Know Your Customer (KYC) in the Banking Industry: A Comprehensive Study



Simão Pedro Custódio Pereira
Master in Information Security
Department of Computer Science
Faculty of Sciences of the University of Porto
2026

Know Your Customer (KYC) in the Banking Industry: A Comprehensive Study

Simão Pedro Custódio Pereira

Dissertation carried out as part of the
Master in Information Security
Department of Computer Science
2026

Supervisor
Prof. João Miguel Maia Soares de Resende, Assistant Professor, University of Porto

Acknowledgements

As this project reaches its conclusion, I want to acknowledge the support that made this research possible. The past few months have been defined by KYC algorithms, smart contracts, and extensive debugging, and I did not navigate this process entirely alone.

First, I thank my supervisor, João Resende. I appreciate your guidance as the scope of this project evolved. Thank you for critically evaluating my work when it lacked rigor and for forcing me to find my own solutions rather than providing them for me. That approach directly improved the quality of this thesis.

I am grateful to my parents for their understanding. Thank you for accommodating my schedule during extended coding stretches and for supporting me without applying pressure regarding a completion date.

To my girlfriend, thank you for your enduring patience. I appreciate your willingness to listen to unsolicited explanations of zero-knowledge identity verification and for your support during the evenings I was focused on unresolved bugs.

Thank you to my friends for providing necessary breaks from the work. I appreciate the regular check-ins and your tolerance for listening to endless discussions about KYC protocols.

Completing this thesis required independent focus, but it also relied on technical conversations that unblocked ideas and reminders to step away from the terminal. I am proud to present this research, recognizing that it is the result of both my own effort and your essential support.

Resumo

Esta tese aborda o problema do conhecimento remoto do cliente (*Know Your Customer*, KYC) no setor bancário a partir de duas perspetivas complementares: uma avaliação empírica da resistência dos mecanismos atuais de biometria e deteção de ataques de apresentação (PAD, do inglês *Presentation Attack Detection*) face a conteúdos sintéticos gerados por inteligência artificial (IA), e a proposta de uma arquitetura assente em blockchain para suportar auditabilidade, confiança entre instituições e reutilização controlada de identidade, adequada ao enquadramento legal português.

No plano experimental, cinco ferramentas de código aberto (DeepFace FASNet; DeepFace rPPG, que recorre à fotopletismografia remota, rPPG; a implementação do algoritmo *Plane-Orthogonal-to-Skin*, POS, no rPPG-Toolbox; InsightFace ArcFace; e Silent-Face Anti-Spoofing) foram submetidas a uma metodologia de *bypass* controlada, utilizando duas famílias de vídeo sintético. A primeira família compreendeu três vídeos gerados pela plataforma Hedra (motor Google Veo 3, geração *audio-driven* para sincronização labial), abrangendo duas identidades de origem distintas. Os detetores baseados em textura e em domínio da frequência aceitaram os três vídeos como reais em $\geq 96\%$ dos fotogramas amostrados; apenas a deteção de vida biológica por rPPG classificou corretamente os três vídeos como *spoof* (pontuações de *peakiness* de 1.23/1.36, 1.59 e 1.27, todas abaixo do limiar de 1.70). A segunda família compreendeu quatro vídeos gerados diretamente através da interface Google Veo 3 a partir de uma fotografia estática (sem restrição de áudio). As quatro verificações de vida, incluindo ambas as *pipelines* rPPG, emitiram veredicto REAL em todos os vídeos (*peakiness* de 1.92–2.55, acima do limiar de 1.70), e o InsightFace detetou um rosto em todos os fotogramas, demonstrando um *bypass* completo de toda a *pipeline*. A conclusão conjunta é que a resiliência dos sistemas PAD depende da *pipeline* de geração completa e não apenas do modelo subjacente.

No plano arquitetural, foi implementado um fluxo de *onboarding* em quatro passos, exposto como serviço web e suportado por três contratos inteligentes em Solidity 0.8.20 (`DIDRegistry`, `VCRegistry` e `AuditTrail` apenas-anexação), em conjunto com módulos de IA *off-chain* para autenticação de documentos (validação da zona de leitura ótica de passaportes e cartões de identificação segundo o Doc 9303 da Organização da Aviação Civil Internacional, ICAO, incluindo o Cartão de Cidadão português), correspondência facial 1:1 com ArcFace, deteção de vida por rPPG-POS e classificação de risco baseada em regras. A arquitetura mantém na cadeia apenas *hashes* `keccak256`, deixando

os dados pessoais em armazenamento cifrado *off-chain*, e disponibiliza, através de um Registo de Confiança global, uma camada técnica para o recurso a terceiros previsto no Artigo 41.º da Lei n.º 83/2017, que continua a exigir termos acordados entre as instituições e a responsabilidade da instituição que recorre a terceiros. Sete execuções de validação cobrem o conjunto completo de casos de aceitação e rejeição, incluindo uma aprovação *end-to-end* com emissão de Credencial Verificável em cadeia e uma verificação de evidência de adulteração independente.

A tese conclui que o KYC remoto seguro não pode depender isoladamente nem de controlos biométricos nem de mecanismos de confiança em blockchain, e defende uma estratégia de defesa-em-profundidade, combinando vida biológica com uma arquitetura que mantém os dados pessoais *off-chain* e evidências *on-chain*, como resposta ao atual ambiente de ameaças com IA generativa.

Palavras-chave: KYC remoto, onboarding bancário, deteção de vida, deteção de ataques de apresentação, rPPG, deepfakes, IA generativa, blockchain, identidade auto-soberana, credenciais verificáveis

Abstract

This thesis examines secure remote Know Your Customer (KYC) in the banking sector from two connected perspectives: an empirical evaluation of how current biometric and Presentation Attack Detection (PAD) mechanisms cope with contemporary synthetic media generated by artificial intelligence (AI), and an architectural proposal for a blockchain-enabled trust, audit, and reusable-identity layer suited to the Portuguese banking sector.

On the empirical side, five open-source verification components (DeepFace FASNet; DeepFace rPPG, which uses remote photoplethysmography, rPPG; the rPPG-Toolbox implementation of the Plane-Orthogonal-to-Skin, POS, algorithm; InsightFace ArcFace; and Silent-Face Anti-Spoofing) were subjected to a controlled bypass methodology using two families of synthetic video. The first family comprised three videos generated via the Hedra platform (Google Veo 3, audio-driven for lip-sync), spanning two distinct source identities. Texture-based and frequency-domain detectors accepted all three as real on $\geq 96\%$ of sampled frames; only rPPG-based biological liveness correctly classified them as spoofs (peakiness 1.23/1.36, 1.59, and 1.27, all below the 1.70 threshold), exemplifying the *Generalisation Failure* of PAD layers built on convolutional neural networks against audio-constrained generative content. The second family comprised four videos generated directly via Google's Veo 3 interface from a static photograph (no audio constraint). All four liveness checks, including both rPPG pipelines, issued a REAL verdict on every video (peakiness 1.92–2.55, above the 1.70 threshold), and InsightFace found a face in every frame, demonstrating a full pipeline bypass. The combined finding is that PAD resilience depends on the complete generation pipeline rather than the underlying model alone: the same generative model yields detectable and undetectable synthetic output depending on how it is driven.

On the architectural side, a four-step onboarding flow was implemented end-to-end as a web service against three Solidity 0.8.20 smart contracts (`DIDRegistry`, `VCRegistry`, and an append-only `AuditTrail`), combined with off-chain AI modules for document authentication (machine-readable zone validation for passports and identity cards under Doc 9303 of the International Civil Aviation Organization, ICAO, including the Portuguese Cartão de Cidadão), ArcFace 1:1 face matching, rPPG-POS liveness, and rule-based risk scoring. The architecture stores only `keccak256` hashes on chain, keeping all raw personally identifiable information (PII) off-chain, and provides, through a global Trust Registry, a technical layer for third-party reliance under Article 41 of Law n.º 83/2017; that reliance still

requires agreed terms between the institutions, and the relying bank remains responsible. Seven validation runs exercise the full taxonomy of acceptance and rejection cases, including a fully approved onboarding with on-chain Verifiable Credential issuance and an independently verifiable tamper-evidence check.

The thesis concludes that secure remote KYC cannot rely on either biometric controls or blockchain-based trust mechanisms in isolation, and argues for defence-in-depth that combines biological liveness with an off-chain-PII / on-chain-evidence architecture as a response to the current generative-AI threat environment.

Keywords: remote KYC, bank onboarding, liveness detection, presentation attack detection, rPPG, deepfakes, generative AI, blockchain, self-sovereign identity, verifiable credentials

Table of Contents

List of Tables xi
List of Figures xii
List of Abbreviations xiv

1. Introduction 1
   1.1. Motivation 2
   1.2. Goals and objectives 3
   1.3. Outline 4
2. Background and Technical Foundations 5
   2.1. Security Requirements and guidelines 5
   2.2. General Data Protection Regulation (GDPR) 6
   2.3. Electronic Identification and Trust Services (eIDAS 2.0) 7
   2.4. The EU AI Act and Biometrics 8
   2.5. Strong Customer Authentication (SCA) under PSD2 9
   2.6. Anti-Money Laundering Directives (AMLD) 11
      2.6.1. Key Provisions of AMLD 11
      2.6.2. Recent Enhancements in AMLD 13
   2.7. Cybersecurity Frameworks (NIS2 and DORA) 14
      2.7.1. NIS2 Directive 14
      2.7.2. Digital Operational Resilience Act (DORA) 14
   2.8. Record-Keeping and Audit Trails 15
   2.9. ISO 27001 Summary 17
   2.10. Blockchain Technology 18
      2.10.1. Core Characteristics 19
      2.10.2. Smart Contracts 19
   2.11. Decentralized Identity and SSI 19
      2.11.1. The Trust Triangle 19
      2.11.2. Verifiable Credentials (VCs) and DIDs 20
   2.12. Distributed Storage: IPFS 20
   2.13. Face Recognition Technology 20
      2.13.1. Deep Convolutional Neural Networks (CNNs) 21
      2.13.2. Matching and Loss Functions 21
      2.13.3. Presentation Attack Detection (PAD) 21
3. Related work 22
   3.1. Know Your Customer (KYC) and Anti-Money Laundering (AML) 22

3.2. AML Compliance in the European Union and Portugal 22
3.3. Technological Advancements in KYC 23
3.3.1. Document Authentication and Biometrics 23
3.3.2. Remote KYC and Digital Onboarding 24
3.4. Sanctions Screening and Transaction Monitoring 25
3.4.1. Sanctions List Checks 25
3.4.2. Transaction Monitoring and Source of Funds Verification 25
3.5. Cybersecurity and Identity Provider Challenges 26
3.5.1. Multi-Factor Authentication and Customer Security 26
3.5.2. Centralized and Decentralized Identity Architectures 26
3.5.3. KYC Bypass Attempts and Fraudulent Activities 27
3.6. Distributed Ledger Technology and Blockchain in KYC 27
3.7. Generative AI Threats to Identity Verification 28
3.7.1. Threats to KYC and Security 28
3.7.2. Evolution of the Identity Threat Landscape (2024 - 2026) 31
3.8. AI Generators and KYC Bypass Requirements 32
3.8.1. Comparison of AI Generative Architectures 32
3.8.1.1. Cloud-Based Probabilistic Models (Latent Diffusion & Transformers) 33
3.8.1.2. Local Deterministic Transformation Tools (Autoencoders) 33
3.8.2. Technical Quality Requirements for Bypass Experiments 34
3.8.3. Content Provenance and Watermarking as a Detection Paradigm 34
4. Algorithm Architecture and Selection 37
4.1. Tested Algorithms: Architecture and Analysis 37
4.1.1. DeepFace 37
4.1.1.1. Architecture and Pipeline 37
4.1.1.2. Strengths 38
4.1.1.3. Weaknesses and Limitations 39
4.1.2. InsightFace (buffalo_l) 39
4.1.2.1. Architecture and Pipeline 39
4.1.2.2. Strengths 40
4.1.2.3. Weaknesses and Limitations 41
4.1.3. MiniFASNet / Silent-Face Anti-Spoofing 41
4.1.3.1. Architecture 41
4.1.3.2. Strengths 42
4.1.3.3. Weaknesses and Limitations 42
4.1.4. rPPG-Toolbox and the POS Algorithm 43
4.1.4.1. POS Algorithm 43

4.1.4.2. Strengths 44
4.1.4.3. Weaknesses and Limitations 45
4.1.5. Hybrid Pipeline: DeepFace + rPPG 45
4.1.5.1. Strengths 45
4.1.5.2. Weaknesses and Limitations 46
4.2. Algorithm Comparison 46
5. Experimental Results 48
5.1. Experimental Setup and Methodology 48
5.2. Testing Methodology for KYC Bypass Experiments 48
5.2.1. Step-by-Step Testing Process 49
5.2.2. Mobile-Specific Vulnerabilities: Rooting and Hooking 50
5.2.3. Threat Model Scope and Attack Vectors Evaluated 50
5.2.4. Open-Source Frameworks for KYC Biometric Security Research . 52
5.3. Preliminary Experimental Results 54
5.3.1. Remote Photoplethysmography (rPPG) Stability Tests 54
5.3.2. DeepFaceLive Authentication and Spoofing Analysis 56
5.3.3. Comparative Presentation Attack Analysis: Hedra-Generated Synthetic Media 57
5.3.3.1. Choice of Generative Platform 57
5.3.3.2. MiniFASNet Passive Liveness Detection (DeepFace FASNet) 58
5.3.3.3. rPPG-Based Biological Liveness (DeepFace rPPG and rPPG-Toolbox) 60
5.3.3.4. InsightFace: Identity Matching Without Liveness 61
5.3.3.5. Frequency-Domain Passive PAD (Silent-Face Anti-Spoofing) 62
5.3.3.6. Comparative Summary and Attack Vector Analysis 63
5.3.3.7. Quantitative Detection Metrics: APCER, BPCER, and ACER 64
5.3.4. Cross-Platform Generalisation: Google Veo 3 Direct (Videos 5–8) 66
5.3.4.1. MiniFASNet Passive Liveness Detection (DeepFace FASNet) 67
5.3.4.2. rPPG-Based Biological Liveness (DeepFace rPPG and rPPG-Toolbox) 67
5.3.4.3. InsightFace: Face Detection on Synthetic Input 69
5.3.4.4. Frequency-Domain Passive PAD (Silent-Face Anti-Spoofing) 70
5.3.4.5. Full Bypass Summary and Cross-Generator Comparison 71
5.3.5. Discussion: Are There Detection Approaches That Work? 71
5.3.5.1. What Worked, and Against What 72
5.3.5.2. What About SynthID? 72
5.4. Blockchain Prototype Evaluation 73
5.4.1. Experimental Configuration 73

5.4.2. End-to-End Run Outcomes ... 74
5.4.3. Anchored On-Chain Events ... 75
5.4.4. Component Latency ... 75
5.4.5. Tamper-Evidence Guarantee ... 76
5.4.6. Discussion ... 77
6. A Blockchain-Anchored Architecture for Remote KYC Onboarding ... 79
6.1. Proposed Architecture ... 79
6.1.1. Prototype Implementation ... 79
6.1.1.1. Smart Contracts (Hardhat / Solidity 0.8.20) ... 79
6.1.1.2. Backend Service (FastAPI + web3.py) ... 80
6.1.1.3. Why Blockchain Rather Than a Centralized Signed Log? ... 81
6.1.2. Architectural Overview ... 81
6.1.3. Key Components ... 82
6.1.3.1. Frontend (User Interface) ... 82
6.1.3.2. Onboarding Microservices Layer ... 82
6.1.3.3. Verification Engine ... 83
6.1.3.4. Blockchain Integration Layer ... 84
6.1.3.5. Data Storage ... 85
6.1.3.6. Data Retention and Regulatory Archiving ... 86
6.1.3.7. Security and Compliance ... 86
6.1.4. The Reusable Identity Model ... 87
6.1.5. Selective Disclosure and Zero-Knowledge Proofs (ZKPs) ... 88
6.1.6. Credential Revocation and Lifecycle Management ... 89
6.1.7. Comparison: Traditional vs. Proposed Architecture ... 89
6.1.8. Differentiators: A Paradigm Shift in Identity Assurance ... 89
6.1.9. Workflow Overview ... 90
6.1.10. Empirical Validation of the Prototype Pipeline ... 92
6.1.11. Extended Verification Pipeline: Session, Temporal, Audio, and Contextual Modules ... 95
6.1.12. Compliance and Data Management ... 99
6.1.12.1. Information Updates and Deletions ... 100
6.1.12.2. Compliance with Regulations (GDPR and Beyond) ... 101
6.2. Current Landscape and Adoption ... 102
7. Conclusion ... 104
7.1. Summary of Contributions ... 104
7.2. Answering the Thesis Questions ... 105
7.3. Limitations ... 107
7.4. Future Work ... 107

7.5. Closing Remarks 108

Bibliography 109
Appendix A: Prototype Implementation 117

List of Tables

3.1. Defensive Benchmarks for Presentation Attack Detection (PAD) 29
3.2. Comparison of AI Generators and KYC Bypass Requirements 34

4.1. Computational complexity comparison of MobileFaceNet and MiniFASNet variants. 41
4.2. Comparative overview of the evaluated KYC algorithms. 46

5.1. Mapping of EBA Remote Onboarding Requirements (EBA/GL/2022/15) to Biometric Algorithm Clusters 52
5.2. Open-Source Tools Selected for KYC Biometric Security Testing 53
5.3. Preliminary Experimental Results: Liveness and Authentication Metrics 57
5.4. Bypass Capability: Video A (no glasses) vs. Video B (with glasses) vs. Video C (independent subject) 63
5.5. APCER per tool on the three Hedra/Veo 3 attack videos. Lower is better; an effective PAD layer should approach 0%. 65
5.6. APCER per tool on the Hedra/Veo 3 attack corpus (Videos A–C). BPCER and ACER are omitted: the bona fide sample is a single video clip ($n=1$), making any derived rate statistically uninformative. 65
5.7. Detection outcomes on Google Veo 3 direct synthetic videos (Videos 5–8). All five tools produce false acceptances; this is a full bypass of every detection layer, including rPPG. 71
5.8. rPPG peakiness across both generator families. Threshold $=$ 1.70. Hedra/Veo 3 output (Videos A–C) falls below and is correctly classified as synthetic; Google Veo 3 direct output (Videos 5–8) exceeds the threshold and is incorrectly classified as real. 71
5.9. Outcomes of the seven end-to-end runs against the prototype 74
5.10. On-chain audit events fired per run 75
5.11. Approximate wall-clock latency per onboarding step on the prototype (local Hardhat, CPU-only Python) 76

6.1. Comparison of traditional Banking KYC vs. Proposed Architecture 89
6.2. End-to-end validation runs of the four-step KYC pipeline. Cosine distance ≤ 0.40 is an ArcFace match; rPPG peakiness > 1.70 is a live capture. 92
6.3. On-chain vs. Off-chain Data Partitioning in the Prototype 100

List of Figures

5.1. rPPG Detection on Synthetic Media: 152 frames correctly identified with stable bounding boxes. 54
5.2. rPPG Detection Failure: 40 frames failed due to artifact-induced motion blur in the generative output. 55
5.3. DeepFace Bypass Success: 188 synthetic frames successfully identified as the target identity. 56
5.4. DeepFace Detection Anomalies: 4 frames where the swap artifacts triggered a safety flag. 56
5.5. FASNet on Video A (no glasses): sampled frames classified as REAL (30/39 frames, score > 0.5). 59
5.6. FASNet on Video B (with glasses): sampled frames classified as REAL (28/39 frames, score > 0.97). 59
5.7. FASNet on Video A (no glasses): only 1 frame flagged as SPOOF out of 39 samples (PASS verdict). 59
5.8. FASNet on Video B (with glasses): zero spoof frames detected across all 39 samples (PASS verdict). 59
5.9. FASNet on Video C (independent female subject): sampled frames classified as REAL, zero spoof flags out of 39 samples (PASS verdict). 59
5.10. DeepFace rPPG on Video A (no glasses): pulse signal absent, SPOOF verdict (peakiness 1.2253, HR 90.0 BPM). 60
5.11. DeepFace rPPG on Video B (with glasses): pulse signal absent, SPOOF verdict (peakiness 1.5898, HR 96.44 BPM). 60
5.12. rPPG-Toolbox on Video A (no glasses): SPOOF (peakiness 1.3585, HR 73.1 BPM). 60
5.13. rPPG-Toolbox on Video B (with glasses): SPOOF (peakiness 1.5898, HR 96.44 BPM). 60
5.14. DeepFace rPPG on Video C (independent female subject): SPOOF (peakiness 1.2748, HR 79.42 BPM). 60
5.15. rPPG-Toolbox on Video C (independent female subject): SPOOF (peakiness 1.2748, HR 79.42 BPM). 60
5.16. InsightFace: stable ArcFace face detection on Video A (no glasses) across all 150 sampled frames. 61
5.17. InsightFace: stable ArcFace face detection on Video B (with glasses) across all 192 sampled frames. 61
5.18. InsightFace: stable ArcFace face detection on Video C (independent female subject) across all 192 sampled frames, zero detection failures. 62
5.19. Silent-Face on Video A: all 192 frames classified as REAL (PASSED). 62
5.20. Silent-Face on Video B: all 192 frames classified as REAL (PASSED). 62
5.21. Silent-Face on Video C (independent female subject): all 192 frames classified as REAL (PASSED). 63
5.22. FASNet on Video 5: 33 REAL / 0 SPOOF / 16 no-face, PASS (false acceptance). 67
5.23. FASNet on Video 6: 46 REAL / 0 SPOOF / 3 no-face, PASS (false acceptance). 67
5.24. FASNet on Video 7: 38 REAL / 0 SPOOF / 11 no-face, PASS (false acceptance). 67

5.25. FASNet on Video 8: 49 REAL / 0 SPOOF / 0 no-face, PASS (false acceptance). 67
5.26. DeepFace rPPG on Video 5: peakiness 1.9178, HR 84.38 BPM, REAL (false acceptance). 68
5.27. DeepFace rPPG on Video 6: peakiness 1.9867, HR 78.75 BPM, REAL (false acceptance). 68
5.28. DeepFace rPPG on Video 7: peakiness 2.5519, HR 56.25 BPM, REAL (false acceptance). 68
5.29. DeepFace rPPG on Video 8: peakiness 2.4903, HR 61.88 BPM, REAL (false acceptance). 68
5.30. rPPG-Toolbox on Video 5: peakiness 1.9178, HR 84.38 BPM, REAL (false acceptance). 68
5.31. rPPG-Toolbox on Video 6: peakiness 1.9867, HR 78.75 BPM, REAL (false acceptance). 68
5.32. rPPG-Toolbox on Video 7: peakiness 2.5519, HR 56.25 BPM, REAL (false acceptance). 68
5.33. rPPG-Toolbox on Video 8: peakiness 2.4903, HR 61.88 BPM, REAL (false acceptance). 68
5.34. InsightFace on Video 5: face detected in all processed frames (false acceptance). 69
5.35. InsightFace on Video 6: face detected in all processed frames (false acceptance). 69
5.36. InsightFace on Video 7: face detected in all processed frames (false acceptance). 70
5.37. InsightFace on Video 8: face detected in all processed frames (false acceptance). 70
5.38. Silent-Face on Video 5: 240/240 frames classified as REAL, PASSED (false acceptance). 70
5.39. Silent-Face on Video 6: 240/240 frames classified as REAL, PASSED (false acceptance). 70
5.40. Silent-Face on Video 7: 240/240 frames classified as REAL, PASSED (false acceptance). 70
5.41. Silent-Face on Video 8: 240/240 frames classified as REAL, PASSED (false acceptance). 70

List of Abbreviations

ACER Average Classification Error Rate. 64, 65, 72

AES Advanced Encryption Standard. 6

AI Artificial Intelligence. 2–4, 8, 9, 12, 22, 28, 30–32, 43, 49, 50, 54, 57, 62, 66, 72, 75, 85, 89, 91, 92, 95, 102, 103, 105, 106, 109

AMA *Agência para a Modernização Administrativa*. 27, 88, 90, 105

AML Anti-Money Laundering. 13–16, 22, 25, 26, 101

AMLA Anti-Money Laundering Authority. 11, 23

AMLD Anti-Money Laundering Directives. 5, 6, 11–15

AMLR Anti-Money Laundering Regulation. 11, 22, 80, 85, 86, 101, 102, 106, 109

APCER Attack Presentation Classification Error Rate. 64–66

API Application Programming Interface. 35, 38, 50, 51, 62, 73, 83, 86, 99, 107

AUC Area Under the Curve. 66

BPCER Bonafide Presentation Classification Error Rate. 64–66

BPM Beats Per Minute. 55, 61, 93

BRDF Bidirectional Reflectance Distribution Function. 29

C2PA Coalition for Content Provenance and Authenticity. 30, 31, 35

CC Cartão de Cidadão. 74

CDD Customer Due Diligence. 11–13, 22, 25, 28

CE Conformité Européenne. 8

CFT Combating the Financing of Terrorism. 26

CID Content Identifier. 20

CMD *Chave Móvel Digital*. 27, 89

CMOS Complementary Metal-Oxide-Semiconductor. 30

CNN Convolutional Neural Network. 21, 23, 29, 30, 38, 59, 63

CPU Central Processing Unit. 40, 41, 74, 75

DB Database. 100

DFT Discrete Fourier Transform. 42

DID Decentralized Identifier. 20, 27, 32, 76, 80, 85, 87, 88, 90, 99, 103, 117, 118

DLT Distributed Ledger Technology. 18, 27

DORA Digital Operational Resilience Act. 14, 15

DPA Data Protection Authority. 7

DPI Dots Per Inch. 42

EBA European Banking Authority. 5, 16, 23, 52, 64, 95, 96, 98, 106

EBSI European Blockchain Services Infrastructure. 27, 103

EDD Enhanced Due Diligence. 12, 22, 25, 75, 84, 91

eIDAS Electronic Identification, Authentication and Trust Services. 4, 7, 23, 24, 32, 82, 86, 102, 103, 106, 109

EU European Union. 3–5, 7, 8, 11, 13, 15, 22–24, 27, 30, 31, 88, 99, 102, 103, 106–108

EUDI European Digital Identity. 7, 24, 25, 32, 77, 81, 82, 90, 99, 102, 103, 107–109

EVM Ethereum Virtual Machine. 73, 79, 99

FAR False Acceptance Rate. 29, 49, 60, 66, 107, 108

FATF Financial Action Task Force. 28

FFT Fast Fourier Transform. 29, 35, 44, 45, 55, 73, 75

FIU Financial Intelligence Unit. 12, 16

FLOPs Floating-Point Operations. 41

FPR False Positive Rate. 42

FR Face Recognition. 37, 46

FRR False Rejection Rate. 66, 107, 108

FRVT Face Recognition Vendor Test. 39

GAN Generative Adversarial Network. 28, 31, 33

GDPR General Data Protection Regulation. 4–9, 14, 77, 79, 85, 86, 88, 99–102, 109

GPAI General-Purpose Artificial Intelligence. 30

GPU Graphics Processing Unit. 39–41, 75

HMAC Hash-based Message Authentication Code. 81

HR Heart Rate. 46, 55, 61, 93

HTTP Hypertext Transfer Protocol. 20

IBI Inter-Beat Interval. 61

ICAO International Civil Aviation Organization. 80, 84, 91–94, 105, 108

ICT Information and Communication Technology. 14, 15

IdP Identity Provider. 26

IEC International Electrotechnical Commission. 17

IFD Inter-Frame Difference. 96

IP Internet Protocol. 20

IPFS InterPlanetary File System. 20

ISMS Information Security Management System. 17

ISO International Organization for Standardization. 17

IT Information Technology. 5

JSON JavaScript Object Notation. 75–77, 80, 81, 95, 100

KYC Know Your Customer. 1–9, 11, 14, 15, 18–20, 22, 24, 25, 27, 28, 30, 32, 35, 37, 38, 41–43, 45, 47–51, 53, 56–58, 64, 73, 87–89, 91, 98, 101, 102, 104–108

LCD Liquid Crystal Display. 29, 42, 63

LDM Latent Diffusion Model. 31, 33, 59, 63, 108

MFA Multi-Factor Authentication. 9, 26

MIME Multipurpose Internet Mail Extensions. 74, 92

MRZ Machine Readable Zone. 77, 80, 81, 84, 91–94, 98, 105, 108

NeRF Neural Radiance Field. 28

NIS Network and Information Security. 14

NIS2 Network and Information Security 2. 5, 14

NIST National Institute of Standards and Technology. 39, 40, 61, 62

NLP Natural Language Processing. 83

OCR Optical Character Recognition. 23, 83, 84, 93, 94

OFAC Office of Foreign Assets Control. 25, 83, 99, 107

ONNX Open Neural Network Exchange. 40, 41

OS Operating System. 49, 50

P2P Peer-to-Peer. 19

PA Presentation Attack. 28

PAD Presentation Attack Detection. 2, 3, 9, 24, 29, 31, 34–37, 41, 46, 48, 51–54, 57, 58, 60, 62–66, 69, 71–73, 83, 88, 95, 104, 106, 108

PCM Pulse-Code Modulation. 96

PDF Portable Document Format. 20

PEP Politically Exposed Person. 12, 99, 107

PII Personally Identifiable Information. 28, 76, 81, 84, 88, 90, 95, 101, 105, 106

PIN Personal Identification Number. 10, 23, 26

PKI Public Key Infrastructure. 24

POS Plane-Orthogonal-to-Skin. 43–46, 53, 54, 60, 61, 67, 69, 74, 95, 104

PSD2 Payment Services Directive 2. 4, 5, 9–11, 23, 26, 86

PSD3 Payment Services Directive 3. 11

PSR Payment Services Regulation. 11

QES Qualified Electronic Signature. 24

RBA Risk-Based Approach. 22

REST Representational State Transfer. 73, 80, 90, 92, 99, 105, 107

RGB Red, Green and Blue. 29, 43, 44, 46

RIDP Remote Identity Proofing. 24

RMS Root Mean Square. 96, 97

ROC Receiver Operating Characteristic. 66

ROI Region of Interest. 43, 44

rPPG Remote Photoplethysmography. 29, 32, 43–48, 51, 53, 54, 56, 60, 61, 64–69, 71, 72, 74, 75, 77, 80, 83, 91–93, 95, 96, 104–106, 108

RTS Regulatory Technical Standards. 26

SaaS Software as a Service. 33

SAR Suspicious Activity Report. 12, 16

SCA Strong Customer Authentication. 5, 9–11, 26

SD-JWT Selective Disclosure JSON Web Token. 88

SoC System on a Chip. 42

SPoF Single Point of Failure. 26

SRM Spatial Rich Model. 31

SSI Self-Sovereign Identity. 1–4, 19, 26, 27, 82, 85, 87, 101–103, 106

SSS Subsurface Scattering. 29

TLS Transport Layer Security. 6, 26

TPM Trusted Platform Module. 30

TPR True Positive Rate. 42

TSP Trust Service Provider. 24

TTS Text-to-Speech. 96

UBO Ultimate Beneficial Owner. 13

UN United Nations. 99, 107

URI Uniform Resource Identifier. 85

UX User Experience. 31

VC Verifiable Credential. 20, 24, 27, 32, 48, 75–77, 80, 81, 85, 87–91, 95, 100, 103, 108

VGA Video Graphics Array. 95

ViT Vision Transformer. 30

W3C World Wide Web Consortium. 20, 80, 81, 85, 91, 99, 117

ZKP Zero-Knowledge Proof. ix, 27, 88

1. Introduction

This chapter introduces the problem the thesis addresses: how a bank can verify a customer's identity remotely when synthetic video can imitate a live face, and how the evidence of that verification can be trusted, audited and reused across institutions. It first sets out the context and the core argument, and then explains the motivation for the work. It goes on to state the goals and the two research questions that structure the thesis, and closes with an outline of the remaining chapters.

