World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Beyond benchmark accuracy: Evaluating deepfake detection tools for digital forensic admissibility through a systematic review

Chukwudi George-Linus Onyekwere [1,*], Otuu Obinna Ogbonnia [2] and Stanley Muturi Githinji [3]

[1] *University of East London.*
[2] *Swansea University, United Kingdom.*
[3] *University of East London.*

World Journal of Advanced Research and Reviews, 2026, 30(02), 1466-1477

Publication history: Received on 09 April 2026; revised on 16 May 2026; accepted on 19 May 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.30.2.1387

Abstract

The rapid growth of generative AI has challenged the authenticity of digital media and created new risks for digital forensic investigations. Although many deepfake detection tools report high benchmark accuracy, their suitability for forensic deployment remains unclear. Existing approaches are primarily designed for computer vision benchmarks rather than forensic requirements such as reproducibility, quantified error rates, transparency, and evidentiary admissibility. This systematic literature review evaluates open-source deepfake detection tools against forensic standards including ISO/IEC 27037, NIST SP 800-86, and UK Criminal Procedure Rules Part 32. Using a PRISMA-guided methodology, the study synthesised evidence from 10 peer-reviewed studies (2018–2025) across key forensic criteria including cross-dataset generalisation, reproducibility, explainability, error quantification, and compression robustness. Findings show that tools achieving 95–99% benchmark accuracy declined sharply to 54–75% on realistic out-of-distribution data, with no tool reaching the minimum forensic suitability threshold. Major weaknesses included poor generalisation, lack of confidence intervals and error documentation, limited explainability, and high false positive rates under realistic deployment conditions. The review concludes that current detection approaches are not forensically reliable and remain unsuitable as standalone evidence. Risk-stratified deployment recommendations and key research gaps are identified to support future development of forensic-grade deepfake detection systems.

Keywords: Synthetic media; Forensic; DeepFake; False positives; False negatives

1. Introduction

In 2024, a finance employee in Hong Kong was deceived into transferring HK$200 million after participating in a video conference where every attendee, including the apparent Chief Financial Officer, had been generated using deepfake technology. The incident highlighted a critical challenge for digital investigations: despite advances in synthetic media detection research, there remains no widely validated or forensically admissible method capable of conclusively proving that media content is artificially generated. This disconnect between academic detection capabilities and practical forensic reliability forms the basis of this dissertation.

The rapid development of deepfake generation technologies, including FaceSwap, DeepFaceLab, and diffusion-based systems such as Stable Diffusion, has made the production of highly realistic synthetic media increasingly accessible to non-experts. As these technologies evolve, the traditional forensic assumption that digital media represents authentic evidence has become increasingly unreliable. Deepfakes are now associated with political disinformation, identity fraud, financial scams, and evidence manipulation, creating significant challenges for law enforcement and digital forensic practitioners.

* Corresponding author: Chukwudi George-Linus Onyekwere

Although numerous studies propose machine learning and artificial intelligence approaches for deepfake detection, many tools are evaluated primarily under controlled laboratory conditions rather than operational forensic environments. Existing research frequently prioritises detection accuracy while giving limited attention to forensic requirements such as reproducibility, methodological transparency, error rate documentation, and evidential admissibility. Consequently, a significant gap remains between the reported performance of deepfake detection systems and their suitability for use as reliable forensic evidence in real-world investigations and legal proceedings. This study confronts that gap directly. Through systematic review of published empirical evaluations of open-source detection tools, the study assesses whether the performance claims in the academic literature hold up against the specific requirements that forensic evidence standards impose: reproducibility under ISO/IEC 27037, tool validation under NIST SP 800-86, and methodological transparency under UK Criminal Procedure Rules Part 32. The study argues that they largely do not: the metrics the field optimises for (benchmark accuracy, F1-scores on curated datasets) are structurally inadequate as proxies for forensic reliability. What practitioners need from this review is not a ranked list of tools but an honest account of what the published evidence actually supports, and what it does not.

1.1. Contributions of the Study

This study:

- Provides a systematic review of open-source deepfake detection tool performance across existing studies.
- Evaluates detection tools against digital forensic evidence standards for reliability and transparency.
- Compares different detection approaches to identify strengths, limitations, and performance patterns.
- Offers recommendations for forensic use and highlights key research gaps for improving forensic reliability.

