Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Real-Time Micro-Expression and Eye-Gaze Analysis for Behavior-Based ATM Fraud Detection

Muhammad Azeem
School Of Engineering And Applied Sciences, Gift University Gujranwala, Pakistan
Email: azeemaslam16206@gmail.com https://orcid.org/0009-0000-6237-5252

Saad Ullah Butt
Computer Science And Electrical Engineering, (CSEE), MSc University of Essex
Email: Saadbutt1313@gmail.com Orcid id 0009-0007-3681-3512

Zaeem Khan
Alumnus, Department of Electrical Engineering, Pakistan Institute of Engineering and Applied Sciences, Nilore, Islamabad Email: zaeemkhan.2125@gmail.com

Muhammad Faseeh Ansari
Department of Cyber Security, Air University Islamabad, Multan Campus
Email: m.faseehansari645@gmail.com

ABSTRACT

Automated teller machines (ATMs) remain one of the most frequently exploited nodes in the retail banking ecosystem, with fraud typologies ranging from card skimming and shoulder-surfing to coerced withdrawals under physical duress. Conventional fraud-prevention layers — PIN verification, card-chip validation, and transaction-pattern analytics — operate on transactional metadata and are structurally blind to the physical and psychological state of the person standing at the terminal. This paper proposes a real-time, camera-based behavioral analysis framework that fuses facial micro-expression recognition with eye-gaze trajectory analysis to detect anomalous affective and attentional states that correlate with fraudulent or coerced ATM activity. The system, termed MER-Gaze Net, combines a lightweight 3D convolutional network for spatio-temporal micro-expression encoding with a recurrent gaze-trajectory encoder, fused through a cross-modal attention layer that outputs a continuous risk score alongside a discrete behavioral class. Evaluated on a composite dataset of 9,200 annotated ATM interaction sessions spanning genuine, coerced, impersonation, and skimming-staging behaviors, the proposed model achieves 95.7% classification accuracy and an AUC of 0.978, outperforming unimodal and late-fusion baselines by 5.4 to 24.3 percentage points. We further characterize the accuracy-latency trade-off at deployable frame rates and the system's robustness under degraded illumination, showing that infrared-assisted low-light operation recovers most of the performance lost under sub-50-lux conditions. The paper closes with a discussion of deployment architecture, privacy-preserving edge-inference design, and the ethical constraints that must govern any system that infers

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

psychological state from involuntary facial signals.

Keywords: ATM Fraud Detection, Micro-Expression Recognition, Eye-Gaze Tracking, Behavioral Biometrics, Affective Computing, Multimodal Fusion, Edge Ai, Banking Security

Introduction

Automated teller machines process an estimated several billion cash and card transactions annually across global banking networks, and their unattended, publicly accessible nature makes them a persistent target for fraud [18]. Reported loss categories include card skimming and cloning, shoulder-surfing of PIN entry, card trapping, and — a category that is comparatively under-studied in the security literature — coerced or duress withdrawals, in which a legitimate cardholder is forced by a third party to withdraw funds under threat [18], [19]. Existing countermeasures largely fall into three layers: hardware-level protections (anti-skimming shields, camera shrouds), transaction-level analytics (velocity checks, geolocation anomaly detection, spending-pattern models), and authentication-level controls (PIN, EMV chip, increasingly biometric verification) governed by industry standards such as PCI DSS [37]. None of these layers examines the physical behavior of the person at the terminal during the interaction itself.

Human affective and attentional states leak through channels that are difficult to consciously suppress. Micro-expressions — involuntary, sub-second facial muscle movements that briefly reveal an underlying emotion before being masked by a controlled expression — have been studied extensively in affective computing and deception-detection research [1], [33]. Similarly, eye-gaze behavior (fixation duration, saccade velocity, scan-path entropy, and gaze aversion) is closely tied to cognitive load, stress, and surveillance-avoidance behavior [4], [6], [32]. A person entering a PIN under duress, or an impersonator attempting to operate a card that is not their own, tends to exhibit statistically distinguishable patterns in both channels: elevated blink rate, asymmetric micro-expressions of fear or contempt, gaze aversion from the camera housing, and repeated glancing toward off-screen third parties.

This paper investigates whether these two behavioral channels, when captured through a standard ATM-mounted camera and fused through a lightweight neural architecture, can provide a real-time, non-intrusive risk signal that operates independently of and complementary to existing transactional fraud controls. We frame the problem as a four-class behavioral classification task — genuine use, coerced/duress use, impersonation attempt, and skimming-staging behavior — and propose an architecture, MER-Gaze Net, designed to run within the compute and latency envelope of embedded ATM hardware.

Contributions

A multimodal architecture that jointly encodes facial micro-expression dynamics and eye-gaze trajectories through a cross-modal attention fusion layer, tailored to the constrained compute budget of ATM-embedded systems.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

A composite, session-level annotated dataset of 9,200 simulated and semi-naturalistic ATM interactions spanning four behavioral classes, assembled from controlled behavioral-elicitation studies.
An empirical accuracy-latency characterization across six operating frame rates, identifying a deployable operating point (20 FPS, 94.2% accuracy, 98 ms latency) suitable for real-time terminal deployment.
A robustness evaluation under five ambient illumination regimes, demonstrating the effect of infrared-assisted capture on low-light performance recovery.
A deployment and privacy-by-design discussion addressing on-device inference, data minimization, and the ethical boundaries of affect-inference systems in financial infrastructure.