The banking industry is undergoing a rapid digital transformation driven by the need to improve security, operational efficiency, and customer trust. At the center of this transformation lies the process of Know Your Customer (KYC), through which financial institutions verify customer identity, prevent fraud, and comply with increasingly demanding regulatory obligations. In the remote channel, these requirements have become significantly more complex, because institutions must establish identity assurance without direct physical presence while still maintaining strong protection against impersonation, presentation attacks, and document fraud.

Traditional KYC procedures remain highly dependent on manual review, repeated document submission, fragmented institutional databases, and operationally expensive compliance workflows. These limitations create friction for customers and cost for institutions, while also leaving the non-face-to-face onboarding channel exposed to new attack vectors. In particular, the rise of synthetic media, deepfake facial animation, and video injection attacks has increased pressure on banks to combine document authentication, biometric verification, liveness detection, and governance controls within a single defensible onboarding process.

In parallel, blockchain-based identity architectures and Self-Sovereign Identity (SSI) models have emerged as promising mechanisms for improving auditability, interoperability, and data minimization in KYC. Rather than treating blockchain as a replacement for biometric verification, this thesis approaches it as a potential trust, provenance, and audit layer that can complement remote identity proofing. This is especially relevant in a banking context, where the problem is not only verifying a customer once, but also ensuring that verification evidence can be governed, reused under strict trust rules, and retained in a way that remains compliant with regulatory and privacy constraints.

This thesis therefore examines secure remote KYC from two connected perspectives. First, it analyzes the technical and regulatory requirements of modern remote onboarding, with particular attention to biometric verification, Presentation Attack Detection (PAD), liveness detection, and the threat posed by synthetic media. Second, it proposes a blockchain-enabled architecture for trust, auditability, and reusable identity in banking KYC. The resulting contribution is not framed as a fully deployed banking platform, but as a study that combines state-of-the-art review, empirical security evaluation of biometric defenses, and the design of a blockchain-supported architecture intended to address trust, compliance, and interoperability challenges.

By doing so, the thesis aims to bridge the gap between traditional KYC workflows and a more resilient digital onboarding model. The core argument is that modern banking KYC cannot rely on either biometric controls or blockchain-based trust mechanisms in isolation: secure remote onboarding requires both robust identity-proofing controls against contemporary attack methods and a governance layer capable of supporting auditability, controlled reuse, and regulatory accountability.

1.1. Motivation

The motivation for this work arises from the convergence of two pressures on financial institutions. The first is operational and regulatory: banks must improve onboarding efficiency while meeting increasingly strict expectations around customer due diligence, risk-based control, auditability, and remote identity proofing. The second is adversarial: the growing realism and accessibility of generative Artificial Intelligence (AI) make it easier to produce synthetic identities, deepfake videos, and spoofing attempts capable of undermining weak remote onboarding pipelines.

These pressures make KYC a useful lens through which to study both technical security and institutional trust. On one side, there is a clear need to understand how well current biometric and liveness mechanisms resist contemporary attacks. On the other, there is a need for architectures that reduce data silos, support verifiable audit trails, and enable controlled reuse of verified identity information across institutions. This thesis is motivated by the view that blockchain and SSI can contribute meaningfully to the second problem, but only if they are grounded in the practical security realities of remote biometric onboarding.

1.2. Goals and objectives

The primary objective of this study is to investigate how remote KYC in banking can be strengthened through a combination of modern biometric security controls and a blockchain-enabled trust architecture. Rather than treating KYC solely as a document-processing problem or solely as a blockchain problem, the thesis addresses it as an end-to-end remote identity assurance problem that includes regulation, technical controls, adversarial pressure, and governance.

More specifically, the goals of the thesis are fourfold. First, it reviews the state of the art in banking KYC, regulatory requirements, blockchain-based identity models, and the technical mechanisms used in document verification, face recognition, and liveness detection. Second, it examines the threat posed by synthetic media to remote onboarding and evaluates selected open-source biometric and PAD frameworks against those attack scenarios. Third, it proposes a blockchain-enabled architecture for auditability, trust management, and reusable identity in banking onboarding. Fourth, it analyzes how such an architecture can support compliance objectives such as data minimization, traceability, controlled third-party reliance, and secure evidence retention.

The intended contribution of the thesis is therefore twofold: an empirical contribution in the form of the security analysis of current biometric and liveness mechanisms, and an architectural contribution in the form of a blockchain-supported KYC model for banking. This framing keeps the thesis aligned with the work actually presented in the following chapters, while preserving the broader research objective of exploring how blockchain can support more secure and interoperable KYC processes.

These contributions are organised around two research questions that structure the empirical and architectural tracks of the thesis:

RQ1 To what extent do current open-source biometric verification and PAD mechanisms resist synthetic video generated by contemporary generative AI platforms, and does PAD resilience depend on the underlying generative model or on the complete generation pipeline through which that model is accessed?

RQ2 Can a blockchain-enabled SSI architecture address the trust, auditability, and controlled identity-reuse challenges of remote banking KYC while remaining compatible with current Portuguese and European Union (EU) regulatory obligations?

Chapter 5 addresses RQ1 empirically; Chapter 6 addresses RQ2 architecturally. Chapter 7 revisits both questions in light of the evidence assembled in the preceding chapters.

1.3. Outline

Chapter 2 presents both the regulatory and technical background required to understand banking KYC: the first half covers the institutional framework (GDPR, eIDAS 2.0, the EU AI Act, PSD2, and anti-money laundering obligations), and the second half introduces the technical foundations of the thesis, covering blockchain, SSI, decentralized identifiers, verifiable credentials, and the core principles of face recognition and presentation attack detection.

Chapter 3 surveys the literature on remote onboarding, identity verification, blockchain-based identity systems, and the evolving threat landscape associated with generative AI and deepfakes. Chapter 4 examines the architecture, strengths, and known limitations of each biometric and liveness algorithm selected for evaluation, providing the technical grounding for the experimental results that follow. Chapter 5 presents the methodology and results of the empirical evaluation, focusing on how current open-source verification pipelines respond to synthetic media attacks.

Chapter 6 presents the proposed blockchain-enabled onboarding architecture and discusses how auditability, reusable identity, compliance, and trust governance could be structured in a banking setting. Chapter 7 concludes the thesis: it summarises the contributions, answers the two research questions, and sets out the limitations and directions for future work. Together, these chapters support the thesis argument that secure remote KYC requires both strong technical defenses against modern biometric attacks and a trust architecture capable of supporting compliant digital identity reuse.

2. Background and Technical Foundations

This chapter establishes the two bodies of knowledge the rest of the thesis depends on. The first half sets out the institutional framework that constrains any remote KYC system operating in the European Union, and in Portugal specifically: the data-protection, authentication, anti-money-laundering and operational-resilience obligations that determine what a compliant onboarding flow is permitted to do with a customer's data and what it must be able to prove afterwards. The second half introduces the technical foundations on which the architecture proposed in Chapter 6 is built: blockchain and smart contracts, decentralised identity and verifiable credentials, distributed storage, and the face-recognition and presentation-attack-detection principles that the experiments in Chapter 5 put under adversarial pressure. Readers already familiar with either half may treat the corresponding sections as reference material.

2.1. Security Requirements and guidelines

Banks in Europe are required to adhere to stringent security guidelines to ensure the confidentiality, integrity, and availability of customer data collected during KYC processes. A primary requirement is compliance with the General Data Protection Regulation (GDPR) [1], which mandates explicit customer consent for data collection and ensures that sensitive information is encrypted both in transit and at rest. Furthermore, banks must implement Strong Customer Authentication (SCA) under Payment Services Directive 2 (PSD2) [2], combining at least two factors, such as something the customer knows (e.g., a password), something they possess (e.g., a smartphone), or something they are (e.g., biometrics). To detect and prevent financial crimes, the EU Anti-Money Laundering Directives (AMLD) [3] require robust customer due diligence procedures, ongoing transaction monitoring, and risk-based assessments. These guidelines are complemented by cybersecurity frameworks like the Network and Information Security 2 (NIS2) Directive [4], which enforces the adoption of secure Information Technology (IT) systems and mandates the timely reporting of cyber incidents. Additionally, banks must maintain comprehensive records of KYC activities for at least five years, ensuring their availability for audits by regulatory bodies such as the European Banking Authority (EBA). Non-compliance with these requirements can lead to severe financial penalties, reputational damage, and operational disruptions, emphasizing the critical need for strict adherence to these guidelines.

2.2. General Data Protection Regulation (GDPR)

The GDPR [1] is a cornerstone of data protection within the European Union, ensuring the ethical handling of personal data. For a bank onboarding customers remotely, its requirements concern how data may be collected, how it must be protected, what the customer may ask of the bank afterwards, and what the bank must be able to show the supervisor.

Banks are obligated to obtain clear and informed consent from customers before collecting or processing their data. Consent must be freely given, specific, and unambiguous, with customers having the option to withdraw consent at any time. For example, during the KYC onboarding process, banks must provide a transparent explanation of how customer data will be used, ensuring there are no “pre-checked” boxes or vague language that could mislead customers. Failure to obtain proper consent can lead to both regulatory penalties and reputational damage. The data collected must also be limited to what is strictly necessary for KYC processes, reducing the risk of misuse or overexposure.

Once collected, the data must be protected: GDPR requires appropriate technical and organisational security measures, and Article 32 names encryption of personal data as one of them. Encryption ensures that even if data is intercepted or accessed unlawfully, it remains unreadable without the appropriate decryption key. For instance, when customers upload identification documents or proof of address during the KYC process, banks must encrypt this data end-to-end to prevent exposure during transmission or storage on servers. Strong encryption protocols, such as Advanced Encryption Standard (AES) with 256-bit keys and Transport Layer Security (TLS) 1.2 or higher, are considered industry standards.

Under GDPR, customers have the right to access the data banks hold about them and can request its correction or deletion under certain circumstances. A customer may, for example, request a copy of their submitted KYC documents to verify how their data has been processed. Once the customer relationship has ended and the retention period mandated by other regulations (e.g., AMLD) has passed, they can also request that their data be deleted; banks must comply unless legal obligations, such as ongoing investigations, require the data to be retained.

GDPR also emphasizes accountability, requiring banks to maintain detailed records of their data processing activities. They must document how personal data is collected,

stored, and shared, and ensure that third-party service providers, such as KYC technology vendors, comply with GDPR as well. Additionally, in the event of a data breach, banks are required to notify the relevant Data Protection Authority (DPA) within 72 hours of discovering the breach and inform affected customers when the breach poses a high risk to their rights and freedoms.

Non-compliance with GDPR can result in significant fines of up to €20 million or 4% of a bank's annual global turnover, whichever is higher. Beyond financial penalties, non-compliance can lead to reputational harm, loss of customer trust, and operational disruptions due to regulatory scrutiny. High-profile data breaches in the financial sector have shown how mishandling sensitive customer data can lead to customer attrition and long-term reputational damage.

By adhering to GDPR, banks not only avoid these risks but also build customer trust and demonstrate their commitment to ethical and secure data practices. GDPR compliance is not just a legal obligation but a strategic advantage in fostering transparency and reliability in the banking industry.

2.3. Electronic Identification and Trust Services (eIDAS 2.0)

While GDPR protects data privacy, the Electronic Identification, Authentication and Trust Services (eIDAS) Regulation (EU) No 910/2014 and its 2024 revision, eIDAS 2.0, establish the legal framework for electronic identification and trust services across the EU. This regulation is fundamental to modern KYC as it ensures that digital identities and electronic signatures are legally equivalent to their physical counterparts.

A key introduction of eIDAS 2.0 is the European Digital Identity (EUDI) Wallet. This wallet allows citizens to store digital credentials (such as national ID cards, driving licenses, and university diplomas) on their smartphones. For banks, this represents a paradigm shift in KYC:

- *Mandatory Acceptance:* Large platforms and banks are required to accept the EUDI Wallet for strong user authentication.
- *High Assurance:* Identity data shared via the wallet is cryptographically signed by the issuing government, significantly reducing the risk of document forgery compared to traditional photo uploads.

- *User Control:* The wallet architecture is designed to give users full control over what data they share (selective disclosure), aligning with GDPR principles.

2.4. The EU AI Act and Biometrics

As of 2024, the European Union Artificial Intelligence Act (EU AI Act) [5] has introduced the world's first comprehensive legal framework for AI. Unlike GDPR, which focuses on data privacy, the AI Act focuses on product safety and fundamental rights. For KYC systems relying on face recognition, this regulation is transformative.

The Act follows a risk-based approach. Annex III, point 1(a), lists remote biometric identification systems as high-risk, but expressly excludes "AI systems intended to be used for biometric verification the sole purpose of which is to confirm that a specific natural person is the person he or she claims to be" [5]. The face check in remote KYC onboarding is of this kind: it compares the live capture with the portrait on the customer's own identity document (a 1:1 comparison) to confirm a claimed identity, rather than searching a database to establish who an unknown person is. On that basis, a 1:1 onboarding check is most likely not a high-risk system under Annex III, and does not by that route require a conformity assessment (with its Conformité Européenne (CE) marking) before entering the EU market. This does not settle the question for the whole onboarding pipeline. Automated decisions that deny a customer a service may need separate assessment, both under the AI Act (for example, where the same system also evaluates creditworthiness, which Annex III, point 5(b), lists as high-risk) and under GDPR Article 22, which restricts decisions based solely on automated processing that significantly affect the person [1].

Where a KYC system does fall within the high-risk category, banks (as "deployers") and software vendors (as "providers") must adhere to strict requirements:

- *Data Governance:* Training, validation, and testing datasets must be relevant, representative, and free of errors. This is a direct legislative response to algorithmic bias, requiring developers to prove their models work equally well across different demographics (e.g., skin tones, ages).
- *Technical Documentation and Logging:* The system must automatically generate logs of events (system functioning) to ensure traceability. In the event of a false rejection or a security breach, the bank must be able to trace exactly why the AI made a specific decision.

- *Transparency and Information:* Users must be informed that they are interacting with an AI system. The system must be accompanied by instructions for use that clearly state its capabilities and limitations (e.g., “this system may struggle in low light”).
- *Human Oversight:* The concept of “Human-in-the-Loop” is mandatory. An automated KYC rejection cannot be the final word; there must be a mechanism for a human compliance officer to review the biometric data and override the algorithm if necessary.
- *Accuracy, Robustness, and Cybersecurity:* High-risk systems must be resilient against errors and third-party attacks. The Act does not name specific techniques, but for a biometric system PAD is the practical way to meet this requirement against spoofing (e.g., deepfakes), which makes security testing more than a best practice.

Article 5 of the Act bans AI practices that pose an unacceptable risk. While banking KYC is allowed, the Act strictly prohibits real-time remote biometric identification in publicly accessible spaces for law enforcement purposes (with narrow exceptions for terrorism or missing persons), AI systems for social scoring or those that exploit vulnerabilities of specific groups, and emotion recognition systems used in workplaces or educational institutions. For KYC, the distinction is critical: the prohibition concerns real-time remote identification in public spaces for law enforcement, whereas onboarding verification is a 1:1 check of a customer who has chosen to identify themselves, and is therefore neither prohibited nor, on the reasoning above, remote biometric identification.

The enforcement mechanism is even stricter than GDPR. Non-compliance with the prohibition of AI practices can lead to fines of up to €35 million or 7% of total worldwide annual turnover, whichever is higher. Violating the obligations for High-Risk AI systems (such as data governance or transparency) carries fines of up to €15 million or 3% of turnover.

2.5. Strong Customer Authentication (SCA) under PSD2

The PSD2 [2] is a regulatory framework introduced by the European Union to enhance the security of electronic payments and customer authentication in the financial sector. A core component of PSD2 is SCA, which requires a Multi-Factor Authentication (MFA) process to verify a customer’s identity when accessing financial accounts or making electronic payments. The purpose of SCA is to protect against unauthorized access, reduce fraud, and increase consumer trust in digital banking services.

To comply with SCA, banks and payment service providers must implement at least two out of the following three authentication factors:

- *Something You Know:* Such as a password or Personal Identification Number (PIN).
- *Something You Have:* For example, a smartphone or hardware token.
- *Something You Are:* Biometric identifiers, like fingerprints or facial recognition.

For instance, when a customer logs into their online banking account or approves a large transaction, the system must prompt them for additional verification steps. This multi-layered security reduces the risk of unauthorized access and fraud.

SCA is mandatory for many types of transactions under PSD2, particularly in the following scenarios:

- *Accessing Online Banking Accounts:* Customers must verify their identity with SCA when logging into their accounts.
- *Approving Payments:* SCA is required for electronic payments, including online purchases and large money transfers.
- *Making Changes to Sensitive Data:* If a customer updates personal information (e.g., address, phone number) or changes their security settings, additional authentication steps are enforced.

For example, a customer making an online purchase over €30 must authenticate using SCA before the payment is processed. However, PSD2 also introduces exemptions for certain low-risk transactions, such as contactless payments under €50 or recurring transactions (e.g., subscriptions) below a predefined threshold. Banks and payment providers can apply these exemptions if they meet strict fraud monitoring and transaction risk analysis criteria.

By enforcing SCA, PSD2 has fundamentally strengthened the security of electronic transactions, paving the way for safer digital banking experiences across Europe. However, its success depends on continuous technological innovation to address emerging threats and maintain a balance between security and user convenience.

Evolution to PSD3: It is important to note that while PSD2 established the foundation for SCA, the regulatory landscape is evolving toward the Payment Services Directive 3 (PSD3) and the Payment Services Regulation (PSR). Proposed in 2023 and finalized around 2025, these new frameworks aim to further reduce fraud by tightening exemptions for SCA and improving the exchange of fraud-related data between financial institutions. For the purpose of this thesis, the technical requirements for biometric authentication established under PSD2 remain the operational standard.

2.6. Anti-Money Laundering Directives (AMLD)

The AMLD [3] are a series of EU regulations designed to prevent financial institutions from being exploited for money laundering and terrorist financing. These directives form the backbone of anti-financial crime legislation in Europe, ensuring transparency and accountability across banking operations. Each successive iteration, from the 4th AMLD to the 6th AMLD, has introduced stricter measures to address evolving threats. Banks are required to implement comprehensive frameworks that detect and prevent illicit financial activity while maintaining compliance with these regulations.

The New AML Authority (AMLA): To address fragmentation in how directives were applied across different member states, the EU established the Anti-Money Laundering Authority (AMLA) in Frankfurt, which began operations in mid-2025. Unlike previous directives which required national transposition, the new Anti-Money Laundering Regulation (AMLR) applies directly to all member states, creating a single rulebook. This ensures that KYC standards, including the technical requirements for digital identity verification, are consistent across the entire Eurozone.

2.6.1 Key Provisions of AMLD

Customer Due Diligence (CDD): Banks must perform thorough identity verification processes for all customers to ensure their legitimacy. This includes collecting and validating official identification documents, such as passports, national IDs, or driver's licenses. Additionally, banks often require proof of address, such as utility bills or lease agreements.

- *Purpose:* CDD ensures that financial institutions know their customers and prevents anonymous or fraudulent accounts from being created.

- *Applicability:* CDD is mandatory for new customers, existing customers undergoing significant transactions, or those displaying unusual activity.
- *Technology Integration:* Many banks now employ AI-driven tools to automate document verification, cross-reference data with government databases, and ensure accuracy.

Enhanced Due Diligence (EDD): High-risk customers, such as Politically Exposed Persons (PEPs), require additional scrutiny under EDD. This involves:

- Conducting deeper background checks to assess the customer's source of funds and wealth.
- Monitoring financial activities more closely to detect unusual or high-risk transactions.
- Maintaining detailed records of the measures taken and any identified risks.
- For example, if a customer is a high-ranking government official, the bank may investigate their financial history, international transactions, and connections to high-risk jurisdictions.

Ongoing Monitoring: AMLD mandates continuous scrutiny of customer accounts and transactions to identify suspicious activities. Banks use advanced transaction monitoring systems to:

- Flag unusual patterns, such as large cash deposits, transactions to high-risk jurisdictions, or frequent small transfers.
- Analyze customer behavior against predefined risk profiles to detect deviations.
- Generate Suspicious Activity Reports (SARs), which are submitted to national Financial Intelligence Units (FIUs) for further investigation.

Monitoring is not a one-time process; it is dynamic and evolves as customer behavior changes or new risks emerge. For instance, a sudden increase in transaction volume for a dormant account would trigger alerts for further investigation.

Ultimate Beneficial Ownership (UBO): For corporate clients, banks must go beyond the surface and identify the Ultimate Beneficial Owners (UBOs), the individuals who ultimately own or control the business.

- *Transparency:* UBO identification prevents bad actors from hiding behind shell companies or layers of ownership.
- *Process:* Banks must collect detailed information about the company's ownership structure, ensuring that no entity exploits anonymity to launder money.
- *Public UBO Registers:* In line with AMLD requirements, several EU member states maintain public registers of UBOs to enhance transparency and facilitate interbank cooperation.

2.6.2 Recent Enhancements in AMLD

The 5th AMLD and 6th AMLD introduced additional measures to combat new threats:

- *5th AMLD:*
  - Extended CDD to cover virtual currencies and prepaid cards, addressing risks posed by cryptocurrency transactions.
  - Required banks to screen customers against publicly available UBO registers.
  - Enhanced monitoring for high-risk third countries identified by the EU.
- *6th AMLD:*
  - Introduced a standardized definition of money laundering across all EU states to streamline prosecution.
  - Expanded the list of predicate offenses to include cybercrime and environmental crime.
  - Emphasized personal liability for individuals facilitating money laundering, including bank employees and directors.

Failure to comply with AMLD can result in severe consequences for financial institutions. Regulatory authorities have the power to impose substantial fines, such as the €775 million penalty levied against ING Bank in 2018 [6] for inadequate Anti-Money Laundering (AML) controls in the Netherlands. Beyond financial penalties, non-compliance can lead

to significant operational disruptions, including restrictions on banking activities or even the suspension of operating licenses in extreme cases. The reputational damage associated with AML failures is also considerable, as banks involved in such incidents often face intense public scrutiny, loss of customer trust, and widespread negative media coverage. Furthermore, the 6th AMLD emphasizes personal liability, holding employees and executives accountable for their involvement in money laundering activities, which can result in criminal charges and imprisonment.

By implementing AMLD, banks not only prevent financial crime but also align with the broader objectives of maintaining a secure and transparent global financial system. The combination of technology, regulatory enforcement, and collaboration ensures that banks remain vigilant against evolving threats while maintaining customer trust.

2.7. Cybersecurity Frameworks (NIS2 and DORA)

Where GDPR governs what may be done with customer data and the AMLD govern what must be verified, two further instruments govern whether the systems performing that verification can be relied upon at all. NIS2 sets a baseline for critical services generally; Digital Operational Resilience Act (DORA), applicable from January 2025, imposes a stricter and more specific regime on the financial sector. Both matter directly to remote onboarding, because a KYC pipeline depends on Information and Communication Technology (ICT) systems and, increasingly, on third-party verification providers whose failure would be the bank's responsibility.

2.7.1 NIS2 Directive

The NIS2 Directive [4], introduced by the European Union, is a key regulatory framework aimed at bolstering cybersecurity for essential and critical services. It builds on the original Network and Information Security (NIS) Directive by introducing more stringent requirements for risk management, incident reporting, and supply chain security. Given the sensitive nature of data handled in banking, particularly during KYC processes, the directive ensures that banks adopt robust cybersecurity measures to safeguard their systems against evolving cyber threats.

2.7.2 Digital Operational Resilience Act (DORA)

While NIS2 applies to general critical infrastructure, the DORA, applicable from January 2025, is specifically designed for the financial sector. DORA consolidates and upgrades

ICT risk requirements for banks, ensuring they can withstand, respond to, and recover from all types of ICT-related disruptions and threats.

Key pillars of DORA relevant to KYC systems include:

- *ICT Risk Management:* Banks must have resilient systems to protect the integrity of customer data during onboarding.
- *Incident Reporting:* A streamlined EU-wide channel for reporting major cyber incidents.
- *Third-Party Risk Management:* Since many banks use third-party providers for Face Recognition (e.g., Jumio, Onfido), DORA mandates strict oversight and contractual arrangements with these ICT third-party service providers to ensure they meet security standards.

2.8. Record-Keeping and Audit Trails

Under the AMLD and other regulatory frameworks, banks are required to maintain detailed and accurate records of all processes related to KYC, transactions, and customer communications. These requirements are critical for ensuring transparency, enabling regulatory compliance, and assisting in the prevention, detection, and investigation of financial crimes. Record-keeping and audit trails not only serve as a foundation for effective AML practices but also demonstrate a bank's accountability in managing financial risks.

Record-keeping obligations bear on three separate questions: how long records must be kept, in what form, and who must be able to retrieve them on demand. Each is treated in turn below, since each imposes a different constraint on the storage design proposed in Chapter 6.

Retention Period: Banks must retain KYC and transaction-related records for a legally specified duration, typically a minimum of 5 years after the end of the customer relationship. In certain jurisdictions or under specific circumstances, this period can extend to 10 years to support longer investigations or align with national legal requirements.

- For example, if a customer closes their account in 2024, the bank is obligated to securely store all associated records until at least 2029, or 2034 if extended retention applies.

- This retention period ensures that sufficient historical data is available for investigations into fraud, money laundering, or terrorist financing that may emerge years after the customer relationship ends.

Comprehensive Logs: Maintaining detailed logs is an essential component of AML compliance. These logs must include:

- *Customer Identification:* Copies of official documents (e.g., passports, national IDs, and proof of address) used during the onboarding process.
- *Risk Assessments:* Documentation of customer risk profiles, including factors such as occupation, transaction behaviors, and country of origin.
- *Transaction Histories:* Detailed records of all financial transactions, including amounts, timestamps, and counterparties, to track patterns and detect anomalies.
- *SARs:* Comprehensive records of any flagged activities reported to FIUs, including the rationale and evidence for suspicion.
- *Communications:* Logs of customer communications, such as inquiries about account details, requests for information, or updates to personal data, which can support dispute resolution or investigations.

By documenting every aspect of customer interaction and transaction activity, banks create a robust database that ensures operational transparency.

Audit Readiness: To comply with regulatory requirements, banks must ensure their records are readily accessible for inspection by supervisory authorities such as the EBA, local financial regulators, or external auditors.

- *Ease of Access:* Records must be organized systematically, allowing auditors to retrieve specific information promptly during inspections.
- *Accuracy and Completeness:* Any gaps or inconsistencies in records can result in fines or further scrutiny, so data accuracy is paramount.
- *Digital Transformation:* Many banks are transitioning to digital record-keeping systems to facilitate seamless audits and ensure data is protected with encryption and secure access controls.

2.9. ISO 27001 Summary

ISO/IEC 27001 [7] is an international standard for Information Security Management Systems (ISMSs), providing a systematic approach to managing sensitive company information to ensure its confidentiality, integrity, and availability. It is published by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC). The standard is designed to be applicable to organizations of all sizes and industries, making it a widely adopted framework for information security governance.

The standard outlines the requirements for establishing, implementing, maintaining, and continuously improving an ISMS. It follows a risk-based approach, helping organizations identify, assess, and mitigate security risks. Compliance with ISO 27001 demonstrates an organization's commitment to securing information assets, managing risks effectively, and complying with legal and regulatory requirements. Additionally, it fosters a culture of security awareness within the organization by embedding security best practices into daily operations.

A key component of ISO 27001 is risk assessment and treatment, which involves identifying and evaluating risks to determine appropriate security controls. Organizations must implement a structured process to assess threats, vulnerabilities, and potential impacts on information assets. Leadership and commitment are essential, requiring management involvement in defining security policies and ensuring the necessary resources for ISMS implementation. Effective leadership fosters a security-conscious culture and drives continuous improvements.

Organizations must also focus on asset management by identifying and protecting information assets against threats. Maintaining an inventory of assets and classifying them based on their importance and sensitivity helps ensure effective protection. Access control measures must be implemented to ensure that only authorized personnel have access to critical information. This includes user authentication, role-based permissions, and monitoring access activities to detect unauthorized access attempts.

Other essential aspects of ISO 27001 include cryptography and data protection, which involve implementing encryption and cryptographic controls to safeguard sensitive data from unauthorized access and breaches. Physical and environmental security measures ensure that locations where information is stored or processed remain protected from unauthorized access, theft, or environmental hazards. Human resource security focuses

on establishing security measures related to employees, contractors, and third parties, including background checks, security training, and clearly defined responsibilities.

Incident management is another crucial aspect, requiring organizations to establish processes for detecting, reporting, and responding to security incidents. This includes incident response planning, forensic analysis, and lessons learned to prevent recurrence. Additionally, organizations must implement business continuity and disaster recovery measures to ensure that operations can continue in the event of security breaches, system failures, or disasters. Compliance with legal, regulatory, and contractual security requirements is also essential, requiring regular internal audits and reviews to maintain adherence to ISO 27001.

A fundamental component of ISO 27001 is Annex A, which provides a comprehensive set of 93 security controls divided into four main categories: organizational controls, people controls, physical controls, and technological controls. Organizational controls include policies, risk assessments, and governance structures that help define the security framework. People controls focus on employee training, awareness programs, and access management. Physical controls involve securing facilities, controlling physical access, and protecting against environmental risks. Technological controls cover encryption, network security, intrusion detection, and system monitoring to prevent cyber threats.

Organizations seeking ISO 27001 certification must undergo a formal audit process conducted by an accredited certification body. The certification process typically involves a gap analysis, a readiness assessment, and a final certification audit. Achieving certification enhances trust among stakeholders, improves risk management, strengthens overall cybersecurity resilience, and provides a competitive advantage in industries where data protection is critical.

2.10. Blockchain Technology

The following sections establish the technical foundations necessary to understand the proposed KYC framework, covering blockchain architecture, decentralised identity, distributed storage, and the core principles of face recognition and presentation attack detection.

Blockchain is a Distributed Ledger Technology (DLT) that enables the secure recording of transactions across a decentralized network of computers. First introduced by Satoshi Nakamoto in the context of Bitcoin [8], it has since evolved into a foundational technology

for secure data sharing. Unlike traditional centralized databases used by banks, where a single authority controls the data, blockchain operates on a Peer-to-Peer (P2P) network where every participant (node) maintains a copy of the ledger.

2.10.1 Core Characteristics

Three properties distinguish a blockchain from the centralised databases banks already operate, and it is these three, not the technology's association with cryptocurrency, that make it relevant to KYC record-keeping.

- *Decentralization:* No single entity has control over the entire network, reducing the risk of a single point of failure.
- *Immutability:* Once a block of data is added to the chain, it is cryptographically linked to the previous block. Altering a record would require recalculating the hash of that block and all subsequent blocks, which is computationally infeasible in a robust network.
- *Transparency and Traceability:* In public blockchains, all transactions are visible to network participants, ensuring auditability. In private or permissioned blockchains (often used in banking), visibility is restricted to authorized entities.

2.10.2 Smart Contracts

Smart contracts are self-executing contracts with the terms of the agreement directly written into code. They run on the blockchain and automatically enforce rules without the need for intermediaries. In the context of KYC, a smart contract could automatically grant access to a customer's data when a bank requests it, provided the customer has digitally signed the authorization.

2.11. Decentralized Identity and SSI

SSI is a user-centric approach to digital identity that gives individuals full control over their data. It shifts the model from a centralized one (where Google or a Bank holds your identity) to a decentralized one.

2.11.1 The Trust Triangle

The SSI model relies on three key roles:

1. *Issuer:* An entity that attests to a claim (e.g., the Government issues a National ID).
2. *Holder:* The user (Customer) who receives the credential and stores it in their digital wallet.
3. *Verifier:* The entity (e.g., a Bank) that requests proof of the user's identity.