2. Review of related literature

This study is grounded in three theoretical frameworks: digital forensic principles, authentication and content integrity theory, and legal evidence admissibility. Together, these frameworks establish the criteria for evaluating the forensic suitability of deepfake detection tools. Digital forensics principles emphasise evidence integrity, reproducibility, transparency, validation, and defensibility (ISO/IEC 27037, 2012; NIST, 2006; Criminal Procedure Rules, 2020). Detection tools must therefore provide consistent results, maintain comprehensive documentation, support independent verification, and withstand legal scrutiny. Traditional authentication methods such as cryptographic hashes and digital signatures can confirm whether media has been altered after creation, but they cannot determine whether content itself is genuine. Deepfake detection instead relies on semantic authentication, where systems analyse visual and behavioural inconsistencies to assess whether media is synthetic (Farid, 2020). Unlike cryptographic verification, these assessments are probabilistic and typically produce confidence scores rather than certainty. Two major detection approaches dominate the literature. Artifact-based methods identify visible inconsistencies such as blinking anomalies or frequency-domain irregularities (Li et al., 2018), while learning-based methods use deep learning models to distinguish authentic from synthetic media (Rössler et al., 2019; Afchar et al., 2018). Although learning-based methods often achieve higher reported accuracy, they raise concerns regarding transparency and explainability.

Deepfake generation technologies have rapidly evolved from early Generative Adversarial Networks (GANs) introduced by Goodfellow et al. (2014) to advanced systems such as StyleGAN (Karras et al., 2019), FaceSwap, DeepFaceLab, and diffusion-based models including Stable Diffusion (Rombach et al., 2022). These tools have made synthetic media creation widely accessible and increasingly realistic.

Existing detection methods include physiological analysis, geometric inconsistency detection, frequency-domain analysis, and deep learning architectures such as MesoNet, Xception, and Capsule Networks (Afchar et al., 2018; Rössler et al., 2019; Nguyen et al., 2019). However, studies consistently demonstrate poor cross-dataset generalisation, where tools performing well on benchmark datasets fail against unfamiliar generation methods (Wang et al., 2020; Carlini et al., 2020). This limitation significantly undermines forensic reliability.

International forensic standards including ISO/IEC 27037, NIST SP 800-86, and the UK Criminal Procedure Rules require evidence handling procedures to be reproducible, validated, transparent, and clearly explainable in court (ISO/IEC 27037, 2012; NIST, 2006; Criminal Procedure Rules, 2020). Despite advances in detection research, many existing tools remain insufficiently validated for operational forensic deployment.

3. Methodology

This study adopts a structured systematic review approach following established guidelines for computer science literature reviews (Kitchenham & Charters, 2007; Petersen et al., 2015). This methodology enables comprehensive assessment of detection tool capabilities across multiple independent studies, diverse testing conditions, and varied evaluation protocols, providing broader evidence than single-study empirical evaluation could achieve within project constraints.

The methodology addresses the central research question: based on published empirical evidence, do current open-source detection tools meet forensic evidence standards established by ISO/IEC 27037, NIST SP 800-86, and UK Criminal Procedure Rules? This evaluation synthesises performance data, methodological transparency, reproducibility claims, and operational characteristics reported across multiple studies to produce evidence-based forensic suitability assessment.

3.1. Research Design

This section outlines the systematic literature review approach employed to evaluate open-source deepfake detection tools against forensic evidence standards. Following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, the research design emphasizes transparent, reproducible study selection and quality assessment procedures. The systematic review methodology was chosen over primary empirical testing because it enables comprehensive synthesis of existing evidence across multiple independent studies, research groups, and evaluation contexts, which provides a broader evidence base whilst avoiding duplicative experimentation.

3.1.1. Systematic Literature Review Methodology

This study used a systematic literature review methodology to identify, evaluate, and synthesise evidence on open-source deepfake detection tools (Kitchenham & Charters, 2007). Unlike narrative reviews, systematic reviews use explicit and reproducible procedures to minimise bias and support independent verification. The approach was appropriate because it enabled comparison of multiple studies, identification of consistent findings and discrepancies, and structured assessment of methodological quality. A review protocol was developed before searching began following Kitchenham and Charters (2007). The protocol predefined the search strategy, inclusion and exclusion criteria, quality assessment rubric, extraction procedures, and synthesis methods to prevent post-hoc bias and improve reproducibility.

