ARTIFICIAL INTELLIGENCE IN FORENSIC SCIENCE: A COMPARATIVE ASSESSMENT OF CAPABILITY, RELIABILITY AND EVIDENTIARY READINESS

Simran Peswani, MCA, IIS (Deemed to be University), Jaipur
Research guide: Dr. Preeti Tiwari, IIS (Deemed to be University), Jaipur

Abstract*— Artificial intelligence (AI) is entering forensic science while the discipline is still establishing the validity of its older methods. This paper compares six forensic domains (face recognition, latent fingerprints, firearms and toolmarks, DNA mixture interpretation, digital evidence triage and forensic pathology) on five criteria applied identically to each: technical maturity, casework deployment, validation evidence, courtroom acceptance, and bias and ethical risk. The study is a structured secondary analysis of a 33-observation workbook transcribed from 34 published sources, each observation carrying a source identifier and a High, Medium or Low confidence code, supported by author-coded maturity scores. No model was trained and no experiment was run. Published benchmarks show large capability gains: the best face-identification miss rate fell from 5.0% to 0.08% and the best latent-print rank-1 rate rose from 67.2% to 98.6%, but the benchmark rounds use different datasets and show direction, not a controlled trend. Human baselines exist but are rarely comparable like for like with algorithmic figures. Two DNA software platforms reported likelihood ratios that differ by roughly five to six orders of magnitude on one item of evidence, which shifts attention from accuracy to reproducibility. Author-coded scores place the widest gap between demonstrated and deployed capability in digital triage, and legal rules for machine-generated evidence remain unsettled. The paper separates capability, operational reliability and evidentiary readiness, and concludes that validation standards and governance, more than model performance, currently limit forensic AI.

***Keywords*— Artificial Intelligence, Forensic Science, Evidentiary Reliability, Probabilistic Genotyping, Face Recognition, Admissibility, Validation**

I. INTRODUCTION

Comparing a latent print to a reference print, a bullet to a test fire, or a mixed DNA profile to a suspect's profile is the foundation of forensic science. Trained examiners conducted these comparisons during the majority of the 20th century, reporting categorical judgments with no documented error rate. Aside from nuclear DNA analysis, no forensic technique had been thoroughly demonstrated to link evidence to a particular source with a high degree of certainty, according to a 2009 National Academy of Sciences assessment [1]. For a number of feature-comparison techniques, foundational validity was not yet proved, according to the 2016 PCAST assessment [2], which looked at it discipline by discipline. Machine learning entered the field at the same time as extensive black-box investigations of examiners [3], [4].

A large portion of the forensic workflow now uses AI. Face galleries comprising well over a million photos are searched by deep networks [5], [6]. Although the identification is made by the examiner, automated latent-print search has long been the norm [7]. Software for probabilistic genotyping assesses DNA combinations that are difficult to understand using manual techniques [8], [9]. Examiners can score seized devices using machine-learning technologies [10], and generative models can now produce artificial images, audio, and video that can be presented as evidence [11]. The published accuracy numbers are derived from several various types of research, and adoption varies significantly between these purposes.

Evidence is not useable by capability alone. A court must ascertain whether other laboratories arrive at the same conclusion, how frequently a method fails on casework-like material, and who is responsible when a conclusion is incorrect, as well as whether the other side can investigate the approach. Thus, this essay distinguishes three concepts that are frequently used in tandem. Performance on a specified benchmark task is known as capability. In routine casework, operational reliability refers to precision and repeatability. The degree to which an output satisfies requirements for admissibility, disclosure, and accountability is known as evidentiary ready. A system may receive a high score on the first and a low score on the other two.

In India, the question has practical significance. While laboratory pendency is said to be high [13], the Bharatiya Nagarik Suraksha Sanhita (BNSS), 2023 mandates that a forensic expert visit the scene of offenses punishable by seven years or more, from a date each State notifies within five years [12].

This research examines what published data indicates AI has changed in forensic science, what hasn't, and if evidence safeguards have kept up. No experiment was conducted, and no model was trained; it is an organized secondary analysis. The technical development of the employed model families is tracked in Section II. Seven papers are reviewed in Section III, which concludes with the research gaps. The dataset and methodology are explained in Section IV. Section VI provides the conclusion, ramifications, limitations, and future scope, while Section V provides evidence-based answers to twenty analytical questions. The references are listed in Section VII.

II. TECHNICAL GROWTH OF COMPUTATIONAL MODELS IN FORENSIC SCIENCE

The technical lineage comes first since how a technique is constructed determines what can be tested about it. The architecture of the four families discussed in this work is contrasted in Fig. 1.

II.1 Architectures of model families

manual comparison (until around 1980). After visually comparing the evidence, the examiner draws a categorical decision. Performance can only be assessed by testing examiners, which was not done on a large scale until the 2010s [3]. This is because the task model is based on the examiner's training and judgment.

Matching statistics (1980s–2000s). DNA databases and automated fingerprint identification systems encode human-selected information, like fingerprint minutiae or short tandem repeat alleles, and rank a database based on a query [7]. Instead of being learning systems, they are algorithmic search systems that establish a division of labor in which the examiner makes the final decision after the machine suggests candidates.

learned representations (starting about 2012). Convolutional neural networks link a face or print to an embedding vector after learning features directly from pixels [14], turning identification into a nearest-neighbor search throughout a gallery [15]. The effect was noted by NIST's independent face benchmarks [5], [6]. Image classification in forensic pathology, including diatom analysis and gunshot-wound categorization, is now supported by the same family [16]. Although these models increase accuracy, they do not provide an examiner with a way to testify about the rationale of a judgment.

Genotyping with probability (2010s). Electropherogram peak data is modeled by programs like STRmix and TrueAllele, Genotyping with probability (2010s). Markov chain Monte Carlo sampling is used by programs like STRmix and TrueAllele to model electropherogram peak data, weigh potential genotype combinations, and return a likelihood ratio [9]. It is not a neural network; rather, it is Bayesian statistical modeling. It is included because it is the most obvious repeatability test and the most important computational method in casework.