2.11.2 Verifiable Credentials (VCs) and DIDs

The World Wide Web Consortium (W3C) has standardized the core components of this architecture:

- *Decentralized Identifiers (DIDs):* Defined by the W3C DID specification [9], a DID is a new type of identifier that enables verifiable, decentralized digital identity. A DID refers to a DID Document, which contains public keys and service endpoints necessary to interact with the identity owner.
- *Verifiable Credentials (VCs):* As defined in the W3C VC Data Model [10], these are digital equivalents of physical credentials (like a passport). They are cryptographically signed by the Issuer. The Verifier can check the signature on the blockchain to confirm the credential is valid and has not been revoked, without needing to contact the Issuer directly.

2.12. Distributed Storage: IPFS

While blockchain is excellent for recording transactions and proofs, it is inefficient for storing large files like images or Portable Document Format (PDF) documents due to high costs and scalability issues. The InterPlanetary File System (IPFS) [11] addresses this by providing a peer-to-peer hypermedia protocol.

Unlike Hypertext Transfer Protocol (HTTP), which addresses content by its location on a server identified by an Internet Protocol (IP) address, IPFS addresses content by its hash, the Content Identifier (CID). When a file is uploaded to IPFS, a unique cryptographic hash is generated. If the file content changes, the hash changes. This ensures data integrity. In a blockchain-based KYC system, the blockchain stores the *hash* (CID) of the KYC documents, while the actual files are stored off-chain on IPFS.

2.13. Face Recognition Technology

Modern KYC relies heavily on biometric verification. This section details the technical pipeline of face recognition systems.

2.13.1 Deep Convolutional Neural Networks (CNNs)

Traditional computer vision methods have been largely superseded by Deep Learning [12]. Convolutional Neural Networks (CNNs) are designed to process grid-like data (images) by learning spatial hierarchies of features.

- *Feature Extraction:* The network processes the input image through multiple layers (convolution, pooling) to extract high-level features (edges, textures, facial landmarks).
- *Embedding Generation:* The final output of the network is a feature vector (embedding), typically of 128 or 512 dimensions. This vector represents the face in a mathematical space.

2.13.2 Matching and Loss Functions

To verify identity, the system calculates the distance (e.g., Euclidean or Cosine similarity) between the embedding of the live face and the embedding of the stored ID photo.

- *ArcFace Loss:* ArcFace [13] is a state-of-the-art loss function that maximizes the angular margin between classes. This ensures that embeddings of the same person are tightly clustered, while embeddings of different people are pushed far apart, significantly improving accuracy in security-critical applications like banking.

2.13.3 Presentation Attack Detection (PAD)

To prevent spoofing (using photos or videos of a victim), Liveness Detection algorithms are employed [14]. These can be:

- *Active:* Asking the user to blink, smile, or turn their head.
- *Passive:* Analyzing texture, depth maps, or light reflection patterns to distinguish skin from screens without user interaction.

3. Related work

This chapter reviews the work the thesis builds on. It begins with banking KYC practice and its regulatory drivers in the EU and Portugal. It then turns to the technologies that have automated it (document authentication, biometrics and remote onboarding), to sanctions screening and transaction monitoring, and to the security weaknesses of identity providers. The later sections cover the two developments reshaping remote KYC: distributed ledgers, which can make verification portable and auditable across institutions, and generative AI, which undermines the assumption that a face in front of a camera belongs to a live person. The chapter closes by comparing the generative tools considered for the bypass experiments of Chapter 5 and by discussing watermarking and content provenance.

3.1. Know Your Customer (KYC) and Anti-Money Laundering (AML)

In recent years, banks have implemented robust Identity Assurance and KYC frameworks to ensure accurate identity verification and compliance with regulations aimed at preventing financial crime. Traditional KYC processes require deterministic validation of government-issued identification documents (such as national IDs, passports, or driver's licenses) and, in some instances, additional documentation like proof of address. To deepen this verification, banks utilize CDD, which employs a Risk-Based Approach (RBA) to assess customer risk profiles based on occupation, transaction type, and country of origin, with EDD mandated for high-risk entities. As part of their AML practices, banks utilize heuristics and pattern recognition to monitor the movement of funds and detect anomalies aligned with illegal activities. These comprehensive procedures aim to protect financial institutions and their clients, albeit through process-intensive workflows that can often be prone to human error and inefficiency.

3.2. AML Compliance in the European Union and Portugal

In the European Union, KYC obligations have evolved from Directives (which require national transposition) to Regulations (which apply directly), creating a harmonized legislative landscape. Adopted in 2024, the new AMLR is paving the way for a "single rulebook" for the EU, significantly reducing regulatory arbitrage, with full application set for

July 2027. Under this framework, the Frankfurt-based AMLA began its supervisory activities in 2025. In Portugal, while Law No. 83/2017 [15] remains the foundational framework, it is now interpreted alongside these direct EU regulations. Additionally, Aviso n.º 1/2022 [16] from the Banco de Portugal sets out how the financial entities it supervises must carry out their preventive anti-money-laundering and counter-terrorist-financing duties, and revokes Aviso n.º 2/2018. For remote identification it admits videoconference as an alternative procedure (Article 21(6) and Annex I): the session must take place in real time and be recorded, must include questions that vary from session to session, and is complete only once the customer enters a one-time code sent during the call. The Aviso does not regulate automated liveness detection; that requirement comes from the EBA guidelines discussed below.

3.3. Technological Advancements in KYC

The regulatory obligations described above were, until recently, discharged largely by hand. Two developments changed that: computer vision made document authentication and biometric matching accurate enough to automate, and eIDAS 2.0 made a government-backed digital identity wallet the intended default for proving who a customer is. The two subsections that follow trace each in turn, since the architecture proposed in this thesis depends on both.

3.3.1 Document Authentication and Biometrics

Document verification processes in banks have also seen technological advancements, with institutions increasingly adopting Computer Vision techniques. These include Optical Character Recognition (OCR) for text extraction and CNNs to authenticate customer-provided documents, including passports and government IDs, by analyzing pixel-level discrepancies. Many banks cross-reference these documents with national databases and leverage third-party solutions, such as Jumio [17], to detect forgeries, tampering (e.g., photo substitution), and expiration, reducing both human error and processing times. For mobile banking security, banks have turned to biometric verification (e.g., facial recognition via liveness detection, fingerprint scans, and PIN-based authentication) to enact “Inherence” and “Knowledge” factors of authentication [18]. Such measures streamline transactions while enhancing security, though they require significant investment in compatible infrastructure, especially to ensure compliance with standards such as PSD2 [2].

Modern biometric engines leverage sophisticated loss functions such as ArcFace [13] and RetinaFace [19] for high-precision identity matching even in uncontrolled environments.

3.3.2 Remote KYC and Digital Onboarding

A major development in KYC compliance is the adoption of Remote KYC, or Remote Identity Proofing (RIDP), which allows customers to complete identity verification without physically visiting a bank. While synchronous video-based onboarding was the standard in the early 2020s, 2025 and 2026 mark the official transition toward the European Digital Identity Wallet (EUDI Wallet) under the eIDAS 2.0 regulation (Regulation (EU) 2024/1183) [20].

The current landscape includes:

- *EUDI Wallet:* Mandated by eIDAS 2.0 [20], this framework enables the transmission of VCs directly from the user's edge device to the Relying Party (bank). Member States are required to provision these wallets by late 2026, leveraging cryptographic signing by Trust Service Providers (TSPs) to ensure data integrity and non-repudiation, far exceeding the assurance levels of optical verification.
- *Biometric Authentication:* Implementation of PAD standards (ISO/IEC 30107) [21] to verify liveness and thwart spoofing attempts.
- *Videoconferencing (Legacy/Fallback):* Retained as a fallback mechanism for "Unattended Remote" failures, utilizing synchronous human-in-the-loop verification.
- *Qualified Electronic Signatures (QESs):* Utilization of asymmetric cryptography within a Public Key Infrastructure (PKI) to generate legally binding signatures equivalent to handwritten consent under the eIDAS Regulation [20].

The European Banking Authority's Guidelines on the use of Remote Customer Onboarding Solutions (EBA/GL/2022/15) [22] operationalize these requirements across the EU, mandating that institutions implement identity proofing measures commensurate with the risk of the non-face-to-face channel, and providing explicit guidance on the use of biometric verification, document authentication, and liveness detection as acceptable technical controls. Where an onboarding solution is unattended, that is, where the customer does not interact with an employee during verification, the guidelines require liveness detection (Section 4.4, paragraph 41(c)). The Banco de Portugal further issued practical guidance

on video-based remote onboarding [23] for Portuguese financial institutions, establishing baseline quality standards for synchronous video sessions used prior to the adoption of EUDI Wallet-based flows. Solutions such as LOQR's KYC platform [24] ensure compliance with Portuguese financial regulations, including LINCE certification and the integration of these new wallet-based standards.

3.4. Sanctions Screening and Transaction Monitoring

Verifying that a customer is who they claim to be is only half of the obligation. The other half is establishing that this particular person may lawfully be onboarded, and that the funds they subsequently move are legitimate. Both are screening problems rather than identity problems, both are irreducibly probabilistic, and both consequently carry a manual-review burden that automation has reduced but not removed.

3.4.1 Sanctions List Checks

Sanctions list checks represent another critical area within KYC frameworks, where banks must ensure customers are not on any international sanctions lists from organizations like the U.S. Treasury's Office of Foreign Assets Control (OFAC), the European Union, or the United Nations. Given the linguistic variability of global naming conventions, this process presents significant compliance challenges. Banks employ heuristic matching and fuzzy logic algorithms (e.g., Levenshtein distance, Soundex) to identify potential matches while attempting to minimize false positive rates. However, due to the probabilistic nature of these algorithms, manual review is often required for disambiguation. This rigorous screening is essential for legal compliance and ethical banking practices, yet it demands substantial computational and human resources.

3.4.2 Transaction Monitoring and Source of Funds Verification

To verify the legitimacy of funds, banks rely on CDD and EDD to request customer disclosures regarding primary income sources. For significant transactions, such as property sales, banks may ask for legal documentation, like sales contracts, as evidence of fund origins. To further enhance compliance, banks employ transaction monitoring systems to track and analyze customer activity for patterns that could indicate suspicious activity or money laundering. This layered approach strengthens banks' AML defenses, though it increases operational complexity and requires sophisticated monitoring technologies.

Understanding the actual source of funds can be complex, particularly when customers use layered transactions, shell companies, or intermediaries to disguise fund origins. Banks are required to follow both national and international AML and Combating the Financing of Terrorism (CFT) frameworks to detect these schemes. Failure to meet AML and CFT obligations can lead to severe penalties, emphasizing the need for diligent compliance. Navigating these frameworks presents a significant challenge, as banks must carefully balance operational efficiency and regulatory requirements.

3.5. Cybersecurity and Identity Provider Challenges

The verification mechanisms discussed so far assume a trustworthy channel and a trustworthy identity provider. Neither assumption holds unconditionally. This section examines the two places where that assumption breaks: the customer endpoint, which the bank does not control, and the identity architecture itself, where concentrating verified identity data in a single provider creates a target whose compromise is far more damaging than any individual account takeover.

3.5.1 Multi-Factor Authentication and Customer Security

Another challenge banks face is securing the customer endpoint and the communication channel against interception (Man-in-the-Middle attacks). While banks utilize TLS and end-to-end encryption to limit vulnerabilities, they have limited control over the operational security of user devices. To mitigate compromise, particularly for high-privilege actions like fund transfers, banks implement SCA via MFA. This requires the combination of at least two independent factors: Knowledge (e.g., PIN), Possession (e.g., registered device), and Inherence (e.g., biometrics). Adherence to PSD2 Regulatory Technical Standards (RTS) for SCA is mandatory but introduces significant friction and integration costs.

3.5.2 Centralized and Decentralized Identity Architectures

The integration of external Identity Providers (IdPs), whether centralized (e.g., *Chave Móvel Digital* [25]) or decentralized (e.g., World ID [26]), introduces complex trade-offs in the Security-Interoperability-Trilemma. Centralized architectures offer high availability but create Single Points of Failure (SPoFs) and attractive “honeypots” for attackers, as a single breach can compromise millions of identities. Conversely, decentralised identity frameworks empower SSI but currently suffer from fragmentation and a lack of standardized resolving mechanisms, creating integration friction for legacy banking systems.

These structural dichotomies highlight the difficulty in establishing a unified, resilient verification ecosystem.

3.5.3 KYC Bypass Attempts and Fraudulent Activities

Malicious actors continuously devise sophisticated methods to circumvent KYC protocols, exploiting vulnerabilities in verification processes. Common tactics include the use of forged or stolen identification documents, synthetic identity fraud (combining real and fake information to create new identities), and impersonation through various means, including the increasing threat of deepfake technologies. These attempts aim to open fraudulent accounts, facilitate money laundering, or engage in other illicit financial activities. The 2024 ransomware attack on the *Agência para a Modernização Administrativa* (AMA) in Portugal, which disrupted the *Chave Móvel Digital* (CMD) system, highlights the persistent threat landscape and the inherent risks of centralized identity infrastructures, emphasizing the need for decentralized and dynamic security measures.

3.6. Distributed Ledger Technology and Blockchain in KYC

DLT has emerged as a paradigm shift for secure digital identity management. Its core architectural attributes (decentralized consensus, cryptographic immutability, and audit transparency) make it particularly suitable for mitigating the “siloed data” problem in KYC [27]. Blockchain-based identity solutions allow for:

- *SSI:* Enabling users to manage DIDs [9] without reliance on a central certificate authority.
- *Immutable Audit Trails:* Creating tamper-evident records of KYC checks, preventing retroactive data manipulation.
- *Privacy-Preserving Interoperability:* Facilitating cross-institutional validation via Zero-Knowledge Proofs (ZKPs) and VCs [10], allowing a bank to verify a customer’s status (e.g., “Over 18”, “KYC Cleared”) without exposing the underlying raw data.

Protocols such as Ripple [28] and Stellar [29] leverage these cryptographic primitives for anti-fraud mechanisms and cross-border settlement, demonstrating feasibility in high-value financial services. Additionally, initiatives like the European Blockchain Services Infrastructure (EBSI) [30] and Alastria [31] are standardizing SSI frameworks to ensure cross-border regulatory compliance within the EU.

3.7. Generative AI Threats to Identity Verification

Rapid advancements in deep generative models have led to the democratization of sophisticated video synthesis and audio-driven facial animation. These tools, leveraging architectures such as Generative Adversarial Networks (GANs) and Neural Radiance Fields (NeRFs), have profound implications for the integrity of digital identity systems. Models such as Hedra [32] and Kling AI [33] utilize deep learning to synthesize photorealistic lip synchronization and facial micro-expressions from arbitrary audio inputs. While technologically impressive, these capabilities represent a potent vector for "Deepfake" Presentation Attacks (PAs), specifically threatening the liveness detection mechanisms pivotal to remote KYC compliance. The Financial Action Task Force (FATF) has explicitly flagged artificial intelligence and deepfakes as an emerging horizon risk for anti-money laundering and counter-terrorism financing compliance [34], identifying video injection attacks and synthetic identity construction as the primary vectors requiring regulatory attention from financial institutions.

3.7.1 Threats to KYC and Security

Hedra and Kling AI are state-of-the-art AI models known for their highly accurate lip-sync capabilities and realistic facial animations. However, their ability to generate convincing deepfake content introduces severe risks in identity fraud and financial security.

Vector Analysis & Threat Modeling:

- *PAs:* The deployment of synthetic media on virtual camera drivers to bypass "Active Liveness" challenges (e.g., nodding, blinking) often used in legacy video onboarding.
- *Video Injection Attacks:* Utilizing virtual devices to inject pre-rendered, high-fidelity deepfakes directly into the application stream, effectively bypassing the physical camera sensor entirely and neutralizing sensor-pattern analysis noise checks.
- *Synthetic Identity Construction:* The combination of valid Personally Identifiable Information (PII) with AI-generated faces to create "Frankenstein identities" that pass initial CDD checks but have no physical existence.
- *Biometric Spoofing and False Acceptance:* The erosion of confidence in heuristic-based liveness detection (e.g., texture analysis), as modern diffusion models can

now accurately replicate sub-surface scattering and skin micro-texture, increasing the False Acceptance Rate (FAR).

TABLE 3.1: Defensive Benchmarks for Presentation Attack Detection (PAD)

Detection Method	**Target Signal**	**Scientific Principle**
rPPG	Blood flow / Pulse	Spectral analysis of RGB fluctuations.
Material CNN	Skin Reflectance	Light diffusion and subsurface scattering.
Temporal ViT	Micro-expressions	Analysis of frame-to-frame consistency.
Hardware Attestation	Video Source	Verification of the physical sensor's data path.

To effectively mitigate the risks posed by generative models, the scientific community has established several defensive benchmarks that analyze specific biological and technical signals:

- *Remote Photoplethysmography (rPPG):* This technique exploits the fact that the human cardiac cycle causes subtle, periodic fluctuations in skin color due to changes in blood volume. While invisible to the naked eye, these signals can be captured by standard Red, Green and Blue (RGB) sensors. Scientifically, rPPG-based detection algorithms perform spectral analysis of the green color channel (which has the highest absorption rate for hemoglobin), typically via a Fast Fourier Transform (FFT). Most generative models, particularly those based on frame-by-frame autoencoders (e.g., DeepFaceLive), fail to replicate the complex, spatially-coherent heartbeat signals across the face, making rPPG a highly effective "biological Turing test". This technology has been successfully deployed by providers such as Sumsub [35] and BioID [36] as a secondary layer of validation against synthetic media.

- *Material Classification (Material CNN):* Human skin exhibits unique optical properties, such as Subsurface Scattering (SSS), where light enters the epidermis, scatters, and then exits at a different point. Modern PAD systems utilize deep CNNs to classify materials based on their Bidirectional Reflectance Distribution Function (BRDF). This allows the system to distinguish the "biological reflectance" of real skin from the "specular highlights" of a Liquid Crystal Display (LCD) screen or the flat texture of a printed mask. This principle is utilized by providers such as iProov via their patented Flashmark technology [37], which uses controlled light reflection to map the material properties and 3D geometry of the subject.

- *Temporal Vision Transformers (ViTs):* Moving beyond frame-by-frame analysis, Temporal ViTs leverage self-attention mechanisms to evaluate the semantic and physical consistency of motion over long sequences. They are specifically tuned to identify anomalies in non-linear facial movements, such as the micro-tremors of eye saccades or the synchronous movement of the larynx during speech. Deepfakes generated by Latent Diffusion often exhibit "temporal hallucination", where micro-details appear and disappear between frames, creating a high-frequency noise signature that ViTs can detect with high precision. Modern forensic engines, such as Onfido's Atlas [38], employ these transformer-based models to catch cross-frame artifacts that traditional CNNs overlook.

- *Hardware Attestation and Sensor Forensics:* This defensive layer shifts from analyzing the *content* of the video to the *integrity* of the capture device. By utilizing Hardware Root of Trust (e.g., Trusted Platform Modules (TPMs) or Secure Enclaves), the system can cryptographically verify that the video stream originated from a physical Complementary Metal-Oxide-Semiconductor (CMOS) sensor rather than a virtual camera driver used for injection attacks. This aligns with the Coalition for Content Provenance and Authenticity (C2PA) standard [39], ensuring that every frame is digitally signed at the point of capture, creating an immutable link between the physical user and the digital session. Implementation of Android Play Integrity [40] and iOS App Attest [41] has become a mandatory standard for mobile banking to neutralize "Man-in-the-Middle" video injection tactics.

Regulatory Countermeasures and Provenance (C2PA): To combat these threats, the EU AI Act (2024) [5] mandates that providers of generative AI systems embed provenance data in their outputs. The technical standard for this is the C2PA specification [39], which cryptographically binds the asset's history to the file. Since August 2025, strict rules for General-Purpose Artificial Intelligence (GPAI) models have enforced this tagging. However, for KYC security, banks cannot rely solely on C2PA, as malicious actors utilize open-source models (e.g., SadTalker, localized Stable Diffusion) that purposefully strip or omit these watermarks.

Defensive Latency and Adversarial Asymmetry:

- *Detection Latency versus User Experience (UX):* The high computational overhead required for deep forensic analysis (e.g., frame-by-frame spectral analysis) introduces latency that conflicts with the "frictionless" user experience demanded by digital banking.
- *Adversarial Evolution:* A definitive arms race exists where generative models evolve faster than detection algorithms [42]; as PAD systems learn to identify specific artifacts (e.g., lack of blinking or heart rate), generative models are retrained via adversarial feedback to eliminate these cues.
- *Generalization Failure:* Detection models trained on specific datasets (e.g., FaceForensics++ [43]) often suffer from poor cross-dataset generalization [44, 45], failing to identify artifacts from novel architectures like Google Veo or Kling AI without retraining.

3.7.2 Evolution of the Identity Threat Landscape (2024 - 2026)

The period between 2024 and 2026 has witnessed a fundamental paradigm shift in the digital identity landscape, transitioning from a "Detection-Centric" to a "Provenance-Centric" security model.

- *2024 - The Artifact Detection Era:* Identity security was primarily reactive, focusing on identifying the "uncanny valley" of GAN-based deepfakes. Defensive mechanisms relied on detecting physiological anomalies (e.g., irregular blinking or gaze) and frequency-domain artifacts found with Spatial Rich Model (SRM) filters. Presentation attacks mainly involved physical spoofs, such as high-resolution displays or 3D masks.
- *2025 - The Generative Explosion & Injection Shift:* The maturation of Latent Diffusion Models (LDMs) and high-fidelity video transformers (e.g., Sora, early Veo) effectively eliminated most visual artifacts detectable by the human eye. Fraudsters pivoted from physical spoofs to Digital Injection Attacks, using virtual camera drivers to bypass the physical sensor entirely. The EU AI Act (2024) [5] and the emergence of C2PA standards established the first regulatory and technical requirements for content provenance, though adoption by malicious actors remained non-existent.

- *2026 - The Pivot to Cryptographic Assurance (Current State):* As of 2026, heuristic detection of synthetic media is considered a secondary defense due to the near-perfect visual fidelity of generative models (e.g., Google Veo 3 [46]). The industry has converged on a Zero-Trust Identity framework where forensic signals move from the pixel domain to the biological and hardware domains:

    1. *Transition to EUDI Wallets:* Under eIDAS 2.0 [20], the focus has shifted from "looking at the user" to "verifying the credential." Verifiable Credentials (VCs) and DIDs have begun to replace optical document scans, significantly reducing the attack surface for deepfakes.
    2. *Biological and Hardware Anchorage:* For applications requiring real-time presence (e.g., account recovery), defenders have shifted from texture analysis to Hardware Attestation [40] and Bio-Signal Analysis (rPPG) [35]. As evidenced in the preliminary results of Section 5.3, while visual artifacts are vanishing, the underlying physiological signals (heartbeat) and temporal consistency remain the current "unmaskable" forensic frontier.

Both Hedra and Kling AI exemplify the dual-use nature of AI-powered lip-syncing and video generation: while they offer valuable innovations, their misuse presents significant security challenges. As AI deepfake technology continues to evolve, financial institutions must adopt more advanced fraud detection tools, regulatory measures, and AI-driven countermeasures to safeguard KYC processes and prevent identity-based cybercrimes.

3.8. AI Generators and KYC Bypass Requirements

To evaluate the robustness of biometric authentication in banking applications, it is essential to define the technical requirements for synthetic media and analyze the capability of current generative AI models.

3.8.1 Comparison of AI Generative Architectures

Several generative models were elected for testing based on their current availability and performance in creating realistic facial animations. From a scientific perspective, these tools differ fundamentally in their underlying neural architectures, inference methodologies, and resulting artifact signatures.

3.8.1.1 Cloud-Based Probabilistic Models (Latent Diffusion & Transformers)

Google Veo 3, Hedra, and Kling AI operate as Software as a Service (SaaS) platforms, leveraging large-scale compute clusters to run computationally intensive models, primarily based on LDMs and Video Transformers.

- *Stochastic Generation and Latent Space Sampling:* Unlike deterministic algorithms, these models function probabilistically. They generate video content by sampling from a high-dimensional latent space conditioned on the user's prompt (text, image, or audio). Mathematically, the model approximates the conditional probability distribution $P(\text{Video}|\text{Prompt})$. This inherently stochastic process introduces variance; identifying specific high-frequency spectral artifacts is difficult because the "noise" is often hallucinated detail rather than compression artifacts.
- *Google Veo 3* [46]: Utilizes advanced 3D-aware temporal attention mechanisms. This allows it to maintain high temporal coherence (consistency of identity across frames) and generate high-fidelity textures that mimic genuine optical characteristics (e.g., sub-surface scattering on skin), making it highly effective against texture-based Liveness Detection.
- *Hedra* [32]: Specializes in Audio-Visual Grounding, likely employing a derivative of Wav2Lip or equivalent phoneme-viseme mapping architectures. It minimizes the "lip-sync error" cost function, ensuring precise synchronization between the audio waveform and the geometric deformation of the mouth region. This directly targets "Active Liveness" challenges that require users to speak a specific phrase.
- *Kling AI* [33]: Demonstrates capabilities akin to "World Models," integrating physics simulation into the generation process. This allows for the realistic rendering of complex interactions, such as hair movement and lighting changes (specular highlights), which are traditionally difficult for standard GANs to synthesize without artifacts.

3.8.1.2 Local Deterministic Transformation Tools (Autoencoders)

In contrast, DeepFaceLive [47] represents an Autoencoder-based architecture (specifically DeepFake or LIAE structures) executed locally on consumer hardware (Edge/GPU).

- *Frame-by-Frame Inference:* Unlike diffusion models that generate a sequence holistically, DeepFaceLive typically performs inference on a per-frame basis. It utilizes a

shared-weight encoder to compress the input face into a latent vector and a specific decoder to reconstruct it as the target identity.

- *Quantization and Temporal Jitter:* Because the process must occur in real-time (detect $\rightarrow$ align $\rightarrow$ swap $\rightarrow$ merge) within an approximate 33ms window (for 30fps), the model often relies on simplified architectures or lower precision (e.g., FP16). This computational constraint leads to two primary scientific flaws detectable by PAD systems:

  1. *Landmark Jitter:* High-frequency instability in the facial landmark detection (68-point mesh) causes the overlay mask to vibrate or "flicker" at the boundaries.
  2. *Temporal Inconsistency:* Due to the lack of deep temporal attention layers (which are computationally expensive), the model struggles to maintain consistent lighting and texture between adjacent frames, generating "temporal noise" that differs significantly from the natural sensor noise of a camera.

3.8.2 Technical Quality Requirements for Bypass Experiments

Table 3.2 compares the capabilities of these models against the critical requirements needed for a successful substitution attack on a banking application.

TABLE 3.2: Comparison of AI Generators and KYC Bypass Requirements

Requirement	**Google Veo 3**	**Hedra**	**Kling AI**	**DeepFaceLive**
Video Quality	Very High	High	High	Medium
Noise Level	Very Low	Low	Low	High (Pixel jitter)
Audio Accuracy	High (Stable)	Excellent (Sync)	High	N/A
Temporal Stability	High	High	High	Low (Flickering)
KYC Suitability	Recommended	Recommended	Recommended	Not Recommended

As demonstrated in the comparison, the primary challenge for attackers is minimizing the "noise" and artifacts that distinguish deepfakes from real camera feeds. Models that lack stability or produce high levels of visual noise, such as DeepFaceLive [47], fail even against basic liveness detection mechanisms.

3.8.3 Content Provenance and Watermarking as a Detection Paradigm

A fundamentally different approach to detecting synthetic media is to shift responsibility from the detector to the generator: instead of training a classifier to distinguish real from fake after the fact, the generator itself embeds an imperceptible signal that downstream verifiers can look for. Google DeepMind's SynthID [48] is the most prominent commercial

implementation of this paradigm. SynthID embeds resolution-dependent carrier frequencies and fixed phase templates directly into the pixel values of generated images; the system was extended to audio, text, and (most relevant to this thesis) video generation via Veo in 2024.

The appeal of watermarking for KYC is clear: if every video produced by a known high-risk generator carries a detectable signature, a verifier could reject any watermarked input without requiring a trained PAD layer. However, three structural limitations prevent watermarking from being a sufficient standalone control:

- *Generator dependency.* A watermark can only be detected in content that passed through a watermarked generator. Video produced by generators that do not implement SynthID (Kling AI, Runway, open-source latent diffusion models, or any future tool) carries no detectable signature.
- *Removal feasibility.* Because the watermark is a signal embedded in pixel or frequency space rather than a cryptographic commitment tied to the content, it can in principle be removed by anyone with knowledge of the encoding scheme. Denny [49] demonstrated this through signal processing alone: by discovering that SynthID uses resolution-dependent carrier frequencies with >99% phase coherence across images from the same Gemini model, a seven-stage pipeline combining elastic deformation and FFT-domain spectral subtraction reduced phase coherence by 91.4% and defeated the detector with visually lossless output quality.
- *Verification dependency.* Checking for a SynthID watermark requires either access to Google's proprietary detection Application Programming Interface (API) or replication of the watermark codebook, neither of which is available to an independent KYC provider as a self-contained on-premise control.

Content provenance initiatives such as the C2PA coalition address the third limitation by standardising a cryptographically signed metadata layer that is independent of any single vendor. However, C2PA metadata is stripped by standard re-encoding operations and requires the entire media pipeline (camera, editing software, hosting platform) to maintain the chain of custody, an assumption that holds for professional broadcast workflows but not for the user-uploaded selfie video characteristic of remote KYC. Watermarking and provenance mechanisms are therefore best understood as complements to PAD rather than replacements: they raise the cost of attack for an adversary who is unaware of or

unable to remove the watermark, but they do not provide the independent, generator-agnostic assurance that a properly layered PAD pipeline aims for.

The preceding review establishes the literature and threat-model context for the empirical part of this thesis. The next chapter details the specific algorithms selected for evaluation, and Chapter 5 presents the experimental setup and results obtained from the tested toolchain.

4. Algorithm Architecture and Selection

This chapter documents the five open-source verification components selected for adversarial evaluation, and explains why each was chosen. The selection deliberately spans the three detection paradigms available to an open-source KYC deployment (texture-based classification, frequency-domain analysis, and physiological signal recovery), so that the experiments of Chapter 5 test a defence-in-depth stack rather than a single model. For each component the chapter sets out its architecture and processing pipeline, then states its strengths and limitations as reported in the literature and as encountered in this work. Those stated limitations matter: several of them predict, in advance, the failure modes that the bypass experiments go on to confirm.

4.1. Tested Algorithms: Architecture and Analysis

This section provides an in-depth examination of each algorithm and framework evaluated during this research. The evaluation covered two distinct but complementary problem domains: *Face Recognition (FR)* (verifying that the person in front of the camera matches a reference identity) and PAD / liveness detection (verifying that the biometric sample originates from a live human being rather than an artefact such as a photograph, video replay, or mask). Throughout this chapter, *PAD* and *liveness detection* are used interchangeably to refer to the anti-spoofing step, while *KYC* refers to the broader remote onboarding process in which both identity verification and liveness checks are performed.

4.1.1 DeepFace

DeepFace [50, 51] is a lightweight, open-source Python framework developed by Sefik Ilkin Serengil that acts as a unified wrapper around multiple state-of-the-art face recognition models, including VGG-Face, FaceNet, OpenFace, DeepID, ArcFace [13], SFace, and GhostFaceNet. Rather than implementing a single fixed architecture, DeepFace abstracts the entire recognition pipeline (detection, alignment, normalization, embedding, and verification) into a single callable interface, allowing the underlying backbone to be swapped with a configuration parameter.

4.1.1.1 Architecture and Pipeline

The framework follows the five-stage modern face recognition pipeline:

1. *Detection:* A pluggable face detector locates and crops the face from the input frame. Supported backends include OpenCV Haar Cascades, MTCNN, RetinaFace [19], Dlib, and MediaPipe. RetinaFace is generally preferred for KYC applications due to its superior accuracy under occlusion and pose variation.

