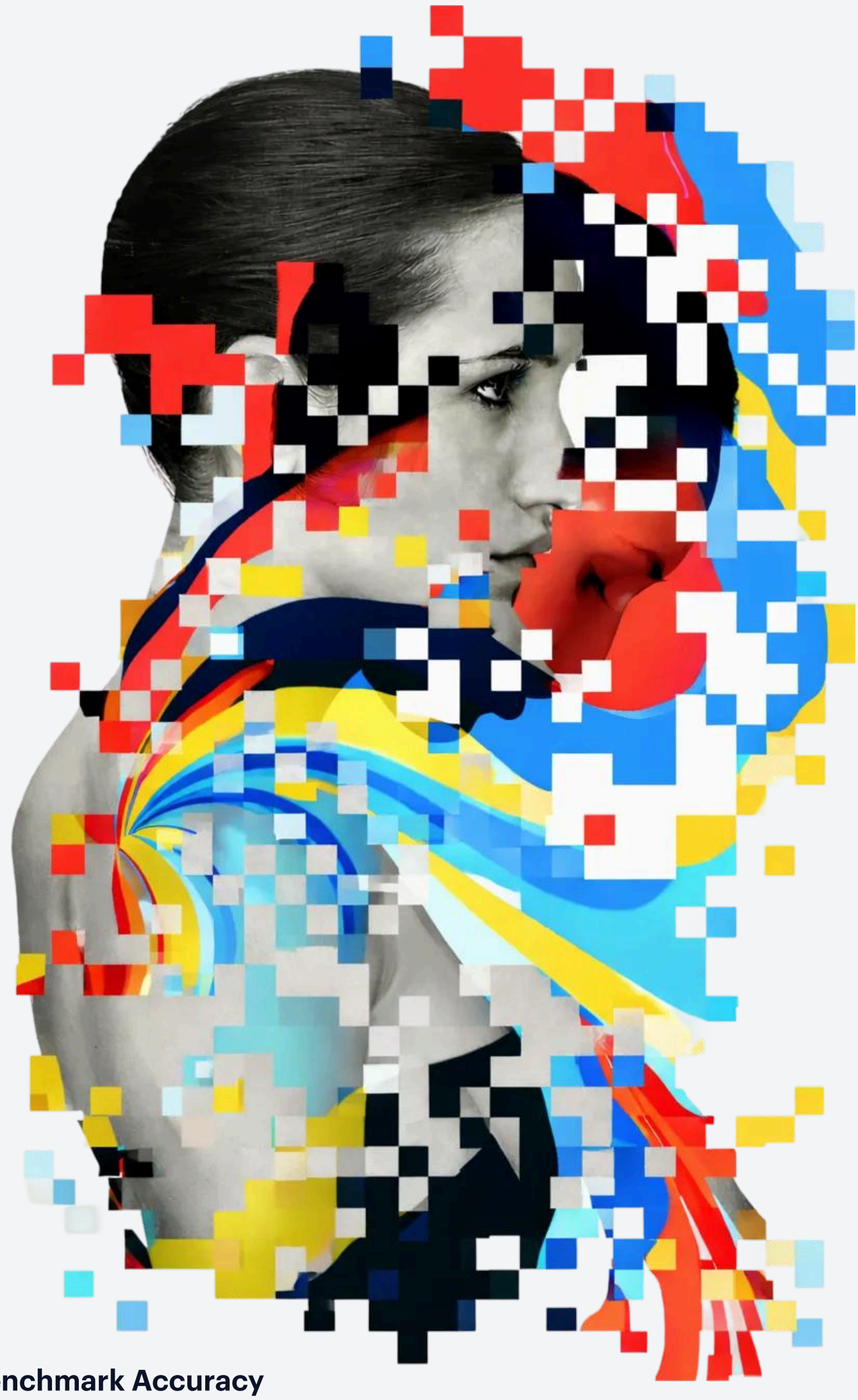
WHITEPAPER

Shufti



Beyond Benchmark Accuracy

Making Deepfake Detection Work for IDV Systems

Trust the Proxy

We live in a world where identity happens in a second. A face scan unlocks a phone, a selfie clears KYC, and a camera confirms who’s on the other side.

Deepfakes don’t break these flows with noise. They slip in quietly as a perfect proxy at the exact moment trust is assumed and decisions move fast. When they succeed, the failure is silent. It passes in real time. It’s discovered only after loss, escalation, or regulatory scrutiny.

This paper challenges the confidence industry's place in benchmark-led deepfake detection. High scores on controlled datasets do not mean real protection in live systems.

The real world is hostile to detectors. Media gets compressed, re-encoded, downsampled, and captured under inconsistent lighting, bandwidth, and device conditions. And those conditions destroy the very signals many detectors rely on. What looks “state-of-the-art” in a lab can become fragile in production.

Contents.

55.5%

Average accuracy of humans detecting deepfakes.[1]

56papers

Evidence base covering 86,155 participants.[2]

15%

People report exposure to harmful deepfakes.[3]

50.2%

Celebrities are the most common deepfake targets.[4]

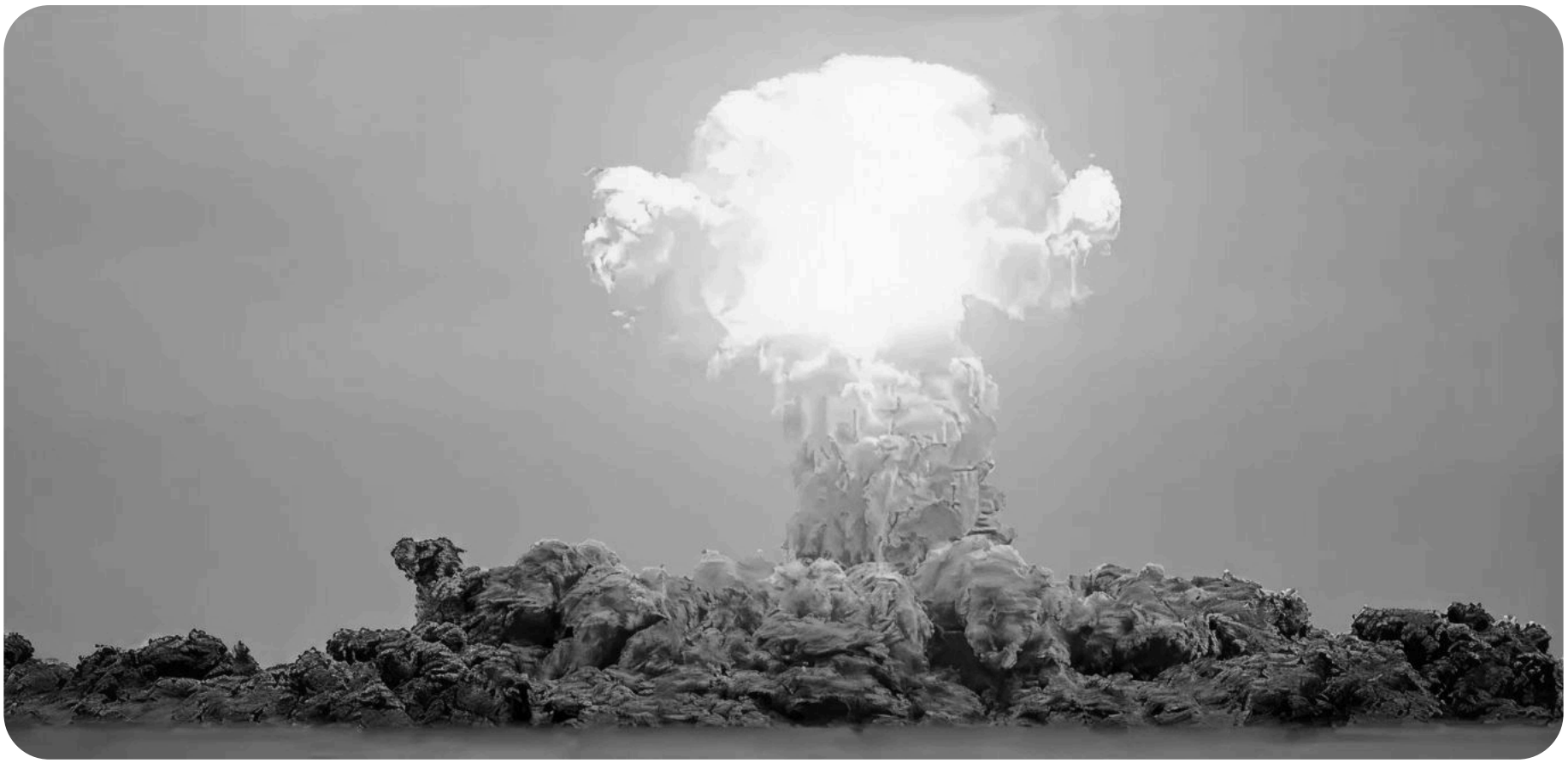
The Problem with Industry's Deepfake Detection Claims

Most detection claims sound stronger than they are because models are trained and tested on controlled datasets, and the results are presented in a generalized context to real identity workflows.

