Real-World AI Systems

https://journals.cypedia.net/rwas

ARTICLE

A Multi-Layered Framework for Robust Real-Time Fraud Detection Against Deepfake Attacks in Digital Banking

Michele Trifiletti

Department of Social Sciences, Universidad Católica San Antonio de Murcia (UCAM), 30107 Murcia, Spain

ABSTRACT

The rapid evolution of generative artificial intelligence has significantly transformed the security risks faced by digital banking systems, particularly through the emergence of highly realistic synthetic audio and video capable of enabling identity fraud. These deepfake-driven threats challenge traditional biometric authentication and liveness detection mechanisms, creating critical vulnerabilities in real-time financial environments where detection systems must operate under strict latency and regulatory constraints. This paper presents a multi-layered, system-oriented framework for real-time fraud detection that reconceptualizes deepfake detection as a robustness problem rather than a standalone classification task. The proposed architecture combines physiological signal analysis using remote photoplethysmography, continuous behavioral biometrics for session-level trust evaluation, and adversarially trained neural models designed to detect artifacts associated with generative systems. The framework explicitly incorporates regulatory considerations relevant to high-risk artificial intelligence applications, including explainability, data minimization, and edge-based processing. By integrating technical resilience with practical deployment and governance requirements, this work offers a deployable model for strengthening digital banking security against emerging synthetic media attacks. The proposed framework is evaluated through a robustness-oriented analytical approach, focusing on system-level resilience rather than isolated detection accuracy. Evaluation dimensions include attack resistance, detection latency, and cross-layer consistency under adversarial conditions. The results indicate that combining heterogeneous signals reduces single-point failure risks and increases attacker complexity. Although the study does not report benchmark-based empirical validation, the proposed evaluation logic provides a structured basis for assessing deployability in

*CORRESPONDING AUTHOR:

Michele Trifiletti, Department of Social Sciences, Universidad Católica San Antonio de Murcia (UCAM), 30107 Murcia, Spain;
Email: mtrifiletti@alu.ucam.edu

ARTICLE INFO

Received: 5 October 2025 | Revised: 2 November 2025 | Accepted: 11 November 2025 | Published Online: 17 November 2025
DOI: https://doi.org/10.64797/rwas.v1i2.107

CITATION

Trifiletti, M., 2025. A Multi-Layered Framework for Robust Real-Time Fraud Detection Against Deepfake Attacks in Digital Banking. Real-World AI Systems. 1(2): 60–73. DOI: https://doi.org/10.64797/rwas.v1i2.107

adversarial, latency-constrained, and regulation-intensive financial settings. These findings support the feasibility of robustness-driven, system-level fraud detection strategies for digital banking applications.

Keywords: Deepfake Detection; Digital Banking Security; Real-Time Fraud Detection; Remote Photoplethysmography; Behavioral Biometrics; Adversarial Machine Learning; Identity Fraud Prevention; Trustworthy Artificial Intelligence

1. Introduction

Artificial intelligence has become a core component of modern digital banking, enabling automated customer onboarding, biometric authentication, transaction authorization, and fraud prevention at scale [1,2]. As these systems evolve, ensuring their resilience against emerging synthetic media attacks has become a primary objective for the industry [3]. The shift toward non-face-to-face interactions has increased efficiency and accessibility, but it has also expanded the attack surface of financial systems, particularly in identity verification processes. Prior work in financial technologies and digital identity ecosystems has emphasized that AI-driven decision systems increasingly operate as critical trust infrastructures rather than auxiliary automation tools [4,5].

Recent advances in generative artificial intelligence have significantly lowered the cost and complexity of producing realistic synthetic identities. Deep generative models are now capable of synthesizing highly realistic facial imagery, voice patterns, and video sequences that challenge traditional assumptions underlying biometric security mechanisms [6,7]. Empirical studies demonstrate that modern deepfake techniques achieve high perceptual realism while introducing artifacts that are often subtle, model-dependent, and difficult to detect using conventional visual inspection or static classifiers [8,9]. As a result, identity fraud is evolving from isolated attack scenarios into a systemic risk affecting the integrity of digital financial infrastructures.

Artificial intelligence (AI) systems deployed in financial services increasingly operate under conditions of incomplete information, asymmetric incentives, and adversarial pressure. Unlike traditional Information Technology (IT) security mechanisms, which are often rule-based and reactive, AI-driven decision systems must generalize across heterogeneous user behaviors while remaining resilient to deliberate manipulation. This tension between generalization and robustness has been extensively discussed in adversarial machine learning literature, which shows that models optimized for predictive performance may exhibit brittle behavior when exposed to adversarial inputs or distributional shifts [10–12]. The challenge is particularly pronounced in identity-centric applications, where errors propagate downstream into financial, legal, and reputational consequences.

In digital banking environments, identity verification does not merely function as an access control mechanism but as a foundational trust primitive upon which subsequent automated decisions depend. Processes such as credit scoring, transaction monitoring, and customer risk profiling implicitly inherit the assumptions established during onboarding and authentication. Consequently, weaknesses in identity verification pipelines amplify systemic exposure, transforming local vulnerabilities into institution-wide risk vectors. Prior research in biometric security and presentation attack detection has consistently highlighted that attackers exploit implicit trust assumptions embedded in sensing and acquisition mechanisms rather than only classifier weaknesses [13,14].

From a regulatory perspective, fraud detection and identity verification systems deployed in digital banking are increasingly classified as high-risk applications. European regulatory frameworks—including the Fifth Anti-Money Laundering Directive [15], the European Banking Authority Guidelines on remote customer onboarding [16], the General Data Protection Regulation [17], and the Artificial Intelligence Act—impose strict requirements concerning robustness, transparency, explainability, and data protection [16–18]. These regulatory developments reinforce the necessity of designing fraud detection mechanisms that are not only accurate but also resilient, auditable, and governance-compatible [19].