Generative models and triage classifiers. Devices and files are ranked for examiner attention by machine-learning classifiers [10]. Instead, generative models produce artificial media, making AI both a source of dubious evidence and a tool for its analysis [11].

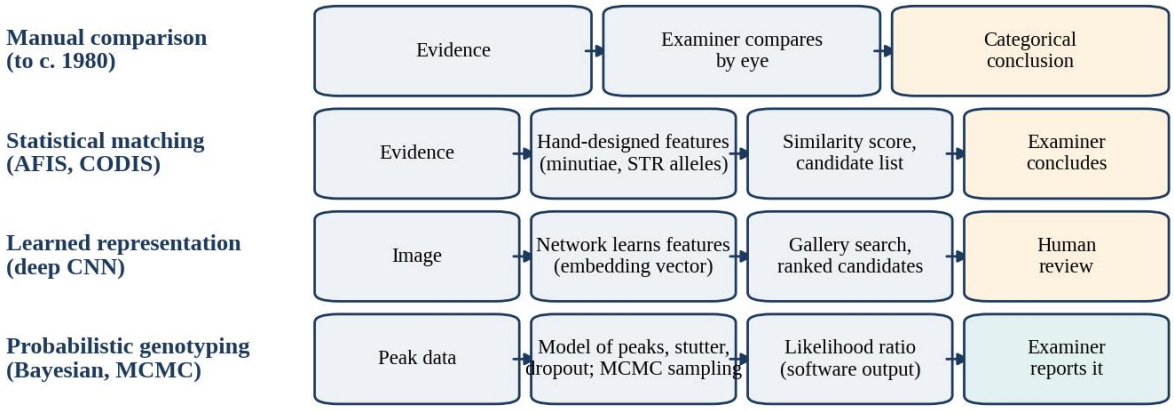

Figure shows a forensic comparison of the four model families' architectural designs. An orange box represents a person's conclusion, and a teal box represents a statistic produced by software.

In this instance, the conclusion's establishment is the architectural point that counts. In the first three families, someone says. The program in probabilistic genotyping produces the statistic that the examiner reports. The model output itself is the contested item in generating and detecting systems. Section V investigates if courtroom acceptance follows this trend.

III. LITERATURE REVIEW

Ulery et al. (2011). Ulery, Hicklin, Buscaglia, and Roberts [3] conducted a blind black-box study in which 169 latent print examiners evaluated 17,121 comparisons with hidden ground truth. They reported a false positive rate of 0.1% and a false negative rate of 7.5%, finding that 85% of examiners made at least one false negative. It was the first substantial published error-rate estimate for latent print evaluation and serves as the human baseline employed in this work. Because the participants were volunteers, the rates might not fairly represent all practicing examiners or casework scenarios.

Butler, Kline and Coble (2018). Butler, Kline, and Coble discuss two NIST interlaboratory tests on the interpretation of DNA mixtures [8]. 108 labs in MIX13 examined the same electropherogram data for five case scenarios with two to four contributors. Just seven labs successfully eliminated the reference profile that did not contribute to the four-person 1:1:1:1 mixture, 74 laboratories included a reference profile that was not a contributor and only seven excluded it correctly (counts as recorded in the workbook). The study shows how widely protocols varied on complex mixtures. It does not test probabilistic software, and the mixture was built to be difficult, so it is not an average casework error rate.

Monson, Smith and Peters (2023). Monson, Smith and Peters [4] of the FBI Laboratory tested 173 firearms examiners in a black-box design with 8,640 comparisons of bullets and cartridge cases. False positive rates were estimated at 0.656% for bullets and 0.933% for cartridge cases, and false negative rates at 2.87% and 1.87%; the 95% interval for bullet false positives was 0.305–1.42%. Most errors came from a limited number of examiners, so a pooled rate hides individual variation. The study gives firearms a human baseline, but participants volunteered and no open benchmark for automated comparison accompanies it.

Grother, Ngan and Hanaoka (2019). NISTIR 8280 [17] evaluated 189 face recognition algorithms from 99 developers to measure demographic effects. Its main result is that false positive differentials are much larger than false negative ones, with false positive rates often varying by factors of 10 to beyond 100 depending on the algorithm and group, and that some algorithms show far smaller differentials. NIST stresses that false positives in one-to-many search matter because they can place the wrong person on a candidate list. The report describes the problem and informs policy; it sets no reporting rules.

Thompson (2023). Thompson [9] compared two probabilistic genotyping programs, STRmix and TrueAllele, on the same item of low-template DNA evidence in a federal criminal case. STRmix reported a likelihood ratio of 24 in favour of the non-contributor hypothesis; TrueAllele reported between 1.2 million and 16.7 million, depending on the reference population. The author traces the difference to modelling parameters, analytic thresholds and mixture ratios. It is a single case, and a response published for the TrueAllele developers disputes the interpretation [18], so it shows that divergence can occur, not how often.

Orsini et al. (2025). Orsini and colleagues [16] reviewed AI in forensic pathology under PRISMA guidelines: 65 articles were identified and 18 met the inclusion criteria. The workbook records accuracies of about 88–98% for gunshot-wound classification, 70–94% for post-mortem neurological analysis and a recall of 0.95 for AI-enhanced diatom testing. The authors conclude that AI is best treated as an enhancement to human expertise, not a replacement, and note small samples and variable performance. The limitation is the evidence base itself: wide ranges point to dependence on dataset and task.

Grossman and Grimm (2025). Grossman and Grimm [11] analyse how US evidence law treats AI-generated material. They distinguish acknowledged AI-generated evidence, where all parties agree a system produced it, from unacknowledged evidence, where one side alleges a deepfake, and they discuss rules they urged on the Advisory Committee on Evidence Rules for each case. The work is doctrinal analysis, not an empirical test. It concerns US federal practice and the proposals were still under committee consideration, so its transfer to Indian proceedings is untested.