Real-world verification looks far different. In practice, it's compressed, re-encoded, and captured on inconsistent devices in imperfect conditions. However, the question isn't whether a detector can identify manipulation, but whether it survives platform processing and an attacker optimized for one decisive moment.

This gap is why benchmarks are being questioned and why scenario-specific evaluation is becoming necessary. The VCF[5] benchmark is one example of this shift, using video conferencing as a representative high-trust setting. Its approach could also be tested against scenarios such as remote onboarding and live identity checks, where even minor variations in capture conditions may influence detection performance.

[1] https://www.sciencedirect.com/science/article/pii/S2451958824001714
[2] https://www.sciencedirect.com/science/article/pii/S2451958824001714
[3] https://www.turing.ac.uk/news/publications/behind-deepfake-8-create-90-concerned
[4] https://www.turing.ac.uk/news/publications/behind-deepfake-8-create-90-concerned
[5] https://isprs-archives.copernicus.org/articles/XLVIII-2-W9-2025/169/2025/



The Arms Race of Deepfake Detection

Deepfake detection often fails as the threat evolves in response to detection systems with improved generation techniques that adapt to avoid the cues those systems learn to recognize. This dynamic turns protection into an ongoing contest rather than a one-time capability.

Regulators and risk bodies increasingly describe deepfakes as an arms race. For instance, Australia's eSafety[6] Commissioner has compared staying ahead of deepfake abuse to "fighting an arms race," while the World Economic Forum[7] highlights an asymmetric dynamic in which advances in generation often outpace detection. The implication is consistent. Static controls and fixed models degrade over time because deepfakes behave like a moving target.

[6] "Fighting the deepfakes arms race | eSafety Commissioner," eSafety Commissioner, Oct. 11, 2019. https://www.esafety.gov.au/newsroom/blogs/fighting-deepfakes-arms-race

[7] B. Colman, "Detecting dangerous AI is essential in the deepfake era," World Economic Forum, Jul. 07, 2025. https://www.weforum.org/stories/2025/07/why-detecting-dangerous-ai-is-key-to-keeping-trust-alive/

Understanding the Line Between Creation and Impersonation

Synthetic media and deepfakes are often created using the same underlying AI models. That is why they are frequently talked about as if they are the same thing. The difference is not the technology. It is what the media is doing and how it is used.

Think of it like this

The same camera can take a family photo or a forged document. The tool is the same, except for the risk.

What does "synthetic media" mean?

Synthetic media, as described in UNESCO's policy primer, refers to digital content, including images, audio, and video, that is created or modified using artificial intelligence and related technologies. It encompasses both benign and malicious uses of AI-generated content and forms the broader category within which deepfakes exist.[8]

Synthetic media refers to any image, video, or audio that is created or altered using AI, regardless of intent. This includes content such as AI-generated faces or voices that do not represent a real person, virtual avatars used for presentations, text-to-speech systems, automated dubbing or translation, and media created for marketing, training, entertainment, or accessibility.

[8] https://unesdoc.unesco.org/ark:/48223/pf0000392181

In most cases, synthetic media is not designed to deceive and does not involve identity at all. It is simply a method of content creation or enhancement, widely used across legitimate digital applications.

What makes something a deepfake?

Deepfakes, according to Europol's law-enforcement reporting, are a form of synthetic media that can be used to impersonate real individuals in ways that may deceive people or systems and facilitate criminal activity.[9]

Examples:

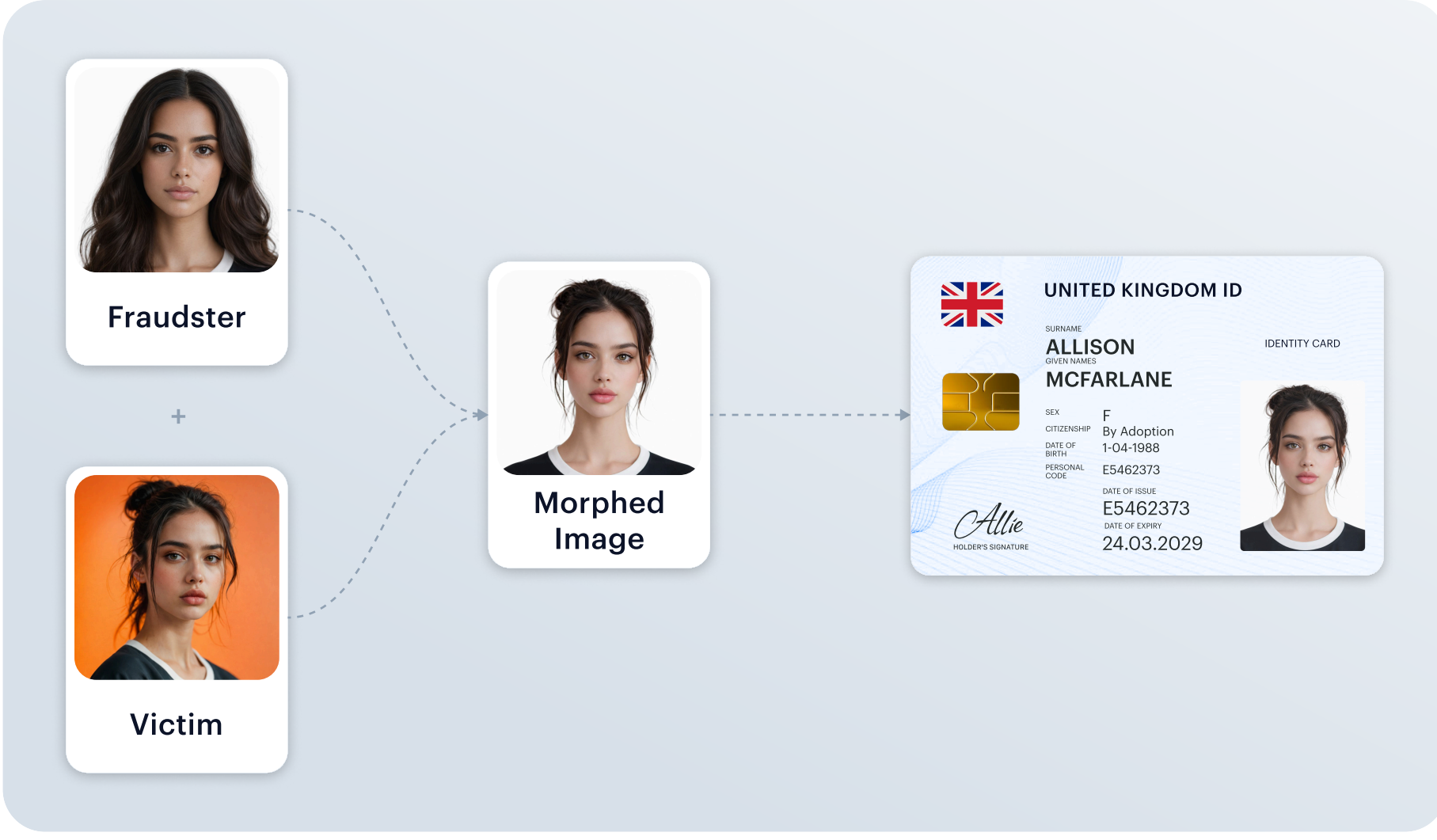
- A synthetic face used to pass identity verification
- A cloned voice used to authorise a payment
- A generated video used to impersonate someone in a live call

At that point, the media is no longer just synthetic. It is acting as a proxy for a real identity.

