported BPCER should be interpreted as the performance observed under the specific participant pool, operating point, laboratory conditions, and interaction protocol used in the present evaluation. Future eval-

Table 3. PAD Level 2 (B) results by PAI species (50 trials per species)

PAI species (Level B)	#Tests	APCER	APNRR
Balaclava mask (face-printed wearable)	50	0/50	0/50
Selfie mounted on mobile cover	50	0/50	0/50
Selfie mounted on iPad cover	50	0/50	0/50
Selfie printed with IR ink	50	0/50	0/50
Half-face (top) mask with IR ink	50	0/50	0/50
Deepfake video on iPad	50	0/50	0/50

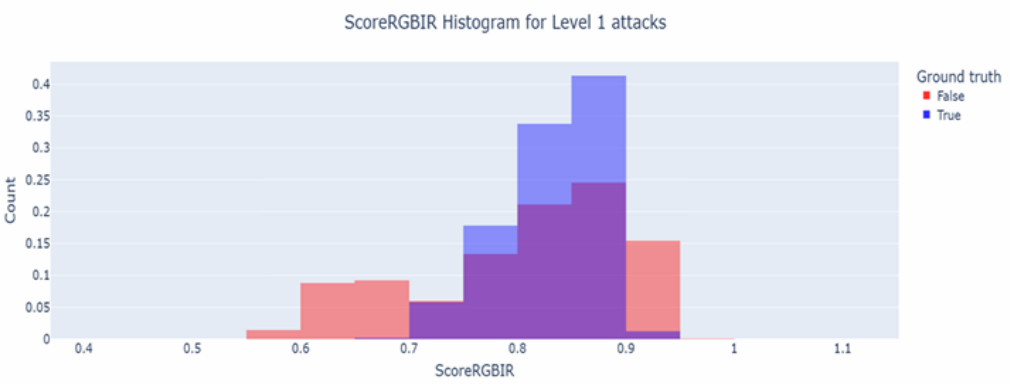


Fig. 5. Histogram of ScoreRGBIR values by ground-truth class for the Level 1 (A) evaluation. Source: BixeLab additional analysis – Test Report Level A (identifier withheld for review), v1.01

Table 4. Additional analysis (Level A): AUC of internal scores computed by BixeLab from labeled transaction logs. Source: BixeLab additional analysis – Test Report Level A (identifier withheld for review), v1.01

Internal score	AUC (Level A)
Primary liveness score (S1)	0.97
RGB–IR consistency (S2)	0.54
Secondary liveness score (S3)	0.92

Table 5. Additional analysis (Level B): AUC of internal scores computed by BixeLab from labeled transaction logs. Source: BixeLab additional analysis – Test Report Level B (identifier withheld for review), v1.0

Internal score	AUC (Level B)
Primary liveness score (S1)	0.96
RGB–IR consistency (S2)	0.93
Secondary liveness score (S3)	0.92

uation should include a larger and more diverse participant population together with broader environmental and operational conditions.

5. Conclusions

This paper presented a complete on-device RGB–IR face presentation attack detection system designed for practical biometric access-control deployment and independent third-party evaluation. The system integrates dual-camera RGB–IR acquisition with temporal frame pairing, validity and quality gating, three complementary PAD signals, multi-frame transaction-level aggregation, and auditable frame-level and transaction-level logging. Rather than introducing a new standalone PAD classifier, the work focuses on the system-level integration required to obtain stable, traceable, and reproducible PAD decisions on an embedded Android platform.

The complete hardware–software system was independently evaluated by BixeLab under ISO/IEC 30107-3-based test protocols at both Level A and Level B. At each level, all 300 presentation attacks across six PAI species were rejected without non-responses (APCER and APNRR of 0/300), while 10 of 50 bona fide presentations were incorrectly rejected (BPCER of 20%). These results demonstrate strong resistance to the evaluated PAI species under the tested configurations and laboratory conditions, while also revealing a clear security–usability trade-off associated with the conservative operating points.

The additional score-level analysis further shows that the contribution of the individual PAD signals depends on the characteristics of the evaluated attacks. In particular, the RGB–IR consistency signal exhibited limited discrimination for the Level A PAI species but substantially stronger discrimination for the more sophisticated Level B attacks. This observation reinforces the value