Despite the rapid expansion of deepfake detection research, the majority of existing approaches remain model-centric and benchmark-driven. Prior studies primarily focus on improving classification accuracy, feature engineering, or artifact detection under controlled experimental conditions [7,8]. Comparatively less attention has been devoted to the systemic properties required for deployment in adversarial, latency-constrained, and highly regulated environments such as digital

banking.

This gap highlights a critical limitation in current research: improvements in detection accuracy have not been matched by equivalent advances in robustness under adversarial, real-time, and regulated deployment conditions.

Current literature provides limited guidance on how heterogeneous detection mechanisms should be orchestrated at the architectural level to mitigate adaptive attackers capable of exploiting model-specific weaknesses. Financial identity systems operate under constraints that differ substantially from typical computer vision benchmarks, including strict response-time requirements, asymmetric error costs, and regulatory obligations concerning explainability, auditability, and data minimization [5,16].

This gap suggests that detection accuracy alone is insufficient as a design objective for real-world fraud prevention systems. Instead, robustness must be examined as an emergent system-level property shaped by signal diversity, failure-mode decoupling, and adversarial cost asymmetries [12].

While existing literature has predominantly addressed deepfake detection as a standalone classification problem, the present study departs from this perspective by framing deepfake-driven identity fraud as a system-level robustness challenge. The methodological contribution of this work does not lie in proposing a new isolated detection algorithm, but in the design of a multi-layered trust assessment architecture intended for adversarial, real-time digital banking environments.

In contrast, this study explicitly shifts the focus from model-centric detection to system-level robustness, where fraud resilience is treated as an emergent property of architectural design rather than classifier performance alone.

The contribution of this work does not reside in proposing a new standalone deepfake detection algorithm. Instead, it lies in reconceptualizing deepfake resilience as a system-level robustness problem and formalizing an architectural framework designed for adversarial financial environments.

Specifically, the novelty of the proposed framework is threefold.

First, the study introduces a robustness-oriented threat model that explicitly accounts for injection attacks and adaptive generative adversaries, extending beyond traditional presentation attack assumptions commonly adopted in biometric security research [12,13].

Second, the framework defines a multi-layered trust assessment architecture that integrates heterogeneous and partially independent signal classes—physiological dynamics, behavioral interaction patterns, and adversarial artifact detection—thereby reducing correlated failure risks and increasing attacker complexity. This design operationalizes robustness principles discussed in adversarial machine learning literature within identity-centric financial systems [11,20].

Third, operational and regulatory constraints are elevated to primary design variables rather than treated as secondary implementation considerations. By embedding explainability, data minimization, and edge-based processing into the core architecture, the framework aligns technical resilience with governance requirements for high-risk AI systems [18,19].

Under this perspective, fraud resilience is conceptualized not as a property of individual models but as an emergent characteristic of architectural composition under adversarial pressure. This approach differs from prior work that primarily optimizes model-level detection accuracy, instead conceptualizing fraud resilience as an emergent property of system design and signal diversity.

Taken together, these contributions redefine deepfake fraud detection as a robustness engineering problem, aligning system design with adversarial threat models and regulatory requirements.

This paper addresses the gap between deepfake detection research and real-world deployment by proposing a system-level framework for real-time fraud detection that enhances robustness against adversarial threats while remaining compatible with operational and regulatory constraints.

The relevance of deepfake technologies in this context extends beyond their immediate capacity to deceive biometric classifiers. Their broader impact lies in the erosion of epistemic certainty within digital identity systems. When synthetic media becomes indistinguishable from authentic biometric input at the perceptual level, the notion of ground truth shifts from observable appearance to latent consistency across multiple dimensions. This transition challenges not only technical detection mechanisms but also the conceptual foundations of trust in remote interactions, a concern increasingly reflected in both media forensics and AI security research [7,8].

Moreover, deepfake-enabled fraud exhibits characteristics of economic scalability that distinguish it from earlier attack classes. Once a generative pipeline is established, marginal attack costs approach zero, enabling repeated and par-

allelized impersonation attempts. Adversarial learning theory suggests that such asymmetries fundamentally alter defender–attacker dynamics, requiring defensive strategies that constrain attack scalability rather than merely improving per-instance detection accuracy[12,20].

Recent scholarship (2023–2025) has shown that high benchmark accuracy alone is insufficient for deepfake detection in adversarial and real-time settings. Current research increasingly stresses multimodal verification, cross-dataset generalization, resilience to evolving manipulations, and evaluation criteria that reflect deployment constraints such as computational efficiency, latency, and robustness in security-sensitive environments[21,22]. These developments reinforce the central premise of the present study: effective fraud prevention in digital banking requires not only accurate detection components, but also architectural strategies capable of integrating heterogeneous evidence under operational and regulatory constraints.

2. Threat Landscape: From Presentation Attacks to Injection Attacks

Early forms of biometric fraud primarily involved presentation attacks, such as the use of printed photographs or prerecorded videos presented to a camera sensor. These attacks could often be mitigated through basic liveness detection techniques, including blink detection or prompted head movements.

It is important to note that presentation attacks implicitly preserved a causal linkage between the attacker and the biometric trait being presented. Even when artifacts were used, the attack surface remained constrained by physical interaction requirements, environmental conditions, and sensor characteristics. This linkage enabled defenders to exploit discrepancies between expected physiological or behavioral responses and observed signals, forming the basis for early liveness detection techniques.

As a result, the defensive posture of biometric systems evolved incrementally, focusing on improving sensor fidelity and refining challenge–response protocols rather than questioning the integrity of the data acquisition pipeline itself. Modern attack vectors have evolved into injection attacks, in which synthetic audio or video streams are injected directly into the software pipeline of an application, bypassing the physical sensor entirely. In such scenarios, the assumption that the camera or microphone provides trustworthy input no longer holds. Generative models can simulate anatomically coherent facial movements, eye blinking, and speech in real time, rendering traditional motion-based liveness checks ineffective.

The sophistication of contemporary generative models lies not only in their visual or acoustic realism, but in their ability to approximate higher-order temporal dependencies. Facial motion, speech prosody, and micro-expressions can now be synthesized in ways that satisfy short-term coherence constraints, thereby undermining detection strategies that rely on isolated frames or brief temporal windows.