[9] https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes

Traditional Morph Attacks Labelled as Deepfake

Morphing is one example to illustrate a common form of impersonation that is often mistaken for deepfakes, even though it relies on real media or simple manipulation rather than synthetic generation.



Morphing is not always a deepfake. Traditional morphs rely on image manipulation, while AI-generated morphs use face synthesis and therefore qualify as deepfakes.

What traditional morphs look like and are considered deepfakes:

- It’s well-documented in identity fraud
- It has appeared in passport and ID document attacks
- It produces convincing results without AI generation

At that point, the media is no longer just synthetic. It is acting as a proxy for a real identity.

The examples described above refer to traditional morph attacks created through image manipulation such as blending, warping, retouching, pixel averaging, and landmark alignment. AI-generated morphs exist as well, but they differ in appearance and are classified as deepfakes.

But the real question is:

Can detection models built for photo-edited morphs reliably identify morphs that are generated using AI?

Types of Deepfake and Synthetic Media Used in Identity Spoofing

In remote identity verification (IDV) and onboarding, deepfake-enabled fraud tends to concentrate into four dominant forms, such as face swaps, voice cloning, lip-sync/face reenactment, and synthetic face generation, each of which undermines trust in a different part of the onboarding pipeline.

1. Face-swap deepfakes

Threat intensity: Very high

A face-swap deepfake replaces one person's face with another in an image or video while preserving lighting, pose, and context.

2. Voice cloning/audio deepfakes

Threat intensity: High

Voice cloning generates speech that imitates a specific person's vocal characteristics from limited audio samples.

3. Lip-sync and face reenactment deepfakes

Threat intensity: Very high

Lip-sync and face reenactment manipulate facial motion so a subject appears to speak specific words or display expressions never performed.

4. Synthetic face generation

Threat intensity: High

Synthetic face generation creates photorealistic faces that do not correspond to real individuals.

Different Threat Scenarios Involving Deepfakes

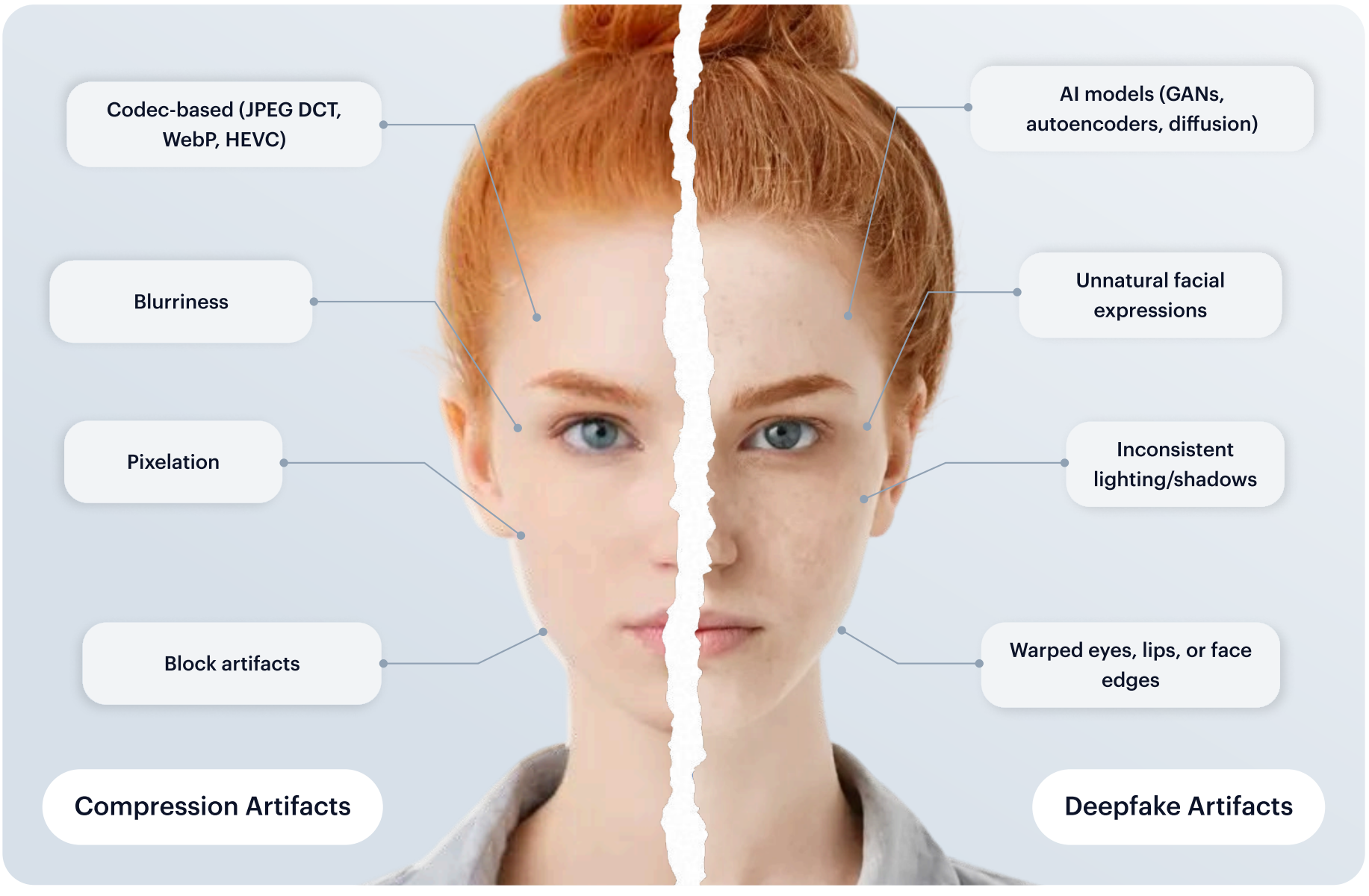
There are multiple threat scenarios involving deepfakes, and the risks vary depending on the attacker's goal, the delivery channel, and the trust context in which the media is used. Common scenarios include:

- Deepfakes as disinformation weapons
- Non-consensual deepfakes and reputation attacks
- Deepfakes as an organised crime enabler
- Live deepfake impersonation and executive fraud in Video Meeting
- Photo ID morphing via deepfake face swaps to bypass identity checks
- Deepfake spoofing attacks on facial biometrics

While these threat scenarios differ in intent and execution, a key question remains: can deepfake detectors trained on specific generators reliably detect deepfakes across these threat scenarios?

Why Compression Artefacts Are a Core Challenge for Deepfake Detection

Compression artefacts are not deepfakes, but they can affect deepfake detection performance. Many detection methods rely on fine-grained visual patterns in images and video, such as texture detail, edge consistency, and noise characteristics. These patterns can be altered or removed when media quality decreases due to compression or repeated re-encoding.



In identity and onboarding workflows, media is often compressed or reprocessed during capture, upload, transmission, and storage. The moment it reaches a detection system, some visual detail may already be reduced or distorted, which can make forensic cues harder to measure consistently and may contribute to both false positives and false negatives, depending on the model, thresholds, and capture conditions.

What do deepfake detectors typically depend on?

- **Fine facial detail:** Natural skin texture and micro-variation
- **Clean transitions:** Consistent edges around eyes, lips, and hairlines
- **Natural noise:** Patterns that come from real camera sensors
- **Frame-to-frame stability:** How details behave across video frames

Compression alters these same cues in ways that can resemble synthetic media:

- **Over-smoothing** can mimic artificial skin texture
- **Blockiness and banding** can look like generated image artefacts
- **Edge ringing** can resemble poorly blended edits
- **Texture loss** removes the natural “real camera” signature

The Critical Breakpoint of Models

Once compression crosses a threshold, the model isn't really spotting manipulation anymore. It's making a call based on media that's already been degraded. Compression disrupts the exact signals the model depends on.