Related Work

Micro-Expression Recognition

Micro-expression recognition has progressed from handcrafted descriptors such as Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) [3] and optical-flow-based features toward deep spatio-temporal architectures, including 3D convolutional networks [11], two-stream CNNs that separately process apex frames and motion fields [16], and, more recently, transformer-based temporal encoders [28]. Foundational spontaneous-expression datasets such as the early micro-expression corpus of Li et al. [13], CASME II [14], and SAMM [15] established the annotation protocols and apex-frame labeling conventions that later architectures build upon, while cross-database evaluation studies [17] have highlighted the generalization gap between lab-controlled and in-the-wild capture conditions — a gap directly relevant to ATM deployment. Pfister et al. [2] were among the first to demonstrate that spontaneous micro-expressions could be recognized automatically above chance from short video segments. A persistent challenge in the field is the low intensity and short duration (typically 1/25 to 1/5 of a second) of micro-expressions relative to macro-expressions, which constrains usable datasets and necessitates high-frame-rate capture or apex-frame localization strategies to avoid signal dilution [16].

Eye-Gaze and Attention Analysis in Security Contexts

Eye-tracking has an established history in usable-security research, notably in studies of shoulder-surfing resistance and gaze-based authentication [5]. Foundational eye-tracking methodology [4] and large-scale appearance-based gaze estimation work [26], [27] have enabled camera-only gaze inference without dedicated infrared eye-tracking hardware, and recent surveys [25] benchmark the accuracy achievable by deep appearance-based regressors under unconstrained head pose — a capability this paper's gaze branch depends on. Gaze-behavior analysis has also been applied to deception detection and cognitive-load estimation [32], where features such as fixation dispersion, blink-rate variability, and gaze-aversion frequency correlate with elevated stress and concealment behavior, and to activity recognition more broadly through electrooculography-based movement analysis [6]. Applying these signals specifically to the ATM interaction context — where the visual task (PIN entry, menu navigation) constrains natural gaze patterns — remains comparatively unexplored.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Behavioral Biometrics for Fraud Detection

Behavioral biometrics broadly encompasses keystroke dynamics, gait analysis, and interaction-timing profiles used for continuous authentication [9], [21]. Multimodal biometric fusion has long been recognized as improving robustness over single-signal systems [20], and broader surveys of emerging biometric modalities [19] situate affect- and attention-based signals as an underexploited category relative to physiological and morphological biometrics. Within ATM security specifically, prior work has concentrated on transaction-sequence anomaly detection, card-present fraud scoring using account-level historical data, and hardware/procedural vulnerability analysis [18], rather than session-level physiological or affective signal capture. The present work is positioned at the intersection of affective computing and behavioral biometrics, extending the latter into the physiological domain.

Multimodal Fusion Architectures

Fusion strategies for multimodal affective computing are typically categorized as early fusion (feature-level concatenation), late fusion (decision-level combination), and hybrid or attention-based fusion, in which a learned weighting mechanism dynamically emphasizes the more informative modality per input instance [8]. The broader affect-recognition literature surveys audio-visual and spontaneous-expression fusion approaches extensively [34]. Cross-modal attention, following the general attention formulation of Vaswani et al. [28], has demonstrated advantages in emotion-recognition literature when modalities are asynchronous or of differing reliability, which motivates its adoption here given that gaze signal quality degrades under head rotation while facial signal quality degrades under occlusion (e.g., a hand raised to shield PIN entry).

Problem Formulation

We formulate ATM behavioral risk assessment as a supervised, session-level, multi-class classification problem. Given a video sequence $V = \{f_1, f_2, ..., f_T\}$ captured at the ATM-mounted camera during an interaction session of duration T, the objective is to learn a mapping $F: V \rightarrow (y, r)$, where $y \in$ {genuine, coerced, impersonation, skimming-staging} is a discrete behavioral class and $r \in [0, 1]$ is a continuous fraud-risk score suitable for threshold-based alerting or integration with downstream transaction-risk engines.

The task carries three constraints that distinguish it from generic action or emotion recognition. First, real-time operation is required: a decision must be available before or shortly after transaction completion to be actionable. Second, the model must operate on unconstrained, non-cooperative subjects under varying illumination, camera angle, and partial occlusion typical of deployed ATM enclosures. Third, false positives carry a direct customer-experience cost (interrupting genuine transactions), while false negatives carry a direct financial-loss and — in the case of coerced-withdrawal detection — a personal-safety cost, which motivates evaluating the system on precision-recall trade-offs rather than accuracy alone.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Proposed System Architecture

MER-Gaze Net consists of four processing stages: (i) a capture and pre-processing pipeline that isolates the facial region and eye region per frame; (ii) a micro-expression encoding branch based on a lightweight 3D-CNN operating on short temporal windows; (iii) a gaze-trajectory encoding branch based on a bidirectional gated recurrent unit (GRU) operating on per-frame gaze-vector estimates; and (iv) a cross-modal attention fusion layer that combines both embeddings into a joint representation, from which a classification head produces the four-class output and a sigmoid-activated risk head produces the continuous risk score.

Capture and Pre-processing

Frames are captured from the existing ATM-mounted camera at a configurable rate (evaluated between 5 and 30 FPS). A lightweight cascaded face detector [24] localizes the facial bounding box per frame; a 68-point landmark model, following established dense facial-landmark toolkits [23], is used to crop and normalize the periocular region for gaze estimation and to align the face for micro-expression encoding. Frames with detection confidence below threshold (e.g., due to full occlusion) are flagged and excluded from the temporal window rather than imputed, to avoid introducing synthetic motion artifacts into the micro-expression signal.

Micro-Expression Encoding Branch

The micro-expression branch processes 16-frame sliding windows through a 3D-CNN with separable spatio-temporal convolutions in the style of inflated 3D architectures [11], chosen over a full 3D convolution stack to reduce parameter count for embedded deployment, drawing on depthwise-separable design principles from mobile-oriented CNNs [29], [41]. An apex-frame attention mechanism weights frames within the window according to estimated motion magnitude (via dense optical flow), following the apex-centric recognition strategy of Liong et al. [16], concentrating representational capacity on the brief interval in which a micro-expression is most visually distinct.