This development blurs the boundary between synthesis and simulation. Rather than producing static forgeries, generative systems increasingly function as adaptive simulators capable of responding to prompts, environmental cues, and interaction dynamics in real time. Such capabilities significantly complicate the task of distinguishing between authentic user behavior and algorithmically generated responses.

Injection attacks fundamentally alter the threat model of digital identity systems and necessitate detection strategies based on signals that are difficult to reproduce coherently through generative processes.

Injection attacks also introduce new challenges for forensic analysis and post-incident attribution. Because synthetic streams may never traverse a physical sensor, traditional indicators such as device tampering, environmental inconsistencies, or hardware artifacts may be absent. This lack of tangible evidence complicates both internal investigations and regulatory reporting, increasing reliance on probabilistic assessments derived from behavioral and physiological inconsistencies.

Furthermore, injection-based impersonation enables attackers to decouple identity fraud from specific devices or locations. A single synthetic identity can be instantiated across multiple sessions, platforms, or institutions, raising concerns about cross-channel and cross-institutional risk propagation.

3. Materials and Methods

The proposed framework is not evaluated on a single benchmark dataset due to the constraints of real-world financial environments, where access to representative fraud data is limited. Instead, the design is informed by established datasets and experimental findings reported in the literature on deepfake detection, biometric authentication, and adversarial machine learning.

The system assumes multimodal inputs, including facial video streams, behavioral interaction data, and audio signals where applicable. Preprocessing is performed at the edge to ensure data minimization and compliance with regulatory requirements.

3.1. System-Level Design Rationale

The proposed framework is conceived as a real-world AI system rather than a standalone detection model. Its design is guided by the assumption that deepfake-based fraud is an adaptive, adversarial phenomenon targeting not only individual classifiers but the entire identity verification pipeline.

Unlike conventional approaches that focus on improving individual detection models, the proposed framework adopts a system-oriented design perspective. The objective is not to maximize classifier performance in isolation, but to enhance robustness under adaptive adversarial conditions through architectural decomposition and heterogeneous signal fusion. This distinction is methodologically significant, as it shifts the analytical focus from model accuracy to system resilience, failure-mode decoupling, and trust dynamics over time.

Accordingly, the system is structured as a multi-layered trust assessment architecture, where heterogeneous signals are evaluated independently and fused at decision time. This approach mitigates the risk of single-point failure and increases resilience against model inversion, overfitting, and targeted evasion strategies.

Table 1 compares traditional fraud detection approaches with the proposed multi-layered framework, highlighting differences in threat coverage, robustness, and regulatory alignment.

Table 1. Comparison between Traditional Fraud Detection Systems and the Proposed Framework.

Dimension	Traditional Systems	Proposed Multi-Layered Framework
Verification Paradigm	One-time authentication	Continuous trust assessment
Threat Model	Presentation attacks	Injection & deepfake attacks
Detection Signals	Visual appearance only	Physiological, behavioral, adversarial
Liveness Dependency	High	Low
Robustness to Adaptation	Limited	High
Explainability	Often opaque	Explicit reason codes
Regulatory Alignment	Partial	Embedded by design
Failure Mode	Single-point failure	Defense-in-depth

From an operational perspective, the framework is designed to:

- Operate under strict real-time latency constraints typical of digital banking workflows;
- Support deployment on heterogeneous consumer devices;
- Enable explainability and auditability at the system level;
- Minimize the processing and transmission of sensitive biometric data.

These constraints reflect the requirements imposed on high-risk AI systems in regulated financial environments.

From a systems engineering perspective, these requirements necessitate a departure from monolithic model design in favor of modular architectures capable of isolating failure modes. In adversarial settings, the assumption that any single analytical component may be compromised must be treated as a baseline condition rather than an edge case. System robustness, therefore, emerges from redundancy, diversity of signals, and controlled interaction between components rather than from maximizing the performance of individual classifiers.

This design philosophy aligns with established principles in safety-critical system engineering, where defense-in-depth strategies are employed to mitigate low-probability, high-impact failures. Applied to fraud detection, this approach emphasizes resilience under attack, graceful degradation, and the ability to surface partial confidence rather than binary outcomes. These properties are particularly relevant in financial contexts, where conservative decision-making under uncertainty is often preferable to brittle optimization.

3.2. Physiological Signal Analysis via Remote Photoplethysmography

Remote photoplethysmography (rPPG) exploits the interaction between light and human skin to estimate cardiovascular activity from video streams [23]. Subtle, periodic variations in skin chrominance correspond to blood volume changes associated with the cardiac cycle and are difficult to reproduce

coherently through generative models.

Within the proposed framework, rPPG serves as a physiological ground-truth signal, independent of visual realism [24,25]. The system continuously extracts rPPG features from facial regions of interest and evaluates their temporal consistency, signal-to-noise ratio, and coherence with observed facial motion.

Unlike traditional liveness detection techniques, rPPG-based analysis does not rely on user interaction or prompted gestures. This characteristic is particularly relevant in injection attack scenarios, where synthetic video streams may perfectly simulate voluntary movements but fail to reproduce physiologically plausible pulse dynamics over time.

Anomalies in rPPG signals are not treated as definitive evidence of fraud but as high-confidence risk indicators contributing to the overall trust assessment.

Unlike appearance-based biometric features, rPPG signals are rooted in physiological processes governed by cardiovascular dynamics and light–tissue interaction. While generative models can approximate surface-level visual realism, reproducing temporally stable and physiologically plausible pulse signals over extended durations remains a non-trivial challenge. This asymmetry introduces a form of implicit liveness that persists even when voluntary user actions are perfectly simulated.

Importantly, the proposed framework does not assume rPPG extraction to be infallible. Variability in illumination, camera quality, and skin tone can affect signal quality, particularly in uncontrolled consumer environments. For this reason, rPPG-derived features are evaluated probabilistically and contextualized within the broader trust assessment pipeline, rather than being used as deterministic accept–reject criteria.