2. *Alignment:* Detected facial landmarks (eye corners, nose tip, mouth corners) are used to apply an affine transformation, rotating and scaling the face crop so that the eyes are horizontally aligned and centred within the fixed input grid.

3. *Normalization:* Pixel intensities are scaled and channel-normalized according to the requirements of the selected backbone.

4. *Embedding:* The backbone CNN produces a dense embedding vector (typically 128 or 512 dimensions) that encodes the identity-discriminative features of the face.

5. *Verification:* The cosine or Euclidean distance between the embedding of the probe face and the gallery reference is compared against a learned threshold to produce a binary same-person / different-person decision together with a confidence score.

DeepFace also provides a built-in anti-spoofing module based on FasNet (Face Anti-Spoofing Network). FasNet is a lightweight binary classifier that operates on the face crop and predicts whether the sample originates from a live person (`is_real = True`) or a printed/replayed artefact. The module returns an `antispoof_score` in the range $[0, 1]$, where higher values indicate greater confidence in a genuine presentation.

4.1.1.2 Strengths

DeepFace's appeal for this study is that it collapses an entire recognition pipeline behind one call, which makes it the realistic choice for a team prototyping a KYC flow without computer-vision specialists on staff.

- *Versatility:* Provides a single API for eight different recognition backbones, enabling rapid experimentation and comparison without engineering effort.

- *Ease of integration:* Installation via `pip install deepface` and a single Python call make it the lowest-barrier entry point for prototyping.

- *Active maintenance:* The library is actively maintained, with frequent releases and a large user community.

- *Combined pipeline:* Face detection, alignment, and verification are handled automatically, reducing the risk of pre-processing mismatches.

4.1.1.3 Weaknesses and Limitations

The same abstraction that makes DeepFace convenient also obscures what it is doing, and its bundled anti-spoofing module proved to be the weakest link in practice.

- *FasNet reliability:* The built-in FasNet anti-spoofing module was found to exhibit instability across DeepFace versions. The field names in the returned dictionary (`is_real`, `antispoof_score`) changed between releases, requiring version-specific key resolution logic in the test harness.
- *Speed:* When using heavier detectors (e.g., RetinaFace) combined with a large backbone (e.g., ArcFace 512-d), inference time per frame is prohibitive for real-time video streams without dedicated Graphics Processing Unit (GPU) acceleration.
- *Black-box anti-spoofing:* FasNet is a shallow network designed for speed; it lacks robustness against high-resolution digital replays and adversarial attacks.
- *Model downloading:* Weights are downloaded lazily at runtime, creating a dependency on external servers and making air-gapped deployments problematic.

4.1.2 InsightFace (buffalo_l)

InsightFace [52] is an open-source 2D and 3D face analysis library developed and maintained primarily by Jiankang Deng and Jia Guo from Imperial College London. It is widely regarded as a production-grade framework: the `buffalo_l` model package achieved *Rank-1* on the National Institute of Standards and Technology (NIST) Face Recognition Vendor Test (FRVT) 1:1 benchmark (Visa track) in 2021, making it one of the most rigorously validated public face recognition systems available [53].

4.1.2.1 Architecture and Pipeline

InsightFace is built around two core models within the `buffalo_l` pack:

- *SCRFD (Sample and Computation Redistribution Face Detector):* An efficient anchor-based single-shot face detector that hierarchically redistributes the number

of training samples and computational cost across detection scales. It produces bounding boxes and five facial landmarks used for subsequent alignment [54].

- *ArcFace recognition model:* A 100-layer ResNet (ResNet-100) trained with the ArcFace loss [13]. The model maps a 112×112 aligned face crop to a 512-dimensional L_2-normalized embedding. Identity verification is performed by computing the cosine similarity between two embeddings and comparing against a fixed threshold (typically 0.3–0.4 for the buffalo_l model).

The ArcFace loss deserves particular attention as it is the foundation of InsightFace's accuracy. Given a face embedding $\mathbf{x}_i$ and its ground-truth class label y_i, the ArcFace loss introduces an additive angular margin penalty m in the softmax formulation:

$$\mathcal{L}_{ArcFace} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cdot(\cos(\theta_{y_i}+m))}}{e^{s\cdot(\cos(\theta_{y_i}+m))}+\sum_{j\neq y_i}e^{s\cdot\cos(\theta_j)}} \tag{4.1}$$

where θ_{y_i} is the angle between the embedding $\mathbf{x}_i$ and the weight vector of its class, s is the feature scale, and m is the additive angular margin (typically $m = 0.5$ radians). This formulation ensures that intra-class angular distances are minimized while inter-class angular distances are maximized, producing a well-separated embedding space that generalizes reliably to unseen identities.

4.1.2.2 Strengths

InsightFace occupies the opposite end of the trade-off from DeepFace: it is harder to deploy but is the component in this study with genuine production credentials.

- *State-of-the-art accuracy:* NIST-FRVT Rank-1 performance provides an independent, government-audited accuracy guarantee.
- *Speed on CPU:* The SCRFD detector and the models exported in Open Neural Network Exchange (ONNX) format are computationally efficient, enabling real-time frame-by-frame processing on a standard Central Processing Unit (CPU) without GPU acceleration.
- *Robustness:* Performs reliably under challenging conditions including partial occlusion, off-axis pose, and low-quality webcam imagery.

- *ONNX deployment:* All models are distributed as ONNX files, enabling deployment on heterogeneous hardware (x86 CPU, ARM edge devices, NVIDIA GPUs) via the ONNX Runtime.

4.1.2.3 Weaknesses and Limitations

InsightFace's limitations follow from its scope. It is a face *recognition* library, and the strength of its identity matching says nothing about whether the face it matched was real, a distinction that proves decisive in Chapter 5.

- *No native anti-spoofing:* The `buffalo_l` package is a face recognition and analysis system only; it does not include a liveness detection component, meaning a separate PAD module is mandatory for any security-critical deployment.
- *Licensing restrictions:* The pre-trained models are released under a non-commercial research license. Commercial deployment requires separate licensing negotiation with the InsightFace team.
- *Large model size:* The `buffalo_l` pack (ResNet-100 backbone) requires several hundred megabytes of disk space, which may be a constraint for edge deployments.

4.1.3 MiniFASNet / Silent-Face Anti-Spoofing

The Silent-Face Anti-Spoofing project [55] by Minivision AI is a passive liveness detection system. *Passive*, or *silent*, detection is significant in a KYC context because it imposes no friction on the user: the system analyses a short video clip without asking the subject to blink, nod, or speak.

4.1.3.1 Architecture

The MiniFASNet architecture is built upon a pruned variant of MobileFaceNet [56], achieving a dramatic reduction in Floating-Point Operations (FLOPs):

TABLE 4.1: Computational complexity comparison of MobileFaceNet and MiniFASNet variants.

Model	FLOPs	Parameters
MobileFaceNet	0.224 G	0.991 M
MiniFASNetV1	0.081 G	0.414 M
MiniFASNetV2	0.081 G	0.435 M

The key insight is the use of Fourier Spectrum Map Auxiliary Supervision. The model is trained with two parallel branches:

1. *Classification branch (main):* A binary cross-entropy branch that predicts *real* or *spoof*.
2. *Fourier spectrum branch (auxiliary):* A regression branch that predicts the 2D Discrete Fourier Transform (DFT) of the input crop. Because printed or screen-displayed faces exhibit characteristic high-frequency repetition artifacts (halftone dot patterns, LCD pixel grids) that are conspicuous in the frequency domain, this auxiliary branch guides the network to learn frequency-sensitive features during training, even though the auxiliary output is discarded at inference.

At runtime, the system applies the model to face crops at two different spatial scales, and the final decision is obtained by combining the per-frame predictions across multiple frames with a majority vote.

4.1.3.2 Strengths

Silent-Face was included because passive detection is the only form of liveness checking that imposes no friction on the customer, which makes it the variety banks are most likely to deploy at scale.

- *Ultra-low latency:* At 0.081 GFLOPs per inference, the model achieves approximately 20 ms per frame on a mobile System on a Chip (SoC) (Kirin 990), making it suitable for on-device deployment without a cloud round-trip.
- *User friction-free:* No active gestures are required, directly improving the KYC conversion funnel.
- *Effective against print and replay attacks:* Achieves a True Positive Rate (TPR) of 97.8% at a False Positive Rate (FPR) of 10^{-5} on the published benchmark.
- *Frequency-domain awareness:* The Fourier auxiliary branch provides robustness against attacks that are perceptually indistinguishable in the spatial domain but detectable in the frequency domain (e.g., moire patterns on screens with a high Dots Per Inch (DPI) count).

4.1.3.3 Weaknesses and Limitations

Its weaknesses are those of any supervised classifier facing a shifting distribution: the model can only reject the artefact classes it was trained on.

- *Limited robustness to 3D attacks:* The model was primarily evaluated on 2D print and screen replay attacks. Its performance against 3D silicone masks or high-quality deepfake video injected directly into the video stream may degrade significantly.
- *Camera dependency:* Published metrics were collected with specific consumer-grade cameras; performance can vary significantly across different sensor characteristics and compression codecs.
- *Dataset recency:* The model was released in 2020, and the anti-spoofing threat landscape has evolved considerably since then, particularly with respect to AI-generated synthetic faces (deepfakes).
- *No identity binding:* MiniFASNet confirms liveness but does not verify identity; it must be paired with a separate recognition system in a KYC pipeline.

4.1.4 rPPG-Toolbox and the POS Algorithm

rPPG is a non-contact physiological sensing technique that estimates the cardiac pulse signal by detecting subtle, periodic colour changes in facial skin caused by the rhythmic variation in blood volume per heartbeat [57]. Because a printed photograph or static video replay does not exhibit a biological pulse, the presence of a detectable heart rate signal constitutes strong evidence of a live subject.

The rPPG-Toolbox [57] (NeurIPS 2023) is an open-source benchmarking platform that supports both unsupervised signal-processing methods and supervised deep learning methods. In this work, the Plane-Orthogonal-to-Skin (POS) algorithm was evaluated as the unsupervised baseline, selected for its computational simplicity and lack of training data requirements [58].

4.1.4.1 POS Algorithm

The POS algorithm, proposed by Wang et al. [58], models skin colour in a normalized RGB colour space and extracts the rPPG signal by projecting the colour trajectory onto a plane orthogonal to the skin-tone vector, where the residual signal corresponds to blood-volume-induced chrominance variation. The processing pipeline is as follows:

1. *Face detection and Region of Interest (ROI) extraction:* A face detector (in the evaluated implementation, DeepFace [50]) is applied to each frame to obtain a facial

bounding box. The mean RGB value of the face ROI is computed per frame, yielding a $T \times 3$ signal matrix $C(t)$, where T is the number of frames.

2. *Temporal normalization:* A sliding window of L frames is applied. Within each window, each channel is divided by its temporal mean:

$$C_n(t) = \frac{C(t)}{\bar{C}(t)} - \mathbf{1} \tag{4.2}$$

3. *Projection:* Two projection axes $\mathbf{p}_1 = [0, 1, -1]$ and $\mathbf{p}_2 = [-2, 1, 1]$ are applied to $C_n(t)$. The POS signal is computed as:

$$S(t) = \mathbf{p}_1 \cdot C_n(t) + \alpha \cdot \mathbf{p}_2 \cdot C_n(t) \tag{4.3}$$

where the scalar $\alpha = \sigma(\mathbf{p}_1 \cdot C_n)/\sigma(\mathbf{p}_2 \cdot C_n)$ scales the second axis to match the first.

4. *Frequency analysis:* An FFT is applied to $S(t)$, and the dominant peak frequency in the physiologically plausible heart rate band ($[0.75, 2.5]$ Hz, i.e., 45–150 bpm) is extracted as the estimated heart rate.

5. *Liveness decision:* If a dominant spectral peak is detected with sufficient power, the subject is classified as *live*; an absence of a coherent pulse signal triggers a *spoof* classification.

4.1.4.2 Strengths

POS is the only component evaluated here that attempts to measure a property of the subject rather than of the image, which gives it a fundamentally different and, as the experiments show, sometimes decisive error profile.

- *Physiologically grounded:* The signal being measured, a heartbeat, is an inherent property of biological life, making it conceptually the most robust type of liveness signal.
- *No training required:* POS is a deterministic algorithm, making it immune to adversarial retraining and deployable with zero data collection overhead.
- *Complementary attack surface:* High-quality deepfake videos that fool texture-based detectors may still be detected if they do not faithfully reproduce blood-flow signals.
- *Passive:* Like MiniFASNet, rPPG requires no active cooperation from the user.

4.1.4.3 Weaknesses and Limitations

Recovering a pulse from skin colour variation is a low signal-to-noise problem, and the method's practical limitations all stem from that.

- *Sensitivity to motion and illumination:* Head movement, variable ambient lighting, and video compression artefacts all corrupt the subtle chrominance signals that rPPG relies on, producing false-positive liveness rejections under real-world conditions.
- *Latency:* Reliable heart rate estimation requires a sustained video clip (typically 10–30 seconds), which conflicts with the sub-second interaction times expected in a digital KYC onboarding flow.
- *Not robust against advanced video attacks:* A high-fidelity synthetic video generated with knowledge of rPPG could theoretically embed a plausible pulse signal, defeating the detector.
- *Fragile face detection dependency:* The signal quality is critically dependent on stable face tracking across all frames. Frames where detection fails produce discontinuities that corrupt the FFT output, as observed experimentally during testing.

4.1.5 Hybrid Pipeline: DeepFace + rPPG

Given the complementary failure modes of texture-based anti-spoofing (e.g., MiniFASNet or FasNet) and physiological liveness detection (POS rPPG), a custom hybrid pipeline was also evaluated. In this configuration, DeepFace is used solely for its face detection and crop extraction capability (leveraging its multi-backend support), while the rPPG POS algorithm is applied to the resulting sequence of face crops to estimate the heart rate. This decoupling allows the face detection quality to be improved independently of the rPPG algorithm.

4.1.5.1 Strengths

The hybrid configuration was constructed specifically to test whether combining an appearance-based detector with a physiological one yields coverage that neither achieves alone.

- *Modular:* Each component (detector, liveness algorithm) can be upgraded independently.
- *Complementary detection:* Attacks that defeat one modality may still be caught by the other.

4.1.5.2 Weaknesses and Limitations

Layering two detectors compounds their costs as well as their coverage, and introduces a decision-fusion question that a single model does not face.

- *Cumulative latency:* The pipeline inherits the processing overhead of both DeepFace detection and the rPPG temporal buffer.
- *Detection fragility propagates:* As seen experimentally, any frames where DeepFace fails to detect a face cause missing entries in the RGB signal matrix, introducing noise into the POS algorithm and potentially preventing a valid Heart Rate (HR) estimate from being produced.

4.2. Algorithm Comparison

Table 4.2 summarises the key characteristics, strengths, and limitations of all evaluated approaches across both problem domains. The task column distinguishes FR from liveness / PAD.

TABLE 4.2: Comparative overview of the evaluated KYC algorithms.

System	Task	Strengths	Weaknesses	GPU Required
DeepFace (FasNet)	FR + PAD	Unified API, multiple backbones, easy prototyping	FasNet instability, slow without GPU, shallow spoof model	Recommended
InsightFace buffalo_l	FR only	NIST Rank-1 accuracy, ONNX/CPU-capable, robust under occlusion	Non-commercial license, no liveness, large model size	Not required
MiniFASNet (Silent-Face)	PAD only	Ultra-low latency, passive, frequency-domain robustness	Weak vs. 3D masks and deepfakes, camera-dependent	Not required
rPPG POS	PAD only	Physiologically grounded, training-free, deepfake-complementary	High latency ($\geq$10 s clip), motion-sensitive, compression artefact fragility	Not required
DeepFace + rPPG (hybrid)	PAD only	Modular, complementary attack surfaces	Cumulative latency, detection gaps corrupt rPPG signal	Recommended

From a KYC deployment perspective, the pre-experimental analysis suggested that MiniFASNet could serve as a fast first-stage liveness gate, with InsightFace buffalo_l providing identity verification and rPPG reserved for high-risk asynchronous re-checks. The experimental evaluation in Chapter 5 substantially revises this picture: texture-domain detectors (MiniFASNet, FasNet) failed completely against diffusion-based generative content, accepting synthetic video on virtually every sampled frame, while rPPG-based liveness proved the only layer capable of distinguishing Hedra/Veo 3 output from genuine captures. However, rPPG itself was bypassed by direct Veo 3 output generated through a different platform interface, demonstrating that no single algorithm in the evaluated set provides pipeline-universal liveness coverage. The practical implication is that a defence-in-depth posture, layering complementary detection signals rather than relying on any one approach, is not merely preferable but necessary in the current generative-AI threat environment.

5. Experimental Results

This chapter reports the experiments of the thesis. It first describes the testing methodology, the attack vectors in and out of scope, and the open-source tools used. It then presents the bypass results: the preliminary rPPG and DeepFaceLive tests, the evaluation of three Hedra/Veo 3 videos against all five tools with standard presentation-attack error metrics, the cross-platform test with four videos generated directly through Google Veo 3, and a discussion of which detection approaches work and of SynthID watermarking. The final section evaluates the blockchain prototype of Chapter 6: the outcomes of seven end-to-end runs, the on-chain events they produced, the latency of each step, and the tamper-evidence check.

5.1. Experimental Setup and Methodology

This chapter presents the experimental setup adopted to evaluate the resistance of current KYC verification pipelines against synthetic media, followed by the results obtained from the selected open-source frameworks. Whereas Chapter 3 reviewed the literature and threat landscape, and Chapter 4 explained the internal operation of the evaluated algorithms, the present chapter focuses on how those tools were used in practice and what results they produced.

The chapter is organised in two clearly separated experimental tracks. The first track (Sections 5.2.4–5.3.3) covers the biometric and PAD experiments against generative-AI synthetic media. The second track (Section 5.4) reports the measurable behaviour of the blockchain-anchored onboarding prototype introduced in Chapter 6: which on-chain events fire under which inputs, what wall-clock latency each verification component contributes, and how the tamper-evidence guarantee of `VCRegistry.verifyCredential` behaves under a controlled VC mutation. Architectural commentary (threat model, design rationale, regulatory mapping) remains in Chapter 6; this chapter reports only what was measured.

5.2. Testing Methodology for KYC Bypass Experiments

To evaluate the actual performance of biometric verification systems against synthetic media, a systematic testing methodology is required. This process involves testing directly within production-grade banking applications to assess real-world security robustness.

5.2.1 Step-by-Step Testing Process

Testing involves a systematic approach to validate face recognition in KYC contexts, performed directly in a banking app (e.g., via its biometric onboarding or verification features).

1. *Environment Setup:*
    - *App Selection:* Identification of banking applications with high-assurance face recognition.
    - *Hardware:* Use of standard smartphones equipped with high-resolution cameras to ensure capture quality matches user norms. For advanced "Digital Injection" tests, rooted devices are utilized to access the camera stream at Operating System (OS) level.
    - *Datasets:* Preparation of source media, including high-fidelity generated videos and real user capture for control groups.
2. *Implementation Steps:*
    - *Enrollment:* Establishing a baseline by enrolling the authentic user's face in the banking application.
    - *Verification / Liveness Phase:* During the authentication challenge, the AI-generated synthetic video is presented to the system.
    - *Presentation Method:*
        - *Virtual Camera Injection:* On a rooted device, the camera stream is intercepted using hooking frameworks (e.g., LSPosed) to inject the video file directly into the application's processing pipeline.
3. *Reporting and Analysis:*
    - *Outcome Logging:* Recording the success or failure of each bypass attempt, including specific error codes or security flags triggered by the app.
    - *Metric Calculation:* Deriving the FAR for various generative models to determine which architectures pose the greatest threat to current KYC implementations.

5.2.2 Mobile-Specific Vulnerabilities: Rooting and Hooking

Banking applications typically employ sophisticated integrity checks (e.g., Google's Play Integrity API and Apple's App Attest) to prevent execution on compromised devices. To conduct thorough security testing, it is necessary to bypass these protections:

- *Rooting (Magisk/Zygisk):* Essential for gaining administrative control over the mobile OS to manipulate input streams.
- *Method Hooking (LSPosed):* Used to hide the device's rooted state from the application and to simulate a live camera feed using pre-rendered AI videos (using tools like OBS Studio or custom virtual drivers).

This experimental framework allows for a rigorous assessment of how current liveness detection mechanisms respond to state-of-the-art synthetic media, providing empirical data to support the technical findings in the literature.

5.2.3 Threat Model Scope and Attack Vectors Evaluated

The taxonomy of remote KYC attack vectors is broader than what this thesis evaluates empirically, and it is important to be explicit about which vectors were actually exercised and which were considered but excluded. Three families of presentation attack on a remote face-verification challenge are commonly discussed in the literature: (i) pre-rendered synthetic video generated by a text-to-video or image-to-video model and submitted to the verifier; (ii) live virtual-camera injection on a rooted mobile device, in which a pre-rendered video is fed into the application's camera stream through an OS-level hook so that the banking app sees what appears to be a live capture; and (iii) real-time face-swap, in which an attacker drives a target identity on a webcam stream using a live face-replacement tool such as DeepFaceLive. This thesis evaluates the first family in depth and excludes the second and third for reasons of attack reliability and signal quality, both of which are documented below so the reader can interpret the bypass figures correctly.

(i) Pre-rendered synthetic video submission (evaluated). This is the attack vector exercised in Section 5.3.3. A talking-head video is generated off-device on a cloud platform (in this work, Hedra with the Google Veo 3 backend), exported as an MP4 without watermarks, and submitted as the biometric upload to the KYC pipeline. From the verifier's perspective, the upload is a normal file submission and is processed frame-by-frame

by the PAD layers under test. This vector is realistic against any remote KYC workflow that accepts video uploads, and it captures the dominant production-quality threat from current diffusion-based generative models.

(ii) Virtual-camera injection on a rooted mobile device (considered but excluded). The mobile-specific scenario described in Section 5.2.4 (Magisk/Zygisk for root, LSPosed for method hooking, custom virtual camera driver) was prototyped but not used to produce the bypass figures reported in this chapter. Two reliability problems made it unsuitable as a controlled experimental track. First, Play Integrity API responses on the test device alternated between `MEETS_DEVICE_INTEGRITY` and `MEETS_BASIC_INTEGRITY` across reboots, depending on Zygisk module ordering, and a non-trivial subset of banking applications refused to launch or silently dropped the biometric step when integrity downgraded. Second, the LSPosed camera-hook approach is sensitive to per-application camera-API version (Camera1 vs. Camera2 vs. CameraX), and the injection succeeded inconsistently across applications even after per-app hook adjustments. Because the experimental goal was to compare PAD layers on identical inputs, not to demonstrate that a hook can be made to work on one specific banking app, the unreliable injection path was discarded in favour of direct video submission. The qualitative threat that virtual-camera injection represents is preserved, however, because once a pre-rendered video is delivered to the verifier, the downstream PAD pipeline is identical under either delivery mechanism.

(iii) Real-time face-swap (considered but excluded). Tools such as DeepFaceLive (tested in the preliminary results of Section 5.3) drive a target identity in real time on a webcam stream. In principle this is a stronger threat than file upload because it can respond to dynamic challenge prompts. In practice, however, real-time face-swap quality remains visibly below that of pre-rendered diffusion-based synthesis: the swap boundary at the temples and chin produces low-frequency texture discontinuities that even simple PAD layers can sometimes detect; subsurface light scattering is not modelled; and the latency budget of real-time inference forces the use of lighter models that propagate temporal artefacts forward. The DeepFaceLive run reported in Section 5.3 achieved a 188/192 frame bypass against a pure identity matcher precisely because no PAD layer was applied; against the FASNet and rPPG layers analysed in Section 5.3.3, pre-rendered Hedra/Veo 3 output is consistently the harder adversary. Real-time face-swap was therefore retained as a baseline indicator but not as the principal adversarial input.

Scope of the bypass figures. The False Acceptance Rates and quantitative metrics reported in Section 5.3.3 and Section 5.3.3.7 therefore describe the response of the selected PAD layers to *vector (i) pre-rendered synthetic video submission*. They do not generalise without further evidence to virtual-camera injection or live face-swap scenarios, although the lower bound on the threat is the same in all three cases (since the verifier sees pre-rendered or live synthetic frames either way) and any PAD layer that is bypassed by vector (i) would by construction also be bypassed by vector (ii) if the injection succeeds.

5.2.4 Open-Source Frameworks for KYC Biometric Security Research

To support transparent and reproducible validation of the attack surfaces described above, several well-established open-source frameworks are publicly available that directly implement the defensive algorithms listed in Table 3.1. The selection of tools for this thesis follows two criteria derived from the EBA Guidelines on Remote Customer Onboarding [22]: (i) the mechanism must be practically deployable without specialised infrastructure, and (ii) it must correspond to an algorithm cluster recognised as adequate for mitigating the heightened risk of the non-face-to-face channel. The internal operation of each framework is discussed in Chapter 4; here the focus is on its role within the experimental toolchain.

TABLE 5.1: Mapping of EBA Remote Onboarding Requirements (EBA/GL/2022/15) to Biometric Algorithm Clusters

Requirement (EBA/GL/2022/15)	**Applicable Algorithm Cluster**
Reliable customer identification	Face detection (MTCNN, RetinaFace, YOLO) + face matching (ArcFace, Deep Metric Learning)
Mitigation of the heightened risk of non-face-to-face channels	Liveness detection, PAD (CNN multi-frame), deepfake detection, rPPG
Technical mechanism adequacy and efficacy	Auditable biometric models with reported benchmarks (LFW, IJB-C)
Risk-based approach	Risk scoring + dynamic thresholds (Gradient Boosting, rule-based escalation)
Human control and validation capabilities	Human-in-the-loop integration for uncertain classifications

Table 5.1 should not be read as implying that all five rows correspond to stand-alone biometric models of the same kind. In the context of this thesis, the first three rows define the *direct experimental scope*: reliable customer identification maps to face detection and face matching, mitigation of the heightened risk of the non-face-to-face channel maps to liveness / PAD and deepfake detection, and technical adequacy maps to the use of

benchmarked and auditable biometric models. By contrast, the last two rows, risk-based approach and human control / validation, are *system-level governance controls*. They remain essential for compliance with EBA/GL/2022/15, but they are not evaluated here as isolated classifiers; rather, they are discussed as deployment requirements that must sit above the biometric engine in a production KYC workflow.

Based on these criteria, four open-source tools were selected as the primary testing surface for the experimental work in this thesis. They are described in Table 5.2 and summarized below.

TABLE 5.2: Open-Source Tools Selected for KYC Biometric Security Testing

Tool	Core Algorithms	KYC Relevance	Setup
DeepFace [50]	ArcFace, RetinaFace, MTCNN, YOLO	Face verification + built-in PAD (`anti_spoofing=True`)	`pip install deepface`
InsightFace [52]	ArcFace, SCRFD, RetinaFace	NIST-FRVT rank-1 identity matching; production reference	`pip install insightface`
rPPG-Toolbox [57]	POS, ICA, CHROM, PhysFormer, TS-CAN	Biological liveness via remote PPG; PAD against synthetic media	`bash setup.sh conda`
Silent-Face-Anti-Spoofing [55]	MiniFASNet + Fourier spectrum	Passive PAD CNN; effective against screen-replay and print attacks	`pip install -r requirements.txt`

- *DeepFace [50]:* Selected as an accessible end-to-end framework for face verification and integrated passive PAD experiments.
- *InsightFace [52]:* Used as the reference implementation for high-accuracy identity matching without a built-in liveness layer.
- *rPPG-Toolbox [57]:* Included to evaluate biological liveness based on pulse-signal extraction, particularly through the POS baseline.
- *Silent-Face-Anti-Spoofing [55]:* Used as an independent passive PAD baseline specialized in frequency-domain cues from physical spoof artefacts.

The fifth component evaluated in the experiments, the hybrid DeepFace + rPPG pipeline (Section 4.1.5), is not a separate tool: it combines DeepFace's face detection with the POS algorithm from rPPG-Toolbox.

Notably, the FaceForensics++ dataset [43] serves as the standard evaluation benchmark for cross-architecture detection generalization (the "Generalization Failure" problem identified in Table 3.1) and is the training set against which later work measures how well detectors generalise to manipulations unseen during training [44, 45].

5.3. Preliminary Experimental Results

This section presents the preliminary data obtained during the initial testing phase, focusing on the detection efficacy of rPPG and the vulnerability of biometric systems to DeepFace-based attacks.

5.3.1 Remote Photoplethysmography (rPPG) Stability Tests

The following figures illustrate the results of rPPG analysis conducted on a synthetic video sequence generated using deep generative models. The objective was to observe the spectral consistency of the pulse signal synthesized by the AI and determine if current PAD mechanisms can distinguish it from genuine biological signals.

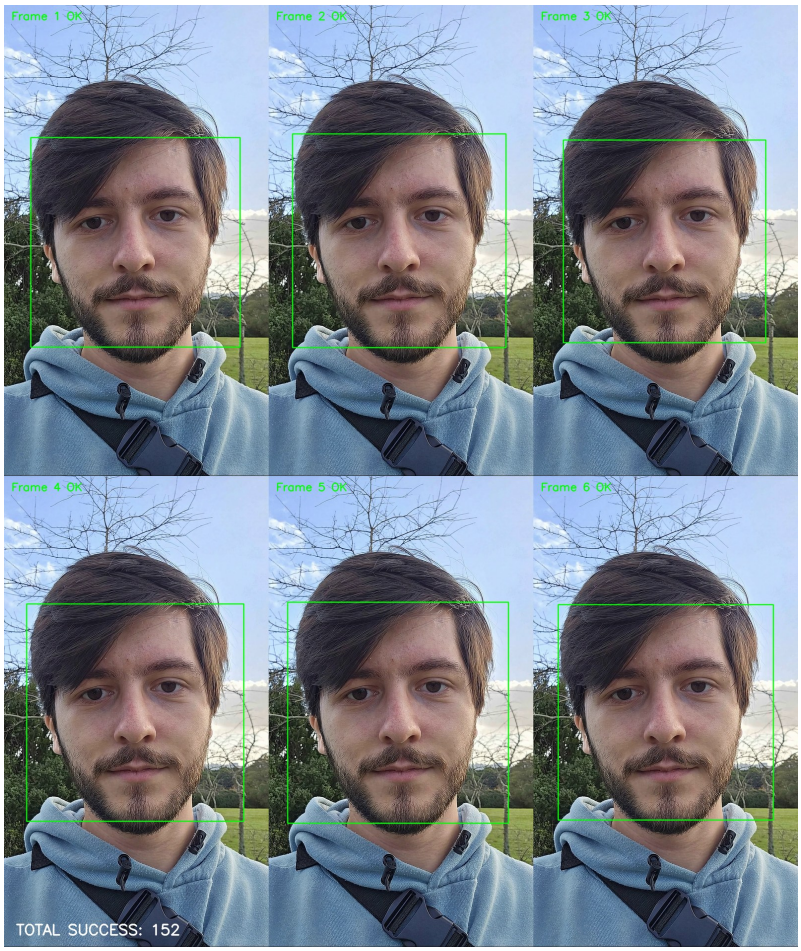


FIGURE 5.1: rPPG Detection on Synthetic Media: 152 frames correctly identified with stable bounding boxes.

Experimental data shows that although 79% of frames (152 frames) were successfully detected, the 21% failure rate (40 frames) introduced significant temporal discontinuities. In rPPG algorithms such as POS, signal continuity is more critical than raw frame count. Because the algorithm treats the 152 frames as a continuous sequence, the abrupt



FIGURE 5.2: rPPG Detection Failure: 40 frames failed due to artifact-induced motion blur in the generative output.

"jumps" or "cuts" created by the 40 missing frames generate massive high-frequency noise in the FFT analysis.