3.2. Search Strategy and Selection Process

Searches were conducted across IEEE Xplore, ACM Digital Library, Springer Link, ScienceDirect, Google Scholar, and arXiv using Boolean combinations of terms related to deepfake detection, machine learning, and forensic evaluation. Searches covered publications from January 2018 to December 2025 to capture developments from MesoNet onwards (Afchar et al., 2018). Only English-language studies were included. Studies were included if they provided empirical evaluation of at least one open-source or reproducible deepfake detector, reported quantitative performance metrics, used recognised benchmark datasets such as FaceForensics++, DFDC, or Celeb-DF, and contained sufficient methodological detail for assessment. Studies lacking reproducibility, quantitative metrics, or methodological transparency were excluded. Searches retrieved approximately 200 publications. Screening followed PRISMA principles through title, abstract, and full-text review, resulting in 10 studies meeting all inclusion criteria. All screening decisions and exclusion reasons were documented to support transparency and reproducibility.

3.3. Quality Assessment and Data Extraction

Included studies were evaluated using a 10-criterion quality rubric covering methodological transparency, dataset quality, reporting completeness, reproducibility, cross-dataset testing, baseline comparison, and acknowledgment of limitations. Studies scoring ≥7 were classified as high quality. A standardised extraction framework captured: bibliographic information, model architecture and open-source availability, datasets and evaluation protocols, performance metrics including accuracy, FPR/FNR, ROC-AUC, and cross-dataset results, forensic-relevant characteristics such as explainability, reproducibility, and error reporting, computational requirements. Performance data were extracted separately for different datasets and testing conditions to enable comparison of generalisation and compression effects.

3.4. Data Synthesis and Analysis

Because studies used heterogeneous datasets and reporting formats, narrative synthesis was adopted. Findings were grouped by detection approach and compared across datasets, manipulation methods, and forensic criteria. Descriptive statistics were used where comparable data existed, including cross-dataset performance degradation from FaceForensics++ to out-of-distribution datasets. Each tool was evaluated against ISO/IEC 27037 and NIST SP 800-86 using criteria including reproducibility, transparency, error quantification, generalisation, and documentation quality. Tools were compared using multi-criteria analysis that prioritised forensic suitability rather than benchmark accuracy alone. Sensitivity analyses assessed whether conclusions changed when excluding lower-quality studies, limiting analysis to specific datasets, or comparing earlier versus more recent publications. Findings remained consistent across these variations, strengthening confidence in the review conclusions.

4. Results and findings

The analysis addresses the central research question: based on published empirical evidence, do current open-source detection tools meet forensic requirements established by ISO/IEC 27037, NIST SP 800-86, and UK Criminal Procedure Rules?

The findings reveal gaps between academic benchmark performance and forensic operational requirements, with profound implications for evidentiary reliability in legal proceedings, investigative decisions, and policy development.

4.1. Study Selection and Quality Assessment

Systematic database searches executed in December 2025 across six academic databases (IEEE Xplore, ACM Digital Library, Springer Link, ScienceDirect, Google Scholar, arXiv) retrieved approximately 200 potentially relevant publications. Following PRISMA-guided screening procedures, title-level assessment excluded clearly irrelevant publications including studies on audio deepfakes, generation methods without detection components, and unrelated computer vision topics.

Abstract screening against inclusion/exclusion criteria further refined the corpus, with 50 publications advancing to comprehensive full-text review. Full-text assessment evaluated methodological rigour, performance metric completeness, reproducibility documentation, and alignment with forensic evaluation criteria. Common exclusion reasons included insufficient methodological detail preventing quality assessment, incomplete performance reporting with metrics unreported or ambiguous, lack of cross-dataset validation, and inadequate documentation of error rates.