3.3. Continuous Behavioral Biometrics as Persistent Authentication

Behavioral biometrics constitute the second defensive layer and address a key limitation of static identity verification: the assumption that trust, once established, remains valid throughout the session.

In real-world banking applications, sessions may persist for extended periods and involve multiple sensitive operations. Behavioral biometrics enable continuous authentication by modeling user-specific interaction patterns [26], including:

- Keystroke timing and rhythm;
- Touch dynamics and pressure profiles;
- Device orientation and micro-movement patterns;
- Temporal correlations between interaction events [27,28].

These signals are inherently dynamic and context-dependent, making them difficult to replicate through deepfake-based impersonation or automated control. Importantly, behavioral models adapt over time, allowing the system to account for natural intra-user variability while remaining sensitive to abrupt deviations indicative of fraud.

From a privacy and compliance standpoint, behavioral features are processed locally where possible, and only abstracted risk scores are transmitted for centralized decision-making.

Behavioral biometrics offer a temporal dimension of security that is absent from static identity verification mechanisms. By continuously observing interaction patterns throughout a session, the system can detect deviations that arise after initial authentication, including those associated with automated control, credential sharing, or delayed impersonation attempts.

A critical advantage of behavioral signals lies in their resistance to direct replay. While synthetic media can replicate visual or auditory traits, accurately emulating fine-grained interaction dynamics—such as keystroke latency distributions or micro-variations in touch pressure—requires precise real-time control and individualized modeling. This increases attacker complexity and reduces the feasibility of scalable impersonation.

3.4. Adversarial Neural Networks and Challenger Models

The third layer of the framework consists of adversarially trained neural networks, referred to as challenger models.

At the implementation level, these models can be instantiated using convolutional or transformer-based architectures, depending on the modality of the input data, and trained on both authentic and synthetically generated samples to improve generalization under adversarial conditions.

These models are explicitly designed to operate under the assumption that attackers continuously improve the realism of generative content.

Challenger models are trained on a diverse and evolving

corpus of synthetic artifacts and focus on identifying subtle inconsistencies introduced during generation, encoding, or real-time manipulation[29]. These include:

- Non-physical lighting behavior in facial regions;
- Irregularities in eye reflections and iris texture;
- Spectral artifacts in compressed or synthesized audio;
- Temporal discontinuities introduced during stream injection[30].

Rather than producing binary classifications, challenger models output calibrated risk estimates. This probabilistic approach reduces brittleness and supports downstream fusion with physiological and behavioral signals.

The role of challenger models within the proposed architecture is deliberately constrained. Rather than serving as authoritative arbiters of authenticity, they function as continuously evolving probes that test the system's exposure to emerging generative artifacts. This positioning mitigates the risk of overfitting detection logic to transient attack characteristics.

By decoupling challenger outputs from final decision authority, the framework accommodates uncertainty and model disagreement. This design choice reflects the empirical observation that adversarial detection models are most effective when integrated as one component within a broader evidentiary framework, rather than when deployed in isolation.

3.5. Signal Fusion and Trust Scoring

The outputs of the three detection layers are combined using a late-fusion strategy, producing a composite trust score that reflects the cumulative evidence available at a given point in time.

Late fusion is preferred over early fusion to preserve interpretability and allow independent evaluation of each signal source. Each layer contributes explainable indicators that can be translated into reason codes, such as:

- Physiological signal incoherence;
- Behavioral deviation beyond adaptive thresholds;
- Detection of generative artifacts.

In practice, each layer contributes an independent risk score that is normalized and aggregated according to predefined policy thresholds. This enables flexible calibration across different operational scenarios, such as onboarding, transaction authorization, and account recovery.

The resulting trust score is continuously updated throughout the session, enabling dynamic risk management and graduated responses, ranging from passive monitoring to step-up authentication or transaction blocking.

Late-fusion strategies are particularly well suited to regulated environments, as they preserve the semantic independence of individual signal sources. Each contributing layer can be evaluated, audited, and adjusted independently, reducing coupling between detection components and simplifying governance processes.

From an operational standpoint, continuous trust scoring enables proportional responses to uncertainty. Rather than forcing binary outcomes, the system can escalate controls gradually, balancing fraud prevention objectives against user experience considerations.

3.6. Evaluation Criteria

The framework is evaluated using robustness-oriented criteria, including:

- Resistance to adversarial attacks[3];
- Detection latency in real-time conditions;
- Reduction of single-point failure risks[31];
- Consistency across independent detection layers;
- Trade-offs between false positives and false negatives.

These metrics reflect operational requirements in high-risk financial environments. Although quantitative benchmarking is not performed, these criteria provide a structured basis for assessing robustness under realistic operational constraints and support future empirical validation.

4. Robustness Analysis

Direct empirical evaluation of fraud detection mechanisms in live digital banking environments presents significant practical, legal, and ethical constraints. Access to representative attack data is inherently limited, while controlled experimentation on production systems is typically infeasible. For this reason, the evaluation is conducted at a robustness and architectural level, consistent with risk-oriented assessment approaches discussed in adversarial machine learning and AI governance literature[12,19].

Rather than measuring benchmark accuracy, the analysis examines the system's resilience properties under plausible

adversarial strategies, focusing on failure-mode independence, signal heterogeneity, and attacker workload escalation. This evaluation perspective reflects the practical constraints of real-world financial systems, where access to fully representative attack data is limited and controlled experimentation on production environments is neither feasible nor ethically acceptable. Therefore, robustness must be inferred from architectural properties, threat coverage, and failure tolerance rather than from isolated performance metrics.

In this context, the primary question is not whether a given detection component can achieve high accuracy under laboratory conditions, but whether the system as a whole can maintain acceptable risk levels under sustained adversarial pressure. This shift in evaluative focus aligns with risk-based frameworks commonly adopted in financial regulation and operational risk management.