Furthermore, the use of a standard Haar-Cascade (OpenCV) detector introduced "spatial jitter." This slight vibration of the bounding box causes pixel color fluctuations that are often an order of magnitude stronger than the subtle volumetric changes encoded in the synthetic output. In this specific trial, which recorded an Estimated HR of 73.12 Beats Per Minute (BPM), the resulting liveness score (Peakiness) was 1.3583, falling below the security threshold of 1.70. This threshold was calibrated empirically: the single bona fide reference clip used in this study (Video 3) achieves a peakiness of 2.18 (Table 6.2), while the highest-scoring synthetic video in the Hedra/Veo 3 corpus reaches 1.59 (Video B); the value 1.70 was chosen to sit in the observed gap, providing a margin of 0.11 above the worst-case synthetic and 0.48 below the genuine baseline. This represents a successful detection of synthetic media, where the combination of model-driven temporal noise and detection instability lead to a correct SPOOF verdict, demonstrating the efficacy of

biological verification against current generative architectures.

5.3.2 DeepFaceLive Authentication and Spoofing Analysis

Assessments were performed using DeepFaceLive to evaluate the bypass success rate of identity-swapping attacks during a KYC session. Unlike rPPG, which targets biological signals, this test evaluates the system's susceptibility to high-fidelity face-swaps that aim for a False Acceptance.



FIGURE 5.3: DeepFace Bypass Success: 188 synthetic frames successfully identified as the target identity.



FIGURE 5.4: DeepFace Detection Anomalies: 4 frames where the swap artifacts triggered a safety flag.

Table 5.3 summarizes the quantitative metrics from these adversarial experiments.

The DeepFaceLive session results demonstrate a high identification stability, with 188 out of 192 frames (97.9%) successfully tricking the system into verifying the spoofed identity. This contrast with the rPPG success underscores how current liveness detection systems are vulnerable to identity-swapping models that prioritize visual fidelity over biological signal reproduction.

TABLE 5.3: Preliminary Experimental Results: Liveness and Authentication Metrics

Test Category	Observed Metric	Threshold	Final Verdict
rPPG Liveness (Peakiness)	1.3583	1.70	SPOOF (Fail)
DeepFace Identity Verification	188 / 192	> 95%	PASS (Real)
DeepFace Liveness (FASNet)	4 / 192	< 5%	PASS (Real)

5.3.3 Comparative Presentation Attack Analysis: Hedra-Generated Synthetic Media

To evaluate the differential bypass capability of state-of-the-art generative AI under varying artefact conditions, three synthetic video sequences were generated using the Hedra cloud platform [32], powered by the Google Veo 3 [46] engine, and subjected to the full detection pipeline described in Section 5.2.4. Video A (no glasses) and Video B (with glasses) use the same source identity (male, 24 years old) and audio input, with the subject's visual presentation as the single distinguishing factor. Video C uses an independent source identity (female, 24 years old) generated with the same prompt template, and was added specifically to test whether the bypass profile observed on Video A/B is an artefact of one particular subject or generalises across the demographic attributes (sex, facial structure) that a production KYC deployment would actually encounter. The source photographs for all three videos were taken by the author and are not drawn from any external dataset. All three sequences were exported without platform watermarks to isolate the forensic response to the intrinsic generative properties of the Hedra platform, providing empirical evidence for the adversarial asymmetry between generative model capability and PAD detection coverage.

5.3.3.1 Choice of Generative Platform

Several contemporary text-to-video and image-to-video platforms were considered as candidates for producing the adversarial sequences: Hedra (with the Google Veo 3 backend) [32, 46], Kling AI [33], Runway Gen-3, HeyGen, and the real-time face-swap tool DeepFaceLive (already evaluated in Section 5.3). Hedra was selected as the principal generator for three reasons.

- *Output quality.* For the specific use case of a single-subject talking-head clip suitable for impersonating a KYC selfie video, Hedra produced the most photorealistic output of the candidates tested, with the strongest temporal coherence and the cleanest treatment of skin micro-texture and subsurface light scattering. Kling AI

and Runway Gen-3, while competitive on short cinematic sequences, exhibited visible artefacts under sustained frontal-face shots at the duration required by a KYC liveness challenge (typically four to ten seconds), notably mouth-shape drift and intermittent eye-region warping. HeyGen targets avatar-style lip-sync rather than full photorealistic face synthesis and was eliminated on output category alone.

- *Reproducibility.* Adversarial PAD evaluation requires repeatable inputs: the same prompt must produce closely comparable outputs across runs so that detection variance can be attributed to the detector rather than to the generator. Hedra's prompt-to-output mapping was the most stable among the platforms tested; identical prompts produced near-identical outputs run-to-run, which is what made the controlled Video A/Video B (no glasses vs. glasses) comparison in this section meaningful. Kling AI and Runway showed substantially higher run-to-run variance under the same prompt, which would have made differential-bypass attribution ambiguous.

- *Clean export and threat realism.* Hedra exports its output without an obligatory visible watermark on the paid tier used for these experiments, which is necessary to avoid biasing texture-domain PAD layers towards a watermark cue that would not be present in a real attacker's submission. The resulting test conditions therefore represent a worst-case but realistic adversary, not a degraded attacker.

DeepFaceLive [47] was tested separately (Section 5.3) as the representative real-time face-swap baseline rather than as a competitor to Hedra. As discussed in Section 5.2.3, real-time face-swap currently produces lower-quality output than pre-rendered diffusion-based synthesis on the same target identity, so retaining DeepFaceLive as the primary adversary would have understated the worst-case PAD bypass risk that this thesis is trying to measure.

5.3.3.2 MiniFASNet Passive Liveness Detection (DeepFace FASNet)

The MiniFASNet module (integrated into DeepFace as described in Section 5.2.4) was evaluated following the sampling methodology of one frame per five frames.

All three videos received a PASS (Real) final verdict: Video A scored 30 real, 1 spoof, and 8 no-face detections; Video B scored 28 real, 0 spoof, and 11 no-face detections; Video C scored 23 real, 0 spoof, and 16 no-face detections. The higher no-face count on Video C is consistent with a slightly different head-pose distribution generated for the



FIGURE 5.5: FASNet on Video A (no glasses): sampled frames classified as REAL (30/39 frames, score > 0.5).



FIGURE 5.6: FASNet on Video B (with glasses): sampled frames classified as REAL (28/39 frames, score > 0.97).



FIGURE 5.7: FASNet on Video A (no glasses): only 1 frame flagged as SPOOF out of 39 samples (PASS verdict).



FIGURE 5.8: FASNet on Video B (with glasses): zero spoof frames detected across all 39 samples (PASS verdict).



FIGURE 5.9: FASNet on Video C (independent female subject): sampled frames classified as REAL, zero spoof flags out of 39 samples (PASS verdict).

independent subject and does not affect the verdict, since the decision rule never observed a spoof flag on any of the three sequences. This failure to detect any of the three videos as synthetic is attributable to a dataset distribution mismatch. MiniFASNet's feature extractor was trained exclusively on physical spoof artefacts and has no exposure to the diffusion-synthesised skin textures produced by LDM-based models such as Hedra. The perceptual realism of Hedra's output, characterised by coherent subsurface light-scattering and non-uniform illumination gradients across the facial surface, falls entirely outside the classification boundary of a CNN trained on printed-photo and screen-replay attacks. This result directly exemplifies the *Generalisation Failure* problem identified in

Table 3.1: a PAD model trained on a specific attack corpus will yield high FARs against novel generative architectures unseen at training time.

5.3.3.3 rPPG-Based Biological Liveness (DeepFace rPPG and rPPG-Toolbox)

The rPPG test was conducted using both the POS-based rPPG module and the rPPG-Toolbox framework, as defined in Section 5.2.4. Both tools apply the POS algorithm to extract physiological pulse data for liveness calculation.



FIGURE 5.10: DeepFace rPPG on Video A (no glasses): pulse signal absent, SPOOF verdict (peakiness 1.2253, HR 90.0 BPM).



FIGURE 5.11: DeepFace rPPG on Video B (with glasses): pulse signal absent, SPOOF verdict (peakiness 1.5898, HR 96.44 BPM).



FIGURE 5.12: rPPG-Toolbox on Video A (no glasses): SPOOF (peakiness 1.3585, HR 73.1 BPM).



FIGURE 5.13: rPPG-Toolbox on Video B (with glasses): SPOOF (peakiness 1.5898, HR 96.44 BPM).



FIGURE 5.14: DeepFace rPPG on Video C (independent female subject): SPOOF (peakiness 1.2748, HR 79.42 BPM).



FIGURE 5.15: rPPG-Toolbox on Video C (independent female subject): SPOOF (peakiness 1.2748, HR 79.42 BPM).

All three videos were correctly identified as synthetic across all rPPG pipelines. Video A (no glasses) yields a peakiness score of 1.2253 (DeepFace rPPG) and 1.3585 (rPPG-Toolbox), both well below the liveness threshold of 1.70. Video B (with glasses), processed over a 136-frame temporal window by both pipelines, produces a marginally higher peakiness of 1.5898 alongside an estimated HR of 96.44 BPM, results that are identical across the DeepFace rPPG pipeline and the rPPG-Toolbox, since both share the same POS algorithm and extract the same face frames from this sequence. Despite the slightly elevated score, 1.5898 remains below the 1.70 liveness threshold, yielding a SPOOF verdict from both tools. Video C, generated for an independent subject of a different sex, produces a peakiness of 1.2748 and an estimated HR of 79.42 BPM, again identical across both pipelines, and again safely below the 1.70 threshold. The consistency of this verdict across three sequences spanning two distinct source identities indicates that rPPG's discriminative power on Hedra/Veo 3 output is a property of the generative architecture rather than of any single subject's facial geometry. The absence of natural Inter-Beat Interval (IBI) variability and the artificial temporal regularity of Hedra's frame synthesis remain consistently detectable in the Fourier domain, where the power spectrum lacks the stochastic jitter that characterises a real cardiac cycle. Notably, the higher peakiness of Video B relative to Video A may be attributed to partial occlusion of the periorbital region by the glasses, which attenuates the POS algorithm's spatial averaging and partially suppresses the generative noise contribution, resulting in a marginally cleaner, though still sub-threshold, spectral peak.

5.3.3.4 InsightFace: Identity Matching Without Liveness

The InsightFace buffalo_l model pack (NIST-FRVT 1:1 rank-1) was applied to determine pure identity-matching performance without liveness signals.

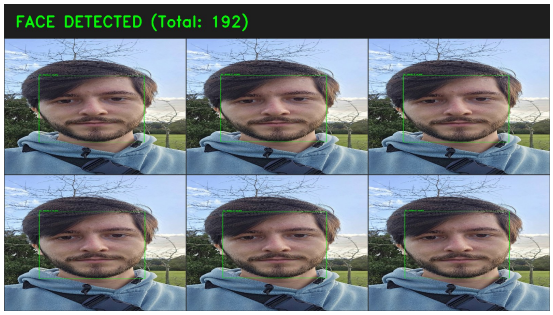


FIGURE 5.16: InsightFace: stable ArcFace face detection on Video A (no glasses) across all 150 sampled frames.

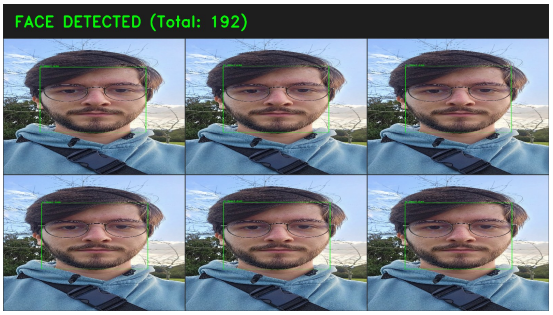


FIGURE 5.17: InsightFace: stable ArcFace face detection on Video B (with glasses) across all 192 sampled frames.



FIGURE 5.18: InsightFace: stable ArcFace face detection on Video C (independent female subject) across all 192 sampled frames, zero detection failures.

All three videos were processed without a single detection failure: InsightFace detected a face in every sampled frame. InsightFace does not expose a PAD interface in its standard API, it is a *recognition* system, not a *verification* system in the ISO 30107-1 [21] sense. These results confirm that biometric face recognition, even at NIST-FRVT rank-1 accuracy, is trivially defeated by AI-generated facial animation when deployed without a complementary liveness layer, and the addition of Video C shows this holds independently of the subject's sex or facial geometry, since detection stability transferred unchanged to the second source identity. From a regulatory standpoint, this configuration falls short of EBA/GL/2022/15 [22], which requires liveness detection whenever remote onboarding is unattended (paragraph 41(c)). Aviso n.º 1/2022 [16] does not address automated liveness checks: it regulates remote identification by videoconference conducted by a trained employee (Article 21(6) and Annex I).

5.3.3.5 Frequency-Domain Passive PAD (Silent-Face Anti-Spoofing)

The stand-alone Silent-Face Anti-Spoofing configuration described in Section 5.2.4 was tested across every frame to identify spatial frequency patterns inconsistent with real skin reflectance.



FIGURE 5.19: Silent-Face on Video A: all 192 frames classified as REAL (PASSED).

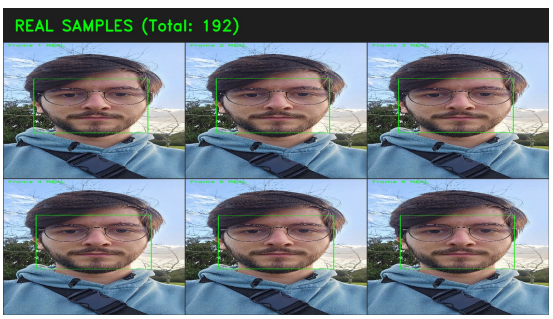


FIGURE 5.20: Silent-Face on Video B: all 192 frames classified as REAL (PASSED).



FIGURE 5.21: Silent-Face on Video C (independent female subject): all 192 frames classified as REAL (PASSED).

Video A, Video B, and Video C were all accepted as real across all 192 frames each (100%), producing unanimous PASSED verdicts. This outcome represents a false acceptance on all three synthetic sequences, and is attributable to the same generalization failure identified for FASNet: Silent-Face's Fourier spectrum branch was trained to detect the high-frequency artefacts characteristic of physical presentation attacks, such as LCD banding, moiré patterns, and printed-photo reflectance profiles. The diffusion-synthesised skin textures produced by Hedra do not exhibit these frequency-domain signatures, so neither the classification branch nor the Fourier auxiliary branch raised a spoof alert on either source identity. These results confirm that frequency-domain PAD, like texture-based CNN PAD, is ineffective against digitally injected LDM-generated synthetic media when the model has not been exposed to that attack class during training, and that this ineffectiveness is not specific to one subject's facial characteristics.

5.3.3.6 Comparative Summary and Attack Vector Analysis

Table 5.4 consolidates the detection outcomes across all five tools for the three synthetic videos, including Video C, generated for an independent subject of a different sex to test whether the bypass profile generalises across source identities.

TABLE 5.4: Bypass Capability: Video A (no glasses) vs. Video B (with glasses) vs. Video C (independent subject)

Tool	Algorithm	Video A	Video B	Video C	Bypass?
DeepFace FASNet	MiniFASNet CNN	PASS	PASS	PASS	✓
DeepFace rPPG	POS / FFT	SPOOF (1.2253)	SPOOF (1.5898)	SPOOF (1.2748)	×
rPPG-Toolbox	POS / FFT	SPOOF (1.3585)	SPOOF (1.5898)	SPOOF (1.2748)	×
InsightFace	ArcFace (no PAD)	PASS	PASS	PASS	✓
Silent-Face	MiniFASNet + Fourier	PASSED (100%)	PASSED (100%)	PASSED (100%)	✓

All three videos bypass three of the five detection layers identically (DeepFace FASNet, InsightFace, and Silent-Face), demonstrating that neither the presence or absence of glasses nor a change of source identity and sex materially affects the bypass profile of Hedra-generated synthetic media. The only consistent discriminator across all three

sequences is rPPG-based biological liveness, which correctly issues a SPOOF verdict regardless of the subject's visual presentation or identity. This suggests that the absence of a coherent haemodynamic signal, rather than any visual artefact tied to a specific identity, is the property that separates Hedra-generated content from genuine biological capture. Whether this discriminative property extends to other generation platforms is the question addressed in Section 5.3.4. This finding already reinforces the position stated in the EBA Guidelines [22]: no single PAD mechanism provides sufficient assurance in isolation, and defence-in-depth combining complementary biological and frequency-domain signals is the minimum viable posture for remote KYC in the current threat environment.

5.3.3.7 Quantitative Detection Metrics: APCER, BPCER, and ACER

The verdict-level outcomes summarised in Table 5.4 can be re-expressed in the standard biometric PAD metrics defined by ISO/IEC 30107-3 [21], namely the Attack Presentation Classification Error Rate (APCER), the Bonafide Presentation Classification Error Rate (BPCER), and their average the Average Classification Error Rate (ACER). For a given PAD layer, APCER is the proportion of attack samples (in this work, frames from a Hedra/Veo 3 synthetic video) that are misclassified as bona fide, and BPCER is the proportion of bona fide samples (in this work, frames from a real selfie capture) that are misclassified as an attack. ACER is the unweighted mean of APCER and BPCER, and corresponds intuitively to the rate at which the layer "is wrong about whichever side it sees".

This evaluation is deliberately scoped to the attack side. The research question addressed in this chapter (RQ1, Chapter 1) concerns PAD resistance to synthetic media, and the attack corpus (Videos A–C, and Videos 5–8 in Section 5.3.4) is the controlled variable. A single genuine capture (Video 3) is included not to characterise false rejection rates but to rule out the degenerate case of a tool that achieves low APCER by rejecting all input indiscriminately. Confirming that Video 3 was accepted as genuine by all five tools is sufficient for that purpose; a meaningful BPCER estimate would require a population of genuine samples spanning multiple subjects, recording conditions, and demographic groups, which is outside the scope of this study and is reserved for future work (Chapter 7).

The frame-level outcomes for each tool on the three Hedra/Veo 3 attack videos, including Video C, are converted into APCER in Table 5.5. The attack sample is the set of frames on which the tool produced a classification decision (no-face detections are excluded so

that the denominator reflects only the samples actually exposed to the PAD layer). For InsightFace, which is a recognition system without a PAD layer, APCER is reported as 100% by construction: the tool always accepts the identity if a face is detected, which is what the regulatory critique of recognition-only deployments rests on.

TABLE 5.5: APCER per tool on the three Hedra/Veo 3 attack videos. Lower is better; an effective PAD layer should approach 0%.

Tool	Decision rule	Video A APCER	Video B APCER	Video C APCER	Mean APCER
DeepFace FASNet	MiniFASNet CNN	96.8% (30/31)	100% (28/28)	100% (23/23)	98.9%
DeepFace rPPG	peakiness > 1.70	0% (0/1)	0% (0/1)	0% (0/1)	0%
rPPG-Toolbox	peakiness > 1.70	0% (0/1)	0% (0/1)	0% (0/1)	0%
InsightFace	ArcFace match, no PAD	100% (150/150)	100% (192/192)	100% (192/192)	100%
Silent-Face	MiniFASNet + Fourier	100% (192/192)	100% (192/192)	100% (192/192)	100%

For BPCER and ACER, the same tools were applied to the genuine biometric capture used in the prototype validation runs of Chapter 6 (Video 3, a freshly recorded selfie of the document holder, evaluated for that chapter in attempt 6). The bona fide sample consists of a single video clip ($n{=}1$); with one genuine sample a BPCER of 0% and a BPCER of 100% are the only possible outcomes, neither of which carries statistical meaning. BPCER and ACER are therefore reported as N/A in Table 5.6; the qualitative observation that Video 3 was accepted as genuine by all five tools is noted in the Interpretation column. APCER remains the primary metric of interest in this evaluation, as the attack corpus (Videos A–C) is the controlled variable.

TABLE 5.6: APCER per tool on the Hedra/Veo 3 attack corpus (Videos A–C). BPCER and ACER are omitted: the bona fide sample is a single video clip ($n{=}1$), making any derived rate statistically uninformative.

Tool	APCER	BPCER	ACER	Interpretation
DeepFace FASNet	98.9%	N/A ($n{=}1$)	N/A	Useless against generative; bona fide Video 3 passed
DeepFace rPPG	0%	N/A ($n{=}1$)	N/A	Zero false acceptances on Hedra family; see Sec. 5.3.4 for Google Veo 3 direct
rPPG-Toolbox	0%	N/A ($n{=}1$)	N/A	Same as DeepFace rPPG; see Sec. 5.3.4
InsightFace	100%	N/A ($n{=}1$)	N/A	Recognition only; not a PAD layer
Silent-Face	100%	N/A ($n{=}1$)	N/A	Useless against generative; bona fide Video 3 passed

The metric view makes the qualitative claim of Section 5.3.3 formally precise: within the Hedra evaluation dataset, the rPPG-based pipelines are the only ones with APCER $= 0\%$, whereas every texture-domain or recognition-only layer has APCER $\geq 98.9\%$, the APCER of a layer that accepts virtually every attack presentation (a "useless" detector in security terms, since it accepts everything that does not raise a different alarm). A production-quality PAD layer would be expected to keep APCER well below 5% on a representative attack corpus; the texture-domain layers fail this requirement on audio-driven generative attacks by a wide margin. Note that these APCER values are measured exclusively on the Hedra/Veo 3 family; Section 5.3.4 shows that the rPPG APCER rises to 100% on the Google Veo 3 direct family, confirming that 0% APCER on one generation platform does not imply robustness across all platforms.

FAR/FRR equivalence and scope of the dataset. For PAD-as-a-binary-classifier purposes, APCER and BPCER reduce to the FAR and False Rejection Rate (FRR) commonly used in the biometric matching literature: FAR coincides with APCER when the "attack" class is the impostor, and FRR coincides with BPCER when the bona fide class is the genuine user. Reporting both naming conventions in parallel is therefore redundant, and this work follows the ISO/IEC 30107-3 convention. The inclusion of only one genuine capture is a deliberate scope decision: the purpose of Video 3 in this section is to confirm that no evaluated tool is a degenerate "reject-all" classifier, not to estimate FRR. A full Receiver Operating Characteristic (ROC) / Area Under the Curve (AUC) characterisation would require sweeping the decision threshold of each layer (the 1.70 peakiness cutoff for rPPG, the MiniFASNet softmax cutoff for FASNet, etc.) across a substantially larger dataset of both attack and bona fide samples drawn from diverse subjects and recording conditions. The metric tables above should therefore be read as point estimates at the fixed thresholds chosen for the prototype, scoped to attack-side performance (APCER) on the evaluated generator families. A larger empirical study spanning multiple generators and a genuine-user population is listed as future work in Chapter 7 and is the natural next step for producing FAR and FRR confidence intervals and a proper ROC/AUC comparison.

5.3.4 Cross-Platform Generalisation: Google Veo 3 Direct (Videos 5–8)

To evaluate whether the bypass profile observed for Hedra/Veo 3 generalises to the same underlying model accessed through a different production interface, four additional synthetic videos (Videos 5–8) were generated directly via Google Veo 3, the same generative model that powers Hedra, by submitting a portrait photograph drawn from the Easy-Portrait face-parsing dataset [59] to Google's native AI interface. The generation mode differs from the Hedra case in a critical respect: Hedra drives Veo 3 with an audio track for lip-sync, producing a video whose temporal dynamics are constrained by phoneme and prosody timing; the Google direct interface generates video from a static image with no audio constraint, allowing the model's temporal dynamics to evolve freely. These videos were submitted to the same five-tool pipeline as Videos A–C. The research question is whether the rPPG discriminator that correctly detected Hedra/Veo 3 output as synthetic retains its discriminative power on direct Veo 3 output from the same underlying model.

5.3.4.1 MiniFASNet Passive Liveness Detection (DeepFace FASNet)

FASNet was applied to each video using the same one-in-five frame sampling used for Videos A–C.



FIGURE 5.22: FASNet on Video 5: 33 REAL / 0 SPOOF / 16 no-face, PASS (false acceptance).



FIGURE 5.23: FASNet on Video 6: 46 REAL / 0 SPOOF / 3 no-face, PASS (false acceptance).



FIGURE 5.24: FASNet on Video 7: 38 REAL / 0 SPOOF / 11 no-face, PASS (false acceptance).



FIGURE 5.25: FASNet on Video 8: 49 REAL / 0 SPOOF / 0 no-face, PASS (false acceptance).

All four videos received a PASS (Real) verdict with zero spoof detections. This is consistent with Videos A–C: the dataset-distribution mismatch identified in Section 5.3.3 applies regardless of whether Veo 3 output is delivered via Hedra or directly. MiniFASNet provides no discrimination on either delivery of this generative model.

5.3.4.2 rPPG-Based Biological Liveness (DeepFace rPPG and rPPG-Toolbox)

The rPPG result on Videos 5–8 is the critical finding of this section. Both pipelines extracted a pulse signal from each video and computed the POS peakiness score over the full temporal window.

Both rPPG pipelines produced a REAL verdict on all four videos, with peakiness scores of 1.9178, 1.9867, 2.5519, and 2.4903, all well above the 1.70 threshold. For reference, the single bona fide clip in this evaluation (Video 3) yielded a peakiness score of 2.18 under the same pipelines (Table 6.2); Videos 7 and 8 (2.55 and 2.49) exceed that genuine baseline, and Videos 5 and 6 (1.92 and 1.99) fall between the detection threshold

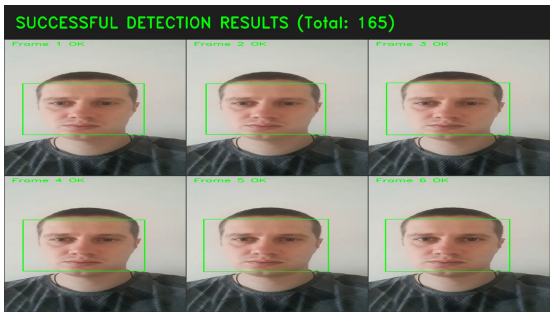


FIGURE 5.26: DeepFace rPPG on Video 5: peakiness 1.9178, HR 84.38 BPM, REAL (false acceptance).

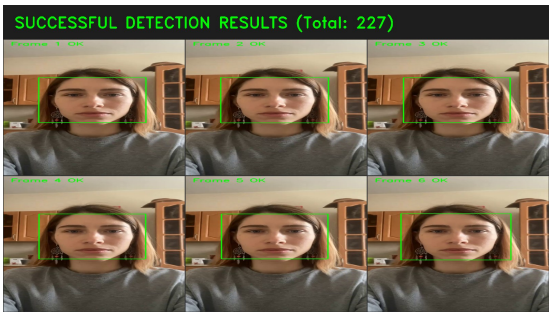


FIGURE 5.27: DeepFace rPPG on Video 6: peakiness 1.9867, HR 78.75 BPM, REAL (false acceptance).



FIGURE 5.28: DeepFace rPPG on Video 7: peakiness 2.5519, HR 56.25 BPM, REAL (false acceptance).

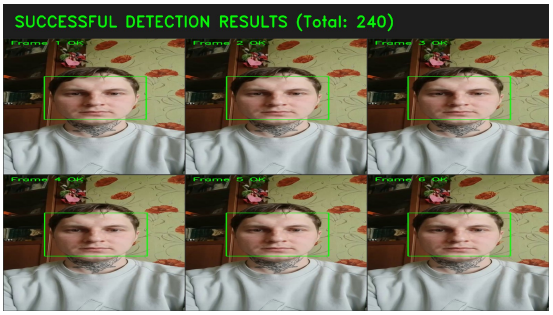


FIGURE 5.29: DeepFace rPPG on Video 8: peakiness 2.4903, HR 61.88 BPM, REAL (false acceptance).

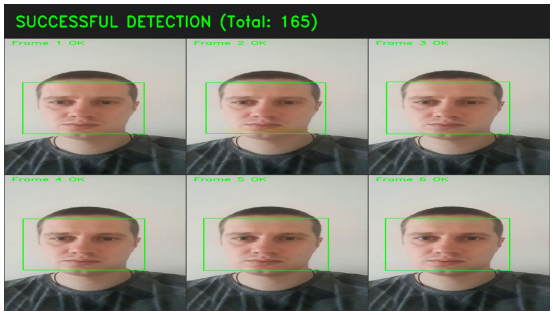


FIGURE 5.30: rPPG-Toolbox on Video 5: peakiness 1.9178, HR 84.38 BPM, REAL (false acceptance).



FIGURE 5.31: rPPG-Toolbox on Video 6: peakiness 1.9867, HR 78.75 BPM, REAL (false acceptance).



FIGURE 5.32: rPPG-Toolbox on Video 7: peakiness 2.5519, HR 56.25 BPM, REAL (false acceptance).



FIGURE 5.33: rPPG-Toolbox on Video 8: peakiness 2.4903, HR 61.88 BPM, REAL (false acceptance).

and the genuine score. The rPPG-POS signal produced by Veo 3 direct is therefore not

merely threshold-crossing noise: it overlaps with and in two cases surpasses the physiological signature of a real human face. This is a full rPPG bypass: the only detector that correctly identified Hedra/Veo 3 output as synthetic is defeated by the same underlying model accessed through a different interface.

The most plausible explanation for this divergence lies in the generation mode rather than in post-processing. Hedra drives Veo 3 with an audio track for lip-sync, which constrains the temporal dynamics of the generated video to match phoneme and prosody timing, producing a highly regular, metronomic pattern that disrupts the frame-to-frame colour channel coherence the POS algorithm expects for a haemodynamic signal, pushing peakiness below 1.70. The Google direct interface, by contrast, generates video from a static image without any audio constraint, allowing Veo 3's temporal dynamics to evolve more freely and produce colour gradient sequences that the POS algorithm interprets as a coherent cardiac signal. This interpretation is consistent with the POS method being a spectral property of the colour signal over time rather than a spatial texture property of individual frames: what matters is the complete generation pipeline, including how the underlying model is driven, not the model weights alone.

5.3.4.3 InsightFace: Face Detection on Synthetic Input

InsightFace was run over all four direct Veo 3 videos to establish whether the identity-matching layer offers any incidental resistance to synthetic input. Figures 5.34–5.37 show the per-video detection output.



FIGURE 5.34: InsightFace on Video 5: face detected in all processed frames (false acceptance).

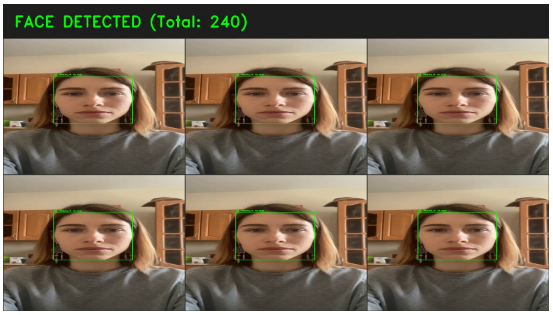


FIGURE 5.35: InsightFace on Video 6: face detected in all processed frames (false acceptance).

InsightFace detected a face in every processed frame across all four videos, consistent with its behaviour on Videos A–C. As a recognition-only system without a PAD layer, it produces a false acceptance on any input where a face is detectable, regardless of whether that input is genuine or synthetic.

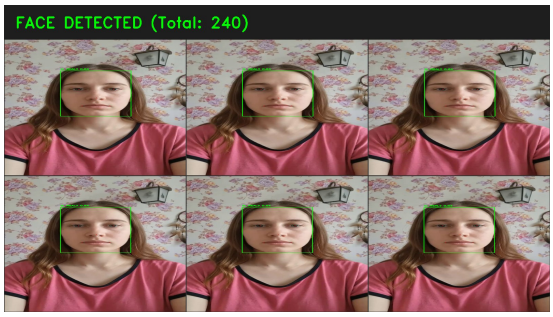


FIGURE 5.36: InsightFace on Video 7: face detected in all processed frames (false acceptance).