TABLE I. COMPREHENSIVE ANALYSIS OF THE REVIEWED LITERATURE

S. No.	Research paper	Researcher and year	Methodology used	Key outcome	Gap left
1	Accuracy and reliability of forensic latent fingerprint decisions	B. Ulery et al., FBI/Noblis, 2011	Blind black-box test; 169 examiners, 17,121 comparisons	False positives 0.1%, false negatives 7.5%	Volunteer sample only
2	NIST interlaboratory studies involving DNA mixtures (MIX05, MIX13)	J. Butler, M. Kline, M. Coble, NIST, 2018	Interlaboratory exercise; 108 laboratories, identical data	74 of 108 laboratories falsely included a non-contributor in one four-person mixture	Does not test software
3	Accuracy of comparison decisions by forensic firearms examiners	K. Monson, E. Smith, E. Peters, FBI Laboratory, 2023	Black-box design; 173 examiners, 8,640 comparisons	False positives 0.656% (bullets), 0.933% (cartridge cases)	No open benchmark for automated comparison
4	FRVT Part 3: Demographic Effects (NISTIR 8280)	P. Grother, M. Ngan, K. Hanaoka, NIST, 2019	Independent benchmark; 189 algorithms	False positive differentials of 10 to beyond 100 times; much smaller false negative ones	Measures; sets no reporting rule
5	Uncertainty in probabilistic genotyping of low template DNA	W. Thompson, 2023	Single case; two software platforms, one item	LR of 24 versus 1.2–16.7 million on one item	One disputed case
6	The application of AI in forensic pathology: a systematic review	F. Orsini et al., 2025	PRISMA review; 18 of 65 articles included	Wide accuracy ranges; AI as enhancement	Small samples; little multi-site validation

S. No.	Research paper	Researcher and year	Methodology used	Key outcome	Gap left
7	Judicial approaches to acknowledged and unacknowledged AI-generated evidence	M. Grossman, P. Grimm, 2025	Doctrinal analysis of US evidence rules	Two evidentiary problems need different rules	US federal focus; no Indian analysis

When combined, the seven studies demonstrate that algorithmic differences between groups may be measured [17], that mistake can be measured for human examiners [3], [4], and that the law is currently debating how to handle machine output [11]. There are still a few gaps. Volunteer samples are the foundation of human baselines. None of the assessed sources provide an open, unbiased benchmark for automated comparison with regard to guns. For face recognition, demographic differences are measured, but the measurement does not lead to a unified reporting standard. One contentious instance is the basis for reproducibility among DNA software platforms. Small studies make up the majority of the pathology evidence. There is little operational data from India, and the evaluated legal literature is American. Lastly, this work seeks to compare domains using a single set of criteria, something that none of the seven do. This claim only applies to the works that have been assessed here; no systematic search was conducted to demonstrate the absence of such a comparison.

IV. ANALYSING RELIABILITY, ACCOUNTABILITY AND EVIDENTIARY READINESS ACROSS CHANGING MODELS

This section explains how the comparison was built. Each domain is assessed against five criteria applied identically: technical maturity, casework deployment, validation evidence, courtroom acceptance, and bias and ethical risk. Evidence comes from a secondary workbook in which every number is traceable to a published source. Three levels are kept apart throughout: capability (benchmark performance), operational reliability (reproducibility and error in routine casework) and evidentiary readiness (whether output meets admissibility, disclosure and accountability requirements). No model was trained and no experiment was run, so the findings describe published evidence, not new measurements made by the author.

A. Dataset description

The workbook has nine worksheets (Table II). Its fact table, Metrics_Timeseries, holds 33 observations in six domains: face recognition (7), latent fingerprints (6), firearms and toolmarks (4), DNA mixture interpretation (5), digital forensics (6) and forensic pathology (5). Each row records a value, unit, year, evaluator, confidence code and Source_ID. Two further sub-disciplines, questioned documents and synthetic-media detection, appear only in the author-coded Domain_Maturity sheet and, for synthetic media, in Adversarial_AI and Risk_Register. Four kinds of content are kept distinct. *Reported* values are transcribed from published studies or official releases. *Derived* values are arithmetic done by the author on reported values. *Coded* values are author judgements against a rubric. *Interpretive* statements are conclusions drawn from the combination.

TABLE II. WORKSHEETS, EVIDENCE TYPE AND USE

Worksheet	Content	Evidence type	Used for (question numbers)
Metrics_Timeseries	33 observations, 6 domains; value, unit, year, era, evaluator, confidence, Source_ID	Reported	Q5–Q9, Q11–Q13
Human_vs_Algorithm	6 paired comparisons with a comparability grade and interpretation	Reported and derived	Q6, Q7, Q10
Domain_Maturity	8 sub-disciplines scored 1–5 on five criteria, with rationale	Coded (author)	Q4, Q10, Q14
India_Context	14 indicators: backlog, referral share, statutory parameters	Reported (secondary sources)	Q19
Adversarial_AI	10 indicators from NCMEC reporting, with caveats	Reported (High/Medium/Low)	Q18
Risk_Register	10 risks scored for severity (1–5) and evidence strength	Coded (author)	Q16, Q17, Q20
Summary_Stats	Formulas counting rows by domain and confidence	Derived	Descriptive statistics
Sources	34 sources with organisation, year, type and URL	–	Traceability
README	Method and caveats	–	Scope and limits

Table III shows a snapshot of the fact table. Confidence coding assigns High to 17 observations (51.5%), Medium to 14 (42.4%) and Low to 2 (6.1%). High means a controlled independent evaluation or official release; Medium means a peer-reviewed study on a non-standard dataset or a vendor result on an independent benchmark; Low means a survey or self-report. Applied strictly, four High-coded rows (M004, M005, M008, M009) report NIST results second-hand, through a policy article or preprints, and would fall to Medium, giving 13, 18 and 2. India_Context carries no confidence code and relies on secondary sources, which conflicts with the stated exclusion of unattributed secondary reporting, so its figures are treated as indicative only.