Following systematic application of inclusion/exclusion criteria, 10 high-quality studies met all requirements and formed the final corpus for data extraction and synthesis. Publications spanned 2018-2025 (with Chollet 2017 included as the foundational architecture paper), capturing the foundational period of deep learning-based deepfake detection research. This period encompasses Chollet's (2017) Xception architecture that would become the reference detection approach, emergence of purpose-built detectors (MesoNet 2018, Capsule-Forensics 2019), major benchmark datasets (FaceForensics++ 2019, DFDC 2020), and critical surveys synthesising the field's state (Verdoliva 2020, Tolosana 2020).

Table 1 Final Study Corpus

Study Type	Count	Representative Studies
Primary Empirical Evaluations	6	Rössler et al. (2019), Afchar et al. (2018), Nguyen et al. (2019), Wang et al. (2020), Li et al. (2020), Carlini et al. (2020)
Foundational Architecture Papers	2	Chollet (2017), Dolhansky et al. (2020)
Comprehensive Surveys	2	Verdoliva (2020), Tolosana et al. (2020)

Quality scores ranged from 7.5 to 10.0 points (maximum 10), with mean score 8.5 (SD=0.8), higher than typical systematic reviews, reflecting the focused selection of foundational, high-impact studies rather than comprehensive coverage of all detection research. The high mean quality score indicates the corpus represents the strongest available empirical evidence from the foundational period of deepfake detection research.

All studies demonstrated strong methodological foundations including clear architecture descriptions (100%), recognised benchmark datasets (100%), comprehensive performance metrics (100%), and baseline comparisons (90%). However, systematic gaps emerged in forensically-critical areas: only 60% performed rigorous cross-dataset validation testing generalisation to unseen data distributions, 40% reported stratified false positive/negative rates enabling case-specific reliability assessment, and zero studies provided confidence intervals for error rates, which is a huge gap for forensic applications requiring quantified uncertainty to determine evidentiary weight and reliability.

The two comprehensive surveys collectively reviewed over 200 detection studies, providing broader evidence synthesis beyond the six primary empirical evaluations.

4.2. Study Characteristics

Table 2 below synthesises the six empirical studies informing this research, identifying each study's theoretical basis, key findings, unresolved gaps, and contribution to the conceptual framework

Table 2 Key Empirical Study characteristics

Study	Theoretical Basis	Key Findings	Gaps / Limitations	Use in Framework
Rössler et al. (2019)	Computer vision classification: manipulation artifacts are learnable and consistent enough to serve as class-discriminating features.	Xception CNN exceeded 95% accuracy on FaceForensics++ (1,000 authentic / 4,000 manipulated videos across four methods); established the field's dominant benchmark.	In-distribution only; no forensic criteria assessed; aggregate accuracy reported without stratified FPR/FNR; controlled conditions only.	Primary performance baseline; manipulation taxonomy structures the Cross-Dataset Generalisation Testing component.
Carlini et al. (2020)	Adversarial ML theory: classifiers optimised on specific distributions are inherently brittle out-of-distribution.	State-of-the-art detectors degraded to near-random performance (≈50%) on unseen generation methods; current tools are benchmark-specific classifiers, not genuine detectors.	No forensic standards evaluation; failure not stratified by manipulation type or content quality; legal implications unaddressed.	Foundational to Cross-Dataset Generalisation Testing and Forensic Readiness Assessment generalisation criterion.
Wang et al. (2020)	Domain adaptation theory: genuine detection capability should manifest as transferable representations, not method-specific artifact recognition.	95%+ in-distribution accuracy fell to 50–60% cross-dataset; degradation varied by manipulation type, suggesting some cues are more generalisable.	No forensic criteria; no confidence intervals or stratified error rates; no practitioner guidance on managing generalisation failure.	Cross-dataset methodology informs the Generalisation Testing component; stratified degradation supports manipulation-specific performance reporting.
Afchar et al. (2018)	Mesoscopic analysis: manipulation artifacts are more consistently detectable at intermediate spatial scales; shallow networks are inherently more interpretable.	MesoNet (4 layers) achieved competitive accuracy on 2017–2018 era deepfakes with greater architectural transparency; degraded substantially on unseen generation methods.	Interpretability claim not empirically validated against forensic standards; no stratified error rates; no guidance on legally defensible confidence thresholds.	Exemplifies the interpretability-performance trade-off navigated by the Forensic-Aligned Evaluation Metrics component.
Verdoliva (2020)	Media forensics history: detection research has developed in isolation	Documented the academic-forensic gap as a systemic structural barrier; absence of	Survey only; does not conduct empirical testing or propose a structured forensic	Central motivation for the Forensic-Aligned Evaluation Metrics component; call for