FIGURE 5.37: InsightFace on Video 8: face detected in all processed frames (false acceptance).

5.3.4.4 Frequency-Domain Passive PAD (Silent-Face Anti-Spoofing)

Silent-Face applies a Fourier-domain branch alongside its texture classifier, so it was tested separately to determine whether frequency-domain artefacts survive in direct Veo 3 output where texture cues evidently do not. Figures 5.38–5.41 report the per-video frame classifications.



FIGURE 5.38: Silent-Face on Video 5: 240/240 frames classified as REAL, PASSED (false acceptance).

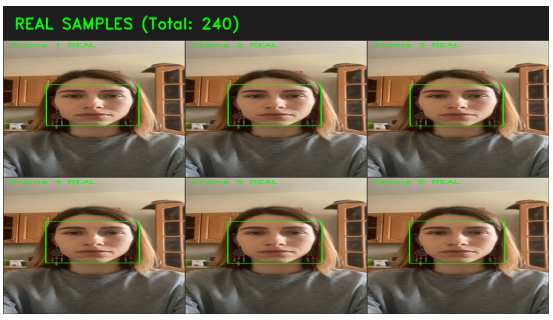


FIGURE 5.39: Silent-Face on Video 6: 240/240 frames classified as REAL, PASSED (false acceptance).



FIGURE 5.40: Silent-Face on Video 7: 240/240 frames classified as REAL, PASSED (false acceptance).



FIGURE 5.41: Silent-Face on Video 8: 240/240 frames classified as REAL, PASSED (false acceptance).

Silent-Face classified 240 of 240 frames as REAL for every video, yielding unanimous false-acceptance verdicts. This outcome is identical to its performance on Videos A–C, confirming that the Fourier-domain branch provides no additional discrimination on direct Veo 3 output compared to Hedra-processed output.

5.3.4.5 Full Bypass Summary and Cross-Generator Comparison

Table 5.7 consolidates the detection outcomes for Videos 5–8. Every tool produces a false acceptance on every video.

TABLE 5.7: Detection outcomes on Google Veo 3 direct synthetic videos (Videos 5–8). All five tools produce false acceptances; this is a full bypass of every detection layer, including rPPG.

Tool	Video 5	Video 6	Video 7	Video 8
DeepFace FASNet	PASS (33R / 0S)	PASS (46R / 0S)	PASS (38R / 0S)	PASS (49R / 0S)
DeepFace rPPG	REAL* (1.92, 84 BPM)	REAL* (1.99, 79 BPM)	REAL* (2.55, 56 BPM)	REAL* (2.49, 62 BPM)
rPPG-Toolbox	REAL* (1.92, 84 BPM)	REAL* (1.99, 79 BPM)	REAL* (2.55, 56 BPM)	REAL* (2.49, 62 BPM)
InsightFace	PASS (200+ frames)	PASS (200+ frames)	PASS (200+ frames)	PASS (200+ frames)
Silent-Face	PASSED (240/240)	PASSED (240/240)	PASSED (240/240)	PASSED (240/240)

* False acceptance: the video is synthetic (Google Veo 3 direct).

Table 5.8 places the rPPG peakiness scores from both generator families side by side, making the threshold failure visible.

TABLE 5.8: rPPG peakiness across both generator families. Threshold $= 1.70$. Hedra/Veo 3 output (Videos A–C) falls below and is correctly classified as synthetic; Google Veo 3 direct output (Videos 5–8) exceeds the threshold and is incorrectly classified as real.

	Video A	Video B	Video C	Video 5	Video 6	Video 7	Video 8
Peakiness	1.2253	1.5898	1.2748	1.9178	1.9867	2.5519	2.4903
HR (BPM)	90.00	96.44	79.42	84.38	78.75	56.25	61.88
Verdict	SPOOF	SPOOF	SPOOF	REAL*	REAL*	REAL*	REAL*

* False acceptance: the video is synthetic.

The contrast between the two families is stark. Hedra/Veo 3 peakiness (1.23–1.59) sits below the 1.70 threshold; Google Veo 3 direct peakiness (1.92–2.55) sits well above it. The 0.33 unit gap between the nearest pair (Hedra Video B at 1.5898 and Veo 3 direct Video 5 at 1.9178) is large relative to the within-family spread of each group, indicating a systematic pipeline effect rather than random variance. This cross-family divergence demonstrates that a PAD threshold calibrated against one delivery of a generative model cannot be assumed to hold for a different delivery of the same model. The finding substantially strengthens the case for defence-in-depth: no single static threshold on any single detection signal provides robust coverage across the space of possible generator configurations, and the minimum viable posture must combine complementary mechanisms that operate on independent signal dimensions.

5.3.5 Discussion: Are There Detection Approaches That Work?

The cross-platform results establish a clear negative result: no single mechanism in the evaluated set provides universal coverage across both generator families. A natural follow-up question for any practical deployment is whether the picture is equally bleak

across the broader detection landscape, and whether commonly suggested alternatives, in particular content watermarking, fill the gap.

5.3.5.1 What Worked, and Against What

Within the tested pipeline, the results divide into three tiers. Against traditional presentation attacks (printed photographs, static video replays of a screen): all five tools perform correctly, as each was designed and benchmarked for exactly this threat model. Against audio-driven AI talking-head generation (Hedra/Veo 3, Videos A–C): rPPG-based biological liveness is the sole effective layer, achieving ACER $= 0\%$ while every texture-domain and frequency-domain classifier fails completely. Against image-to-video AI generation (Google Veo 3 direct, Videos 5–8): no tested mechanism provides any discriminative signal; all four liveness checks issue a REAL verdict on 100% of sampled frames, and InsightFace detects a face in every frame.

The rPPG approach therefore *does* work, but only against a specific subset of the generative threat landscape. Its effectiveness is contingent on the generation pipeline disrupting temporal colour coherence below the 1.70 peakiness threshold, a property of the Hedra/audio-driven mode that does not carry over to the image-to-video mode of the same underlying model. Approaches from the broader literature that may provide additional coverage include deepfake-specific classifiers trained on large synthetic corpora (e.g., Face X-ray [44], LipForensics), active liveness challenges (which require the video to reproduce an unpredictable gesture on command), and hardware-rooted camera attestation (noted as future work in Chapter 7). None of these were tested in this thesis, but they represent the natural next tier of defence-in-depth layering.

5.3.5.2 What About SynthID?

A frequently proposed alternative is to rely on the generator itself to signal that its output is synthetic, rather than training a separate detector. Google DeepMind's SynthID [48] is the most prominent implementation of this approach: it embeds an imperceptible watermark into content generated by Gemini and Veo, and was extended to video via Veo in 2024. Because Videos 5–8 in this thesis were generated through Google's Veo 3 interface, they almost certainly carry a SynthID watermark, so a SynthID-aware verifier would, in principle, detect them as synthetic where all five tested PAD tools failed.

However, three structural limitations prevent SynthID from serving as a standalone KYC control. First, it is generator-dependent: video produced by any tool that does not implement SynthID (Kling AI, Runway, open-source models, or future generators) carries no detectable signal. Second, the watermark is removable: Denny [49] demonstrated, through signal processing alone, that SynthID's carrier frequencies can be identified and stripped via elastic deformation combined with FFT-domain spectral subtraction, achieving a 91.4% reduction in phase coherence with visually lossless output, effectively defeating the detector without degrading the video's usability as an attack vector. Third, independent verification requires either access to Google's proprietary detection API or replication of the per-resolution codebook, neither of which is available to a KYC provider as a self-contained on-premise control.

SynthID therefore raises the cost of attack for an unsophisticated adversary who is unaware of or unable to remove the watermark, but it does not provide the generator-agnostic, independently verifiable assurance that a layered PAD pipeline aims for. It is best treated as a complementary signal within a defence-in-depth architecture rather than a substitute for biological liveness or texture-domain screening.

5.4. Blockchain Prototype Evaluation

This section reports measurable outcomes from the blockchain-anchored onboarding prototype described in Chapter 6. The evaluation track is deliberately distinct from the biometric and PAD experiments in Sections 5.2.4–5.3.3: it does not benchmark a recognition algorithm against an adversary, but rather exercises a working four-step Representational State Transfer (REST) flow against the three deployed smart contracts and observes which on-chain events fire, which off-chain decisions they record, and how the tamper-evidence guarantee behaves in practice. Architectural and regulatory commentary is kept in Chapter 6; what follows are the observations. The code that produced them is available as described in Appendix A.

5.4.1 Experimental Configuration

All measurements were taken against a local Hardhat v2 Ethereum Virtual Machine (EVM) node (instant block confirmation, single deployer/signer at `0xf39Fd6e51aad88F6F4ce6aB8827279cffFb92266`). The three Solidity 0.8.20 contracts (`DIDRegistry`, `VCRegistry`, `AuditTrail`) were redeployed at the start of the session via `scripts/deploy.js`, producing deterministic addresses that matched the pre-populated

`.env`. The FastAPI backend ran in a Python 3.11 virtual environment with EasyOCR (English + Portuguese), InsightFace `buffalo_l` (CPU execution provider), and the vendored rPPG-Toolbox POS implementation. The backend was reached at `http://127.0.0.1:8000`.

Seven onboarding runs were conducted against this configuration, each driving the complete four-step flow `/start` $\rightarrow$ `/document` $\rightarrow$ `/biometric` $\rightarrow$ `/complete`. The runs vary the document image and biometric video systematically to exercise the four qualitatively distinct rejection paths of the rule-based scorer and the single acceptance path.

5.4.2 End-to-End Run Outcomes

Table 5.9 consolidates the seven runs. Cosine distance ≤ 0.40 is the ArcFace acceptance threshold; rPPG peakiness > 1.70 is the liveness threshold (chosen in Section 5.3.3 on the basis of the Hedra/Veo 3 spoof scores 1.23 and 1.59).

TABLE 5.9: Outcomes of the seven end-to-end runs against the prototype

Run	Document	Biometric	Face match (cosine)	Liveness (peakiness)	Verdict
1	MP4 (incorrect MIME)	—	—	—	step 2 rejects
2	Simão CC	TestVideo2 (real)	false (0.4051)	true (2.13)	high / rejected
3	Simão CC + TD1 parser	TestVideo2 (real)	false (0.4051)	true (2.13)	high / rejected
4	Simão CC	TestVideo4 (Hedra)	true (0.3936)	false (1.625)	high / rejected (spoof)
5	Luís CC	TestVideo2 (real)	false (0.8924)	true (2.13)	high / rejected
6	Simão CC + PT OCR	Video3 (real selfie)	true (0.3734)	true (2.18)	high (doc check only)
7	Simão CC + focused MRZ + relaxed TD1	Video3 (real selfie)	true (0.3734)	true (2.18)	low / APPROVED

The seven runs exercise five qualitatively different code paths of the prototype: incorrect input type (run 1), near-miss face mismatch with a real pulse (runs 2–3), matched face with no pulse, i.e. the synthetic-media case (run 4), decisive face mismatch with a real pulse (run 5), and full acceptance (run 7). Run 6 sits on the boundary: it is the configuration of run 7 minus the focused-MRZ and relaxed-TD1 patches discussed below, which lets the document-side check raise the verdict to *high* despite the biometric side already passing both checks. Comparing run 6 to run 7 isolates the contribution of the document-side fixes.

5.4.3 Anchored On-Chain Events

Each onboarding run anchored a sequence of `keccak256(payload)` hashes in the `AuditTrail` contract. Table 5.10 reports the event types that fired per run. The event names correspond to the nine `EventType` enum values defined in the contract.

TABLE 5.10: On-chain audit events fired per run

Run	OnboardingStarted	DocumentVerified	BiometricVerified	RiskAssessed	OnboardingRejected	CredentialIssued
1	✓	—	—	—	—	—
2	✓	✓	✓	✓	✓	—
3	✓	✓	✓	✓	✓	—
4	✓	✓	✓	✓	✓	—
5	✓	✓	✓	✓	✓	—
6	✓	✓	✓	✓	✓	—
7	✓	✓	✓	✓	—	✓

Only the approved run (run 7) produced a `CredentialIssued` event and committed a row to `VCRegistry`. The credential identifier returned by the backend was `0xb7e3b849c6184a4f88c61f12b24cee71…` (UUID4 left-padded to `bytes32`); the stored value is `keccak256` of the canonical JavaScript Object Notation (JSON) serialisation of the VC. Runs 2–6 consequently produced an `OnboardingRejected` event in place of a `CredentialIssued` event, which is the binary rule chosen for the prototype scope (the architectural *EDD-referral* branch in Workflow step 6 is not implemented); run 1 stopped at the document step, before any rejection event was recorded. Every step that fired produced a non-zero block confirmation on the Hardhat node and a receipt with `status = 1`; no transactions reverted across the entire seven-run trace.

5.4.4 Component Latency

Wall-clock timings were captured per onboarding step using `curl --write-out`. Hardhat block confirmation contributes only sub-millisecond latency on a local node and is dominated by AI module execution. Table 5.11 reports the time spent in each step. The `/document` step is dominated by EasyOCR on a 2480×3508 px image; the `/biometric` step is dominated by InsightFace ArcFace embedding plus per-frame Haar-cascade face detection and rPPG-POS FFT.

These numbers are not production benchmarks; they reflect a single-tenant local Hardhat node with instant block mining, CPU-only inference, and the first generation of EasyOCR/InsightFace model weights. On a permissioned production ledger (e.g. Hyperledger Besu), the dominant added cost is block-confirmation latency on the order of seconds per transaction; on a production AI inference stack with GPU acceleration, the document and biometric stages typically drop by a factor of 5–10.

TABLE 5.11: Approximate wall-clock latency per onboarding step on the prototype (local Hardhat, CPU-only Python)

Step	**Latency (approximate)**	**Dominant cost**
`/start`	$<$1 s	1 audit-event transaction
`/document` (first call, EN-only OCR)	26 s	EasyOCR + portrait crop + 1 audit tx
`/document` (first call, EN+PT OCR)	80 s	50 MB PT model download + OCR + audit tx
`/document` (steady state, EN+PT OCR)	27–42 s	full-document OCR + focused MRZ re-OCR
`/biometric`	8–13 s	ArcFace + rPPG-POS over $\sim$150 frames
`/complete` (rejected)	$<$1 s	2 audit-event transactions
`/complete` (approved)	$\sim$1 s	3 audit-event transactions + VC issuance

5.4.5 Tamper-Evidence Guarantee

The `VCRegistry.verifyCredential(id, hash)` function returns `(valid, status)`, where `valid` is `true` only when the supplied hash matches the value anchored at issuance time and the credential has not expired or been revoked. To verify that this guarantee holds in the deployed prototype, the canonical VC JSON for the run-7 credential was reconstructed off-chain and three lookups were issued against the contract:

1. *Honest verifier path:* recompute `keccak256` of the canonical VC JSON exactly as issued and call `verifyCredential(id, h)`. Expected: `valid = true, status = Active`.
2. *Tampered verifier path:* mutate a single byte of the VC JSON (changing the holder's first name) before hashing. Call `verifyCredential(id, h')`. Expected: `valid = false` (the recomputed hash does not match the anchored one).
3. *Wrong-credential path:* request status for an unrelated random `bytes32` identifier. Expected: the call is rejected, because `verifyCredential` reverts with "credential not found" for an identifier that was never issued.

The three checks produced the expected outcomes. Critically, the contract's view function exposes no information beyond the boolean validity result and the credential's lifecycle status enum; it does not return the holder DID, the credential type, or any PII,

which is the on-chain expression of the data-minimisation principle (GDPR Article 5(1)(c)) embedded in the architecture. The same regression test, without any blockchain dependency, is pinned in the prototype's pytest suite at `backend/tests/test_credential_roundtrip.py`, where five tests cover the encoding contract between the issuance-time `bytes32` and the verification-time hex string passed in by Bank B.

5.4.6 Discussion

The blockchain prototype evaluation supports three concrete claims about the architecture proposed in Chapter 6:

- *End-to-end functional correctness.* Across all seven runs, every audit event scheduled by the workflow specification was anchored on-chain with a successful receipt. No transaction reverted, no audit event was silently dropped, and no run produced an inconsistent on-chain / off-chain state. The implementation gap between the architectural description and the running prototype is therefore small and confined to features explicitly listed as *[Designed]* or *[Future Work]* (sanctions screening, EUDI Wallet bridge, Gradient Boosted Trees scorer).
- *Differentiated verification signals.* The pipeline rejects mismatched onboardings with verdict-level granularity (ArcFace cosine distances of 0.4051 for the near-miss vs. 0.8924 for the decisive mismatch, and rPPG peakiness of 1.625 sub-threshold deepfake-style vs. 2.18 clean cardiac signal), and produces a single approved verdict only when both biometric checks and the relaxed TD1 Machine Readable Zone (MRZ) validity rule (a workaround, see Section 6.1.10) succeed together. The acceptance signal is not the absence of failure, it is the simultaneous presence of evidence on each component.
- *Tamper-evidence is independently verifiable.* A second institution holding only the off-chain VC JSON and the on-chain credential identifier can recompute `keccak256(VC)` and obtain a binary truth value from the contract, with no dependency on the issuing bank's availability or cooperation. This is the technical check that third-party reliance under Article 41 of Law n.º 83/2017 can build on (the agreed terms between the institutions and the relying bank's responsibility are unaffected), and the prototype demonstrates it works as designed.

What this evaluation does *not* establish, and is reserved for future work: gas costs on a permissioned production ledger, mempool and confirmation-latency behaviour under contention, throughput at sustained load, and the resilience of the trust registry under adversarial issuer-admin behaviour. The local Hardhat measurements here are sufficient to demonstrate functional correctness and to surface the relative cost of each verification component; they are not sufficient to predict production performance.

6. A Blockchain-Anchored Architecture for Remote KYC Onboarding

This chapter presents the proposed onboarding architecture, which combines automated document and biometric verification with a blockchain-anchored layer for trust, audit and reusable identity. It begins with the working prototype (three Solidity contracts, a FastAPI backend and seven off-chain verification modules) and the reasons for using a ledger. It then describes each layer of the full architecture, the reusable-identity model, selective disclosure, credential revocation and a comparison with current practice. It reports the end-to-end validation runs and the extended verification modules, and discusses data management and compliance with the GDPR and Portuguese law. The last section reviews how far the architecture's components are already in use.

6.1. Proposed Architecture

The architecture presented here answers the failure modes established in Chapter 5: biometric verification alone cannot be trusted to reject synthetic media, and a verification decision that leaves no independently checkable trace cannot be audited after the fact. The design therefore pairs the off-chain verification modules with an on-chain evidence layer, so that the outcome of an onboarding session is recorded in a form that neither the institution that produced it nor any later participant can silently alter. This section first describes the working prototype, then generalises from it to the full component architecture.

6.1.1 Prototype Implementation

To ground the proposals in this chapter in demonstrable evidence, a functional prototype was developed comprising three Solidity smart contracts, a FastAPI backend service with `web3.py` integration, and off-chain verification modules for document authentication, face matching, liveness detection, and risk scoring. The complete source is published; Appendix A gives the repository location, the commit the reported results correspond to, and what is deliberately withheld from it.

6.1.1.1 Smart Contracts (Hardhat / Solidity 0.8.20)

Three contracts were deployed on a local Hardhat EVM node:

- *DIDRegistry.sol:* Anchors W3C-compliant DID Documents on-chain. Each DID maps to a controller address, a public key, and a service endpoint. An `onlyController` modifier ensures that only the original registrant may update or deactivate a given DID.
- *VCRegistry.sol:* Manages the full credential lifecycle. An `authorizedIssuers` mapping (the on-chain Trust Registry) restricts credential issuance and revocation to institutions approved by the registry admin, providing a technical layer for third-party reliance under Article 41.º of Law n.º 83/2017. The contract stores only the `keccak256` hash of the VC JSON; raw credential data is retained off-chain. Each credential carries one of three states: *Active*, *Revoked*, or *Suspended*. Any participant may call the public `verifyCredential` function to check hash integrity and revocation status without contacting the issuer.
- *AuditTrail.sol:* An append-only event log satisfying Article 51.º of Law n.º 83/2017 and AMLR record-keeping obligations. Each entry stores a `keccak256` hash of the off-chain event payload, the subject's DID, and one of nine `EventType` enum values spanning the onboarding lifecycle (*OnboardingStarted* through *CredentialReused*). Entries cannot be modified or deleted.

6.1.1.2 Backend Service (FastAPI + web3.py)

A FastAPI application exposes a four-step REST endpoint sequence under `/api/v1/onboarding/`: `POST /start`, `POST /{session_id}/document`, `GET` and `POST /{session_id}/challenge`, `POST /{session_id}/biometric`, and `POST /{session_id}/complete`. The routing, session management, and blockchain integration calls are fully implemented. The verification modules invoked across the document and biometric steps are: `document_auth.py` (EasyOCR + International Civil Aviation Organization (ICAO) 9303 MRZ validation + portrait extraction), `face_verification.py` (InsightFace ArcFace 1:1 matching, cosine-distance threshold 0.40), `liveness.py` (rPPG-POS peakiness threshold 1.70, the best-performing detector against Hedra/Veo 3 synthetic media in the experimental evaluation, though Section 5.3.4 shows it is not pipeline-universal), `session_quality.py` (frame-rate, resolution, duration, duplicate-fraction, luminance, and frame-count consistency), `temporal_consistency.py` (inter-frame pixel-difference analysis for replay and discontinuity detection), and `audio_consistency.py` (audio-track presence, silence ratio, spectral flatness, energy variance, and lip-audio synchronisation). A knowledge-based contextual challenge

derived from MRZ fields is generated at document upload and answered through the `/challenge` endpoint pair. The risk-scoring module (`risk_scoring.py`) is functional via a rule-based heuristic that aggregates signals from all seven modules. On approval the backend constructs a W3C-compliant VC JSON, hashes it with `keccak256`, and calls `VCRegistry.issueCredential` under an authorized issuer address. Subsequent verifiers may independently recompute the hash from the off-chain document and call `verifyCredential` to confirm authenticity without contacting the original issuer.

6.1.1.3 Why Blockchain Rather Than a Centralized Signed Log?

A conventional alternative is a relational database with server-side Hash-based Message Authentication Code (HMAC) signatures or a certificate-based audit log. The fundamental limitation of that model is that the signing authority and the party under compliance scrutiny are the same entity: an institution could in principle alter or delete records undetected. The blockchain design addresses this: once a hash is written to `AuditTrail` or a credential record is committed to `VCRegistry`, it is cryptographically immutable even for the contract deployer. A blockchain is not the only way to separate the signer from the audited party: signed logs whose hashes are regularly anchored with an independent timestamping authority, or a shared registry operated by the supervisor, would achieve a similar separation. The ledger was chosen here because it combines that separation with a registry that several institutions can read without a central operator; it is one approach among others, not a precondition for cross-bank verification. Furthermore, the on-chain Trust Registry lets Bank B check Bank A's credential by reading shared chain state, without transmitting PII between institutions and without a dedicated point-to-point integration. It does not remove the need for an agreement between the banks: reliance under Article 41.º of Law n.º 83/2017 must still be set out in contractual terms between the relying institution and the third party, and the relying institution remains responsible for the identification as if it had carried it out itself [15]. The ledger is the technical layer within that legal framework, not a substitute for it. The prototype scope is intentionally narrow: no production banking deployment, no EUDI Wallet integration, and no multi-institution pilot are claimed.

6.1.2 Architectural Overview

This section outlines a modular, secure, and scalable architecture for digital onboarding, combining automated document and biometric verification with blockchain for immutable

record-keeping and SSI.

6.1.3 Key Components

The architecture decomposes into seven layers, described in the order a request traverses them: the frontend that collects evidence from the user, the microservices that mediate the session, the verification engine that adjudicates the evidence, the blockchain layer that anchors the outcome, the storage tier that holds the artefacts, the retention policy that governs how long they persist, and the cross-cutting security controls. Each layer is specified below at the level of detail needed to reproduce it; the components marked as prototype notes are those actually implemented and measured in Section 6.1.10, while the remainder are architectural.

6.1.3.1 Frontend (User Interface)

The frontend is deliberately thin: it captures evidence and relays it, but performs no verification of its own, since any check running on a device the attacker controls can be bypassed by the injection attacks described in Chapter 5. Its two responsibilities are collecting the document and biometric samples, and brokering the user's consent to share attributes held in the EUDI Wallet.

- *Mobile/Web Application:* The primary interface for users to initiate onboarding, submit documents, perform biometric scans, and interact with the European Digital Identity Wallet (EUDI Wallet) [1].
- *Identity Wallet Integration:* Seamless connection with the EUDI Wallet for securely sharing government-verified attributes directly from a user's mobile device, in compliance with eIDAS 2.0 regulation.

6.1.3.2 Onboarding Microservices Layer

This layer holds the session state that the frontend cannot be trusted to hold. Splitting registration, document upload, and biometric capture into separate services keeps each one's data in scope for only as long as that step requires, which is what makes the storage-limitation obligations discussed later in this section enforceable in practice rather than merely declared.

- *User Registration Service:* Manages initial user input, temporary account creation, and session management during the onboarding flow.
- *Document Upload Service:* Provides a secure channel for receiving and temporarily storing user-provided identification documents.
- *Biometric Capture Service:* Orchestrates the capture of biometric data, including facial recognition scans for identity matching and liveness detection.

6.1.3.3 Verification Engine

The verification engine is where the evidence is adjudicated, and it is the layer most directly constrained by the experimental findings of Chapter 5. Because no single detector proved robust across generation pipelines, the engine is specified as four independent modules whose signals are aggregated rather than as one classifier: document authentication, facial recognition with presentation attack detection, sanctions screening, and risk scoring. A verdict that depends on several uncorrelated signals degrades gracefully when one of them is defeated, which is precisely the failure mode observed against direct Veo 3 output.

- *OCR and Document Authentication Module:* Employs specialized Computer Vision architectures to extract alphanumeric data and verify the structural integrity (e.g., guilloche pattern analysis, microprinting presence) of submitted documents. The module performs multi-spectral analysis where possible to detect photometric inconsistencies indicative of tampering or digital substitution. This module may integrate with third-party verification APIs (e.g., Jumio) [17].
- *Facial Recognition and PAD Module:* Orchestrates biometric 1:1 matching between the extracted ID portrait and live video frames using Siamese Neural Networks for high-fidelity embedding comparison. It integrates multi-modal PAD, including rPPG for pulse detection and texture-based classification, to neutralize deepfake injection attacks from generative models such as Hedra, Kling AI, or Google Veo 3 [14, 18, 32, 33, 46, 60].
- *Sanctions Screening Module:* Integrates with mandatory international sanctions lists (primarily the consolidated list of the European Union and the United Nations Security Council) alongside global lists such as OFAC (United States) for cross-border compliance. It uses Natural Language Processing (NLP)-driven heuristics and fuzzy

matching algorithms (e.g., Levenshtein distance) to minimize false positives while ensuring high sensitivity to sanctioned identities [3].

- *Risk Scoring Module:* Utilizes ensemble learning (e.g., Gradient Boosted Trees) to aggregate findings from all modules into a multidimensional risk score, automating the trigger for EDD workflows. *Prototype note:* the implemented `risk_scoring.py` substitutes a deterministic rule-based scorer producing the same *low*/*medium*/*high* output schema, pending labelled training data (see Workflow step 4).

MRZ algorithm (prototype detail). The document authentication module parses the MRZ of the submitted identity document according to ICAO Doc 9303 [61]. Two layouts are recognised: the TD3 format (two lines of 44 characters, used by international passports) and the TD1 format (three lines of 30 characters, used by national ID cards including the Portuguese Cartão de Cidadão). Candidate MRZ lines are extracted from the EasyOCR output by filtering on the MRZ character set (`[A-Z0-9<]`), the presence of filler '<' characters, and the expected line length. Once a line set is identified, every field check digit is recomputed using the ICAO 9303 Annex-B algorithm (positional weights 7–3–1, modulo 10) and compared against the OCR'd value: document number, date of birth, expiry date, and (for TD1) the composite check digit covering the document-number, dates and optional-data subfields. A document is considered structurally authentic when all check digits validate and the expiry date is not in the past; partial check-digit failures are reported as the `mrz_checksum_failed` flag for downstream risk scoring rather than as a hard reject. Portrait extraction proceeds in parallel using InsightFace SCRFD with an OpenCV Haar-cascade fallback.

6.1.3.4 Blockchain Integration Layer

The blockchain layer was prototyped with three Solidity smart contracts (see Section 6.1.1 for implementation details). The VCRegistry contract implements the Trust Registry through an `authorizedIssuers` mapping enforced by an `onlyAuthorizedIssuer` modifier; only addresses approved by the registry admin, representing institutions licensed by Banco de Portugal, may issue or revoke credentials. Verification events are anchored on-chain via the AuditTrail contract as `keccak256` hashes of the event payload; raw PII is never written to the ledger. Credential status checks (*Active* / *Revoked* / *Suspended*) are available to any participant through a public read function, enabling real-time revocation without contacting the original issuer.

The blockchain layer serves as the trust anchor for the identity system, transitioning from a model of “data ownership” to “data provenance” using decentralized standards.

- *DIDs:* Standardized by the W3C, DIDs are Uniform Resource Identifiers (URIs) that enable verifiable, decentralized digital identity. Unlike traditional identifiers (e.g., email or usernames), DIDs are not controlled by a central authority. In the proposed system, the user’s DID points to a DID Document on the blockchain which contains public keys for cryptographic authentication and service endpoints for communication [9].
- *VCs:* These are the digital equivalents of physical identity documents. A VC contains specific claims about the user (e.g., name, citizenship, address) that have been verified by the AI engine. The bank signs these claims with its private key, and the user stores the signed VC in their personal wallet. The recipient of a VC can verify its authenticity and integrity using the Issuer’s DID Document on the blockchain [10].
- *SSI and Ledger Integration:* Records cryptographic hashes of key onboarding events and verification outcomes on a permissioned blockchain, ensuring a tamper-proof audit trail for regulatory compliance (e.g., AMLR, Law n.º 83/2017) without storing sensitive personal data directly on the ledger [8, 27].

6.1.3.5 Data Storage

Storage is partitioned by what each tier is allowed to hold. Raw identity documents and biometric samples never reach the ledger; they remain in encrypted off-chain storage, while the chain holds only `keccak256` digests of the corresponding events. This split is what allows the architecture to satisfy the immutability expected of an audit trail without placing personal data somewhere it can never be erased.

- *Secure Document Storage:* Off-chain, encrypted storage for original identification documents and sensitive personal data, compliant with GDPR [1].
- *Operational Database:* Stores user profiles, application statuses, audit logs, and other operational data.
- *Blockchain Ledger:* The decentralized, immutable ledger holding cryptographic proofs and verifiable credentials.

6.1.3.6 Data Retention and Regulatory Archiving

A critical design constraint for Portuguese financial institutions is the mandatory data retention period. According to Article 51.º of Law n.º 83/2017, banks are legally obligated to retain copies of the identification documents and transaction records for a period of seven years following the termination of the business relationship.

The proposed architecture addresses this by:

- *Hybrid Retention Policy:* While the blockchain stores immutable proofs of the verification, the original encrypted artifacts are archived in the Secure Document Storage layer.
- *Encrypted Lifecycle Management:* Data is encrypted with institutional keys, ensuring that even if the bank ends the relationship, the records remain available for regulatory audit (e.g., by *Banco de Portugal*) while being strictly isolated from active production databases in compliance with GDPR “Purpose Limitation” and “Storage Limitation” principles.

6.1.3.7 Security and Compliance

The preceding layers each carry their own controls; the measures collected here apply across all of them and are the ones an auditor would expect to see evidenced independently of any single component. They cover confidentiality in transit and at rest, a single enforced entry point to the service surface, the mapping from architectural features to the specific obligations of AMLR, eIDAS 2.0, PSD2 and the Portuguese regime, and the transition from one-off onboarding checks to continuous monitoring of the relationship.