TABLE III. SNAPSHOT OF THE FACT TABLE (METRICS_TIMESERIES)

Obs_ID	Domain	Metric	Year	Value	Unit	Confidence	Source
M001	Face recognition	Best algorithm 1:N miss rate	2010	5.0	%	High	S01
M004	Face recognition	Best algorithm 1:N miss rate	2020	0.08	%	High*	S04
M008	Latent fingerprint	Best rank-1 identification rate	2012	67.2	%	High*	S07
M011	Latent fingerprint	Best rank-1 hit rate (285 probes)	2026	98.6	%	Medium	S10
M012	Latent fingerprint	Examiner false positive rate	2011	0.1	%	High	S11
M018	DNA mixture	Labs falsely including non-contributor	2013	68.5	%	High	S13
M022	DNA mixture	Divergence between two platforms (as recorded)	2023	6	log10 LR	Medium	S15
M027	Digital forensics	Practitioners rating policy adequate	2024	14.8	%	Low	S18

*All eight rows are Reported values. * Confidence code is High in the workbook; see the sensitivity note above. The M022 value is revisited in Section V.B.4.*

B. Experimental setup

No model training or inference was carried out, because the study analyses published measurements, statutes and rulemaking records. The requirement is a spreadsheet tool for the workbook and a plotting tool for the figures; no GPU or training environment is needed. [Author to confirm the software actually used, for example Excel, Power BI or Python with Matplotlib.]

C. Research methodology

Fig. 2 summarises the method. The five criteria and twenty questions are fixed first. Sources are collected in tiers, from independent government benchmarks and black-box error-rate studies down to practitioner surveys, with vendor marketing and undated blog content excluded. One observation is transcribed per row with its source, rated for confidence and labelled by evidence type. Table IV lists the parameters. A human result and an algorithmic result are set side by side as a comparison only when task, protocol and population match; otherwise the pair is labelled indirect or not comparable. Only descriptive statistics are used, so the paper makes no claim of statistical significance or causation.

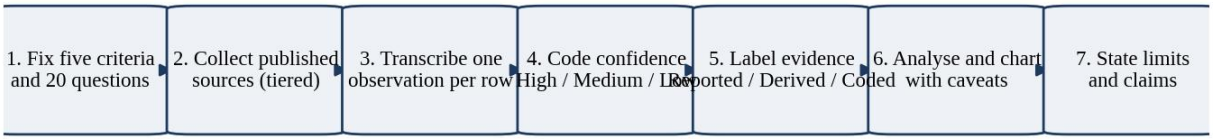

Fig. 2. Research methodology used in this study.

TABLE IV. RESEARCH PARAMETERS

Parameter	Standard applied
Source tiers	(1) Independent government benchmarks; (2) black-box error-rate studies; (3) peer-reviewed systematic reviews; (4) official statistics and statute; (5) practitioner surveys, rated Low
Excluded	Vendor marketing, undated blog content, unattributed secondary reporting
Recency	Performance data from 2010 onward; NAS 2009 retained for historical framing
Traceability	Every observation carries a Source_ID that resolves to a citation
Contested figures	Retained with the dispute noted, not silently resolved
Maturity rubric	1 absent or theoretical; 2 experimental, single laboratory; 3 multi-site validation published; 4 routine casework in several jurisdictions; 5 established, standardised, judicially tested
Bias and ethics risk	Scored 1–5; the workbook gives no rubric, and higher is read as greater risk from the rationale text

V. RESULTS AND DISCUSSION

A. Analytical questions

Table V lists the twenty questions with the main evidence behind each. The classification is the author's for this paper: the presentation refers to a Q_Chart_Map sheet that is absent from the workbook supplied, so its split of 14 data-based and 6 framework answers could not be reproduced. Under the classification here, ten questions rest mainly on reported data, three on coded scores, two on documentary records and five on analytical frameworks.

TABLE V. ANALYTICAL QUESTIONS AND EVIDENCE TYPE

Q	Question	Evidence type	Workbook / source
1	What separates AI from earlier automation?	Framework	Sec. II
2	How did Mayfield, NAS and PCAST reframe error?	Documentary	[1], [2]
3	What did forensics lack that AI supplies?	Framework	Sec. I–II
4	Why does 'machine narrows, human concludes' endure?	Coded	Domain_Maturity
5	How much has performance improved?	Reported, derived	M001–M004, M008–M011

Q	Question	Evidence type	Workbook / source
6	How do human and algorithm errors compare?	Reported, derived	Human_vs_Algorithm
7	What does MIX13 show about manual methods?	Reported	M018–M019
8	Does AI reduce or redistribute error?	Reported (interpretive)	M007, M012–M013
9	What has AI triage done for backlog?	Reported (Medium, Low)	M023–M028
10	How wide is the demonstrated–deployed gap?	Coded	Domain_Maturity
11	What does platform divergence imply?	Reported (one case)	M022
12	Are reported accuracies externally valid?	Reported	M029–M033; R04
13	How should demographic differentials be reported?	Reported, framework	M007; R01
14	Which standard should govern admission?	Framework	R03, R06
15	What does the pace of legal change show?	Documentary	R06
16	Can the defence interrogate the software?	Framework	R02, R03
17	Who is responsible for a wrong conclusion?	Framework	R02, R03, R06
18	Is detection winnable?	Reported (Low), framework	Adversarial_AI; R05
19	Can BNSS s.176(3) be met without automation?	Reported, derived	India_Context
20	What governance model lets AI expand safely?	Coded, framework	Risk_Register

B. Findings

1. Capability gains (Q5). In the NIST face-identification rounds recorded in the workbook, the best 1:N miss rate fell from 5.0% in 2010 to 4.1% in 2014, 0.2% in 2018 and 0.08% in 2020 [5], [6], [19], a relative reduction of 98.4% (derived). Best latent-print rank-1 rates rose from 67.2% in 2012 and 71.4% in 2013 to 93.7% in 2025 and 98.6% in 2026 [20], [21], [22], a gain of 31.4 percentage points (derived). The 2025–26 values are vendor submissions reported by trade press, rated Medium, and the 2026 value rests on 285 probes. Because NIST states that evaluation rounds use different datasets, Fig. 3 shows direction and order of magnitude only. How much of the gain comes from better models rather than changed benchmarks cannot be separated with these data. Best-in-class results are also not median deployed performance.