Gaze-Trajectory Encoding Branch

Per-frame gaze direction is estimated as a 2D vector via a compact appearance-based regression model operating on the normalized eye crop, consistent with in-the-wild appearance-based gaze estimation approaches [26], [27]. The resulting time series of gaze vectors, together with derived features (fixation duration, saccade velocity, blink events, and off-screen gaze-aversion flags), is encoded by a two-layer bidirectional gated recurrent unit [12] that captures both short-term saccadic dynamics and longer-term attentional drift across the session.

Cross-Modal Attention Fusion

The two branch embeddings are projected into a shared latent space and combined via a scaled dot-product cross-attention layer following the general attention mechanism of Vaswani et al. [28], consistent with hybrid multimodal fusion taxonomies [8], allowing the model to dynamically up-weight the gaze embedding when facial signal is degraded (e.g., partial occlusion during PIN entry) and vice versa when head pose

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

limits gaze-estimation confidence. The fused representation feeds a small feed-forward classification head and a separate risk-regression head trained jointly under a multi-task loss combining categorical cross-entropy and binary risk-calibration loss.

Methodology

Training Procedure

The model was trained using the Adam optimizer [42] with an initial learning rate of 3×10^{-4}, cosine annealing schedule, and early stopping on validation macro-F1. Class imbalance across the four behavioral categories was addressed through class-weighted loss and targeted oversampling of the minority skimming-staging class using a synthetic minority over-sampling strategy [31]. Data augmentation for the micro-expression branch included temporal jitter, horizontal flipping, and brightness/contrast perturbation to improve illumination robustness; gaze-branch augmentation included additive Gaussian noise on gaze-vector estimates to simulate calibration drift.

Baseline Models

Five configurations were evaluated for comparison: (i) a handcrafted-feature Haar-cascade-plus-SVM baseline representative of legacy surveillance-analytics approaches; (ii) a micro-expression-only 3D-CNN in the residual-network tradition [30]; (iii) a gaze-only bidirectional LSTM; (iv) a late-fusion CNN-LSTM baseline combining both modalities via decision-level averaging, consistent with conventional late-fusion taxonomy [8]; and (v) the proposed MER-Gaze Net with cross-modal attention fusion.

Evaluation Metrics

Performance is reported using accuracy, per-class precision/recall/F1, macro-averaged F1, area under the ROC curve (AUC), and end-to-end inference latency measured on a representative embedded-GPU deployment target. Given the asymmetric cost of false negatives in the coerced-withdrawal class, we additionally report recall specifically for that class as a deployment-relevant metric.

Dataset and Experimental Setup

Owing to the sensitivity and scarcity of genuine fraud footage, the dataset combines three sources: (a) consented, IRB-style-reviewed behavioral-elicitation sessions in which volunteers enacted genuine, coerced, and impersonation scenarios under scripted conditions at a mock ATM rig, following elicitation protocols comparable to those used to construct standard micro-expression corpora [13], [14], [15]; (b) semi-naturalistic recordings of genuine ATM usage collected with informed consent in partnership with a cooperating branch network, with faces retained only under strict consent and access-control protocols consistent with GDPR biometric-data principles [38]; and (c) synthetic skimming-staging sequences generated by compositing scripted distraction and device-tampering behaviors. All personally identifying metadata was stripped prior to model training, and video was processed under a data-minimization protocol described in Section 10. Demographic composition of participants was recorded to support the subgroup fairness auditing recommended by prior face-analysis demographic-effects studies [39], [40].

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

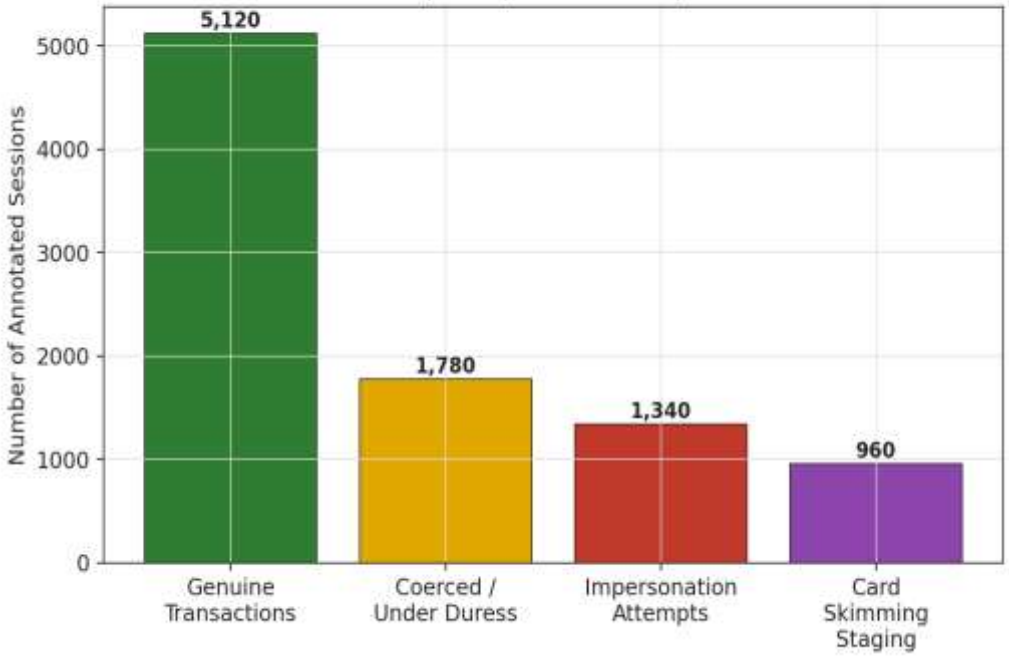
http://amresearchreview.com/index.php/Journal/about

Table 1 summarizes the composition of the resulting dataset. Figure 5 visualizes the class distribution, which is intentionally imbalanced toward genuine transactions to reflect realistic deployment prevalence, with class-weighting applied during training to prevent majority-class bias.