Model Needs	What Compression Does
Fine Texture	Removes it
Natural noise	Replaces it with block noise
Clean edges	Introduces ringing and aliasing
Temporal stability	Creates flicker and inconsistency

This breakpoint often leads to two predictable results:

Over-Blocking (False Positives)

- Real users flagged as deepfakes
- More retries, manual review, and abandonment
- Trust and conversion drop

Under-Blocking (False Negatives)

- Thresholds relaxed to protect UX
- Sophisticated attacks pass by hiding signals in low-quality
- Fraud enters the system and is discovered later

Why Existing Deepfake Detection Approaches Fall Short in Real-World Threat Scenarios

Academic studies consistently show that deepfake detectors achieving strong results on popular benchmarks often degrade when exposed to real-world media. Most benchmark datasets are lab-style, synthetic, or narrowly scoped and do not reflect how deepfakes appear in practice, where content is compressed, re-encoded, reposted, and captured under inconsistent devices, lighting, and network conditions.

As a result, benchmark accuracy may primarily reflect relative performance under controlled conditions, not robustness against adversarial, production-level spoofing attempts. High scores on benchmarks should therefore be interpreted as comparative indicators rather than evidence of deployment readiness in identity verification or onboarding systems.

Why Vendor Accuracy Numbers Without Context Are Misleading

Vendors report deepfake detection performance using different datasets, generators, capture conditions, and decision thresholds, so headline percentages may refer to scenarios that are either not applicable to remote identity proofing or limited to a few deepfake generation engines. In many cases, the datasets used and the operating point used to measure FAR/FRR-type outcomes are not disclosed, which makes independent comparison difficult.

Additionally, some metrics are reported for off-site detection on limited generator sets or framed around iBeta Level 2-style certification contexts, neither of which consistently represents real-world accuracy for detecting deepfakes for threat scenarios like injection attacks during remote onboarding. Where datasets are disclosed, they often reflect limited generator coverage, which may not generalize to newer or more diverse deepfake methods.

Vendor 1

99.98%

Deepfake detection claimed on first attempt
— Vendor 1

Limitation: Benchmarking methodology and performance under production-like attack conditions not disclosed.

Vendor 2

99%

Detection rate for DeepFaceLive-generated deepfakes
— Vendor 2

Limitation: Performance measured against a single deepfake generator and may not generalize across broader deepfake attack methods.

Vendor 3

91.07%
Image

77.27%
Video

Deepfake detection accuracy reported in Purdue PDID results
— Vendor 3

Limitation: Results are benchmark- and dataset-specific and depend on decision thresholds.

Vendor 4

2.3%
BPCER @ 1% APCER (Minimum accepted frauds)

1.7%
Both APCER and BPCER (Equal error rate)

4.4%
APCER @ 1% BPCER (Minimum rejected attempts)

Operational error rates across defined operating points
— Vendor 4

Limitation: Reports error rates rather than accuracy; comparability depends on threat models and capture/injection controls.

*Claims referenced are based on publicly available information and have not been independently verified by Shufti.

Generalizability and Domain Shift

Why Deepfake Detection Breaks Outside the Lab

A core limitation in deepfake detection research is generalizability: performance achieved on controlled benchmarks often fails to transfer to real-world conditions. Most detectors are trained and tested on synthetic datasets that do not reflect how deepfakes appear in practice, where content is compressed, re-encoded, streamed through platform codecs, captured on inconsistent devices, and shaped by environmental noise. Under these conditions, the signals many detectors rely on are weakened or distorted, which can cause significant accuracy degradation.

This gap shifts the real question from

"How accurate is the model on a benchmark?"

to

"Does it hold up in the attack scenarios and capture channels that matter?"

Because deepfake generation and evasion techniques evolve rapidly, robust detection requires scenario-based evaluation and multi-signal (multi-model) systems rather than relying on a single detector trained to recognize a fixed set of artefacts.

How Vendors Commonly Interpret PDID Results

A common pattern can be seen in how commercial vendors present PDID-style results. For example, Incode publicly references its performance in the third-party PDID evaluation conducted by Purdue University, highlighting metrics such as accuracy and a reported False Accept Rate (FAR) of 2.56% as indicators of deepfake detection capability.

While these figures are valid within the scope of the PDID benchmark, they are often elevated to serve as de facto proof of real-world deepfake protection. This interpretation extends beyond what the dataset is designed to measure.

PDID does not capture several threat scenarios critical to identity verification, including injection attacks during live onboarding, controlled capture environments, or adaptive adversaries targeting biometric workflows. Treating benchmark outcomes as a final claim, therefore, overstates real-world detection readiness.

Shufti’s PDID Stress Test Results

99.54%

Shufti’s Accuracy at 0.46% FAR

98.28%
Cheapfake Detection Accuracy

99.63%
Deepfake Detection Accuracy

Benchmark datasets like PDID are useful for understanding how detection systems behave under specific, known conditions. They help identify whether performance collapses under compression, platform noise, or other forms of media degradation.

PDID as a Robustness Stress Test

Benchmark Performance is not Where Deepfake Risk is Ultimately Decided

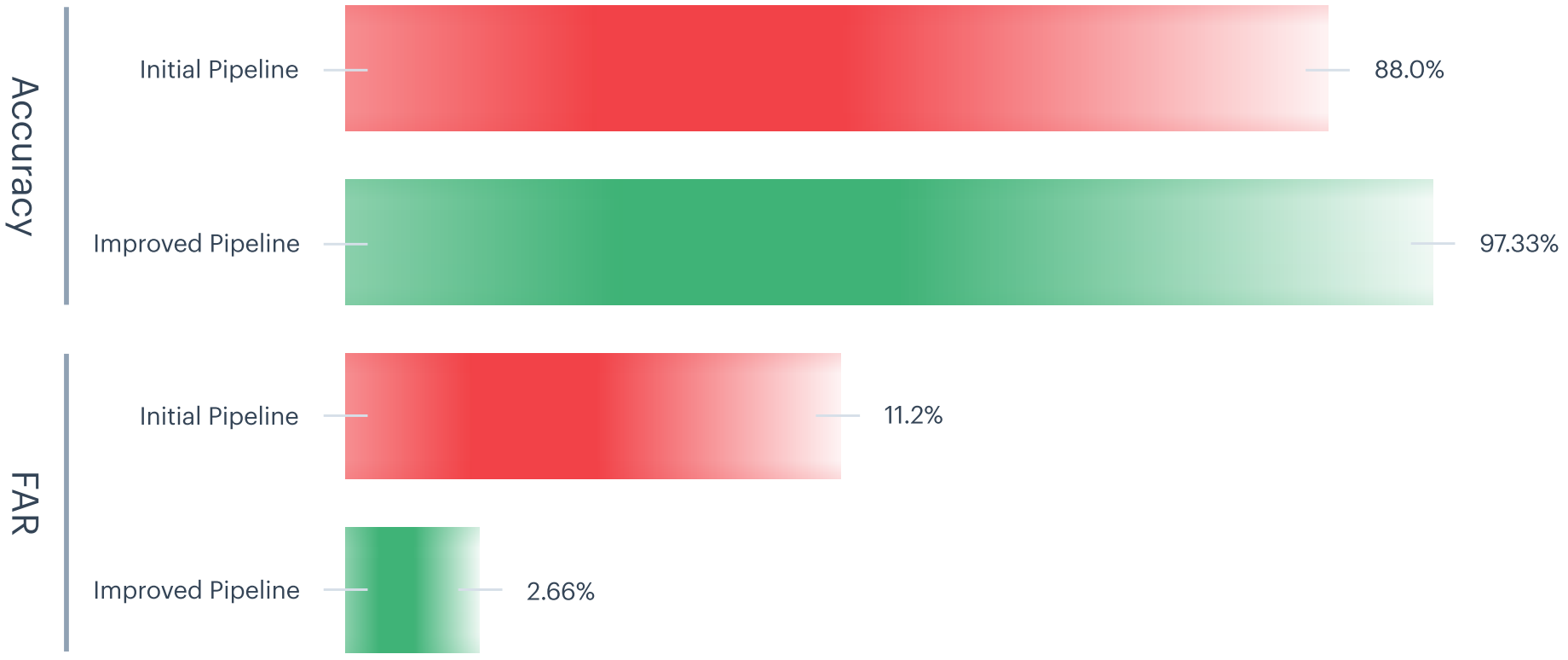
The harder question is whether that performance holds when deepfakes evolve, attacks target live onboarding flows, and adversaries optimize for specific workflows rather than dataset artefacts. Real-world deepfake detection is decided outside benchmarks.