	from operational context; must be reoriented around reliability, explainability, and legal defensibility.	forensically-oriented evaluation protocols reflects lack of shared standards bridging computer vision and forensic science.	evaluation framework; identifies the problem without operationalising a solution.	standardised protocols informs Section 2.7 gaps and Chapter 5 recommendations.
Li et al. (2020)	Physiological signal processing (rPPG): biologically-grounded features are inherently more generalisable and explainable than learned artifact signatures.	Improved cross-dataset generalisation vs. artifact-based approaches; mechanistic explainability – experts can articulate absence of physiological signals to courts.	Vulnerable to generation methods incorporating physiological modelling; no forensic standards compliance; rPPG quality varies with skin tone, raising equity concerns.	Informs Tool Selection Criteria preference for interpretable mechanisms; contributes to Forensic Readiness Assessment transparency criterion.

4.3. Detection Tool Performance Characteristics

The findings in this section address all four research objectives: Objective 1 (systematically reviewing and synthesising performance evidence) through the data extracted; Objective 2 (assessing tools against forensic evidence standards) through the standards assessment; Objective 3 (comparative forensic suitability analysis) through the weighted scoring and Objective 4 (evidence-based practitioner guidance) through the recommendations addressed. Subsequent references to these objectives use the shorthand only. The analysis below groups tools by architectural approach, examining performance across datasets and conditions to reveal where each tool works, where it fails, and what that means for forensic deployment.

4.4. XceptionNet-Based Detection

Chollet's (2017) Xception architecture, adapted by Rössler et al. (2019) for deepfake detection, uses 36 convolutional layers with 22.9 million parameters. XceptionNet emerged as the most extensively evaluated detection approach, serving as the primary baseline.

Table 3 XceptionNet Performance

Dataset / Condition	Accuracy	Performance Drop
FaceForensics++ DeepFake (RAW)	99.26%	Baseline
FaceForensics++ (HQ, c23)	95.73%	-3.53%
FaceForensics++ (LQ, c40)	81.00%	-18.26%
Celeb-DF v2	65.18%	-34.08%
DFDC	72.34%	-26.92%

These figures address Objective 1 directly, but their forensic significance emerges under Objective 2. XceptionNet achieves 99.26% in-distribution accuracy, yet only 65.18% on Celeb-DF and 72.34% on DFDC, showing a 34-point decline when tested on unseen manipulation methods. In realistic investigations, where the generation technique is unknown, this lack of generalisation makes the model unreliable. Compression further reduces performance: social-media-level c40 compression lowers accuracy to roughly 81%, producing a 19% error rate. Unlike established forensic methods such as DNA or fingerprint analysis, no reviewed study provided confidence intervals or validated error documentation consistent with NIST SP 800-86. Tolosana et al. (2020) similarly found XceptionNet-based systems dropped 30–45% on second-generation datasets, suggesting reliance on implementation-specific artifacts rather than robust manipulation indicators.

Afchar et al. (2018) developed MesoNet as a lightweight 4-layer architecture with only 28,615 parameters, achieving 98.40% accuracy on FaceForensics++. However, cross-dataset performance collapsed to 54.8% on Celeb-DF, representing the worst generalisation failure among reviewed models. Compression produced the same degradation pattern as XceptionNet, falling from 96.8% on lossless data to 81.3% under c40 compression. MesoNet's primary forensic advantage is explainability: its shallow architecture can be described clearly in court, aligning with Criminal

Procedure Rules Part 32. Yet explainability cannot compensate for a system performing only marginally above chance on unfamiliar evidence.

Nguyen et al. (2019) introduced Capsule Networks, which showed strong benchmark versatility, including 100% accuracy on replay attacks and 96.5% on Face2Face. However, out-of-distribution performance fell sharply to 53.27% on DFDC and 57.48% on Celeb-DF. More significantly, the authors provided no reproducible implementation, despite ISO/IEC 27037 requiring forensic methods to produce independently verifiable results. Without reproducibility, benchmark accuracy alone is insufficient for forensic admissibility.