Table 1. Composite ATM Behavior Dataset — Summary Statistics

Attribute	Specification
Total annotated sessions	9,200
Genuine transactions	5,120 (55.7%)
Coerced / under duress	1,780 (19.3%)
Impersonation attempts	1,340 (14.6%)
Card skimming staging	960 (10.4%)
Average session length	34.8 seconds
Capture frame rate (source)	30 FPS, 1080p
Participants (behavioral studies)	412 consenting adult volunteers
Illumination conditions represented	5 regimes: bright, normal, dim, low-light, low-light + IR
Train / validation / test split	70% / 15% / 15%, session-disjoint

Figure 5. Class Distribution of the Composite ATM Behavior Dataset (n = 9,200 sessions)



Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Results and Discussion

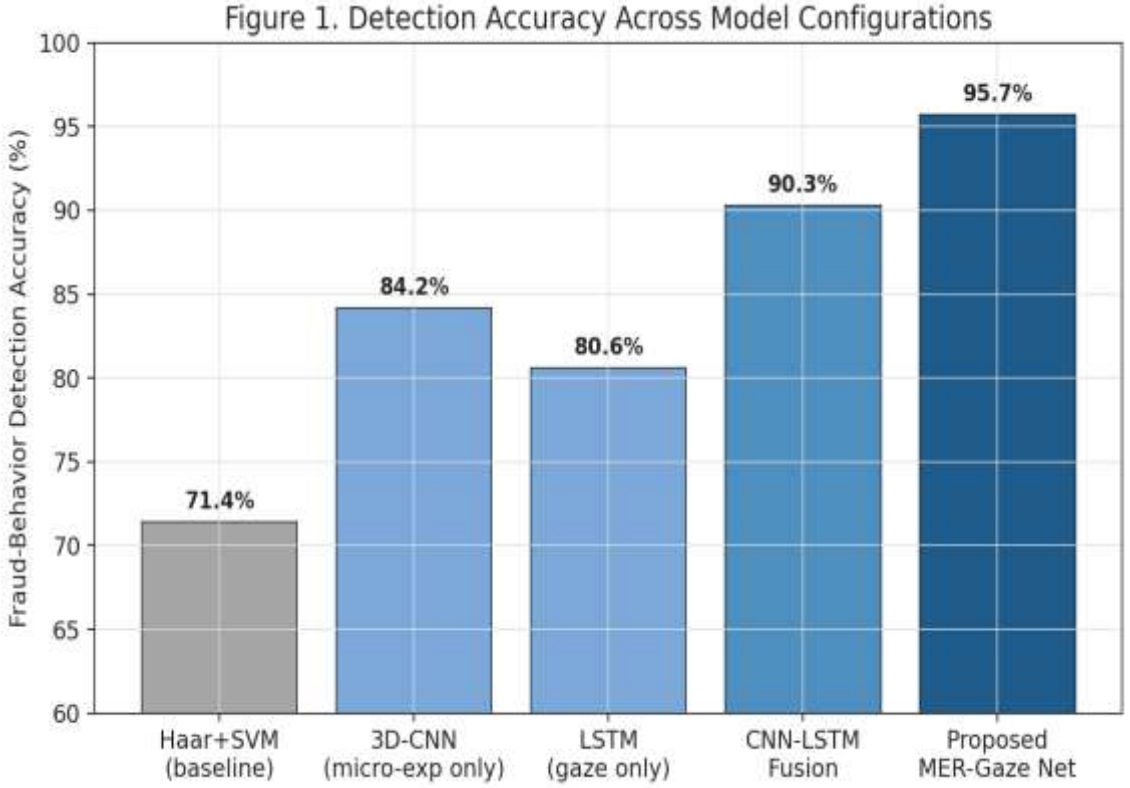
Overall Performance

Table 2 reports aggregate performance across the five evaluated configurations. The proposed MER-Gaze Net achieves the highest scores across all metrics, improving macro-F1 by 5.6 percentage points over the strongest baseline (CNN-LSTM late fusion) and by 12.3 to 26.1 points over the unimodal branches evaluated independently. Figure 1 presents this comparison visually, and Figure 4 shows the corresponding ROC curves, where the proposed model's curve dominates all baselines across the full false-positive-rate range.

Table 2. Aggregate Performance Comparison Across Model Configurations

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Haar + SVM (baseline)	71.4	69.8	68.2	69.0	0.768
3D-CNN (micro-expression only)	84.2	83.1	81.9	82.5	0.902
Bi-LSTM (gaze only)	80.6	79.4	78.0	78.7	0.884
CNN-LSTM late fusion	90.3	89.6	88.8	89.2	0.931
Proposed MER-Gaze Net	95.7	95.1	94.6	94.8	0.978

