ATTESTOR

Forensic media authentication for legal proceedings

Technical White Paper

Methodology, system design, and a proposal for independent academic validation

Version 1.0 — September 2026

Indrevo · ABN 21 979 134 918 · New South Wales, Australia

attestor.online · admin@attestor.online

Abstract

Synthetic and manipulated media now reach investigative files, insurance claims, and court exhibits. Attestor is a forensic examination platform that analyses photographs, video and audio through multiple independent signal families, records every action in a tamper-evident chain of custody, and issues reports whose language is deliberately probabilistic. This paper documents the system's design and examination methodology in full, states its limitations plainly, and sets out a concrete proposal for independent academic validation. **Attestor publishes no accuracy claims in advance of that validation**: we consider third-party evaluation on datasets we do not control, with results published whatever they show, to be a precondition of legitimacy for any tool used in legal settings.

1 The problem

Generative models can now produce photographs of events that never occurred, clone a voice from seconds of sample audio, and re-render video frame by frame. The barrier to producing convincing fabricated media has collapsed from specialist skill to a consumer subscription. Courts, tribunals and investigators face a structural problem: the traditional presumption that a photograph depicts something that happened is no longer safe, and the professionals who must weigh such material rarely have in-house tools to interrogate it.

Two failure modes matter. The obvious one is a fabricated exhibit admitted as genuine. The quieter one — sometimes called the liar's dividend — is genuine evidence dismissed as possibly fabricated because no principled examination was available. A credible examination tool must reduce both errors, which is why Attestor reports findings with stated confidence rather than issuing binary verdicts.

2 Design principles

No single oracle. Authenticity is never decided by one detector. Attestor runs independent signal families and reports each separately. Signals may agree, disagree, or be individually unavailable; the report preserves that structure instead of collapsing it into a score.

Probabilistic language, by construction. Report wording is generated from finding states (*no anomaly / observation / anomaly / not performed*) with an attached confidence. The system is incapable of emitting "this file is fake" as a finding. Absence of detected artefacts is expressly stated not to be proof of authenticity.

Custody before analysis. An exhibit is hashed (SHA-256) on receipt, before anything else touches it. Every subsequent event is appended to a hash-chained log in which each entry incorporates the digest of its predecessor, making after-the-fact alteration of the record detectable.

The examiner stays in the loop. Attestor is an instrument, not a witness. Its reports are designed to support professional judgement and, where required, qualified expert testimony — not to replace either.

Transparency over mystique. This document describes the examination pipeline in enough detail to be criticised. A detection tool whose methods cannot be interrogated has no place in a courtroom.

3 System architecture

3.1 Intake and custody

Uploaded exhibits are hashed with SHA-256, encrypted at rest with AES-256-GCM under a server-held key, and never modified; all analysis operates on verified copies whose digest is re-checked against the intake hash before each examination. The custody log records ingestion, examination, report issue and deletion events. Log integrity is verifiable at any time by recomputing the chain, and the verification routine is exposed to administrators.

3.2 Signal families

An examination currently comprises five families. Families I–IV are transparent, rule-based analyses implemented directly in the platform; family V is a pluggable model-based classifier service.

I	**Provenance & Content Credentials**	Detection and verification of embedded C2PA manifests (JUMBF structures). Where a manifest exists, its asserted generator and edit history are quoted — including AI generators that sign their output.
II	**Metadata consistency**	Structured parsing of EXIF/XMP/IPTC fields: capture device presence, software fields naming generative or editing tools, internal timestamp coherence, and absence patterns typical of stripped or generated files.
III	**Compression & error-level analysis**	Controlled JPEG recompression differencing over a spatial grid to surface regions whose error statistics diverge from their surroundings — a classical indicator of regional editing, reported with its known limitations on lossless formats.
IV	**Generative screening heuristics**	Low-weight structural cues associated with model output (characteristic output dimensions, format/metadata combinations, tonal-variance profile). Deliberately assigned the lowest evidentiary weight of all families and always labelled as heuristic.
V	**Model-based classification**	An external detection-engine interface for learned classifiers (deepfake/GAN/diffusion detectors). Engine findings are merged as a distinct family with their own confidence, never silently blended into other results.

3.3 Finding states and confidence

Each family returns one of four states — **no anomaly**, **observation**, **anomaly**, or **not performed** — with a confidence grade (*high* / *medium* / *low*) that qualifies the finding itself, not the authenticity of the exhibit. An overall summary is derived from the pattern of states using fixed, documented rules; multiple independent anomalies escalate the recommended response to full expert examination.

3.4 Reports

Reports are sealed PDFs comprising a findings summary in plain language, per-family results with their confidence, technical appendices (including the error-level heatmap where applicable), and the complete custody log for the exhibit. On issue, the SHA-256 of the report document itself is written back into the custody log, so a report tendered later can be verified against the record of its issue.

4 Limitations

We state these prominently because a validation partner should test against them, and a user must know them.

No published accuracy figures yet. Families I–IV are established forensic techniques with well-understood failure modes, but Attestor as an integrated system has not yet been benchmarked by an independent party. Until it has, we make no quantitative detection claims — that is the purpose of the programme in §5.

Adversarial robustness is an arms race. A sophisticated actor can strip metadata, avoid characteristic dimensions, and launder compression history. Signal families are chosen so that defeating all of them simultaneously is harder than defeating any one, but no such system can be complete.

Heuristics are heuristics. Family IV in particular can be triggered by legitimate workflows (screenshots, exports from editing software) and is weighted accordingly; it exists to prompt scrutiny, not to accuse.

Format coverage is asymmetric. Pixel-level families apply to still images; video and audio are currently examined at container, metadata and (via family V) model level, with frame-level video analysis on the roadmap.

Provenance standards are young. C2PA adoption is growing but sparse; absence of a manifest is neutral and treated as such.

5 Proposed independent validation programme

We are seeking one or more university research groups in digital media forensics to evaluate Attestor independently. The proposed structure:

Datasets. Evaluation on benchmark and partner-held datasets of authentic and manipulated media that Indrevo does not control and, for held-out sets, does not see. Partner may include novel generators unseen by the system.

Protocol. Blind examination through the same interface available to customers; per-family and overall results scored against ground truth. Metrics co-designed with the partner — at minimum ROC/PR characteristics at stated operating points, calibration of confidence grades, and per-family contribution analysis.

Independence. The partner controls the protocol and retains unconditional publication rights, including negative results. Indrevo's commitment: results are published whatever they show, and material shortcomings are fixed or disclosed in product.

What we provide. Free full access for the research group, engineering support, complete methodology documentation (this paper and beyond), and an API for batch evaluation. Where useful, co-authorship is welcome; where preferred, full arms-length separation is respected.

What the field gains. A documented case study of a commercial forensic tool submitting to open evaluation before making claims — a norm we would like to see become standard for software used in legal proceedings.

6 Privacy, ethics and legal context

Attestor operates under Australian law (Privacy Act 1988 (Cth) and the Australian Privacy Principles). Exhibits are encrypted at rest, are never used to train models, and are deletable by the account holder at any time, with hash-only custody stubs retained so previously issued reports remain verifiable. Access requires verification of professional standing. Reports are framed to reduce both wrongful acceptance of fabricated media and wrongful dismissal of genuine media, and terms of service prohibit use of the platform to launder fabricated material through "clean" reports — an abuse case we actively consider in design.

7 Contact

Research partnerships, evaluation proposals and technical questions: **admin@attestor.online**. Attestor is developed by Indrevo (ABN 21 979 134 918), New South Wales, Australia. This white paper, the platform's report format, and the validation proposal in §5 are current as at September 2026.