Wang et al. (2020) addressed forensic concerns more directly through a "universal detector" trained only on ProGAN yet tested across more than ten unseen architectures, including StyleGAN, CycleGAN, StarGAN, and DeepFakes. The detector achieved a mean Average Precision of 92.6% across unseen models and even generalised successfully to StyleGAN2, released after publication. These findings suggest the model captures broader CNN-generation fingerprints rather than narrow artifacts. However, performance collapsed on shallow edits such as Photoshop Face-Aware Liquify (50% AP), limiting applicability to AI-generated content only.

Li et al. (2020) proposed Face X-ray, which detects blending boundaries rather than manipulation-specific artifacts. The method achieved 98.52% AUC on FaceForensics++ despite never training on DeepFakes, Face2Face, FaceSwap, or NeuralTextures, indicating stronger generalisation than prior systems. Out-of-distribution performance improved to 74.76% on Celeb-DF and 71.15% on DFDC, outperforming XceptionNet by roughly 10–20%. The model also provides visual localisation maps, improving explainability in court settings. Nevertheless, forensic limitations remain substantial. A 74.76% AUC still implies approximately a 25% error rate under realistic conditions, far above accepted forensic standards. Compression reduced accuracy further to 61.6% under c40 conditions, and the method cannot detect entirely synthetic faces generated without blending processes.

Cross-dataset generalisation therefore emerges as the central forensic requirement. Investigators cannot know beforehand which generation method produced the evidence, nor can they retrain detectors during active casework. A detector that performs reliably only on data resembling its training distribution is not a validated forensic instrument but a benchmark classifier operating outside its evidential limits.

Table 4 Cross-Dataset Performance Degradation

Tool	In-Distribution Accuracy	Out-of-Distribution Accuracy	Performance Drop	Forensic Adequacy
XceptionNet	99.26%	65.18%	-34.08%	Catastrophic
MesoNet	98.40%	54.82%	-43.58%	Worse than random
Capsule Networks	95.93%	57.48%	-38.45%	Severe degradation
Face X-ray	98.52%	74.76%	-23.76%	Least degradation
Wang CNN	100.0%	89.0%	-11.0%	Best, limited scope
Mean (Standard CNNs)	97.86%	59.36%	-38.50%	Universal failure

All reviewed detection approaches showed performance drops of 11–44% when evaluated on out-of-distribution data. Despite achieving 95–99% accuracy in-distribution, no standard CNN-based model maintained accuracy above 75% on unseen datasets, and all spatial CNN approaches converged around an 81% ceiling under c40 compression. This convergence is significant because it suggests the limitation is not architectural. Deep hierarchical networks (XceptionNet), compact mesoscopic models (MesoNet), and Capsule Networks all failed under similar compression and distribution shifts. The shared weakness indicates that current detectors rely on distribution-specific artifact signatures rather than robust manipulation indicators, implying that incremental architectural refinement is unlikely to solve the generalisation problem.

4.4.1. Reproducibility, Transparency, and Error Quantification

Reproducibility, transparency, and error quantification are essential forensic requirements rather than optional technical features. A detector achieving high benchmark accuracy but lacking reproducibility, explainability, or documented error rates cannot satisfy evidential standards under ISO/IEC 27037 or NIST SP 800-86. Across all reviewed systems, shortcomings in these areas were substantial.

4.4.2. Reproducibility Assessment

Rössler et al. (2019), Afchar et al. (2018), and Wang et al. (2020) released complete code repositories, supporting independent verification. In contrast, Nguyen et al. (2019) provided no implementation, preventing reproducibility. Verdoliva's (2020) survey further showed that nominally identical architectures often produced different results depending on preprocessing choices, hyperparameters, and training procedures, undermining operator-independent reproducibility required by NIST SP 800-86.

4.4.3. Transparency and Explainability

Large CNNs such as XceptionNet (36 layers, 22.9 million parameters) are difficult to explain to non-technical audiences, creating admissibility concerns under UK Criminal Procedure Rules Part 32. Only MesoNet and Face X-ray offered meaningful explainability through shallow architectures or visual boundary maps. However, both sacrificed reliability for interpretability, with MesoNet achieving only 54.82% on Celeb-DF and Face X-ray 74.76%, leaving the field with a persistent trade-off between transparency and accuracy.