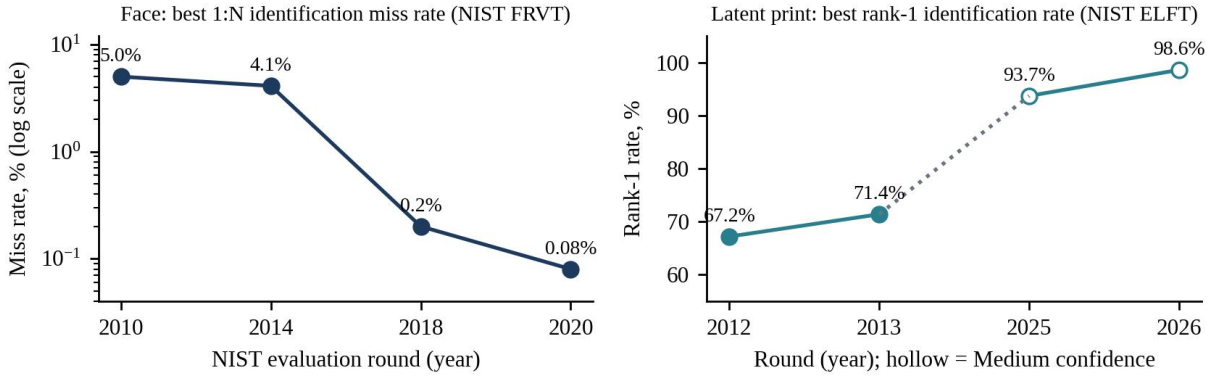

Fig. 3. Best reported results across NIST evaluation rounds (Reported). The rounds use different datasets and are not a controlled time series.

2. Human baselines and comparability (Q6, Q7). Human error is asymmetric in latent prints, with a 0.1% false positive and a 7.5% false negative rate [3], while firearms examiners show false positive rates of 0.656–0.933% and false negative rates of 1.87–2.87% [4] (Fig. 4). Both studies used volunteers, and Monson et al. report that most errors came from a limited number of examiners. Table VI re-examines the six human–algorithm pairs in the workbook. Only the DNA pair is graded Moderate there, and this paper downgrades it to indirect: 68.5% concerns whether laboratories included one reference profile in one deliberately difficult four-person mixture [8], while 0.005% is an upper bound on false positives pooled across validation sets for probabilistic software [23], rated Medium. MIX13 shows that manual binary interpretation varied widely and often failed on that mixture; on the three-person 7:2:1 mixture, 83% of laboratories correctly excluded the non-contributor (M019). It does not give an average casework error rate, nor show that software would score better on the same data.

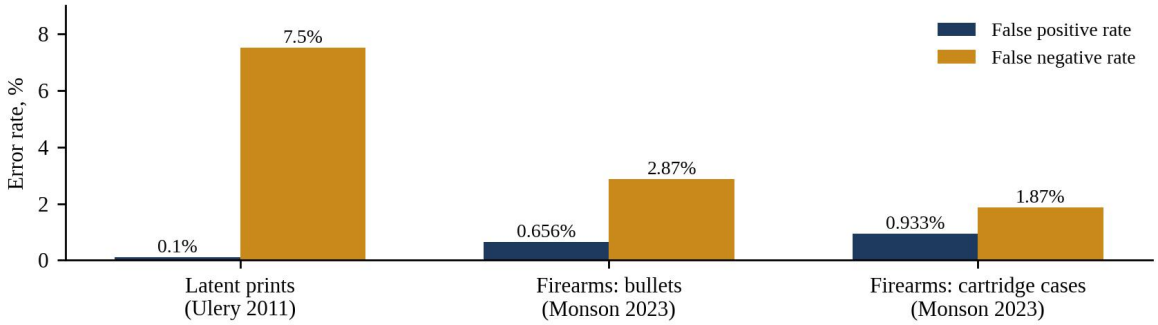

Fig. 4. Human-examiner error rates from two black-box studies (Reported). Protocols and treatment of inconclusive decisions differ, so the studies are not directly comparable.

TABLE VI. HUMAN AND ALGORITHMIC PAIRS RE-GRADED FOR COMPARABILITY

Pair	Human result	Algorithmic result	Workbook grade	Reading in this paper
P1	Latent print false positive 0.1% (2011)	Rank-1 miss 1.4% (2026; derived as 100 − 98.6; 285 probes)	Indirect	Not comparable: decision error versus retrieval miss
P2	Latent print false negative 7.5% (2011)	Same 1.4%	Indirect	Suggestive only: different task, data and year
P3	68.5% of laboratories falsely included a non-contributor (MIX13, 2013)	Probabilistic genotyping false positive bound 0.005% (2026)	Moderate	Indirect: different samples and metrics
P4	No human baseline	Face miss rate 0.08% (2020)	Not comparable	Task exists only at machine scale
P5	Cartridge-case false positive 0.933% (2022)	No independent benchmark	Not comparable	Evidence gap, not a perfect score
P6	Practitioners reporting effective tools 11.5% (2024)	Triage time reduced 45% (2024)	Indirect	Capability versus adoption only

3. Error profile and fairness (Q8, Q13). The evidence shows different failure modes in different domains, not that AI redistributes error. Latent-print examiners mostly miss identifications [3], whereas NIST finds that face algorithms show much larger differentials in false positives than in false negatives [17]. These come from different tasks and cannot be read as a before-and-after, so redistribution remains a hypothesis. What the evidence does support is that an aggregate figure can hide group differences. The headline 0.08% miss rate (M004) and NIST's reported variation of 10 to beyond 100 times in false positive rates [17] describe different metrics in different tests. The workbook records a maximum of 720 times (M007), which could not be traced to NISTIR 8280 and is flagged for verification. Error rates should therefore be reported by group and by error type.