The analysis indicates that:

- Attacks capable of bypassing one detection layer are unlikely to evade all layers simultaneously;
- Physiological signals provide strong resistance against injection attacks;
- Behavioral biometrics enhance detection latency by identifying anomalies early in the session;
- Adversarial models adapt more effectively when integrated as challengers rather than primary decision-makers.

Overall, the framework demonstrates increased resilience to adaptive attacks compared to single-layer or static verification approaches.

An important observation emerging from this analysis is that robustness arises from the interaction between layers rather than from their individual strength. While each detection mechanism exhibits known limitations when considered in isolation, their combination constrains the adversary's feasible attack space. Successful evasion, therefore, requires the simultaneous satisfaction of heterogeneous constraints, significantly increasing attack complexity.

This property is particularly relevant in adaptive threat scenarios, where attackers iteratively probe system boundaries. The presence of multiple, partially independent detection layers reduces the informational value of individual probing attempts, slowing adversarial learning and diminishing the effectiveness of feedback-driven optimization.

Table 2 summarizes the relationship between attack vectors, exploited weaknesses, and the corresponding defensive layers of the proposed system.

Table 2. Attack Vectors and Corresponding Defensive Layers.

Attack Vector	Exploited Weakness	Detection Layer
Photo/Video Replay	Static biometric matching	Challenger models
Deepfake Video Injection	Sensor trust assumption	rPPG analysis
Real-Time Face Synthesis	Motion-based liveness	rPPG + adversarial
Session Hijacking	One-time authentication	Behavioral biometrics
Automated Control	Lack of interaction modeling	Behavioral biometrics
Model Adaptation	Static classifiers	Challenger models

Beyond its descriptive function, this mapping highlights an asymmetry between attacker capabilities and defensive requirements. While attackers may selectively exploit individual weaknesses, defenders must account for the full spectrum of plausible attack vectors. The proposed framework addresses this asymmetry by distributing defensive responsibility across layers that target distinct aspects of the attack surface.

5. Figures, Schemes and System Representation

In real-world AI systems, architectural transparency plays a critical role not only for engineering validation but also for regulatory scrutiny. Visual representations of system logic enable stakeholders - including auditors, regulators, and risk officers - to assess how heterogeneous signals are combined and how decisions emerge from complex AI-driven pipelines.

5.1. System Architecture Representation

Figure 1 depicts the overall architecture of the proposed real-time fraud detection framework. The system is structured around three parallel analytical layers—physiological, behavioral, and adversarial—each operating independently on distinct data modalities. This architectural design reduces correlated failure risks by ensuring independence between detection layers and supports explainable, audit-ready decision processes.

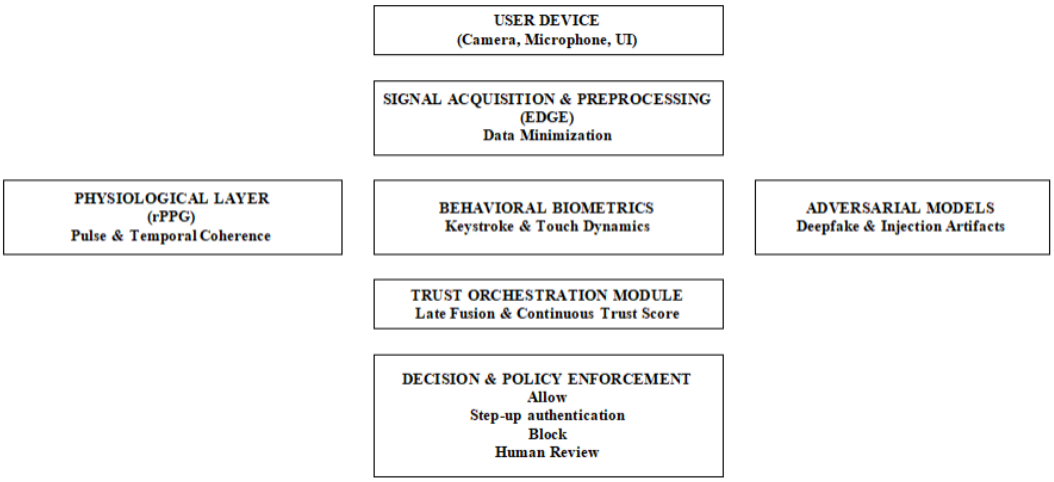

Figure 1. System architecture real-time fraud detection framework.

These layers feed into a centralized trust orchestration module responsible for signal fusion, risk scoring, and decision management.

The architecture explicitly separates:

- Signal acquisition and preprocessing, performed as close as possible to the user device;
- Feature extraction and local inference, executed under strict latency constraints;
- Risk aggregation and policy enforcement, managed by the banking backend.

This separation supports scalability, privacy-by-design, and regulatory compliance, while reducing systemic coupling between detection components.

Beyond its illustrative function, **Figure 1** makes explicit the trust assumptions embedded in the proposed architecture. In particular, the parallel organization of the analytical layers reflects the assumption that no single signal source can be considered authoritative in adversarial environments. Physiological signals, behavioral dynamics, and adversarial artifacts are treated as complementary and partially independent sources of evidence, reducing the risk of correlated failure modes.

The centralized trust orchestration module shown in **Figure 1** should be interpreted as a coordination layer rather than as a monolithic decision engine. Its role is to aggregate heterogeneous risk indicators under policy-defined rules, enabling continuous trust assessment and proportional responses without coupling detection logic to enforcement mechanisms.

From a governance perspective, the architectural decomposition visualized in **Figure 1** supports auditability and accountability by clearly separating data acquisition, local inference, risk aggregation, and policy enforcement. This separation facilitates independent inspection of system components and aligns with regulatory expectations for high-risk AI systems deployed in financial environments.

5.2. Threat Evolution Representation

Figure 2 illustrates the evolution of identity fraud threats in digital banking, from traditional presentation attacks to modern deepfake-driven injection attacks. The figure highlights how the attack surface has shifted from the physical interface (camera, microphone) to the software and data pipeline, invalidating assumptions underpinning classical liveness detection methods. This shift demonstrates the inadequacy of traditional liveness-based approaches and motivates the adoption of continuous, system-level trust assessment strategies.