For this reason, Shufti's approach does not end with achieving higher detection accuracy on limited in-the-wild datasets such as PDID. Instead, it extends to evaluation against emerging deepfake generators. The focus shifts to how detectors perform within the verification pipeline, how they adapt to new attack patterns, and how they are continuously improved in response to emerging threats.

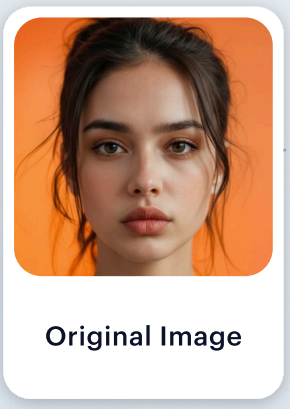
Shufti's Multi-Model Detection Strategy

Shufti's deepfake detection is implemented as a multi-model pipeline, combining multiple independent detection signals rather than relying on a single model or feature. This reduces sensitivity to real-world factors such as compression, capture variability, and evolving generation techniques, where any single signal can degrade or fail.

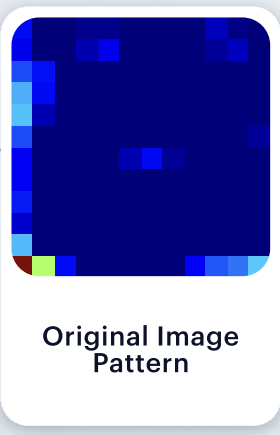
Shufti continuously improves this pipeline through scenario-based testing, observed failure patterns, and iterative updates to models and decision thresholds so performance stays reliable as media conditions and attack methods change.



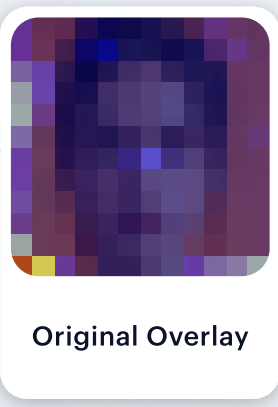
Shufti puts this multi-model approach into practice through a structured forensic evaluation flow. Instead of relying on one model to make a single-pass decision, the system checks several independent signals that are known to fail under real-world conditions such as compression, re-encoding, low-quality capture, and post-processing. Any one signal can weaken on its own, but when multiple signals point to the same outcome, the risk assessment becomes far more reliable.



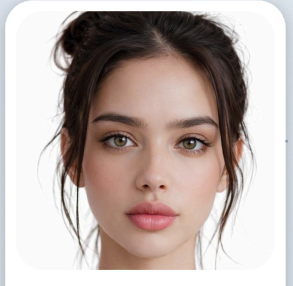
Original Image



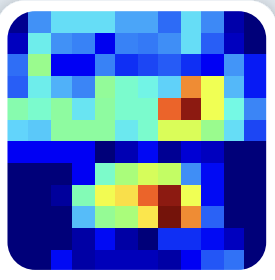
Original Image Pattern



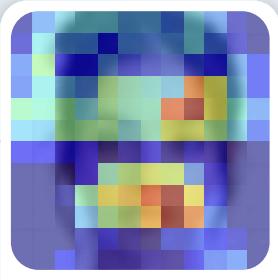
Original Overlay



Deepfake Image



Deepfake Image Pattern



Deepfake Overlay

Why Single-Model Deepfake Detection Fails in Real Pipelines

Deepfake detection is often presented as a clean, single-model problem. In practice, research and real deployments show that no single detector stays reliable across changing generators, channels, and capture conditions.

Compression cues can help in one pipeline but weaken in another. Re-encoding, transcoding, and heavy compression can alter or replace the very features a detector relies on, causing performance to drop outside its original workflow.

Spatial and frequency-based signals can be informative in ideal media. In production capture, low resolution, poor lighting, motion blur, latency, and post-processing can mask or distort these patterns, reducing real-world consistency.

Quantization analysis and camera fingerprints (e.g., PRNU) also have constraints. Editing, recompression, denoising, stabilization, cropping, or resizing can change or suppress the signals needed for these methods to work.

Because each signal family can degrade under routine handling, relying on any one method creates a structural weakness. A multi-model approach cross-checks independent cues (compression, spatial/frequency, quantization, and camera fingerprints) to produce more resilient detection decisions.

The Design Rationale Behind the Seven Gates of Shufti

Shufti's thinking behind the **Seven Gates** is straightforward: deepfakes don't sneak in through a hidden back door. They come through the **same front door as real users**, using normal capture flows and media that look authentic. The real challenge is not just spotting synthetic content but keeping decisions reliable when the input appears clean, passes through legitimate pipelines, and is intentionally engineered to survive verification checks. That risk increases as new-generation engines appear that did not exist in yesterday's training data.

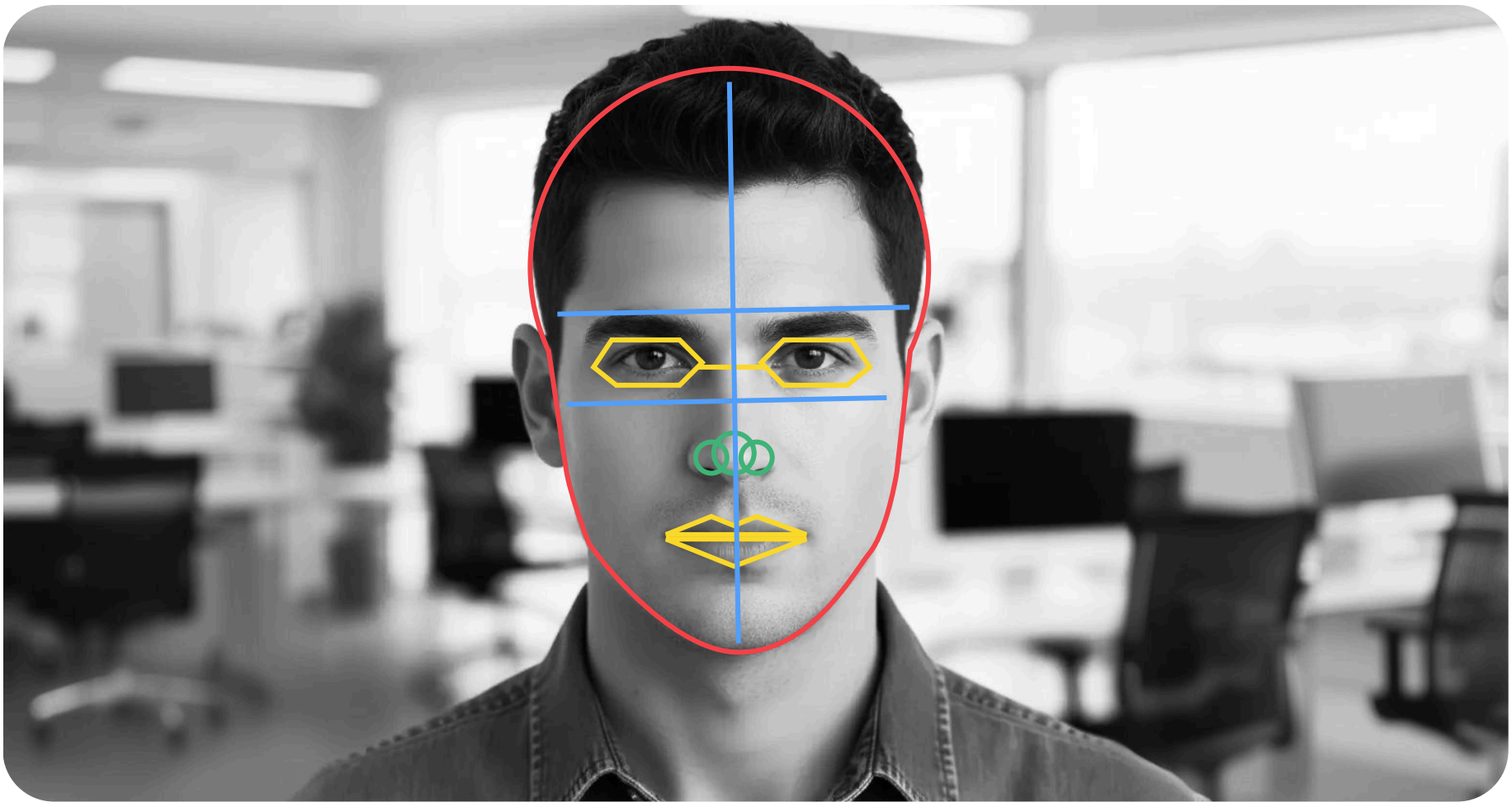
The **Seven Gates** evidence domains reflect the main independent ways authenticity breaks in the real world, including:

- Biometric structure
- Generator artefacts
- Compression history
- Frequency behaviour
- Texture realism
- Robustness under degradation
- Pixel-level coherence

Using fewer than seven creates predictable blind spots because it skips entire classes of evidence. Using more than seven often adds overlap, increasing complexity and tuning risk without meaningfully improving resilience.