- *End-to-End Encryption:* Ensures confidentiality and integrity of all data in transit and at rest [7].
- *API Gateway:* Acts as a single entry point, enforcing security policies, rate limiting, and authentication.
- *Regulatory Compliance Framework:* Embedded mechanisms for continuous adherence to regulations like AMLR, eIDAS 2.0, PSD2, and Portuguese financial regulations (e.g., LINCE certification, Aviso n.º 1/2022) [2, 3, 15, 16].
- *Continuous Fraud Detection:* Integrates with transaction monitoring for ongoing detection and prevention of fraudulent activities [6].

6.1.4 The Reusable Identity Model

The cornerstone of the proposed architecture is the “Reusable Identity” (or “Portable Identity”) model. In traditional banking, every institution must independently verify a customer’s identity, leading to redundant work and friction. The Reusable Identity flow operates through a three-party trust model:

1. *Issuance (Trust Foundation):* After a user successfully completes the KYC process at “Bank A,” the bank acts as an *Issuer*. It generates a VC containing the verified identity attributes and delivers it to the user’s digital wallet.
2. *Storage (SSI Control):* The user maintains a non-custodial digital wallet (the *Holder*) where the VC is stored locally. This wallet is linked to their DID, and the user holds the sole private keys necessary to “sign” proofs of their identity.
3. *Verification (Reuse):* When the user subsequently interacts with “Bank B,” they do not need to repeat the KYC process. Instead, they present a cryptographic proof derived from the VC in their wallet. Bank B (the *Verifier*) validates the proof against the blockchain by verifying the signature using Bank A’s public keys.

For the Reusable Identity model to be effective, “Bank B” must trust that “Bank A” performed the initial KYC with sufficient rigor. In the Portuguese legal framework, specifically under Article 41.º of Law n.º 83/2017, this is known as “Third-Party Reliance” (*Recurso a Terceiros*).

However, it is a critical technical and legal nuance that the ultimate responsibility for the identification remains with the Verifying institution (Bank B), and that the reliance must be set out in contractual terms agreed between the two institutions [15]. The architecture supports this through a Governance Framework (a set of business and legal rules recorded on the blockchain that defines the “Trust Registry”), which complements those agreed terms rather than replacing them. This ensures that all Issuers are authorized entities (e.g., banks licensed by Banco de Portugal) and that their verification processes are audit-compliant with Aviso n.º 1/2022.

6.1.5 Selective Disclosure and Zero-Knowledge Proofs (ZKPs)

A critical advantage of the VC-based approach is the ability to implement Selective Disclosure. Using modern standards such as Selective Disclosure JSON Web Token (SD-JWT) or BBS+ signatures, the system allows the user to prove specific attributes without revealing the underlying sensitive data. This technical capability directly addresses the "Biometric Data Honeypot" problem.

Traditional centralized KYC systems create massive databases of high-resolution biometric data and identification documents. These "honeypots" are primary targets for malicious actors. The 2024 ransomware attack on the AMA in Portugal, while primarily a disruption of service, demonstrated the catastrophic impact of a Single Point of Failure in a centralized identity infrastructure like the *Chave Móvel Digital*.

The proposed architecture prevents this by ensuring that:

- *Decentralized Proofs:* Instead of Bank B receiving and storing the user's raw biometric templates or ID scans, it only receives a ZKP: a cryptographic proof that the data was already verified by a trusted Issuer (Bank A) and that the holder currently possesses the private keys for that identity.
- *Minimal Data Exposure:* The verified attributes are shared only as needed. For instance, a user can provide:
  - *Age Verification:* Proving they are over 18 without disclosing their exact date of birth.
  - *Residency Correlation:* Confirming they reside within the EU without sharing a precise residential address.
  - *Identity Binding:* Proving that the holder of the DID is the same person who passed the liveness check (PAD), without the Verifier ever seeing the original biometric file.
- *GDPR Compliance:* By reducing the volume of personal data (GDPR's PII) processed and stored by subsequent institutions, the "privacy-by-design" approach significantly lowers their liability and minimizes the systemic damage caused by potential local breaches.

6.1.6 Credential Revocation and Lifecycle Management

To maintain the integrity of the trust triangle, the system includes a robust revocation mechanism. If a bank (the Issuer) detects that an identity was obtained fraudulently or if the underlying ID document is reported stolen, the VC must be invalidated.

- *Revocation Lists:* The Issuer maintains a cryptographically signed Revocation List (or a more efficient structure like a *Credential Status List*) on the blockchain.
- *Verification Check:* During any reuse of the identity, the Verifier (Bank B) automatically queries the ledger to ensure the VC is still "Active" before granting access.
- *Dynamic Updates:* Since the blockchain provides a real-time, global status check, revocation is effective across all participating institutions immediately.

6.1.7 Comparison: Traditional vs. Proposed Architecture

Table 6.1 summarizes the improvements provided by the integration of AI and Blockchain over traditional centralized KYC methods.

TABLE 6.1: Comparison of traditional Banking KYC vs. Proposed Architecture

Feature	Current Banking Landscape	Proposed AI-Blockchain Architecture
Trust Model	Hub-and-Spoke: Reliant on central authorities (e.g., CMD/AMA) or siloes.	Multilateral Peer-to-Peer: Decentralized trust between institutions via the ledger.
Biometric Data	Centralized Honeypots: Temples/images stored by banks or 3rd parties.	Privacy-by-Design: Raw biometrics stay in the EUDI Wallet; only ZKPs are shared.
Fraud Defense	Pixel-Centric: Reactive analysis of deepfake artifacts.	Sensor-Centric: Proactive Hardware Attestation (Root of Trust).
KYC Reusability	Non-existent/Siloed: Repeat KYC for every institution.	Instant Portability: One-time verification, cross-institutional reuse.
Data Disclosure	Binary: Full ID submission for every check.	Granular: Selective disclosure (e.g., prove "Over 18" only).

6.1.8 Differentiators: A Paradigm Shift in Identity Assurance

The proposed architecture distinguishes itself from existing solutions (e.g., LOQR, Jumio, or CMD) by addressing the fundamental "Identity Trilemma": the conflict between security, privacy, and scalability.

- *Transition from "Content Analysis" to "Origin Attestation":* Unlike current systems that struggle to detect high-fidelity diffusion-based deepfakes through texture analysis, this architecture prioritizes hardware-level anchorage. By requiring cryptographically signed metadata from the physical camera sensor, it renders virtual-camera injection attacks, the primary vector for modern identity fraud, technically non-viable.
- *Decentralized Third-Party Reliance:* Third-party reliance under Article 41.º of Law n.º 83/2017 requires agreed contractual terms between institutions, and the relying institution remains responsible for the identification. Within that framework, this system adds a Global Trust Registry on the blockchain, where the "Audit Trail" of an identity's genesis is immutable and verifiable by any participant without revealing the PII of other customers, so that each new reliance relationship needs agreed terms but not a new technical integration.
- *Neutralizing the Identity Honeypot:* By leveraging VCs, the proposed system ensures that the bank never possesses the "master key" to the user's biometric identity. This significantly reduces the systemic risk highlighted by the 2024 AMA cyberattack, as a breach at a single bank no longer provides a gateway to millions of reusable biometric templates.

6.1.9 Workflow Overview

The following workflow reflects the scope of the prototype. Steps labelled [Implemented] are demonstrated by the prototype codebase; [Designed] denotes architecture specified but not built within the prototype scope; [Future Work] indicates steps dependent on infrastructure not yet available (e.g., a production EUDI Wallet environment).

1. *Initiation [Implemented]:* A client calls `POST /api/v1/onboarding/start` with the holder's Ethereum address. The backend derives a `did:ethr:<checksum-address>` identifier from that address, creates a session, records an *OnboardingStarted* hash in `AuditTrail` under the derived DID, and returns both the session identifier and the holder DID.
2. *Identity Submission [Implemented / Future Work]:* The REST endpoints for document and biometric upload are implemented. Direct attribute sharing from an EUDI Wallet is *[Future Work]*, pending the availability of production wallet reference implementations.

3. *AI Verification [Implemented]:* Document authentication uses EasyOCR for text extraction and ICAO 9303 MRZ check-digit validation, with InsightFace SCRFD (Haar-cascade fallback) for portrait crop extraction. Face verification performs 1:1 ArcFace embedding comparison (cosine distance ≤ 0.40 = match). Passive liveness detection applies rPPG-POS frequency-domain peakiness (threshold 1.70), the only method in the experimental evaluation that correctly classified Hedra/Veo 3 synthetic videos (Videos A–C) as spoofs; Section 5.3.4 shows that direct Google Veo 3 output (Videos 5–8) bypasses this layer, so rPPG alone is not pipeline-universal. The biometric submission additionally triggers three supplementary verification layers (see Section 6.1.11): session integrity, temporal consistency, and audio consistency. A fourth layer, contextual knowledge-based verification, is evaluated at document upload time and answered by the client prior to the biometric step. All four feed into the risk scorer. The `AuditTrail` recording calls for *DocumentVerified* and *BiometricVerified* execute on completion of the respective steps.

4. *Risk Assessment [Partially Implemented]:* The rule-based heuristic in `risk_scoring.py` is functional and produces a *low* / *medium* / *high* risk level from session data. The full Gradient Boosted Trees model described in the architecture is *[Designed]*, awaiting labelled training data. The result is recorded on-chain (*RiskAssessed*).

5. *Blockchain Integration [Implemented]:* For approved sessions (risk level *low* or *medium*), a W3C-compliant KYC Verifiable Credential is constructed with a three-year validity period, its `keccak256` hash is anchored in `VCRegistry` by an authorized issuer address, and a *CredentialIssued* event is appended to `AuditTrail`. Sessions classified *high* are rejected (*OnboardingRejected*); the prototype does not implement a separate EDD-referral branch.

6. *Account Decision [Designed]:* Based on the risk level, the account is automatically approved, referred for EDD, or declined. The EDD escalation workflow and human review queue are architectural specifications not implemented in the prototype.

7. *Post-Onboarding [Future Work]:* Continuous transaction monitoring and cross-institutional VC reuse (*CredentialReused* is defined in the contract but not piloted with multiple institutions).

6.1.10 Empirical Validation of the Prototype Pipeline

To confirm that the prototype exercises every component of the architecture end-to-end, not merely that its REST endpoints respond, a series of five validation runs was conducted against the local Hardhat node, the FastAPI backend, and the AI verification modules. Each run drives the complete four-step flow (`/start` $\rightarrow$ `/document` $\rightarrow$ `/biometric` $\rightarrow$ `/complete`) with concrete inputs, exercising the EasyOCR pipeline, the ICAO 9303 MRZ parser, InsightFace ArcFace 1:1 matching, rPPG-POS liveness detection, the rule-based risk scorer, and the on-chain audit trail.

The document inputs span two real Portuguese Cartões de Cidadão (the holders' personal identification was used with their consent) and the biometric inputs are the two test clips `TestVideo2.mp4` and `TestVideo4.mp4` drawn from the dataset analysed in Chapter 5. Videos 5–8 (Google Veo 3 direct) were not submitted to the prototype pipeline; based on the peakiness scores documented in Section 5.3.4 (1.92–2.55, all above the 1.70 threshold), the rPPG-POS liveness layer would produce a REAL verdict and would not constitute a detection barrier for that generator family. Table 6.2 summarises the validation runs and the resulting verification signals.

TABLE 6.2: End-to-end validation runs of the four-step KYC pipeline. Cosine distance ≤ 0.40 is an ArcFace match; rPPG peakiness > 1.70 is a live capture.

Run	Document	Biometric	Face match (cosine)	Liveness (peakiness)	Verdict
1	MP4 (incorrect MIME)	—	—	—	step 2 rejects (input not an image)
2	Simão Pereira CC	TestVideo2	false (0.4051)	true (2.13)	high / rejected
3	Simão Pereira CC (after TD1 parser added)	TestVideo2	false (0.4051)	true (2.13)	high / rejected
4	Simão Pereira CC	TestVideo4	true (0.3936)	false (1.625)	high / rejected (deepfake signal)
5	Luís Pereira CC	TestVideo2	false (0.8924)	true (2.13)	high / rejected
6	Simão CC (PT OCR enabled)	Video3 (real selfie)	true (0.3734)	true (2.18)	high / rejected on document check only
7	Simão CC (focused MRZ + relaxed TD1)	Video3 (real selfie)	true (0.3734)	true (2.18)	low / APPROVED, VC issued

Discriminating signals. The six meaningful runs (2–7) cover qualitatively distinct verification cases that the architecture is designed to separate:

- *Near-miss face mismatch with a real pulse* (runs 2–3): the subject of TestVideo2 lies near Simão's ArcFace embedding (cosine 0.4051, marginally above the 0.40 acceptance threshold), but rPPG-POS recovers a sharp cardiac peak (peakiness 2.13, $\sim$91 BPM). The pipeline rejects via face mismatch.

- *Matched face with a missing pulse* (run 4): TestVideo4 matches Simão's portrait (cosine 0.3936) yet rPPG-POS finds no clean cardiac peak (peakiness 1.625, below the 1.70 threshold). This is the textbook deepfake-detection case: the synthetic preserves enough texture and identity features to defeat face matching in isolation, but lacks the spatial-temporal colour modulation produced by real haemoglobin perfusion. The pipeline rejects via liveness, consistent with the Hedra and Google Veo 3 measurements in Chapter 5 (peakiness 1.23 and 1.59, respectively).

- *Decisive face mismatch with a real pulse* (run 5): the same TestVideo2 is matched against a different Cartão de Cidadão (Luís Pereira's), yielding a cosine distance of 0.8924, comfortably in unrelated-faces territory and an order of magnitude clearer than the near-miss in run 2.

- *Both biometric checks pass, the document check blocks acceptance* (run 6): a freshly recorded selfie of Simão (`Video3.mp4`) produces the cleanest signals seen across the runs (cosine 0.3734, peakiness 2.18, estimated HR 96.76 BPM). The pipeline nevertheless rejects with risk = *high*, driven entirely by the document side: the position of the document-number check digit on the Cartão de Cidadão holds the filler character '<'. That character is correct: ICAO Doc 9303 places a filler in the check-digit position when the document number is longer than nine characters, and puts the remaining characters of the number, followed by its check digit, at the start of the optional-data field [61]. The prototype's parser does not implement this rule, so it compares the filler with a check digit computed over the first nine characters, and the resulting MRZ checksum failure pushes the rule-based scorer above the rejection threshold.

- *End-to-end approval and on-chain credential issuance* (run 7): the same Cartão de Cidadão and `Video3.mp4` after two targeted code changes. First, a content-aware MRZ-region localiser that crops the image around the MRZ block detected in the first OCR pass and re-OCRs that strip with an MRZ-only allow-list (`[A-Z0-9<]`). Second, as a workaround for the parser's missing long-number rule, a relaxed TD1 validity rule that requires both temporal check digits (date of birth and expiry) to

validate and at least two of the four TD1 check digits in total, while preserving the strict ICAO-conformant rule as a separate `strict_valid` flag. The proper fix is to implement ICAO's rule for document numbers longer than nine characters, after which the strict rule can be applied. The focused re-OCR raised the per-line MRZ confidence to between 0.888 and 0.969 (vs. 0.328 for the general pass), the relaxed TD1 rule accepted the document, the rule-based scorer returned risk = *low* (score 0.10), the `/complete` step issued a Verifiable Credential on `VCRegistry` under the deployer (authorised issuer) address, and a *CredentialIssued* event was appended to `AuditTrail`.

Document parser behaviour on real Cartões de Cidadão. On both ID cards the TD1 branch of the MRZ parser correctly recovers every visible field (document number, country, nationality, date of birth, expiry, sex, surname, and given names). Two of the four ICAO 9303 check digits (date of birth and expiry) validate on both cards; the document-number check fails because the Cartão de Cidadão's document number is longer than nine characters: the check-digit position correctly holds the filler "<", and the real check digit follows the rest of the number in the optional-data field, a rule the parser does not implement. The composite check digit also fails in these runs; its cause was not isolated. The structural authenticity decision therefore remains conservative (`authentic = false`). Adding the Portuguese language pack to EasyOCR (`easyocr.Reader(["en", "pt"])`) and, if necessary, a second OCR pass on the cropped MRZ region with an alphanumeric+< allow-list, is expected to lift the document-wide OCR confidence above the 0.50 threshold; it cannot fix the document-number check, which requires the parser to implement ICAO's long-number rule.

On-chain audit coverage. Every audit event emitted by the rejected runs (*OnboardingStarted*, *DocumentVerified*, *BiometricVerified*, *RiskAssessed*, and *OnboardingRejected*) was anchored as a `keccak256` hash on the local Hardhat node, exercising the `AuditTrail` contract end-to-end. None of the runs produced a *CredentialIssued* event, because the rule-based scorer (see Workflow step 4) classified every session as *high* risk and the prototype's binary approve/reject branch rejects high-risk sessions outright.

The accepted case. Run 7 is the first end-to-end *approved* session: `risk_level = low`, `approved = true`, with an on-chain `CredentialIssued` event under credential identifier `0xb7e3b849c6184a4f88c61f12b24cee71…`. The Verifiable Credential's `keccak256` hash is anchored in `VCRegistry` for a three-year validity window; the raw VC JSON is not persisted to disk and is dropped from the in-memory session when `/complete` returns. Bank B can independently verify the credential's status and integrity by recomputing the canonical hash from any held off-chain copy of the VC JSON and calling `VCRegistry.verifyCredential(id, hash)`, which is the technical check on which third-party reliance under Article 41 of Law n.º 83/2017 can build; the agreed terms between the institutions and Bank B's own responsibility under that article are unaffected. The prototype demonstrates that this check works as designed without exposing any PII on chain.

6.1.11 Extended Verification Pipeline: Session, Temporal, Audio, and Contextual Modules

Beyond the three core AI checks (document authentication, face matching, and liveness detection), the prototype implements four additional verification layers. These are not mandated individually by any single instrument; rather, they are motivated by the defence-in-depth posture that the EBA Guidelines on Remote Customer Onboarding [22] expect of remote solutions, and designed so that their outcomes could be reported under the evaluation framework of ISO/IEC 30107-3 [21], which governs how presentation-attack detection performance is tested and reported rather than prescribing the checks themselves.

Session Integrity Module (`session_quality.py`). Before analysing the subject, the clip is inspected for six structural properties that distinguish a genuine live capture from a replayed or artificially generated recording. Frame rate must be at least 24 fps (below this the rPPG cardiac signal degrades); resolution at least 640×480, the Video Graphics Array (VGA) minimum for reliable face PAD; duration between 5 s (minimum for the POS algorithm's frequency-domain analysis) and 60 s (clips longer than one minute are almost certainly pre-recorded). A single-pass frame decoder additionally measures the fraction of near-duplicate frames (consecutive frames whose mean absolute pixel difference on a 160×120 greyscale thumbnail is below 0.5; values above 10 % indicate a frozen or looped recording), the mean luminance, sampled every tenth frame (flagged when below 30 or

above 220 on the 0–255 scale, indicating an over-dark or overexposed capture unsuitable for colour-based rPPG), and the consistency between the container-reported frame count and the number of frames that actually decode (a discrepancy larger than 20 % suggests file-level editing). All six checks run in a single O(n) pass through the decoded frame sequence, keeping the overhead bounded by clip duration rather than multiplied by the number of checks.

Temporal Consistency Module (`temporal_consistency.py`). During the full clip the Inter-Frame Difference (IFD) between consecutive greyscale thumbnail pairs is computed. Two failure modes are detected: a clip whose mean IFD falls below 1.0 is flagged as `static_or_frozen_video` (characteristic of replayed or looped media); and clips in which more than 5 % of inter-frame transitions exhibit an IFD greater than four times the clip mean are flagged as `frame_discontinuities`, a signature of stitched or edited video segments, which deepfakes generated by looping shorter clips often exhibit. Applied to `TestVideo2.mp4` the module returns mean IFD 9.85, standard deviation 2.56, and zero discontinuities across 191 transitions: a clean real-world capture.

Audio Consistency Module (`audio_consistency.py`). The audio consistency module addresses a distinct attack surface: the replacement or synthesis of the voice track in an otherwise visually authentic recording. Such attacks are increasingly viable with off-the-shelf voice-conversion and Text-to-Speech (TTS) systems, and fall within the broader category of manipulated remote-onboarding evidence that the EBA guidance expects institutions to guard against [22].

Audio is extracted from the uploaded video using `ffmpeg` (system binary, already required by the rPPG-Toolbox) and decoded to mono float32 Pulse-Code Modulation (PCM) at 16 kHz, a rate sufficient for speech intelligibility analysis without the computational overhead of high-fidelity audio codecs. Five properties are then measured:

1. *Audio track presence.* The video must carry an audio stream. A clip with no audio track receives an immediate `no_audio_track` flag, which both adds the module's +0.20 risk contribution and independently sets the hard-fail flag, because a legitimately recorded identification video always includes ambient sound, however faint.

2. *Silence ratio.* The per-hop Root Mean Square (RMS) energy envelope is computed over 128-sample hops. Hops with RMS below 0.01 are classified as silent. A clip

in which more than 70 % of hops are silent is flagged as `high_silence_ratio`, because an authentic interaction, even one where the subject is not actively speaking, generates ambient noise well above the digital floor.

3. *Spectral flatness (Wiener entropy).* The short-time Fourier transform of the audio is computed using `scipy.signal.stft` with 512-sample frames and 128-sample hops. For each frame the flatness is taken as the ratio of the geometric mean to the arithmetic mean of the spectrum, then averaged over all frames carrying non-trivial energy. The measure is bounded in $[0, 1]$: values near zero indicate energy concentrated in a few frequencies, as in the harmonic structure of voiced speech, while values near one indicate energy spread uniformly, as in noise. Note that the canonical formulation is defined over the power spectrum; this implementation computes it over the magnitude spectrum, so the absolute values reported here are not directly comparable to published power-spectrum figures. The rationale for the check is that vocoder-based synthesis characteristically over-smooths the spectral envelope relative to natural phonation, which raises the measured flatness. The 0.45 decision threshold was set empirically against the reference clips available to this prototype rather than derived from published values; the module therefore reports the raw flatness alongside its verdict, so the threshold can be recalibrated once a labelled corpus of genuine and synthetic onboarding audio is assembled. Clips exceeding it are flagged as `flat_spectrum`.

4. *Short-term energy variance.* The variance of the RMS envelope over the clip duration measures how much the speaker's vocal energy fluctuates naturally. Looped or synthetically generated audio tends to have abnormally low variance. Clips whose energy variance falls below 3×10^{-5} are flagged as `low_energy_variance`.

5. *Lip-audio synchronisation.* For each decoded frame, OpenCV's frontal-face Haar-cascade locates the largest detected face and extracts the lower 35 % of the bounding box as a mouth-region proxy. The per-frame pixel standard deviation of that region (a coarse measure of mouth openness) is resampled to align with the audio RMS envelope, and the Pearson correlation between the two time series is computed. A strong anti-correlation (Pearson $r < -0.30$) is flagged as `poor_lip_sync`, because dubbed or replaced audio produces systematic opposition between lip motion and audio energy that is absent in genuine recordings. Near-zero correlation is not flagged: it is normal when the subject is not actively speaking (the typical case in

a KYC selfie clip). This check is gated on the silence ratio: clips already exceeding the 70 % silence threshold carry too little speech for the correlation to be meaningful, so the module skips it and reports a null correlation rather than a spurious value.

On `TestVideo2.mp4` the module finds: audio present, silence ratio 0.44 (well below the 0.70 threshold), spectral flatness 0.085 (well below the 0.45 threshold, consistent with naturally phonated speech), energy variance 8.2×10^{-5} (above the 3×10^{-5} floor), and lip-audio correlation -0.086 (near-zero, as expected for a clip recorded without active speech). All flags clear; the audio consistency module passes.

The module contributes +0.20 to the risk score when any flag fires, and escalates to a hard-fail when the audio track is absent entirely. All three libraries required (`scipy`, `numpy`, and `cv2`) are already in the prototype dependency set; no additional packages are introduced.

Contextual Verification Module (`onboarding.py`). After the identity document is uploaded and the MRZ is parsed, the backend derives a knowledge-based challenge from fields that only the genuine document holder should know. The primary challenge uses the last three characters of the document number; a fallback uses the numeric birth month extracted from the date-of-birth field. The challenge is served to the client via `GET /{session_id}/challenge` and the answer submitted via `POST /{session_id}/challenge`. A wrong answer adds +0.25 to the risk score and sets a hard-fail flag; a correct answer clears the penalty. The outcome is held in session state rather than written as a separate on-chain entry, because the `AuditTrail EventType` enum defines no challenge-specific value; it reaches the ledger indirectly through the *RiskAssessed* event that follows. This mechanism provides a knowledge-based evidence layer (information that a replay video or deepfake injection, by itself, cannot supply), complementing the biometric checks in the spirit of the multi-factor approach to remote identity assurance set out in the EBA guidance [22].

Prototype Limitations.

- *No production deployment:* All tests run on a local Hardhat node with instant block confirmation. Gas costs and confirmation latency on a production permissioned ledger (e.g., Hyperledger Besu) are not measured.

- *Sanctions and PEP screening not implemented:* The `SanctionsCleared` event type is reserved in `AuditTrail.sol`, but the corresponding screening module described in the Key Components section is architectural only; integration with the EU consolidated list, the United Nations (UN) Security Council list, and OFAC is future work.
- *Risk scoring uses a rule-based heuristic:* The Gradient Boosted Trees model described in the Key Components section is not implemented. `risk_scoring.py` substitutes a deterministic rule-based scorer that produces the same *low*/*medium*/*high* output, pending labelled training data.
- *Cross-institutional reuse not piloted:* The `CredentialReused` event type is defined and the `/api/v1/verification/credential` endpoint is functional, but no second-institution (Bank B) deployment exists; cross-bank reuse is demonstrated only at the API level against the same Hardhat node. A regression test (`tests/test_credential_roundtrip.py`) pins the encoding contract that lets the verifier re-derive the on-chain credential identifier from the string returned at issuance.
- *EUDI Wallet not integrated:* The prototype uses direct REST API calls; integration with the EUDI Wallet holder application is future work, pending the availability of production wallet reference implementations.
- *DID document registration is decoupled from onboarding:* The prototype derives a `did:ethr:<checksum-address>` identifier from the holder's Ethereum address at session start, satisfying the W3C DID syntax for every audit-trail entry. Publishing the corresponding DID Document (public key + service endpoint) to `DIDRegistry` remains a separate step exposed through `/api/v1/identity/did` and is not automatically orchestrated as part of the four-step onboarding flow.

6.1.12 Compliance and Data Management

This section addresses how the proposed architecture handles critical aspects of data lifecycle management, including updates and deletions, while ensuring robust compliance with regulatory frameworks, particularly the GDPR.

Table 6.3 summarizes the data partitioning between the on-chain and off-chain layers, along with the rationale for each decision. All functional tests were performed against a local Hardhat EVM node (instant block mining), which confirms functional correctness;

production latency on a permissioned Ethereum-compatible network (e.g., Hyperledger Besu) is not benchmarked in this work.

TABLE 6.3: On-chain vs. Off-chain Data Partitioning in the Prototype

Data / Operation	Location	Contract / Store	Rationale
DID public key + service endpoint	On-chain	DIDRegistry	Must be publicly resolvable for VC verification; contains no PII
`keccak256`(VC JSON)	On-chain	VCRegistry	Immutable binding between credential and on-chain state; hash alone cannot reconstruct PII
Credential status (Active / Revoked / Suspended)	On-chain	VCRegistry	Real-time revocation check by any verifier without contacting the issuer
`keccak256`(onboarding event payload)	On-chain	AuditTrail	Tamper-evident record for AMLR and Law n.º 83/2017 audit; no PII exposed
Raw VC JSON (verified identity attributes)	Off-chain	Encrypted DB / IPFS	Contains PII; subject to GDPR right to rectification and erasure
Identity documents and biometric data	Off-chain	Secure Document Storage	High-sensitivity PII; 7-year retention required by Article 51.º of Law n.º 83/2017
Event payload JSON (audit context)	Off-chain	Operational Database	Linked to on-chain hash; retained for 7 years then purged

Tamper-evidence check: The `verifyCredential` function accepts a credential identifier and a candidate hash. If the raw VC JSON stored off-chain is altered after issuance, the recomputed `keccak256` will not match the anchored value, and the function returns `valid = false`. This provides an independently executable tamper-detection mechanism that does not rely on the issuer.

6.1.12.1 Information Updates and Deletions

The system is designed to manage data updates and deletions in a manner that respects user rights and regulatory requirements, balancing the immutability of blockchain with the necessity for data modifiability.

- *Off-Chain Personal Data Storage:* Sensitive personal data, such as identification documents and biometric information, is never stored directly on the blockchain. Instead, it resides in secure, encrypted off-chain databases (Secure Document Storage, Operational Database). This design choice is fundamental for compliance with the “right to be forgotten” and the “right to rectification” under GDPR [1].

- *Data Updates:* When a user's information needs to be updated (e.g., a new address, updated ID), the existing records in the off-chain databases are modified or replaced. New cryptographic hashes reflecting the updated data can be recorded on the blockchain, creating a new verifiable credential and an auditable trail of the change. The old hashes on the blockchain remain as a historical record of past verifications, but they point to data that has been superseded or removed off-chain. This maintains the integrity of historical verification events while allowing personal data to be current.

- *Data Deletions and Legal Retention:* While the "right to be forgotten" is a pillar of GDPR, it is legally constrained by AMLR and Portuguese national law. Under Article 51.º of Law n.º 83/2017, financial institutions are required to retain identity and transaction data for seven years following the termination of the business relationship. The architecture accommodates this by archiving the off-chain data in an "immutable-read-only" encrypted state for the duration of the legal limitation period, after which the automated "Right to be Forgotten" trigger purges the raw PII, leaving only the irreversible cryptographic hash as a permanent, non-identifiable record on the blockchain.

6.1.12.2 Compliance with Regulations (GDPR and Beyond)

The architecture integrates multiple layers of control to ensure comprehensive regulatory compliance, with a strong emphasis on GDPR principles [1].

- *Data Minimization:* Only essential data required for KYC and AML compliance is collected and processed, adhering to GDPR Article 5(1)(c) [1].

- *Purpose Limitation:* Data is processed only for specified, explicit, and legitimate purposes (onboarding, identity verification, fraud prevention, AML compliance), in line with GDPR Article 5(1)(b) [1].

- *Storage Limitation:* Personal data is stored for no longer than is necessary for the purposes for which it is processed, with clear data retention policies (GDPR Article 5(1)(e)) [1].

- *Consent Management:* Explicit user consent is obtained for data processing, particularly for sharing verifiable credentials via the SSI module. Users retain control over who accesses their data (GDPR Article 6(1)(a)) [1].

- *Security by Design and Default:* End-to-end encryption, secure data storage, and robust access controls are implemented at all layers to protect personal data against unauthorized processing or accidental loss, destruction, or damage (GDPR Article 25, Article 32) [1, 7].
- *Accountability and Auditability:* The immutable nature of the blockchain provides an unalterable audit trail of verification events and credential issuances. This enhances accountability and facilitates compliance with audit requirements from regulatory bodies (e.g., Banco de Portugal, European Banking Authority) [3, 16].
- *Cross-Border Data Transfers:* Mechanisms are in place to ensure any cross-border data transfers comply with GDPR Chapter V, leveraging appropriate safeguards like Standard Contractual Clauses or adequacy decisions [1].
- *eIDAS 2.0 and AMLR Integration:* The integration of the EUDI Wallet and adherence to its specifications ensure compliance with the eIDAS 2.0 regulation for digital identity verification. Furthermore, the robust identity verification and fraud detection capabilities directly support the stringent requirements of the EU AMLR, set to apply in 2027, and national laws like Law n.º 83/2017 and Aviso n.º 1/2022 [3, 15, 16].
- *Ethical AI Practices:* The machine-learning components of the verification engine are developed with ethical guidelines, focusing on fairness, transparency, and non-discrimination in decision-making, in anticipation of requirements from the EU AI Act (2024), especially regarding high-risk AI systems used in financial services [5].