By juxtaposing attack evolution with corresponding defensive strategies, the figure emphasizes the necessity of moving from static verification to continuous, system-level trust assessment.

Figure 2 highlights that the progression from presentation attacks to deepfake-driven injection attacks constitutes a shift in the underlying threat model rather than a gradual increase in attack sophistication. When synthetic media is injected directly into the software pipeline, the physical sensor ceases to function as a trust boundary, invalidating the assumptions on which classical liveness detection mechanisms are based.

By explicitly visualizing the displacement of the attack surface from the physical interface to the data and software

pipeline, **Figure 2** clarifies why defenses anchored to perceptual cues or short-term motion analysis are insufficient in modern digital banking environments. This representation motivates the adoption of continuous, system-level trust assessment strategies that operate across the entire session lifecycle rather than at a single verification point.

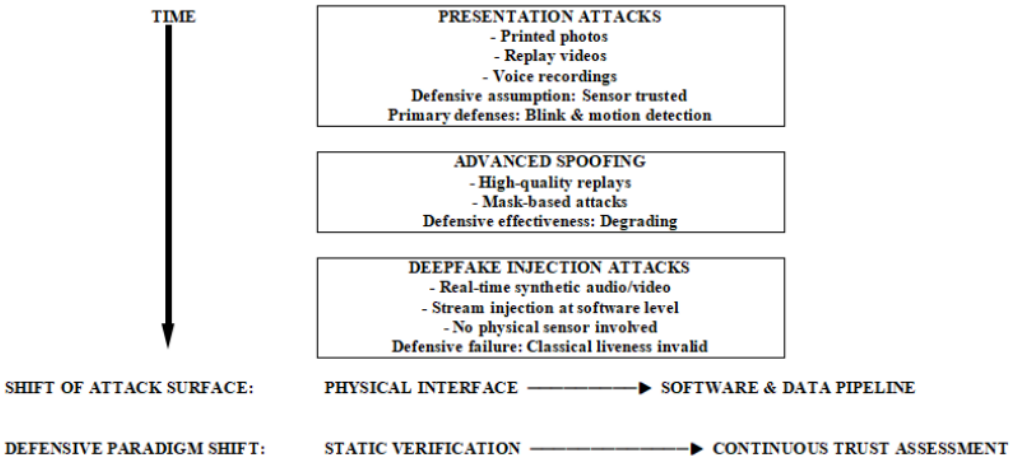

Figure 2. Evolution of identity fraud threats in digital banking.

6. Discussion

The proposed framework can be integrated into existing digital banking infrastructures, particularly within remote onboarding and continuous authentication processes. Its multi-layered design supports risk-based decision-making, enabling dynamic responses such as step-up authentication or transaction blocking, while maintaining compliance with regulatory requirements.

6.1. Deepfake Fraud as a Systemic Risk

The primary contribution of this study resides in the reconceptualization of deepfake-related fraud detection as a robustness engineering problem rather than a purely discriminative modeling task. This distinction has important implications for real-world financial systems, where adversarial adaptation, latency constraints, and regulatory obligations limit the applicability of detection-centric paradigms.

The results of this study reinforce the notion that deepfake-enabled fraud should be treated as a systemic risk, rather than as a sequence of isolated attack events. The increasing realism and accessibility of generative models enable attackers to scale identity fraud operations while continuously adapting to detection mechanisms.

In this context, robustness cannot be achieved through incremental improvements to individual classifiers. Instead, it must emerge from architectural design choices that assume persistent adversarial pressure and partial signal compromise.

The proposed framework embodies this perspective by ensuring that no single modality—visual, physiological, or behavioral—is sufficient on its own to establish trust.

From an institutional perspective, treating deepfake-enabled fraud as a systemic risk implies a shift in governance priorities. Rather than optimizing individual controls, organizations must focus on the resilience of end-to-end processes, including escalation paths, human oversight, and incident response mechanisms. Technical detection thus becomes one component within a broader socio-technical defense strategy.

This systemic framing also has implications for accountability. When trust is established through the aggregation of multiple probabilistic signals, responsibility for outcomes cannot be attributed to a single model or threshold. Instead, accountability resides in architectural design choices and risk management policies.

6.2. From Identity Verification to Continuous Trust Assessment

A key conceptual contribution of this work is the transition from one-time identity verification to continuous trust assessment. Traditional digital onboarding workflows implicitly assume that once a user has been authenticated, subsequent actions can be trusted until session termination.

This assumption no longer holds in adversarial environments where session hijacking, stream injection, and real-time impersonation are feasible. Continuous assessment enables the system to detect anomalies that emerge after initial authentication, thereby reducing dwell time and limiting potential damage.

From an operational standpoint, this paradigm aligns more closely with risk-based approaches mandated by financial regulation, where controls are expected to adapt dynamically to evolving risk profiles.

Continuous trust assessment introduces a temporal dimension to identity that is absent from traditional verification paradigms. Identity is no longer treated as a static attribute confirmed at a single point in time, but as an evolving hypothesis that must be continuously supported by behavioral and physiological evidence.

This temporal perspective enables earlier detection of anomalous behavior and reduces the dwell time available to attackers. In high-value financial contexts, even modest reductions in dwell time can significantly limit potential losses and downstream exposure.

6.3. Explainability and Governance in High-Risk AI Systems

Regulatory frameworks increasingly emphasize that AI systems deployed in sensitive domains must be explainable, auditable, and accountable. In the context of fraud detection, opaque decisions can undermine trust, complicate incident response, and expose institutions to regulatory sanctions [32].

Table 3 maps each detection layer of the proposed framework to key regulatory requirements applicable to high-risk AI systems in digital banking.

Table 3. Mapping Detection Layers to Regulatory and Governance Requirements.