4. Reproducibility and external validity (Q11, Q12). In the case reported by Thompson [9], STRmix gave a likelihood ratio of 24 favouring the non-contributor hypothesis and TrueAllele gave 1.2 to 16.7 million. As log10 values these are 1.38 against 6.08 to 7.22 (Fig. 5), a difference of about 4.7 to 5.8 orders of magnitude (derived). The workbook records the difference as 6 (M022), and the presentation and dossier describe the outputs as opposite conclusions. The case report and the developers' response [18] instead both describe the results as pointing towards exclusion but differing in strength, so this paper states a difference of strength, not direction, pending a check of the full text. The developers attribute the gap to the amount of peak data each program used. One disputed case cannot show how often this happens, but it shows that validating each platform separately does not guarantee agreement on the same evidence. On external validity, reported pathology ranges are wide, 70–94% for post-mortem neurological analysis and about 88–98% for gunshot-wound classification [16], and the workbook stores only upper bounds (M029–M033). Without confidence intervals and multi-site validation, a headline accuracy says little about performance elsewhere.

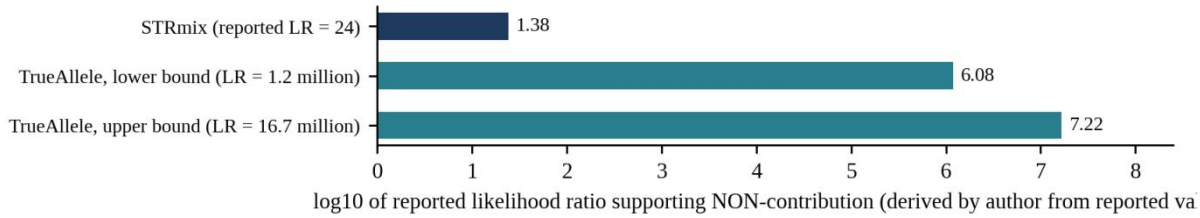

Fig. 5. Strength of support for non-contribution in one casework item, as log10 of the reported likelihood ratio (Derived from Reported values; [9]).

5. Deployment gap (Q9, Q10). The author-coded scores in Fig. 6 give latent-print search 5 for both technical maturity and casework deployment, a gap of zero. Digital triage shows the widest gap, 4 against 2; every other

domain shows a gap of 1. The measured evidence behind the triage gap is mixed in kind. Hybrid machine-learning prioritisation was reported to cut triage time by 45% (Medium) [10], and a two-stage triage design reduced false positives from 20.3% to 7.9% for images and 4.3% for videos [24]. Against this, 14.8% of surveyed practitioners rated their unit's triage policy adequate and 11.5% reported effective tools (Low; sampling frame unpublished) [25]. Experimental results and self-reported adoption come from different samples and years, so the gap is a coded judgement supported by indicative data, not a measured quantity.

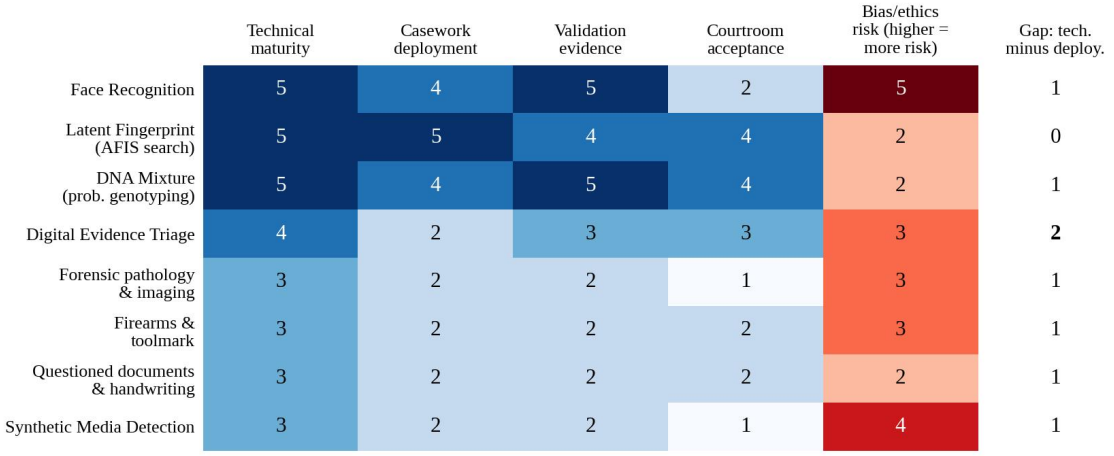

Fig. 6. Author-coded maturity scores for eight sub-disciplines (Coded, not measured). Higher is better in the first four columns; for bias and ethics risk, higher means more risk.