The Seven Gates Forensic Evaluation Framework

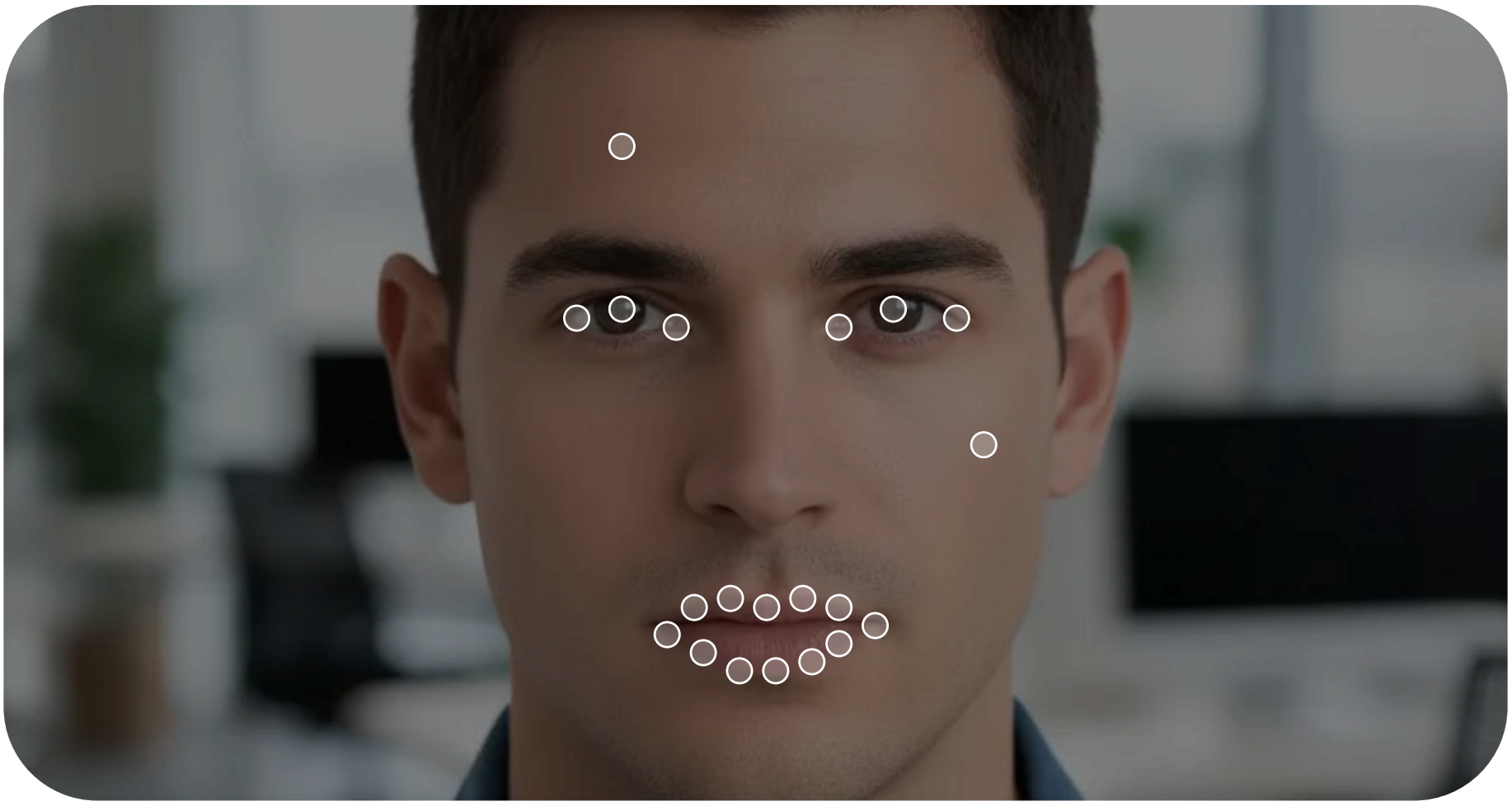
The Seven Gates translates Shufti's multi-model architecture into a sequential forensic evaluation framework. Each gate tests a different hypothesis about authenticity, and deepfake risk emerges only when multiple independent signals align.



Gate One

The Biometric Detective

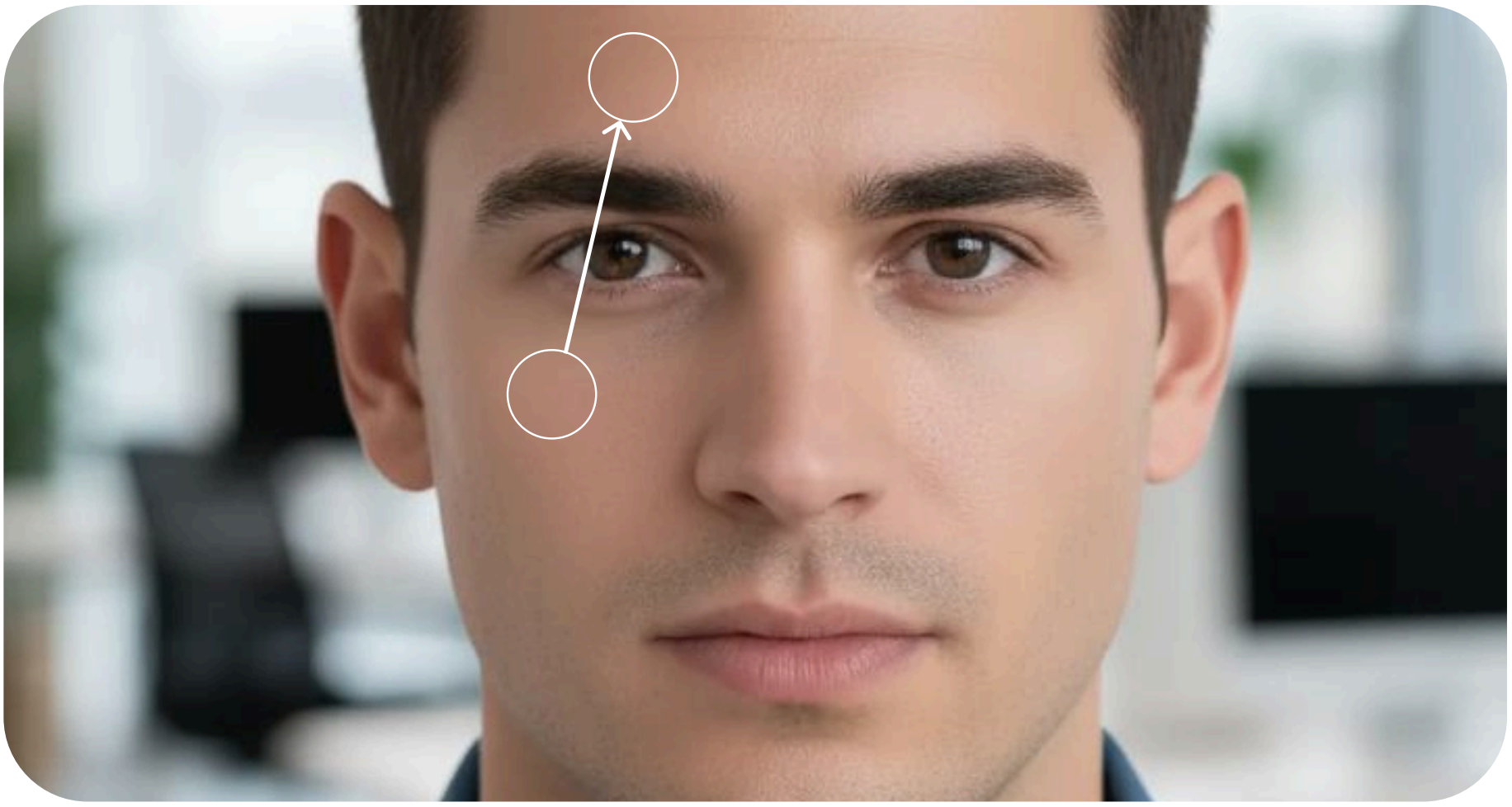
This gate tests whether the face follows real human structure and motion. It checks geometry and consistency across movement and expression, looking for subtle distortions that often appear when synthetic faces try to stay coherent over time.