Regulatory Requirement	rPPG Layer	Behavioral Biometrics	Challenger Models
Risk-Based Approach (AML)	✓	✓	✓
Explainability (XAI)	✓ (physiological coherence)	✓ (behavioral deviation)	✓ (artifact detection)
Data Minimization (GDPR)	✓ (edge processing)	✓ (local modeling)	✓ (feature abstraction)
High-Risk AI Compliance	✓	✓	✓
Auditability	✓	✓	✓
Human Oversight Support	✓	✓	✓

The proposed framework addresses these concerns by design. Each detection layer generates interpretable indicators that can be mapped to human-readable reason codes, such as:

- Physiological signal incoherence;
- Behavioral deviation beyond learned baselines;
- Detection of synthetic artifact patterns.

These reason codes support internal audits, customer dispute resolution, and regulatory inspections, bridging the gap between complex AI inference and governance requirements.

6.4. Privacy, Data Minimization, and Edge Processing

Biometric data processing raises significant privacy concerns, particularly under data protection regulations that mandate minimization and purpose limitation. Centralized storage of raw biometric data amplifies the impact of potential data breaches and increases compliance risk.

To mitigate these concerns, the framework prioritizes edge-based processing, whereby raw biometric signals are analyzed locally on the user's device. Only aggregated risk indicators or trust scores are transmitted to backend systems, reducing exposure of sensitive data.

This design choice aligns with privacy-by-design principles and supports compliance with data protection obligations without sacrificing detection capability.

6.5. Limitations and Practical Constraints

While the proposed framework enhances robustness, it is not without limitations. Physiological signals such as rPPG may be affected by environmental conditions, including lighting variability and camera quality. Behavioral biometrics may exhibit drift over time, requiring adaptive models and periodic recalibration.

Moreover, adversarial models depend on continuous access to updated threat intelligence and synthetic attack data. Maintaining such datasets introduces operational complexity and requires ongoing investment.

These limitations highlight that robustness is not a static property but a continuous process requiring monitoring, adap-

tation, and governance.

This observation underscores the importance of institutional commitment to long-term system stewardship. Robustness cannot be "deployed" once and forgotten; it must be maintained through periodic review, model updates, and governance processes that incorporate emerging threat intelligence.

Future work should therefore consider not only technical enhancements, but also organizational mechanisms that support sustained robustness, including cross-functional collaboration between security, compliance, and data science teams.

Future empirical validation on real-world datasets remains necessary to confirm the effectiveness of the framework under operational deployment conditions.

7. Conclusions

This work demonstrates that effective real-time fraud detection in digital banking cannot rely on static biometric checks or isolated deepfake classifiers. Instead, robustness must be engineered at the system level through the fusion of heterogeneous, partially independent signals and continuous trust assessment.

The proposed multi-layered framework provides a practical and regulation-aware blueprint for addressing deepfake-driven fraud in real-world banking environments. By integrating physiological sensing, behavioral biometrics, and adversarial learning within a governance-conscious architecture, the system raises the cost and complexity of successful attacks while preserving usability and compliance.

Beyond its immediate technical contributions, this work underscores the necessity of interdisciplinary collaboration in addressing emerging fraud threats. Effective system design at the intersection of artificial intelligence, security, and regulation requires continuous dialogue between engineers, risk managers, legal experts, and policymakers.

As generative technologies continue to evolve, the challenge facing financial institutions will not be to eliminate fraud entirely, but to manage it within acceptable risk thresholds while maintaining transparency and accountability. System-level robustness, rather than isolated detection accuracy, will therefore remain the defining criterion of success.

Future research directions include:

- The development of standardized robustness stress tests for deepfake resilience in financial AI systems;
- Longitudinal studies on behavioral drift and adaptive trust thresholds;
- Formal methods for incorporating robustness metrics into regulatory certification processes for high-risk AI systems;
- Exploration of cross-institutional threat intelligence sharing mechanisms to improve adversarial model training.

As generative technologies continue to evolve, maintaining trust in digital financial systems will increasingly depend on the ability to design AI systems that are not only accurate, but resilient, transparent, and governable by design.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

No new data were created or analyzed in this study. Data sharing does not apply to this article.

Conflicts of Interest

The author declares no conflict of interest.

AI Use Statement

During the preparation of this work, the author used AI-based tools to support language refinement, structural editing, and stylistic consistency of the manuscript. After using these tools, the author critically reviewed, edited, and validated all content, and takes full responsibility for the scientific accuracy, originality, and integrity of the published work. No AI system was used to generate scientific claims, research results, or interpretations. The author takes full responsibility for the

final content of the published article.

References

[1] Qureshi, S.M., Saeed, A., Almotiri, S.H., et al., 2024. Deepfake forensics: A survey of digital forensic methods for multimodal deepfake identification on social media. PeerJ Computer Science. 10, e2037. DOI: https://doi.org/10.7717/peerj-cs.2037

[2] Kaur, A., Hoshyar, A.N., Saikrishna, V., et al., 2024. Deepfake video detection: Challenges and opportunities. Artificial Intelligence Review. 57, 159. DOI: https://doi.org/10.1007/s10462-024-10810-6

[3] Khan, A.A., Laghari, A.A., Inam, S.A., et al., 2025. A survey on multimedia-enabled deepfake detection: State-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions. Discover Computing. 28, 48. DOI: https://doi.org/10.1007/s10791-025-09550-0

[4] Arner, D.W., Barberis, J., Buckley, R.P., 2017. FinTech, RegTech, and the reconceptualization of financial regulation. Northwestern Journal of International Law & Business. 37(3), 371–413.