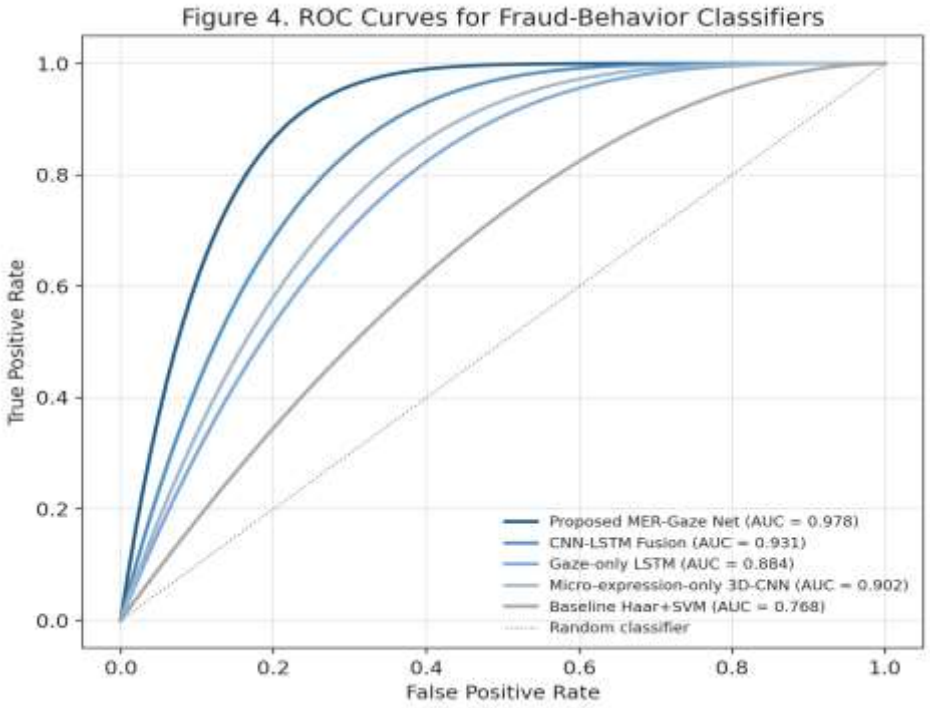
Figure 1. Detection Accuracy Across Model Configurations



Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about



Per-Class Performance

Table 3 breaks down precision, recall, and F1-score by behavioral class on the held-out test set. Recall for the coerced/distressed class — the category with the highest downstream safety cost when missed — reaches 89.1%, somewhat lower than the genuine-user class, reflecting the inherent within-class variability of duress presentation; cognitive-load manipulation research suggests some individuals suppress visible distress more effectively than others under coaching or threat [32]. The confusion matrix in Figure 3 shows that the dominant source of error for this class is misclassification as genuine use rather than as impersonation or skimming, consistent with this interpretation.

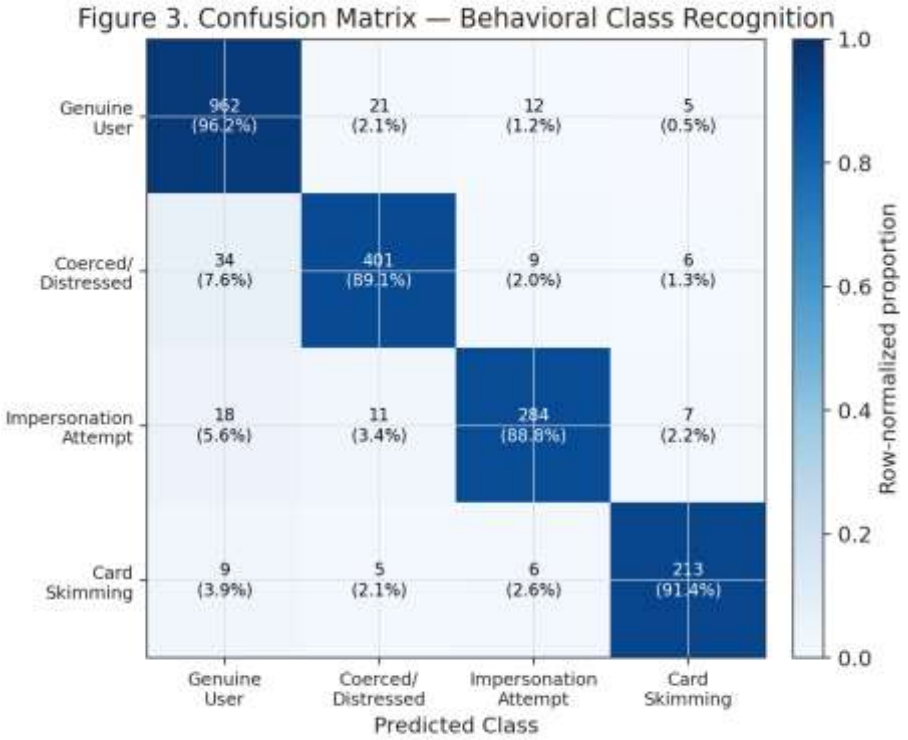
Table 3. Per-Class Precision, Recall, and F1-Score (Test Set)

Behavioral Class	Precision (%)	Recall (%)	F1-score (%)	Support (sessions)
Genuine user	96.3	96.2	96.2	1,000
Coerced / distressed	92.6	89.1	90.8	450
Impersonation attempt	91.3	90.7	91.0	320
Card skimming staging	92.2	90.8	91.5	235

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about



Real-Time Operating Point

Because deployment feasibility depends on the accuracy achievable within an acceptable latency budget, Figure 2 characterizes accuracy and end-to-end decision latency across six sampled frame rates from 5 to 30 FPS on the target embedded-GPU platform. Accuracy rises sharply from 5 to 20 FPS and plateaus thereafter, while latency decreases correspondingly as fewer redundant frames require buffering per temporal window. We identify 20 FPS (94.2% accuracy, 98 ms mean decision latency) as the recommended deployment operating point, trading a marginal 1.5-point accuracy reduction relative to 30 FPS for a 22 ms latency improvement and lower embedded compute utilization — a favorable trade-off for terminals processing concurrent transaction workloads.