6. Evidentiary readiness and accountability (Q4, Q14–Q17). Latent-print search and probabilistic genotyping receive a score of 4, digital triage receives a score of 3, facial recognition, guns, and questioned papers receive a score of 2, and pathology and synthetic media detection receive a score of 1. According to the presentation, acceptance follows architecture, peaking where the examiner retains the conclusion. This is only partially supported by the entire eight-domain coding: probabilistic genotyping receives a score of 4 even if software generates the statistic, while pathology, where a person makes the final decision and AI helps, receives a score of 1. A more reasonable interpretation is that acceptance is lowest when independent validation is weak or detection is unreliable, and maximum when lengthy operational use, multi-site validation, and an established standard coincide. The explanation behind the architecture remains a theory. A benchmark error rate is not a deployment error rate under the Daubert reliability factors [26], and while the basis of a deep network cannot be described, its output can be tested (R03). Regarding rule-making, draft Rule 901(c) on purported deepfakes was never published for comment and was kept in reserve; proposed Federal Rule of Evidence 707 was published for discussion from August 2025 to February 16, 2026; on May 7, 2026, the Advisory Committee rejected to advance it or its redrafts [27]. Therefore, the phrase "not advanced" in the presentation understates the extent of Rule 707's advancement. The Committee planned a mini-conference for October 15, 2026, but the results were not reviewed. As demonstrated by the Alameda County court's imposition of terminating sanctions for false video exhibits [28] According to reports, counsel is required to exercise caution under and Louisiana Act 250 [29]. The access and accountability matrix in the presentation is seen as a framework of questions rather than a conclusion because no source in the workbook specifies the extent to which a defense may examine proprietary software.

7. AI as a source of questioned evidence (Q18). NCMEC reports flagged as involving generative AI rose from 67,000 in 2024 to 485,000 in the first half of 2025, but the flag is a submitter-selected field, and one provider filed 380,000 of the 2025 reports while stating that none concerned AI-generated material [30]. The workbook records 400,000 for the 2025 half-year; this paper uses the cited 485,000. Flag counts are not verified determinations and are rated Low. The workbook contains no measured detector performance, so the claim that detection is structurally losing is an argument supported by moderate-strength evidence (R05), and provenance at capture remains a proposed remedy.

8. Indian context (Q19). BNSS section 176(3) requires the officer in charge to cause a forensic expert to visit the scene of offences punishable with seven years or more, with the process videographed, from a date notified by the State Government within five years, and requires a State without a suitable facility to notify use of another State's [12]. The duty therefore begins State by State, not nationally from 1 July 2024. Reported national pendency is 0.7–0.8 million cases (a 2021 estimate), with 10–12% of cognizable offences referred to a laboratory [13]. Haryana's laboratory reduced pending cases from 16,040 to 14,863 in the first quarter of 2025, a fall of 7.3% (the source headline says 8%), at 2,273 disposals a month [31]. Annualised, that is 27,276 cases

(derived), so the national estimate equals about 26 to 29 years of Haryana's annual disposals. This mixes a 2021 national estimate with a 2025 state rate and is an order-of-magnitude illustration only. The sources are secondary and uncoded, and the workbook holds no evidence on the validated performance of any AI tool in Indian laboratories. The claim that automation is the only route to compliance is therefore an argument; what the data support is that capacity pressure is large and that triage-type tools with demonstrated gains merit pilot evaluation.

9. Governance (Q20). The risk register lists ten risks with a mean severity of 4.2 and four rated 5 (face-recognition differentials, platform reproducibility, synthetic media and capacity against statutory demand). Four controls drawn from the mitigations in the literature are inter-platform proficiency testing with disclosure of software version and parameters (R02, R04), disaggregated error reporting that separates flags from verified counts (R01, R10), human-in-the-loop requirements with explainability reporting (R03, R06), and provenance at capture (R05). They address seven of the ten risks; capacity (R07), the deployment gap (R08) and volunteer-based baselines (R09) need procurement reform, laboratory capacity and representative proficiency testing instead. The presentation's 'eight of ten' could not be reproduced. Each control is a disclosure, testing or procedural requirement, not a technical advance, and severities are author-coded, so the ranking is indicative.

TABLE VII. CAPABILITY, OPERATIONAL RELIABILITY AND EVIDENTIARY READINESS BY DOMAIN

Domain	Capability (reported)	Operational reliability (reported)	Evidentiary readiness (coded)
Face recognition	Miss rate 5.0% to 0.08% across NIST rounds	False positive differentials of 10 to beyond 100 times between groups	Courtroom 2; bias risk 5
Latent print search	Rank-1 rate 67.2% to 98.6% (different datasets)	Examiner false negative 7.5%, false positive 0.1%	Courtroom 4; bias risk 2
DNA mixtures (genotyping)	False positive bound ≤0.005% (pooled validation, Medium)	Likelihood ratios about 4.7–5.8 orders apart in one case	Courtroom 4; bias risk 2
Digital triage	Triage time cut 45% (Medium)	False positives 20.3% to 7.9% and 4.3%; adoption 14.8% and 11.5% (Low)	Courtroom 3; bias risk 3
Forensic pathology	Accuracy 70–98% across tasks	Wide ranges; few multi-site studies	Courtroom 1; bias risk 3
Firearms and toolmarks	No independent benchmark for automated comparison	Examiner false positive 0.656–0.933%	Courtroom 2; bias risk 3
Synthetic-media detection	No measured value in workbook	Detectors described as unreliable (no measured value)	Courtroom 1; bias risk 4

VI. CONCLUSION

Four reasonable conclusions are supported by the facts gathered. First, published benchmarks demonstrate significant improvements in latent-print search and face recognition capabilities, but the rounds are not sufficiently comparable to determine the extent to which these improvements are attributable to improved models. Second, a benchmark figure does not determine performance in casework; two DNA platforms that differ by about five to six orders of magnitude in a single case, broad pathologic accuracy ranges, and the high dependence of face recognition on group and image quality all demonstrate this. Third, evidential preparedness falls behind both, with unresolved legal guidelines for evidence produced by machines and responsibility resting on examiners and attorneys who lack the resources to analyze a model. Fourth, rather than model capabilities, the binding limitations for forensic AI are validation requirements, disclosure obligations, and procedural protections.

The results suggest that validation should be done under deployment conditions rather than benchmark conditions, that errors should be reported by group and kind, that software version and parameters should be disclosed, that proficiency testing should be conducted across platforms, and that the conclusion should be stated in writing. In order to achieve the BNSS timeline, they advise India to pilot triage-type technologies with independent review. There are limitations to the study. There is no experiment; it is a secondary analysis. Different datasets are used in benchmark rounds, and best-in-class outcomes range from usual deployment performance. Volunteers provide human baselines. The bias and risk scale lacks a documented rubric, and the maturity and risk ratings are ordinal assessments made by a single coder without a second-coder check. The DNA divergence is based on a single contested case, several high-coded rows rely on secondary and uncoded Indian data, and the legal status is only up to date as of the checking date. Future work should include a second coder, replace secondary sources with primary information, evaluate inter-platform agreement on shared samples, gather validation data from Indian laboratories, and expand the workbook to further areas.