[5] Basel Committee on Banking Supervision, 2024. Principles for the Sound Management of of Third-Party Risk. Bank for International Settlements (BIS): Basel, Switzerland. Available from: https://www.bis.org/bcbs/publ/d577.pdf

[6] Korshunov, P., Marcel, S., 2018. DeepFakes: A New Threat to Face Recognition? Assessment and Detection. arXiv preprint. arXiv:1812.08685. DOI: https://doi.org/10.48550/arXiv.1812.08685

[7] Verdoliva, L., 2020. Media forensics and deepfakes: An overview. IEEE Journal of Selected Topics in Signal Processing. 14(5), 910–932. DOI: https://doi.org/10.1109/JSTSP.2020.3002101

[8] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., et al., 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion. 64, 131–148. DOI: https://doi.org/10.1016/j.inffus.2020.06.014

[9] Matern, F., Riess, C., Stamminger, M., 2019. Exploiting visual artifacts to expose deepfakes and face manipulations. In Proceedings of the IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 7–11 January 2019; pp. 83–92. DOI: https://doi.org/10.1109/WACVW.2019.00020

[10] Szegedy, C., Zaremba, W., Sutskever, I., et al., 2014. Intriguing properties of neural networks. International Conference on Learning Representations (ICLR). arXiv preprint. arXiv:1312.6199. DOI: https://arxiv.org/abs/1312.6199

[11] Goodfellow, I., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR). arXiv preprint. arXiv:1412.6572. DOI: https://arxiv.org/abs/1412.6572

[12] Biggio, B., Roli, F., 2018. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition. 84, 317–331. DOI: https://doi.org/10.1016/j.patcog.2018.07.023

[13] Galbally, J., Marcel, S., Fierrez, J., 2014. Biometric anti-spoofing methods: A survey in Face. IEEE Access. 2, 1530–1552. DOI: https://doi.org/10.1109/ACCESS.2014.2381273

[14] ISO/IEC 30107-3:2023. 2023. Information Technology—Biometric Presentation Attack Detection—Part 3: Testing and Reporting.

[15] The European Parliament, the Council of the European Union, 2018. Directive (EU) 2018/843 of the European Parliament and of the Council of 30 May 2018 Amending Directive (EU) 2015/849 on the Prevention of the Use of the Financial System for the Purposes of Money Laundering or Terrorist Financing, and Amending Directives 2009/138/EC and 2013/36/EU (Text with EEA Relevance). Official Journal of the European Union: Luxembourg, Luxembourg. Available from: http://data.europa.eu/eli/dir/2018/843/oj

[16] European Banking Authority, 2022. Final Report: Guidelines on the Use of Remote Customer Onboarding Solutions under Article 13(1) of Directive (EU) 2015/849. EBA: Paris, France.

[17] The European Parliament, the Council of the European Union, 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance). Official Journal of the European Union: Luxembourg, Luxembourg. Available from: http://data.europa.eu/eli/reg/2016/679/oj

[18] The European Parliament, the Council of the European Union, 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA Relevance). Official Journal of the European Union: Luxembourg, Luxembourg. Available from: http://data.europa.eu/eli/reg/2024/1689/oj

[19] National Institute of Standards and Technology, 2023. The EU Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology. NIST: Gaithersburg, MD, USA.

[20] Carlini, N., Wagner, D., 2017. Towards evaluating the robustness of neural networks. In Proceedings of the

IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–26 May 2017; pp. 39–57.

[21] Salvi, D., Liu, H., Mandelli, S., et al., 2023. A Robust Approach to Multimodal Deepfake Detection. Journal of Imaging. 9(6), 122. DOI: https://doi.org/10.3390/jimaging9060122

[22] Ramanaharan, R., Guruge, D.B., Agbinya, J.I., 2025. DeepFake video detection: Insights into model generalisation—A systematic review. Data and Information Management. 9(3), 100099. DOI: https://doi.org/10.1016/j.dim.2025.100099

[23] de Haan, G., Jeanne, V., 2013. Robust pulse rate from chrominance-based rPPG. IEEE Transactions on Biomedical Engineering. 60(10), 2878–2886.

[24] Nowara, E.M., Sabharwal, A., Veeraraghavan, A., 2017. PPGSecure: Biometric presentation attack detection using photoplethysmograms. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017.

[25] Hernandez-Ortega, J., Tolosana, R., Fierrez, J., et al., 2020. DeepFakesON-Phys: DeepFakes Detection based on Heart Rate Estimation. arXiv preprint. arXiv:2010.00400. DOI: https://doi.org/10.48550/arXiv.2010.00400

[26] Drozdowski, P., Rathgeb, C., Dantcheva, A., et al., 2020. Demographic Bias in Biometrics: A Survey on an Emerging Challenge. IEEE Transactions on Technology and Society. 1(2), 89–103.

[27] Upadhyaya, S.J., 2017. Continuous authentication using behavioral biometrics. In IWSPA '17: Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics. Association for Computing Machinery: New York, NY, USA. DOI: https://doi.org/10.1145/3041008.3041019

[28] Acien, A., Morales, A., Vera-Rodriguez, R., et al., 2020. Mobile Active Authentication based on Multiple Biometric and Behavioral Patterns. In: Bourlai, T., Karampelas, P., Patel, V.M. (Eds.). Securing Social Identity in Mobile Platforms. Advanced Sciences and Technologies for Security Applications. Springer: Cham, Switzerland. DOI: https://doi.org/10.1007/978-3-030-39489-9_9

[29] Li, Y., Chang, M.-C., Lyu, S., 2018. In Ictu Oculi: Exposing AI-generated fake face videos by detecting eye blinking. arXiv preprint. arXiv:1806.02877. DOI: https://doi.org/10.48550/arXiv.1806.02877

[30] Agarwal, S., Farid, H., Gu, Y., et al., 2019. Protecting world leaders against deepfakes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–20 June 2019.

[31] Dal Pozzolo, A., Bontempi, G., Snoeck, M., et al., 2018. Adversarial drift detection in financial fraud. In DEC '23: Proceedings of the Second ACM Data Economy Workshop. Association for Computing Machinery: New York, NY, USA. DOI: https://doi.org/10.1145/3600046.3600051

[32] Raghavan, M., Barocas, S., Kleinberg, J., et al., 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT), Barcelona, Spain, 27–30 January 2020; pp. 469–481. DOI: https://doi.org/10.1145/3351095.3372828