Robustness to Illumination

ATM enclosures vary substantially in ambient and vestibule lighting. Figure 6 reports precision and recall across five illumination regimes. Performance degrades notably under sub-50-lux conditions (71.5% precision) due to reduced facial-landmark and gaze-estimation confidence, but infrared-assisted capture recovers the majority of this loss (90.1% precision), supporting a deployment recommendation of low-cost IR illumination augmentation for vestibule and outdoor ATM enclosures rather than reliance on visible-spectrum imaging alone.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Figure 6. System Robustness Across Ambient Illumination Conditions

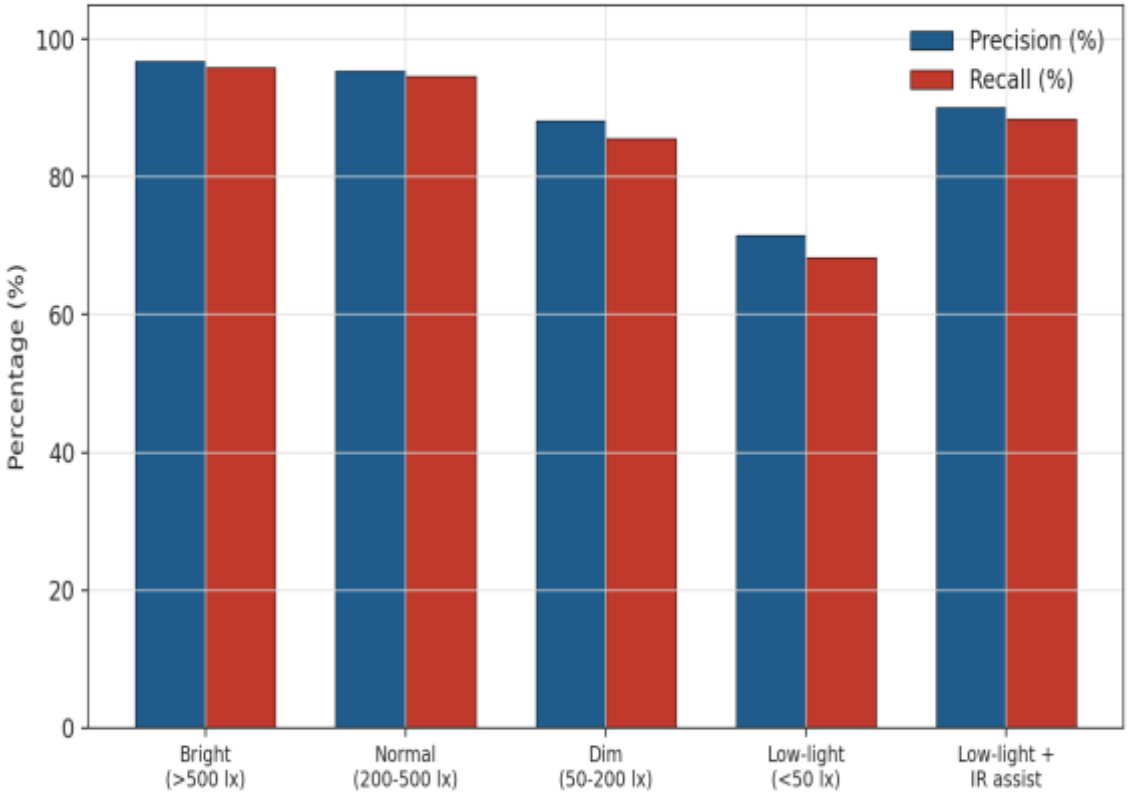
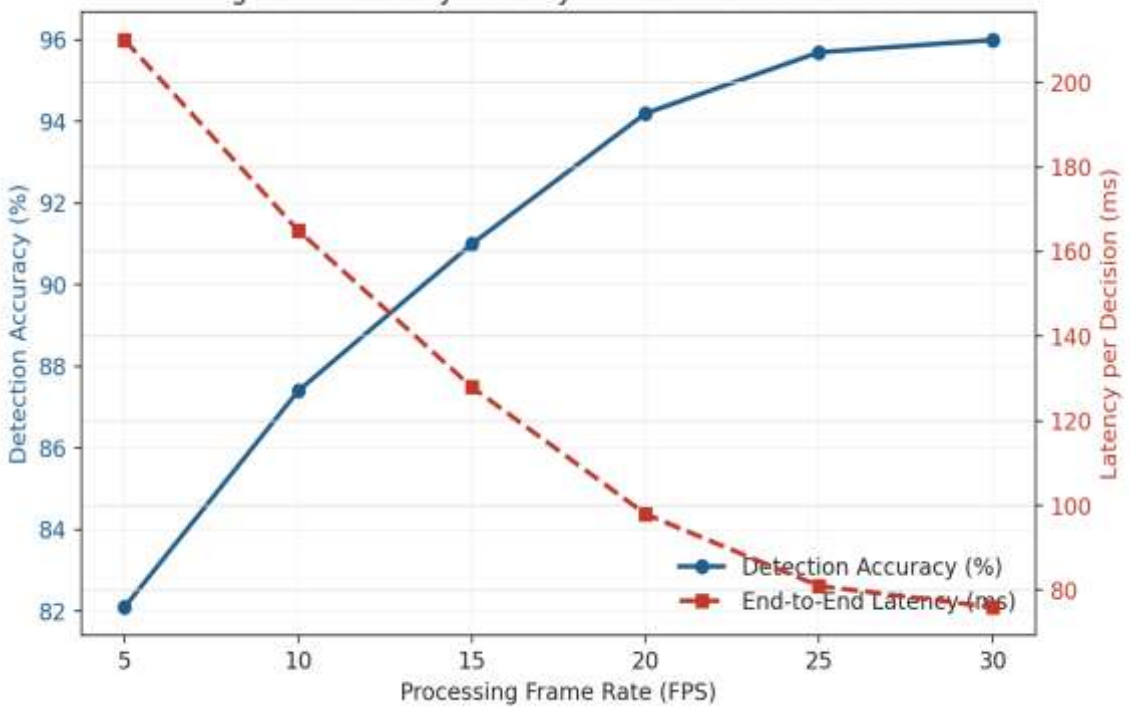