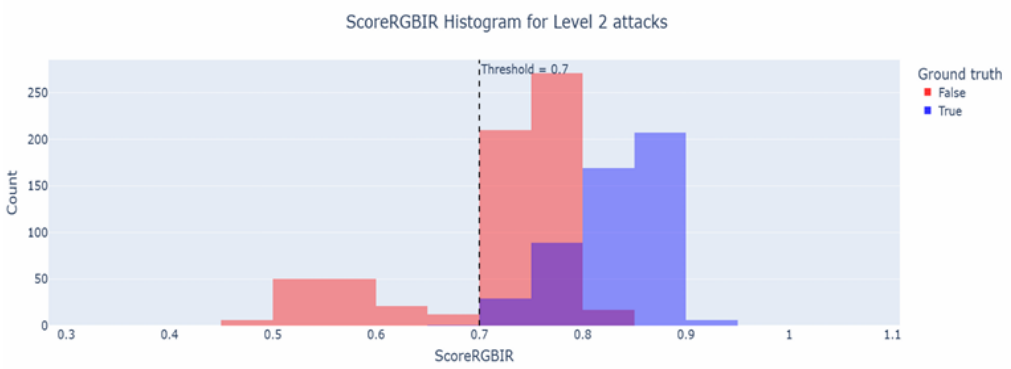


Fig. 6. Histogram of RGB–IR consistency scores (S2) by ground-truth class for the Level 2 (B) evaluation. Source: BixeLab additional analysis – Test Report Level B (identifier withheld for review), v1.0

of analyzing the behavior of individual system components rather than relying solely on the final transaction outcome.

Future work will focus on reducing bona fide rejection while preserving low attack error rates through improved operating-point calibration, clearer separation between acquisition-quality failures and spoof evidence, and more adaptive transaction-level aggregation. Further evaluation across a broader range of environmental conditions, with larger subject populations and additional PAI species will also be necessary to assess the generalizability of the system beyond the conditions considered in the present independent evaluation.

Acknowledgements

The acknowledgements will be updated after the peer-review process.

References

[1] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, OULU-NPU: A mobile face presentation attack database with real-world variations, in Proccedings 12th IEEE Int. Conference Automatic Face & Gesture Recognition (FG), Washington, DC, USA, pp. 612–618, Jun. 2017.

[2] Y. Liu, A. Jourabloo, and X. Liu, Learning deep models for face anti-spoofing: Binary or auxiliary supervision, in Proccedings IEEE Conference Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, Jun. 2018.

[3] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, Deep learning for face anti-spoofing: A survey, IEEE Transactions Pattern Analysis and Machine Intelligence, vol. 45, iss. 5, pp. 5609–5631, May 2023. https://doi.org/10.1109/TPAMI.2022.3215850

[4] S. Zhang, A. Liu, J. Wan, Y. Liang, G. Guo, S. Escalera, H. J. Escalante, and S. Z. Li, CASIA-SURF: A large-scale multi-modal benchmark for face anti-spoofing, IEEE Transactions Biometrics, Behavior, and Identity Science, vol. 2, iss. 2, pp. 182–193, Apr. 2020. https://doi.org/10.1109/TBIOM.2020.2973001

[5] A. George, Z. Mostaani, D. Geissbuhler, O. Nikisins, A. Anjos, and S. Marcel, Biometric face presentation attack detection with multi-channel convolutional neural network, IEEE Transactions Information Forensics and Security, Sep. 2019.

[6] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, CASIA-SURF CeFA: A benchmark for multi-modal cross-ethnicity face anti-spoofing, in Proccedings IEEE/CVF Winter Conference Applications of Computer Vision (WACV), pp. 1179–1187, 2021.

[7] Z. Yu, A. Liu, C. Zhao, K. H. M. Cheng, X. Cheng, and G. Zhao, Flexible-modal face anti-spoofing: A benchmark, in Proccedings IEEE/CVF Conference Computer Vision and Pattern Recognition Workshops, pp. 6346– 6351, Aug. 2023.

[8] Z. Yu, R. Cai, Y. Cui, X. Liu, Y. Hu, and A. C. Kot, Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing, International Journal of Computer Vision, vol. 132, pp. 5217–5238, Jun. 2024. https://doi.org/10.1007/s11263-024-02055-1

[9] X. Lin, S. Wang, R. Cai, Y. Liu, Y. Fu, W. Tang, Z. Yu, and A. Kot, Suppress and rebalance: Towards generalized multi-modal face anti-spoofing, in Proccedings IEEE/CVF Conference Computer Vision and Pattern Recognition (CVPR), pp. 211–221, 2024.

[10] J. Yang, X. Lin, Z. Yu, L. Zhang, X. Liu, H. Li, X. Yuan, and X. Cao, DADM: Dual alignment of domain and modality for face anti-spoofing, in Proccedings IEEE/CVF International Conference Computer Vision (ICCV), pp. 12045–12056, May 2025.

[11] D. Anghelone, C. Chen, A. Ross, and A. Dantcheva, Beyond the visible: A survey on cross-spectral face recognition, Neurocomputing, vol. 611, Jan. 2025, Art. no. 128626. https://doi.org/10.1016/j.neucom.2024.128626

[12] B. Hu, K. Wang, and C. H. Chang, Red bleed: A pragmatic near-infrared presentation attack on facial biometric authentication systems, in Proccedings 34th USENIX Security Symposium, Seattle, WA, USA, Aug. 2025.

[13] ISO/IEC, ISO/IEC 30107-3:2023 Information Technology—Biometric Presentation Attack Detection—Part 3: Testing and Reporting, Geneva, Switzerland: International Organization for Standardization, 2023.