Gate Two

The AI Signature Hunter

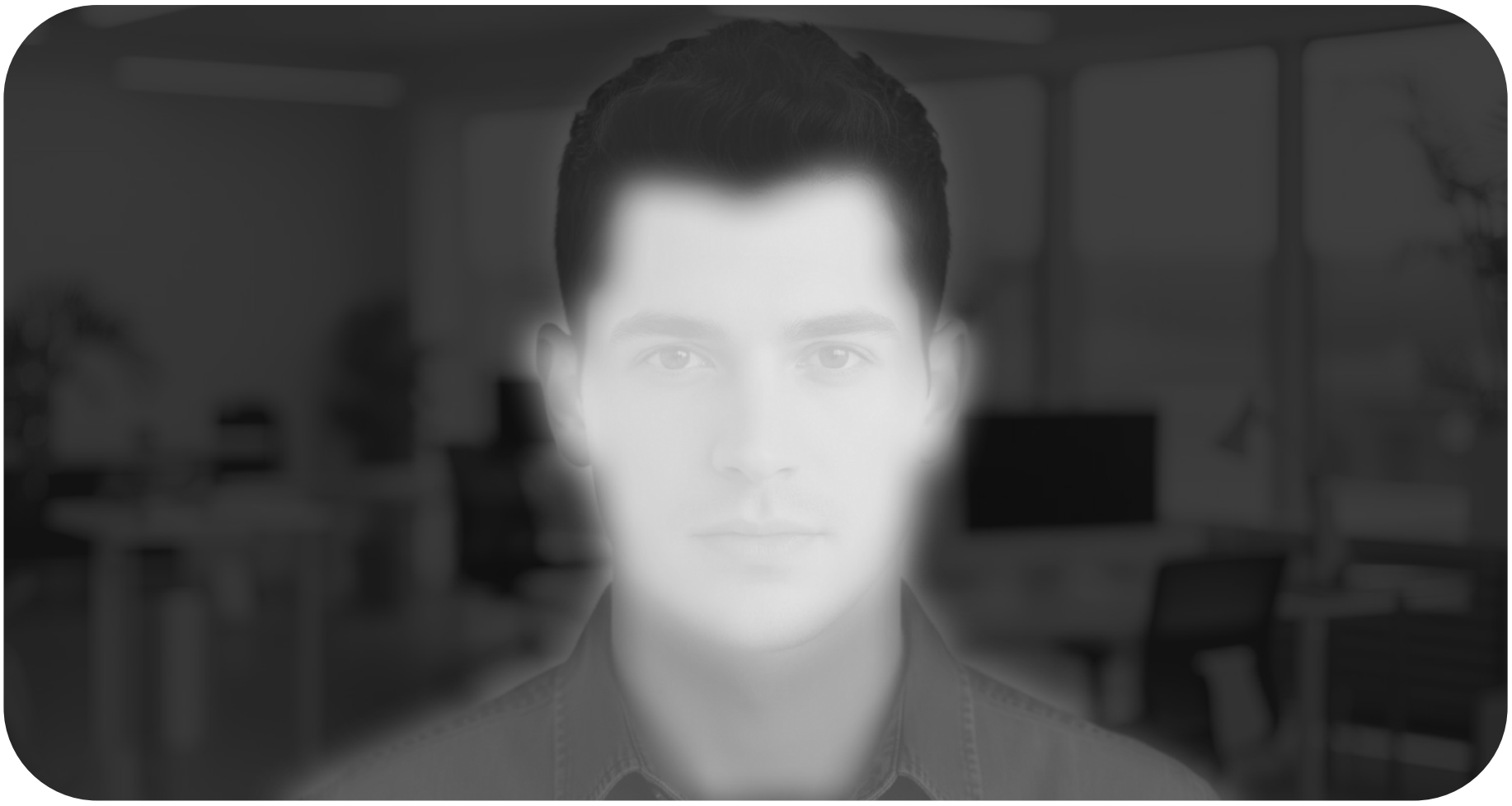
This gate looks for statistical traces left by generative systems. It does not depend on a visible watermark. It evaluates whether the media behaves more like model output than camera capture, using patterns that are difficult to remove without degrading the content.



Gate Three

The Digital Archaeologist

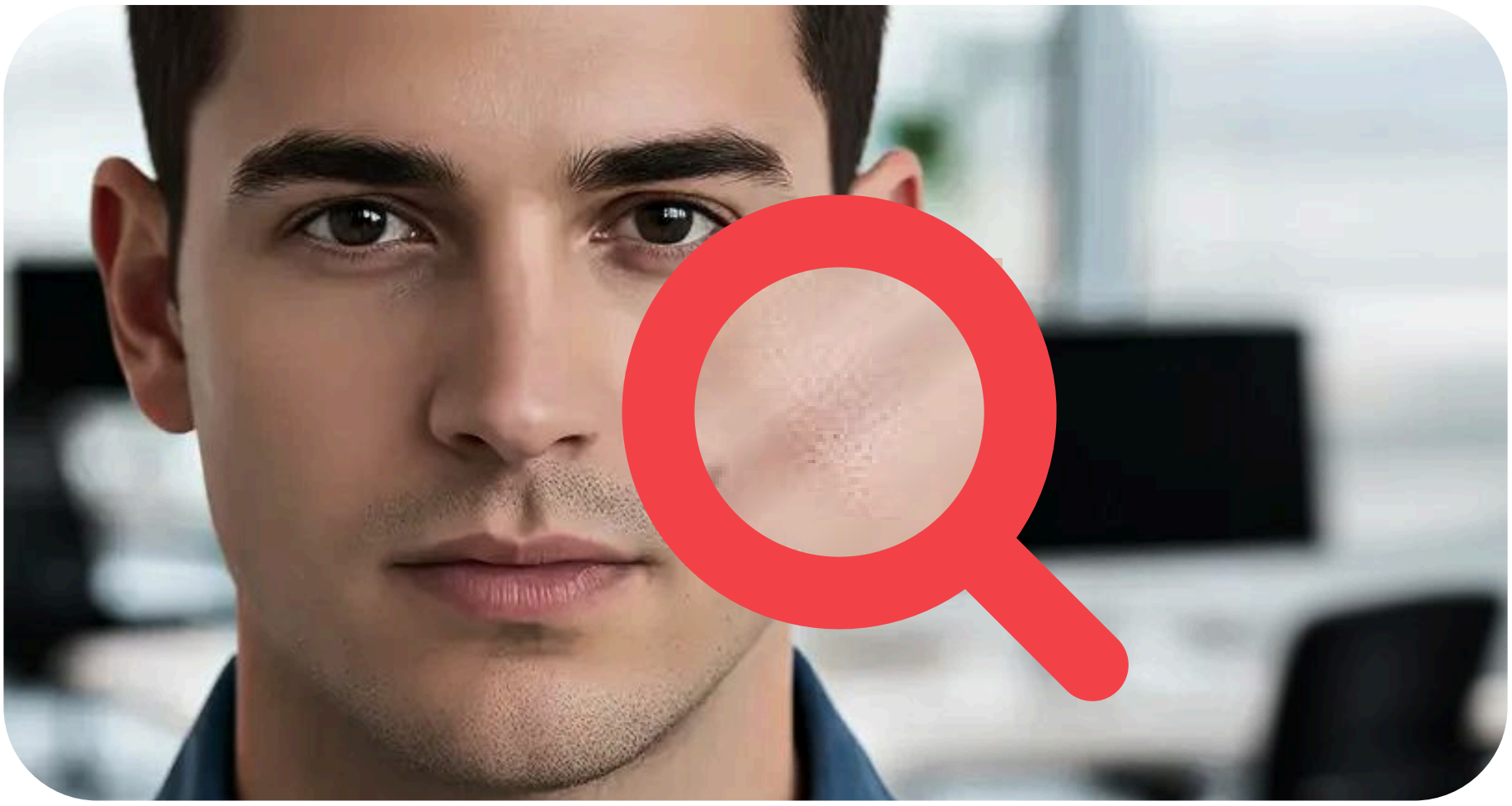
This gate examines the processing history of the media. It looks for inconsistencies in compression and re-encoding that suggest parts of the image or video have been handled differently, which is common when manipulated regions are edited and then cleaned up.