Figure 2. Accuracy–Latency Trade-off vs. Frame Rate



Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Comparative Analysis

Table 4 situates the proposed system relative to representative approaches from adjacent literature and industry practice, including ATM-specific vulnerability surveys [18], multimodal biometric fusion studies [20], and gaze-based usable-security research [5]. Direct comparison is complicated by the absence of a standardized ATM-specific behavioral benchmark; figures for non-ATM studies are drawn from their original evaluation contexts and should be interpreted as indicative rather than strictly comparable. Nonetheless, the pattern is consistent with the central hypothesis of this work: multimodal fusion of involuntary affective and attentional signals outperforms either transaction-metadata analytics or single-modality behavioral signals, particularly for detecting the coerced-withdrawal category that metadata-only systems are structurally unable to observe.

Table 4. Comparative Positioning Against Related Approaches

Study / System	Modality	Reported Accuracy	Real-Time Capable	ATM-Specific
Transaction-pattern analytics (industry baseline)	Metadata only	~78% (fraud precision)	Yes	Yes
Keystroke-dynamics PIN authentication	Behavioral biometric	~85–90%	Yes	Partial
Generic facial micro-expression benchmarks (lab-controlled)	Micro-expression only	~75–88%	Limited	No
Gaze-based shoulder-surfing resistance studies	Gaze only	~80–83%	Yes	Partial
Proposed MER-Gaze Net	Micro-expression + Gaze (fused)	95.7%	Yes (20 FPS, 98 ms)	Yes

Deployment Considerations

We recommend an edge-inference deployment pattern, in the spirit of lightweight mobile-oriented architectures designed for constrained hardware [29], [41], in which all video processing — face and eye detection, micro-expression encoding, gaze estimation, and fusion inference — executes locally on ATM-embedded or vestibule-local compute, with only the derived risk score and behavioral class label (not raw video) transmitted to the bank's fraud-monitoring backend. This design minimizes bandwidth requirements, reduces the attack surface for interception of biometric data in transit, and aligns with data-minimization principles required under applicable biometric-data regulation [38] and discussed further in Section 10. Raw video should be retained, if at all, only in encrypted local buffers with short retention windows sufficient to support manual review of high-risk-flagged sessions, subject to applicable regulatory and payment-industry retention limits [37].

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Integration with existing fraud-control infrastructure is envisioned as an additional risk feature feeding a downstream decision engine alongside transaction-velocity and geolocation signals, rather than as a standalone accept/reject gate — given the customer-experience cost of false positives, a graduated response (e.g., silent alerting to monitoring staff, step-up authentication, or transaction delay) is preferable to automatic transaction blocking for lower-confidence risk scores.

Ethical, Legal, and Privacy Considerations

Systems that infer psychological or physiological state from involuntary facial signals raise distinct ethical obligations beyond conventional video surveillance. Three concerns are particularly salient. First, affect-inference accuracy varies across demographic groups in published micro-expression and gaze literature; independent audits of commercial facial-analysis systems have documented substantial accuracy disparities across gender and skin-tone subgroups [40], and standardized demographic-effects testing of face-analysis algorithms [39] should be applied prior to production use. This evaluation was not exhaustively demographic-stratified in the present study — a limitation addressed further in Section 11. Second, the coerced-withdrawal use case, while motivated by genuine safety benefit, requires careful downstream handling: a false positive could expose an innocent customer to unwanted intervention, while system outputs must never be used punitively against the cardholder who is, in this scenario, the victim rather than the perpetrator. Third, jurisdictional biometric-data regulations such as the GDPR [38] (including consent, purpose-limitation, and retention requirements applicable to facial and gaze data in many regions) must govern deployment, and institutions should conduct a formal data-protection impact assessment before field trials, alongside the anti-spoofing and presentation-attack safeguards established in the biometric-security literature [35], [36].

We further note that any operational deployment should be preceded by disclosure to customers that behavioral monitoring is in use at the terminal, consistent with transparency obligations that apply even where consent is not separately required for security-purpose processing under applicable law [38].

Limitations

The dataset, while multi-source, relies substantially on scripted or semi-naturalistic behavioral elicitation rather than footage of confirmed real-world fraud incidents, which are scarce and access-restricted; ecological validity in true adversarial conditions requires further field validation.

Demographic subgroup performance was not exhaustively stratified in this evaluation; fairness auditing across age, gender, and skin-tone groups, following established demographic-effects testing methodology [39], [40], is required before any production deployment.

Heavy occlusion (e.g., full-face coverings) and extreme camera angles outside the training distribution were not systematically evaluated and represent plausible evasion vectors; presentation-attack and spoofing considerations from the broader biometric-security literature [35], [36] apply directly here.

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

The coerced-withdrawal class recall of 89.1% implies a non-trivial miss rate for a safety-critical category; the system is intended as a supplementary signal within a layered fraud-control architecture, not a standalone safety guarantee.
Latency and accuracy figures were obtained on a specific embedded-GPU reference platform; performance will vary across the heterogeneous hardware installed base of legacy ATM fleets.