6.2. Current Landscape and Adoption

While a fully integrated automated onboarding system leveraging blockchain for SSI is still an emerging concept, research on blockchain-based e-KYC systems has grown quickly; a recent systematic review nonetheless identifies three problems that remain to be solved before such systems can deliver their full benefits: the lack of global standardisation, the risk of fraud during the e-KYC process, and privacy concerns [27]. Several individual components of this architecture are, however, already in use in the financial industry, as the examples below show.

AI in KYC/AML: Artificial Intelligence is widely adopted for document verification, biometric authentication (e.g., facial recognition, liveness detection), and transaction monitoring to combat financial crime and enhance customer due diligence [14, 17, 18, 60].

Solutions from companies like Jumio are already providing AI-driven identity verification services that reduce fraud and streamline processes [17]. In Portugal, companies like LOQR provide "Journey-as-a-Service" platforms that integrate AI-driven document and biometric verification tailored to the local regulatory environment, including LINCE certification [24].

Blockchain for Identity and Data Management: Blockchain technology is increasingly explored for digital identity, particularly with the rise of SSI frameworks. Significant initiatives include the EBSI, a joint project by the European Commission and the European Blockchain Partnership, which aims to provide cross-border public services using blockchain, including identity verification [30]. In the Iberian region, Alastria has developed one of the world's first semi-public blockchain ecosystems, supporting permissioned identity solutions for diverse sectors [31]. Other projects are focused on issuing VCs and using DIDs to give users control over their data [9, 10]. Companies like Ripple and Stellar leverage blockchain for cross-border payments and fraud prevention, showcasing its potential in financial services [28, 29].

Regulatory Push for Digital Identity: Regulations like eIDAS 2.0 in the European Union, mandating the European Digital Identity Wallet (EUDI Wallet), are accelerating the shift towards digital and verifiable identities, which naturally aligns with blockchain-based SSI principles. These regulations provide a legal and operational framework for the adoption of such advanced systems.

Challenges and Future Outlook: Despite the progress, full-scale integration of AI and blockchain into a seamless onboarding experience presents challenges, including interoperability between different blockchain networks, standardization of SSI protocols, the computational demands of advanced AI fraud detection, and the continuous evolution of regulatory landscapes (e.g., the EU AI Act [5]). However, the strong regulatory incentives from the European Commission and the success of local entities like LOQR [24] and Alastria [31] indicate a clear trend towards the eventual widespread adoption of such comprehensive systems for secure, efficient, and compliant digital onboarding.

7. Conclusion

This chapter brings the thesis together. It summarises the empirical, architectural and prototype contributions, answers the two research questions posed in Chapter 1, and states the limitations of the work. It closes with directions for future work and some final remarks on where remote KYC stands.

This thesis examined secure remote KYC from two connected angles: an empirical evaluation of how current biometric and PAD mechanisms cope with contemporary synthetic media, and an architectural proposal for a blockchain-enabled trust, audit, and reusable-identity layer suited to the Portuguese banking sector. The two angles are deliberately complementary, since modern remote onboarding cannot rely on biometric controls or on blockchain-based trust mechanisms in isolation, and the chapters that precede this one together support that argument.

7.1. Summary of Contributions

The work delivered two main contributions, plus a working end-to-end prototype that exercises both.

Empirical contribution. Chapter 5 subjected five open-source verification components (DeepFace FASNet, DeepFace rPPG, the rPPG-Toolbox POS implementation, InsightFace ArcFace, and Silent-Face Anti-Spoofing) to a controlled bypass methodology using two families of synthetic video. The first family comprised three videos generated by the Hedra platform on Google Veo 3 (Videos A, B, and C), spanning two independent source identities. Three of the five layers (FASNet, InsightFace, Silent-Face) accepted all three as real on $\geq 96\%$ of sampled frames; only the rPPG-based pipelines correctly classified them as spoofs, with peakiness scores of 1.23/1.36, 1.59, and 1.27 respectively, all below the 1.70 liveness threshold. The second family comprised four videos generated directly via Google AI (Veo 3, Videos 5–8). These bypassed all four liveness checks, including the rPPG pipelines, which produced peakiness scores of 1.92–2.55 (above the 1.70 threshold) and reported a REAL verdict on every video, while InsightFace found a face in every frame. The combined finding is the empirical core of this thesis: PAD resilience is sensitive to the complete generation pipeline, not only to the underlying model. Texture-domain

and frequency-domain layers exhibit *Generalisation Failure* against both platforms. Biological liveness (rPPG) provides a meaningful security margin against Hedra-processed Veo 3 output but is itself defeated when the same underlying model is accessed through a different interface with a different generation mode. No single tested mechanism provides pipeline-universal coverage across both delivery platforms.

Architectural contribution. Chapter 6 proposed a blockchain-anchored onboarding architecture grounded in three on-chain smart contracts (`DIDRegistry`, `VCRegistry`, `AuditTrail`) and four core off-chain AI modules (document authentication with ICAO 9303 TD3/TD1 MRZ validation, ArcFace 1:1 face matching, rPPG-POS liveness, and rule-based risk scoring), complemented by three further modules for session quality, temporal consistency, and audio consistency. The architecture addresses three specific weaknesses of the current Portuguese KYC landscape: the centralised "biometric honeypot" problem highlighted by the 2024 AMA incident, the lack of a shared technical layer for checking another institution's KYC when relying on it under Article 41.º of Law n.º 83/2017 (reliance still requires agreed contractual terms, and the relying bank remains responsible), and the legal asymmetry between an institution being both the auditor of and the audited party for its own records under Article 51.º. The proposed design places only `keccak256` hashes and credential lifecycle state on chain, and binds raw PII retention to the seven-year off-chain window required by Portuguese law.

Prototype contribution. The chapter included a functional prototype with three Solidity 0.8.20 contracts, a FastAPI backend, the seven off-chain verification modules, and a four-step REST onboarding flow. The prototype is not a production system, but it is more than a sketch: every step of the workflow is implemented end-to-end against a local Hardhat node, and the empirical validation runs in Section 6.1.10 demonstrate that the pipeline produces differentiated and defensible verdicts on real Cartões de Cidadão and on both genuine and synthetic biometric inputs, culminating in a verifiable on-chain credential issuance for a successful onboarding.

7.2. Answering the Thesis Questions

The thesis advanced a single composite claim: *secure remote KYC requires both robust technical defences against modern biometric attacks and a governance/trust architecture capable of supporting auditability, controlled reuse and regulatory accountability.* The

evidence assembled in the previous chapters supports the claim on both legs, one for each research question posed in Chapter 1.

RQ1: empirical track. The two-family generator evaluation in Chapter 5 establishes that the algorithm classes required by regulation and implemented across commercial remote KYC systems (CNN texture classifiers, frequency-domain analysers, and identity matchers) can be bypassed by current generative AI without specialist effort. Biological liveness via rPPG provides an additional security margin that is absent from purely texture-based approaches, but it is not pipeline-universal: it discriminates reliably against Hedra/Veo 3 output and fails against video generated through Google's native Veo 3 interface from the same underlying model. Taken together, these findings answer both parts of RQ1: current open-source PAD mechanisms offer incomplete resistance to modern generative AI, and that resistance depends on the complete generation pipeline rather than on the underlying model weights alone. This result makes the defence-in-depth argument of EBA Guideline 2022/15 more urgent, not less: because no single mechanism, including biological liveness, provides sufficient assurance in isolation, any production deployment must layer complementary signals and continuously adapt as the same underlying models are delivered through new platforms and interfaces.

RQ2: architectural track. The architectural analysis in Chapter 6 showed that the three classic risks of remote KYC (biometric honeypot centralisation, tamper-prone single-party audit, and silo-bound third-party reliance) can be addressed with on-chain hashes plus an SSI credential layer, without storing PII on the ledger. A blockchain is one way of providing that shared, tamper-evident layer among others, such as signed logs anchored with an independent timestamping service or a shared registry operated by the supervisor. The prototype's append-only `AuditTrail`, hash-anchored `VCRegistry`, and globally-resolvable Trust Registry give technical support to the record-keeping and third-party reliance provisions of Articles 41 and 51 of Law n.º 83/2017 and to the eIDAS 2.0 and AMLR frameworks, without replacing the agreed terms and institutional responsibility those provisions require. This supports a partly affirmative answer to RQ2: a blockchain-enabled SSI architecture can address the technical side of the trust, auditability, and controlled-reuse challenges of remote banking KYC within the current Portuguese and EU regulatory landscape, but cross-institutional reuse was not piloted with a second institution, and the legal side of reliance still depends on agreements between banks.

7.3. Limitations

The honest framing of the work requires a clear statement of what was *not* demonstrated:

- *Production ledger.* All on-chain measurements were performed against a local Hardhat node with instant block confirmation. Gas costs, mempool behaviour, and confirmation latency on a permissioned Ethereum-compatible network (e.g. Hyperledger Besu, Quorum) were not benchmarked.
- *Sanctions and PEP screening.* The architecture reserves a `SanctionsCleared` event type, but the integration with the EU consolidated list, UN Security Council list, and OFAC was scoped as future work.
- *Trained risk model.* The proposed Gradient Boosted Trees scorer was substituted by a deterministic rule-based heuristic, pending labelled onboarding data, a constraint that is well established in academic KYC research where production training data is not available outside the institutions themselves.
- *Cross-institutional pilot.* The `CredentialReused` event type and the Bank B verification endpoint are implemented and unit-tested, but no second-institution deployment exists; the reuse path is exercised at the API level against the same Hardhat node.
- *EUDI Wallet integration.* Onboarding currently uses direct REST calls. Coupling to a production EUDI Wallet reference implementation depends on infrastructure not yet generally available.
- *Sample size and generator coverage for the biometric study.* The empirical evaluation covered seven synthetic videos across two generator families (three Hedra/Veo 3 and four Google AI) and a small number of genuine reference captures. While the bypass findings are internally consistent and directionally unambiguous (no tested tool achieved universal coverage across both families), quantitative FAR/FRR figures with confidence intervals would require a substantially larger and more diverse dataset spanning additional generators, subjects, and recording conditions, and are reserved for future work.

7.4. Future Work

The most natural extensions of this work are:

- *Pilot on a permissioned ledger.* Deploy the three contracts on a small-validator Hyperledger Besu network and measure end-to-end onboarding latency, transaction throughput, and gas costs under realistic banking load.
- *EUDI Wallet bridge.* Implement the holder-side wallet flow once a reference EUDI Wallet release is publicly available, and demonstrate selective disclosure (over-18 proof, residency correlation) on top of the existing VC issuance path.
- *Sanctions/PEP integration.* Add the screening module against EU/UN/OFAC consolidated lists, and connect the result to the on-chain `SanctionsCleared` event already reserved in `AuditTrail`.
- *Larger empirical study.* Expand the bypass evaluation to cover multiple generative architectures (Kling AI, Hedra non-Veo backends, Runway Gen-3, open-source LDMs) across at least dozens of subjects, and report FAR/FRR with confidence intervals for each PAD layer.
- *Replace the rule-based risk scorer.* Substitute the deterministic scorer with a trained Gradient Boosted Trees model once labelled onboarding outcomes are available, retaining the rule scorer as a fall-back and as a baseline for measuring model lift.
- *ICAO long document numbers.* The prototype reaches its first *approved* run by combining a content-aware MRZ-region crop with a relaxed TD1 validity rule (Section 6.1.10, attempt 7). The relaxed rule is a workaround: the Cartão de Cidadão's document number is longer than nine characters, and implementing ICAO 9303's rule for such numbers (a filler in the check-digit position, with the remaining characters and the check digit in the optional-data field) would restore strict check-digit conformance on Portuguese Cartões de Cidadão and allow the relaxed rule to be removed.
- *Hardware-rooted camera attestation.* The strongest medium-term mitigation against video-injection attacks is to require cryptographically signed metadata from the physical camera sensor at the moment of capture. Combining hardware attestation with rPPG-POS liveness would close the residual exposure that even the strongest biological PAD layer leaves open.

7.5. Closing Remarks

The current generation of remote KYC stands at an awkward inflection point. Liveness detection has become a regulatory requirement for unattended remote onboarding under

EBA/GL/2022/15, and Aviso n.º 1/2022 sets the conditions for identification by videoconference in Portugal, but generative AI has degraded the assurance of the very controls used to satisfy these requirements. At the same time, customer expectations of frictionless digital experience, GDPR's data-minimisation principle, and the rise of the EUDI Wallet have made the centralised-database model increasingly anachronistic.

The contribution of this thesis is to argue, with experimental evidence on the biometric side and a working architectural prototype on the governance side, that the resolution of this tension is neither "more AI" nor "more blockchain" in isolation. It is the combination of biological liveness, defence-in-depth biometric design, and an off-chain-PII / on-chain-evidence architecture grounded in concrete Portuguese legal anchors (Articles 41 and 51 of Law n.º 83/2017, Aviso n.º 1/2022, eIDAS 2.0, the AMLR). The path from a research prototype to a production deployment is long, but the work presented here suggests that the path is at least clearly visible.

References

[1] European Parliament and Council, “Regulation (EU) 2016/679 (General Data Protection Regulation),” Official Journal of the European Union, 2016, current consolidated version. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679 [Cited on pages 5, 6, 8, 82, 85, 100, 101, and 102.]

[2] ——, “Directive (EU) 2015/2366 (Payment Services Directive 2),” Official Journal of the European Union, 2015. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32015L2366 [Cited on pages 5, 9, 23, and 86.]

[3] E. Parliament and Council, “Directive (EU) 2018/843 (5th Anti-Money Laundering Directive),” Official Journal of the European Union, 2018. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018L0843 [Cited on pages 5, 11, 84, 86, and 102.]

[4] European Parliament and Council, “Directive (EU) 2022/2555 (NIS 2 Directive),” Official Journal of the European Union, 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2555 [Cited on pages 5 and 14.]

[5] ——, “Regulation (EU) 2024/1689 (Artificial Intelligence Act),” Official Journal of the European Union, 2024, accessed: 2025-04-10. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj [Cited on pages 8, 30, 31, 102, and 103.]

[6] Dutch Public Prosecution Service, “ING pays 775 million euros due to serious shortcomings in money laundering prevention,” 2018, accessed: 2025-04-10. [Online]. Available: https://www.prosecutionservice.nl/topics/settlements/news/2018/09/04/ing-pays-775-million-euros-due-to-serious-shortcomings-in-money-laundering-prevention [Cited on pages 13 and 86.]

[7] *ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection — Information security management systems — Requirements*, International Organization for Standardization Std., 2022. [Online]. Available: https://www.iso.org/standard/27001 [Cited on pages 17, 86, and 102.]

[8] S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash system,” 2008, accessed: 2025-04-10. [Online]. Available: https://bitcoin.org/bitcoin.pdf [Cited on pages 18 and 85.]

[9] W3C, “Decentralized identifiers (DIDs) v1.0,” W3C Recommendation, 2022, accessed: 2025-04-10. [Online]. Available: https://www.w3.org/TR/did-core/ [Cited on pages 20, 27, 85, and 103.]

[10] ——, “Verifiable credentials data model v1.1,” W3C Recommendation, Mar. 2022, accessed: 2026-11-17. [Online]. Available: https://www.w3.org/TR/2022/REC-vc-data-model-20220303/ [Cited on pages 20, 27, 85, and 103.]

[11] J. Benet, “IPFS - content addressed, versioned, P2P file system,” 2014, draft 3. [Online]. Available: https://arxiv.org/abs/1407.3561 [Cited on page 20.]

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org [Cited on page 21.]

[13] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699. [Cited on pages 21, 24, 37, and 40.]

[14] Z. Yu *et al.*, “Deep learning for face anti-spoofing: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5609–5634, 2023. [Cited on pages 21, 83, and 102.]

[15] P.-G. D. de Lisboa, “Legislação - PGDLisboa,” Apr. 2025, accessed: 2025-04-10. [Online]. Available: https://www.pgdlisboa.pt/leis/lei_mostra_articulado.php?nid=2750&tabela=leis&so_miolo= [Cited on pages 23, 81, 86, 87, and 102.]

[16] Banco de Portugal, “Aviso do Banco de Portugal n.º 1/2022: deveres preventivos do branqueamento de capitais e do financiamento do terrorismo,” 2022, revoga o Aviso n.º 2/2018 e a Instrução n.º 2/2021. Accessed: 2026-11-17. [Online]. Available: https://www.bportugal.pt/sites/default/files/anexos/cartas-circulares/472858571_3.docx.pdf [Cited on pages 23, 62, 86, and 102.]

[17] Jumio, “Jumio - AI-powered identity verification,” Apr. 2025, accessed: 2025-04-10. [Online]. Available: https://www.jumio.com [Cited on pages 23, 83, 102, and 103.]

[18] S. Minaee *et al.*, “Biometric recognition using deep learning: A survey,” *Artificial Intelligence Review*, vol. 56, pp. 8647–8695, 2023. [Cited on pages 23, 83, and 102.]

[19] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “RetinaFace: Single-shot multi-level face localisation in the wild,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5203–5212. [Cited on pages 24 and 38.]

[20] European Parliament and Council of the European Union, “Regulation (EU) 2024/1183 of the European Parliament and of the Council of 11 April 2024 amending Regulation (EU) No 910/2014 as regards establishing the European Digital Identity Framework,” *Official Journal of the European Union*, vol. L, no. 2024/1183, 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1183/oj [Cited on pages 24 and 32.]

[21] ISO/IEC JTC 1/SC 37 Biometrics, *ISO/IEC 30107-3:2023 Information technology — Biometric presentation attack detection — Part 3: Testing and reporting*, International Organization for Standardization Std., Jan. 2023, second edition. [Online]. Available: https://www.iso.org/standard/79520.html [Cited on pages 24, 62, 64, and 95.]

[22] European Banking Authority, “Guidelines on the use of remote customer onboarding solutions (EBA/GL/2022/15),” Nov. 2022, accessed: 2025-04-10. [Online]. Available: https://www.eba.europa.eu/sites/default/files/document_library/Publications/Guidelines/2022/EBA-GL-2022-15%20GL%20on%20remote%20customer%20onboarding/1043884/Guidelines%20on%20the%20use%20of%20Remote%20Customer%20Onboarding%20Solutions.pdf [Cited on pages 24, 52, 62, 64, 95, 96, and 98.]

[23] Banco de Portugal, “Boas práticas para a realização de videoconferência no âmbito da abertura de conta à distância,” 2020, accessed: 2025-04-10. [Online]. Available: https://www.bportugal.pt/sites/default/files/anexos/documentos-relacionados/boas_praticas_videoconferencia.pdf [Cited on page 25.]

[24] LOQR, “LOQR - the journey-as-a-service provider,” 2025, accessed: 2025-04-10. [Online]. Available: https://loqr.io [Cited on pages 25 and 103.]

[25] Agência para a Modernização Administrativa, “A chave móvel digital,” 2025, accessed: 2025-04-10. [Online]. Available: https://www.autenticacao.gov.pt/a-chave-movel-digital [Cited on page 26.]

[26] Worldcoin, “World ID by World - digital proof of human for the internet,” 2025, accessed: 2025-04-10. [Online]. Available: https://world.org/world-id [Cited on page 26.]

[27] M. A. Hannan, M. A. Shahriar, M. S. Ferdous, M. J. M. Chowdhury, and M. S. Rahman, “A systematic literature review of blockchain-based e-KYC systems,” *Computing*, vol. 105, no. 10, pp. 2089–2118, 2023. [Cited on pages 27, 85, and 102.]

[28] Ripple Labs Inc., “Ripple: Global real-time payments, anywhere,” https://ripple.com, Apr. 2024. [Online]. Available: https://ripple.com [Cited on pages 27 and 103.]

[29] Stellar Development Foundation, “Stellar: Open network for storing and moving money,” https://stellar.org, Apr. 2024. [Online]. Available: https://stellar.org [Cited on pages 27 and 103.]

[30] European Commission, “European Blockchain Services Infrastructure (EBSI),” 2025, accessed: 2025-04-10. [Online]. Available: https://ec.europa.eu/digital-building-blocks/sites/display/EBSI/Home [Cited on pages 27 and 103.]

[31] Alastria, “Alastria - blockchain ecosystem,” 2025, accessed: 2025-04-10. [Online]. Available: https://alastria.io [Cited on pages 27 and 103.]

[32] Hedra, “Hedra - scalable AI infrastructure,” Apr. 2025, accessed: 2025-04-10. [Online]. Available: https://www.hedra.com [Cited on pages 28, 33, 57, and 83.]

[33] K. AI, “Kling AI - global AI solutions,” Apr. 2025, accessed: 2025-04-10. [Online]. Available: https://klingai.com/global/ [Cited on pages 28, 33, 57, and 83.]

[34] Financial Action Task Force (FATF), “Horizon scan: Artificial intelligence and deepfakes,” FATF, Tech. Rep., Jan. 2026, accessed: 2026-11-17. [Online]. Available: https://www.fatf-gafi.org/content/dam/fatf-gafi/reports/Horizon%20Scan%20AI%20and%20Deepfakes.pdf.coredownload.inline.pdf [Cited on page 28.]

[35] Sumsub, “Deepfake detection: AI-powered biometric verification,” 2025. [Online]. Available: https://sumsub.com/deepfake-detection/ [Cited on pages 29 and 32.]

[36] BioID, “BioID liveness detection: Multi-modal PAD with rPPG integration,” 2024, specialized in 3D liveness and spectral analysis for biometric security. [Online]. Available: https://www.bioid.com/liveness-detection/ [Cited on page 29.]

[37] iProov, “iProov Flashmark: Patented one-time biometric liveness detection,” 2025. [Online]. Available: https://www.iproov.com/technology/flashmark [Cited on page 29.]

[38] Onfido (Entrust), “Onfido Atlas: AI-powered identity verification orchestration,” 2025. [Online]. Available: https://onfido.com/atlas-ai/ [Cited on page 30.]

[39] Coalition for Content Provenance and Authenticity (C2PA), “C2PA technical specification v1.3,” C2PA, Tech. Rep., Dec. 2023, defining the standard for cryptographic content provenance and authenticity. [Online]. Available: https://c2pa.org/specifications/ [Cited on page 30.]

[40] Google Developers, “Android Play Integrity API documentation,” 2025, security measure for verifying device and app integrity against injection attacks. [Online]. Available: https://developer.android.com/google/play/integrity [Cited on pages 30 and 32.]

[41] Apple Developer Documentation, “Establishing device integrity with App Attest,” 2025. [Online]. Available: https://developer.apple.com/documentation/devicecheck/establishing_device_integrity_with_app_attest [Cited on page 30.]

[42] iProov, “iProov threat intelligence report 2024: The impact of generative AI on remote identity verification,” iProov Ltd, Tech. Rep., Feb. 2024. [Online]. Available: https://www.iproov.com/reports/iproov-threat-intelligence-report-2024 [Cited on page 31.]

[43] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in *International Conference on Computer Vision (ICCV)*, 2019. [Online]. Available: https://github.com/ondyari/FaceForensics [Cited on pages 31 and 53.]

[44] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, “Face X-Ray for more general face forgery detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5001–5010. [Online]. Available: https://arxiv.org/abs/1912.13458 [Cited on pages 31, 53, and 72.]

[45] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A large-scale challenging dataset for DeepFake forensics,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3204–3213. [Cited on pages 31 and 53.]

[46] G. DeepMind, “Veo: Google’s most capable video generation model,” 2025, accessed: 2025-05-14. [Online]. Available: https://deepmind.google/technologies/veo/ [Cited on pages 32, 33, 57, and 83.]

[47] Iperov, “DeepFaceLive: Real-time face swap for PC streaming or video calls,” 2024, accessed: 2025-04-10. [Online]. Available: https://github.com/iperov/DeepFaceLive [Cited on pages 33, 34, and 58.]

[48] Google DeepMind, “Watermarking AI-generated text and video with SynthID,” Google DeepMind Blog, Oct. 2024, accessed: 2026-09-08. [Online]. Available: https://deepmind.google/blog/watermarking-ai-generated-text-and-video-with-synthid/ [Cited on pages 34 and 72.]

[49] A. Denny, “reverse-SynthID: Reverse engineering Gemini’s SynthID detection,” Software repository, GitHub, Dec. 2025, repository created December 2025; the spectral-subtraction (V3) results cited here were added in February 2026. Accessed: 2026-09-08. [Online]. Available: https://github.com/aloshdenny/reverse-SynthID [Cited on pages 35 and 73.]

[50] S. I. Serengil and A. Ozpinar, “LightFace: A hybrid deep face recognition framework,” in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2020, pp. 23–27. [Cited on pages 37, 43, and 53.]

[51] S. I. Serengil, “DeepFace: A lightweight face recognition and facial attribute analysis framework for Python,” 2026, repository page accessed 2026-04-23. [Online]. Available: https://github.com/serengil/deepface [Cited on page 37.]

[52] J. Deng, J. Guo, X. An, J. Zhu, and S. Zafeiriou, “InsightFace: 2D and 3D face analysis project,” 2021, accessed: 2025-04-10. [Online]. Available: https://github.com/deepinsight/insightface [Cited on pages 39 and 53.]

[53] National Institute of Standards and Technology (NIST), “Face technology evaluations - FRTE/FATE,” Apr. 2025, updated 2025-04-22; includes the ongoing FRTE 1:1 benchmark and latest reports. [Online]. Available: https://www.nist.gov/programs-projects/face-technology-evaluations-frtefate [Cited on page 39.]

[54] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, “Sample and computation redistribution for efficient face detection,” *arXiv preprint arXiv:2105.04714*, 2021. [Online]. Available: https://arxiv.org/abs/2105.04714 [Cited on page 40.]

[55] Minivision AI, “Silent-Face-Anti-Spoofing: A practical face anti-spoofing solution,” 2020, accessed: 2025-04-10. [Online]. Available: https://github.com/minivision-ai/Silent-Face-Anti-Spoofing [Cited on pages 41 and 53.]

[56] S. Chen, Y. Liu, X. Gao, and Z. Han, “MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices,” in *2018 Chinese Conference on Biometric Recognition (CCBR)*, 2018, pp. 428–438. [Online]. Available: https://arxiv.org/abs/1804.07573 [Cited on page 41.]

[57] X. Liu, G. Narayanswamy, A. Paruchuri, X. Zhang, J. Tang, Y. Zhang, Y. Wang, S. Sengupta, S. Patel, and D. McDuff, “rPPG-Toolbox: Deep remote PPG toolbox,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2023, neurIPS 2023 Datasets and Benchmarks Track. GitHub repository: https://github.com/ubicomplab/rPPG-Toolbox. [Online]. Available: https://github.com/ubicomplab/rPPG-Toolbox [Cited on pages 43 and 53.]

[58] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, “Algorithmic principles of remote PPG,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2017. [Cited on page 43.]

[59] K. Kvanchiani, E. Petrova, K. Efremyan, A. Sautin, and A. Kapitanov, “EasyPortrait: Face parsing and portrait segmentation dataset,” arXiv preprint arXiv:2304.13509, 2023, accessed: 2026-09-08. [Online]. Available: https://arxiv.org/abs/2304.13509 [Cited on page 66.]

[60] Y. Mirsky and W. Lee, “The creation and detection of deepfakes: A survey,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021. [Cited on pages 83 and 102.]

[61] International Civil Aviation Organization, *Doc 9303: Machine Readable Travel Documents, Eighth Edition*, ICAO, Montréal, Canada, 2021, part 3: Specifications Common to all MRTDs (Annex B: Check Digits); Part 4: Specifications for Machine Readable Passports (TD3); Part 5: Specifications for TD1 Size Machine Readable Official Travel Documents. [Online]. Available: https://www.icao.int/publications/pages/publication.aspx?docnum=9303 [Cited on pages 84 and 93.]

Appendix A: Prototype Implementation

A.1 Source code availability

The prototype described in Chapter 6 and measured in Section 5.4 is published in full at:

`https://github.com/MrRedPTT/KYC-Thesis-Implementation`

All results reported in this thesis correspond to commit `6dbd3b2`. The repository comprises 41 files: three Solidity contracts, the FastAPI backend, seven off-chain verification modules, the contract and backend test suites, and two end-to-end runner scripts.

A.2 Repository layout

- `blockchain/`: Hardhat project. `contracts/` holds `DIDRegistry.sol`, `VCRegistry.sol` and `AuditTrail.sol`; `scripts/deploy.js` deploys all three and registers the issuer; `test/` contains the Mocha contract tests.
- `backend/`: FastAPI application. `app/routers/` exposes the onboarding, verification and identity endpoints; `app/services/` bridges to the verification modules and, through `web3.py`, to the deployed contracts; `app/utils/credential_id.py` holds the shared `bytes32` encoding used on both sides of a credential check.
- `ai_verification/`: The seven off-chain modules: `document_auth.py`, `face_verification.py`, `liveness.py`, `session_quality.py`, `temporal_consistency.py`, `audio_consistency.py` and `risk_scoring.py`.
- `test_onboarding.ps1`, `kyc_test.py`: End-to-end runners that drive the full session and print a pass/fail line for every check.
- `ARCHITECTURE.md`: Design rationale, component map, per-step request flow, and an explicit statement of what the design does not solve.

A.3 Smart contract interfaces

`DIDRegistry.sol` anchors W3C DID Documents. `createDID`, `updateDID` and `deactivateDID` are guarded by an `onlyController` modifier so that only the original registrant may modify a given DID; `resolveDID` and `exists` are public reads.

`VCRegistry.sol` manages the credential lifecycle and implements the Trust Registry. `authorizeIssuer` and `revokeIssuerAuthorization` are restricted to the registry admin; `issueCredential`, `revokeCredential` and `suspendCredential` require `onlyAuthorizedIssuer`. `verifyCredential` is the function that makes cross-institutional reliance possible: any participant may call it to confirm hash integrity and revocation status without contacting the original issuer.

`AuditTrail.sol` is an append-only log. `recordEvent` writes a `keccak256` hash of the off-chain event payload together with the subject DID and one of nine `EventType` values (*OnboardingStarted*, *DocumentVerified*, *BiometricVerified*, *SanctionsCleared*, *RiskAssessed*, *CredentialIssued*, *OnboardingCompleted*, *OnboardingRejected*, *CredentialReused*). There is no update or delete function; `getEvent`, `getEventsByDID` and `totalEvents` are reads.

A.4 Reproducing a session

The repository README gives the full setup. In summary: start a local Hardhat node, deploy the three contracts, record the resulting addresses in `.env`, clone rPPG-Toolbox into `Tools/` (it is not distributed on PyPI), install the Python dependencies, and start the backend. A complete onboarding session is then driven by a single command:

```
python kyc_test.py --doc <document> --video <clip>
```

The runner exercises the same code path as the validation runs of Section 6.1.10 and reports each verification outcome individually, so a rejection can be attributed to a specific gate.

A.5 What is deliberately not published

The evaluation used the author's own Portuguese Cartão de Cidadão and selfie recordings, together with those of one further consenting subject. These are identity documents and biometric samples of identifiable individuals. Publishing them would be difficult to reconcile with the data-protection reasoning the architecture itself is built on, and they are therefore withheld.

The consequence should be stated plainly: the validation runs cannot be reproduced byte-for-byte from the repository alone. They can be reproduced in structure. Supplying any identity document and a matching video clip exercises the same modules, the same

thresholds and the same contract calls, and emits the same on-chain event sequence. The synthetic videos used in the bypass experiments of Chapter 5 are likewise not redistributed, as they depict identifiable individuals in fabricated recordings.

The repository contains no secrets. The only private key it carries is the well-known Hardhat development key published in that tool's own documentation, present in `.env.example` so that a local node can be driven without further configuration.