Gate Four

The Frequency Analyst

This gate moves beyond visual appearance and checks frequency space. It looks for spectral patterns that real sensors and lenses produce naturally and that generators often smooth out or reproduce incorrectly, especially when the content has been edited or heavily processed.



Gate Five

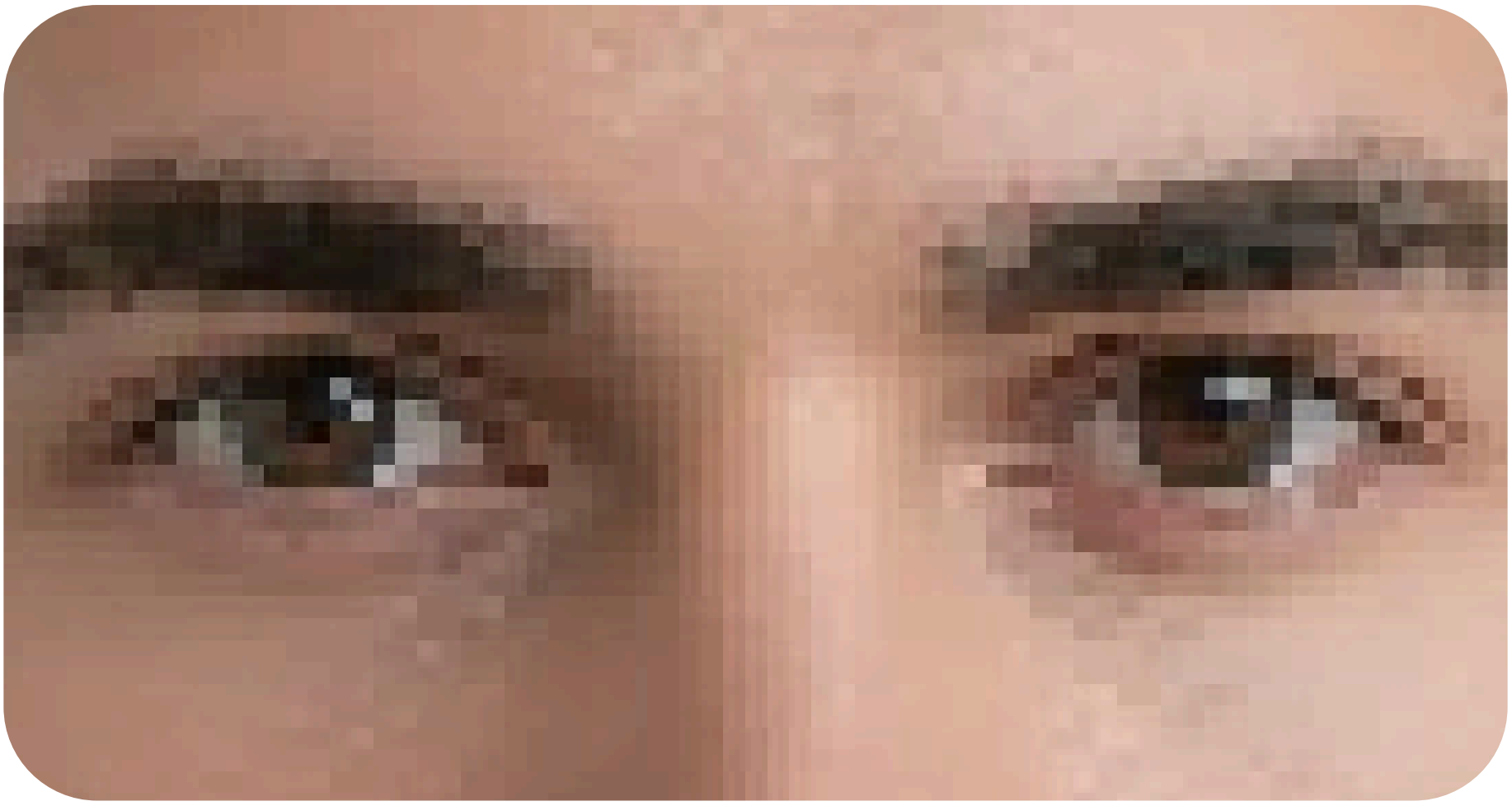
The Texture Specialist

This gate evaluates whether fine detail behaves like real skin and real surfaces. It focuses on micro variation, irregularity, and natural transitions such as hairlines and edges, where synthetic media often become too uniform, repetitive, or averaged.

Gate Six

The Degradation Expert

This gate is designed for low-quality conditions that attackers often exploit. It tests what remains measurable when resolution drops, blur increases, noise is added, or lighting is poor, and it looks for anomalies that persist when clarity collapses.



Gate Seven

The Pixel Inspector

This gate examines high-resolution coherence. It checks pixel-level continuity and local reconstruction patterns that can reveal how an image was assembled, catching subtle discontinuities that only emerge when detail is high and there is less room for synthetic shortcuts.

Keeping Detection Aligned With Evolving Threats

Industry-wide, deepfake detection is still constrained by limited and fragmented datasets. In addition, many solutions are built around off-site capabilities or broad checks rather than real-world threat scenarios.

These standalone detection models often remain tied to a specific environment, pipeline, or data collection method. Therefore, it becomes impractical to generalize solutions across industries, geographies, or use cases.

These standalone detection models often remain tied to a specific environment, pipeline, or data collection method. Therefore, it becomes impractical to generalize solutions across industries, geographies, or use cases.

In contrast, Shufti's deepfake detection solution is being trained to address the diversity of manipulation techniques. Shufti uses:

A multi-model strategy because no single model is enough to detect the full range and intensity of modern deepfakes. At the same time, these models are continuously trained and updated using new data and evolving threat scenarios, ensuring alignment with real adversarial behaviour rather than static or outdated patterns.

To ensure detection remains aligned with real-world threat scenarios, Shufti's approach:

- Deepfake detection aligned with real-world threats through a structured audit and feedback loop with stakeholders.
- Findings are shared across teams to continuously refine training, thresholds, and policies as new deepfake typologies emerge.
- Building the deepfake defense is treated as a threat-specific challenge rather than a fraud problem.

A current limitation is that synthetic media created using newly introduced generation models may differ from the data used to train earlier detection systems. Nevertheless, Shufti's R&D team prioritizes continuous monitoring of emerging fraud typologies and actively retrain models on content produced by these newer generators.

Shufti

Deploy Deepfake Detection Inside Your AWS Environment

Deepfakes can slip through legacy checks and hide inside historic KYC records. Shufti's Deepfake Blindspot Audit runs as a secure AWS AMI in your own cloud, keeping biometric data within your security perimeter while scanning for manipulation and generative AI signals.

Audit historic KYC at scale, flag genuine vs. synthetic identities, and strengthen compliance without moving data off-prem.

These standalone detection models often remain tied to a specific environment, pipeline, or data collection method. Therefore, it becomes impractical to generalize solutions across industries, geographies, or use cases.

Start Your AWS Deepfake Blindspot Audit with Shufti