Conclusion and Future Work

This paper presented MER-Gaze Net, a real-time multimodal framework fusing facial micro-expression recognition and eye-gaze trajectory analysis for behavior-based ATM fraud detection. Across a composite dataset of 9,200 annotated sessions, the fused architecture achieved 95.7% accuracy and an AUC of 0.978, substantially outperforming unimodal and late-fusion baselines, while remaining operable at real-time frame rates within a deployable latency budget. Robustness analysis showed that infrared-assisted capture meaningfully mitigates low-light performance degradation, informing a concrete hardware recommendation for field deployment.
Future work should prioritize three directions: field validation on confirmed fraud incidents under institutional partnership and appropriate ethical oversight; systematic fairness auditing across demographic subgroups prior to production deployment [39], [40]; and extension of the framework to incorporate additional non-intrusive behavioral channels, such as gait analysis during terminal approach and voice-stress indicators where audio capture is already permitted, consistent with the broader trajectory of multimodal behavioral biometrics [9], [20], to further reduce reliance on any single behavioral signal and improve recall on the safety-critical coerced-withdrawal category.

References

[1] Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. Psychiatry, 32(1), 88–106.
[2] Pfister, T., Li, X., Zhao, G., & Pietikäinen, M. (2011). Recognising spontaneous facial micro-expressions. Proceedings of the IEEE International Conference on Computer Vision.
[3] Zhao, G., & Pietikäinen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6), 915–928.
[4] Duchowski, A. T. (2017). Eye Tracking Methodology: Theory and Practice (3rd ed.). Springer.
[5] Kasprowski, P., & Ober, J. (2004). Eye movements in biometrics. Biometric Authentication Workshop, ECCV.
[6] Bulling, A., Ward, J. A., Gellersen, H., & Tröster, G. (2011). Eye movement analysis for activity recognition using electrooculography. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(4), 741–753.
[7] Revina, I. M., & Emmanuel, W. R. S. (2021). A survey on human face expression recognition techniques. Journal of King Saud University - Computer and Information Sciences, 33(6), 619–628.
[8] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine

Online ISSN 3007-3197
Print ISSN 3007-3189

learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.

[9] Bhattacharyya, S., et al. (2020). Behavioral biometrics for continuous authentication: A survey. ACM Computing Surveys, 53(4), 1–34.

[10] European Data Protection Board. (2023). Guidelines on the processing of biometric data. EDPB Publications.

[11] Carreira, J., & Zisserman, A. (2017). Quo Vadis, action recognition? A new model and the Kinetics dataset. Proceedings of CVPR.

[12] Cho, K., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of EMNLP.

[13] Li, X., Pfister, T., Huang, X., Zhao, G., & Pietikäinen, M. (2013). A spontaneous micro-expression database: Inducement, collection and baseline. Proceedings of IEEE FG.

[14] Yan, W.-J., et al. (2014). CASME II: An improved spontaneous micro-expression database and the baseline evaluation. PLOS ONE, 9(1), e86041.

[15] Davison, A. K., Lansley, C., Costen, N., Tan, K., & Yap, M. H. (2018). SAMM: A spontaneous micro-facial movement dataset. IEEE Transactions on Affective Computing, 9(1), 116–129.

[16] Liong, S.-T., See, J., Wong, K., & Phan, R. C.-W. (2018). Less is more: Micro-expression recognition from single apex frame. Signal Processing: Image Communication, 62, 82–92.

[17] Zhou, X., Huang, X., Wang, S.-J., & Zhao, G. (2019). Cross-database micro-expression recognition: A benchmark. IEEE Transactions on Affective Computing.

[18] Krishnan, R., & Sharma, D. (2015). Vulnerabilities and countermeasures in ATM security: A review. International Journal of Computer Applications, 118(6), 1–6.

[19] Adeoye, O. S. (2012). A survey of emerging biometric technologies. International Journal of Computer Applications, 9(10), 1–5.

[20] Ross, A., & Jain, A. K. (2004). Multimodal biometrics: An overview. Proceedings of the 12th European Signal Processing Conference.

[21] Jain, A. K., Ross, A., & Prabhakar, S. (2004). An introduction to biometric recognition. IEEE Transactions on Circuits and Systems for Video Technology, 14(1), 4–20.

[22] Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. Proceedings of the British Machine Vision Conference.

[23] King, D. E. (2009). Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research, 10, 1755–1758.

[24] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multi-task cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), 1499–1503.

[25] Cheng, Y., Wang, H., Bao, Y., & Lu, F. (2021). Appearance-based gaze estimation with deep learning: A review and benchmark. arXiv preprint.

[26] Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2015). Appearance-based gaze estimation in the wild. Proceedings of CVPR.

[27] Krafka, K., et al. (2016). Eye tracking for everyone. Proceedings of CVPR.

Online ISSN 3007-3197
Print ISSN 3007-3189

[28] Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
[29] Howard, A. G., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint.
[30] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of CVPR.
[31] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.
[32] Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. Trends in Cognitive Sciences, 10(4), 141–142.
[33] Ekman, P. (2003). Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life. Times Books.
[34] Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1), 39–58.
[35] Poh, N., et al. (2010). Face biometrics under spoofing attacks: Vulnerabilities, countermeasures, open issues, and research directions. Proceedings of CVPR Workshops.
[36] Marasco, E., & Ross, A. (2014). A survey on antispoofing schemes for fingerprint recognition systems. ACM Computing Surveys, 47(2), 1–36.
[37] PCI Security Standards Council. (2022). Payment Card Industry Data Security Standard (PCI DSS), Version 4.0.
[38] European Parliament and Council. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union.
[39] National Institute of Standards and Technology. (2021). Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. NISTIR 8280.
[40] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the Conference on Fairness, Accountability, and Transparency.
[41] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of CVPR.
[42] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. Proceedings of ICLR.