VII. REFERENCES

[1] National Research Council, Strengthening Forensic Science in the United States: A Path Forward. Washington, DC, USA: National Academies Press, 2009.

[2] President's Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. Washington, DC, USA: Executive Office of the President, Sep. 2016.

[3] B. T. Ulery, R. A. Hicklin, J. Buscaglia, and M. A. Roberts, "Accuracy and reliability of forensic latent fingerprint decisions," Proc. Natl. Acad. Sci. U.S.A., vol. 108, no. 19, pp. 7733–7738, 2011.

[4] K. L. Monson, E. D. Smith, and E. M. Peters, "Accuracy of comparison decisions by forensic firearms examiners," J. Forensic Sci., vol. 68, no. 1, pp. 86–100, 2023, doi: 10.1111/1556-4029.15152.

[5] P. Grother and M. Ngan, "Face Recognition Vendor Test (FRVT): Performance of face identification algorithms," NIST, Gaithersburg, MD, USA, NISTIR 8009, 2014.

[6] National Institute of Standards and Technology, "NIST evaluation shows advance in face recognition software's capabilities" (reporting NISTIR 8238), news release, Nov. 2018.

[7] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, Handbook of Fingerprint Recognition, 2nd ed. London, U.K.: Springer, 2009.

[8] J. M. Butler, M. C. Kline, and M. D. Coble, "NIST interlaboratory studies involving DNA mixtures (MIX05 and MIX13): Variation observed and lessons learned," Forensic Sci. Int. Genet., vol. 37, pp. 81–94, 2018, doi: 10.1016/j.fsigen.2018.07.024.

[9] W. C. Thompson, "Uncertainty in probabilistic genotyping of low template DNA: A case study comparing STRmix and TrueAllele," J. Forensic Sci., vol. 68, no. 3, pp. 1049–1063, 2023, doi: 10.1111/1556-4029.15225.

[10] G. Mohmed et al. (as reported in review), "AI-driven digital evidence triage in digital forensics: A comprehensive review," 2024. Available via ResearchGate (workbook source S16).

[11] M. Grossman and P. Grimm, "Judicial approaches to acknowledged and unacknowledged AI-generated evidence," Sci. Technol. Law Rev., vol. 26, no. 2, 2025, doi: 10.52214/stlr.v26i2.13890.

[12] Government of India, Bharatiya Nagarik Suraksha Sanhita, 2023, s. 176(3) (in force 1 July 2024).

[13] Budding Forensic Expert, "The backlog crisis: India's forensic labs are drowning," Apr. 2026 (secondary analysis citing national estimates; workbook source S21).

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. NeurIPS, 2012.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proc. IEEE CVPR, 2015, pp. 815–823.

[16] F. Orsini et al., "The application of artificial intelligence in forensic pathology: A systematic literature review," Front. Med., vol. 12, Art. no. 1583743, 2025, doi: 10.3389/fmed.2025.1583743.

[17] P. Grother, M. Ngan, and K. Hanaoka, "Face Recognition Vendor Test (FRVT) Part 3: Demographic effects," NIST, Gaithersburg, MD, USA, NISTIR 8280, Dec. 2019, doi: 10.6028/NIST.IR.8280.

[18] M. W. Perlin et al. [full author list and title to be verified], "Commentary on: Uncertainty in probabilistic genotyping of low template DNA: A case study comparing STRmix and TrueAllele," J. Forensic Sci., 2024.

[19] Center for Strategic and International Studies, "How accurate are facial recognition systems – and why does it matter?," 2020 (policy analysis reporting the April 2020 NIST leaderboard; workbook source S04).

[20] A. K. Jain et al., "End-to-end latent fingerprint search," arXiv:1812.10213, 2018 (workbook source S07; authors and publication status to be confirmed).

[21] A. Nagar et al., "Improving automated latent fingerprint identification using extended minutia types," arXiv:1810.09801, 2018 (workbook source S08; to be confirmed).

[22] Biometric Update, "ROC scores 30% latent fingerprint biometrics accuracy gain in NIST's ELFT," Feb. 2025; and "Innovatrics, ROC improve rankings in NIST ELFT," Feb. 2026 (trade press reporting NIST ELFT; replace with the NIST ELFT record).

[23] "DNA mixture deconvolution: A four-strategy framework from physical separation to database searching," Genes, vol. 17, no. 4, Art. no. 434, 2026 (workbook source S14; authors to be added).

[24] "Methods and tools of digital triage in forensic context: Survey and future directions," Symmetry, vol. 9, no. 4, Art. no. 49, 2017 (workbook source S17; authors to be added).

[25] DFPulse, "2024 digital forensic practitioner survey," as reported by Cyacomb (workbook source S18; sampling frame not published).

[26] Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

[27] Advisory Committee on Evidence Rules, Report to the Committee on Rules of Practice and Procedure, Judicial Conference of the United States, May 17, 2026 (as reproduced on uscourts.gov).

[28] Mendones v. Cushman & Wakefield, Inc., No. 23CV028772 (Cal. Super. Ct., Alameda County, Sept. 9, 2025).

[29] State of Louisiana, Act No. 250 (H.B. 178), 2025 (known to the author only through secondary commentary; primary text to be checked).

[30] R. Pfefferkorn, "Letter to NCMEC about AI-CSAM report statistics," Stanford Center for Internet and Society, 2026.

[31] Budding Forensic Expert, "Haryana FSL achieves 8% reduction in pending cases," Apr. 2026 (secondary report; workbook source S22).