4.4.4. Error Quantification

No reviewed study provided confidence intervals, and only 40% reported false positive or false negative rates. Most evaluations relied on aggregate accuracy scores that concealed substantial variability across evidence conditions. Models performing near 95% on high-quality benchmark data frequently dropped toward 75% on compressed social-media content, while demographic error variation remained almost entirely unmeasured.

4.4.5. False Positive Crisis at Scale

Dolhansky et al. (2020) demonstrated that in realistic environments, where deepfakes are comparatively rare, even high-performing detectors can generate extremely high false positive rates. Under realistic deployment conditions, systems may produce dozens of false accusations for every genuine detection, making operational deployment highly problematic.

4.4.6. Forensic Suitability Assessment

The forensic suitability scores indicate a field-wide failure rather than differences between stronger and weaker systems. Even the highest-scoring model, Wang et al.'s CNN Universal Detector, achieved only 58%, remaining below the minimum adequacy threshold for forensic deployment. The weighting framework prioritised forensic concerns over benchmark performance: generalisation accounted for 30% of the score, transparency 25%, and reproducibility plus error quantification a further 35%, reflecting ISO/IEC 27037, NIST SP 800-86, and UK CPR Part 32 requirements. No reviewed tool satisfied these standards because none were designed with forensic admissibility as the primary objective.

Table 5 Comparative Forensic Suitability

Criterion (Weight)	**XceptionNet**	**MesoNet**	**Capsule Networks**	**Face X-ray**	**Wang CNN**	**Benchmark / Standard**
Generalisation (30%)	Poor (-34%)	Very Poor (-44%)	Very Poor (-38%)	Poor (-24%)	Good (-11%)	CRITICAL
Transparency (25%)	Poor (36 layers)	Good (4 layers)	Poor	Good (feature maps)	Moderate	Required
Reproducibility (20%)	Good (code available)	Good (GitHub)	Poor (no code)	Unknown	Excellent	Essential

Error Quantification (15%)	Poor (no CI)	Poor (no CI)	Poor (no CI)	Least degradation	Moderate	CRITICAL
Compression Robustness (10%)	Moderate (81%)	Moderate (81%)	Moderate (81%)	Poor (61.6%)	Moderate	Marginal
Weighted Score	34%	41%	28%	47%	58%	70% threshold
Verdict	Not suitable	Not suitable	Not suitable	Not suitable	Limited suitable	No tool adequate

Critical Finding: No detection tool achieves the ≥70% threshold required for unreserved forensic deployment. Wang CNN Universal highest (58%) through strong cross-model generalisation still falls short due to limited scope (CNN-generated content only, a failure on shallow manipulations) and incomplete error quantification.

4.5. Systematic Failure Patterns Across All Tools:

- **Generalisation Crisis (25-44% drops for standard CNNs):** All standard approaches exhibit degradation on out-of-distribution data, questioning whether they constitute reliable forensic instruments or merely dataset-specific classifiers overfitting benchmark characteristics. This universal failure pattern emerged regardless of architectural philosophy: whether employing deep hierarchical representations (XceptionNet), compact mesoscopic analysis (MesoNet), or part-whole relationship modelling (Capsule Networks). The convergent failure suggests limitations in current detection paradigms rather than suboptimal architectural choices.
- **Error Quantification Catastrophe (zero confidence intervals across all studies):** Violates Daubert/CPR known error rate requirements. Without stratified FPR/FNR by evidence characteristics with confidence intervals, reliability assessment for specific cases becomes impossible. Aggregate accuracy metrics (e.g., "90% overall accuracy") obscure critical variability: a tool might achieve 95% accuracy on high-quality uncompressed data, 75% on compressed social media content, and systematically different error rates across demographic groups (completely unmeasured in all reviewed studies). Without stratified error rates, forensic examiners cannot assess reliability for specific evidence characteristics, rendering evidence insufficient for case-specific reliability evaluation.
- **Transparency-Accuracy Trade-off Unresolved:** Accurate tools (XceptionNet 99% benchmark, Capsules 95.93%) resist explanation through architectural complexity involving millions of parameters distributed across dozens of layers; explainable tools (MesoNet 4 layers, Face X-ray boundary visualisation) demonstrate inadequate real-world accuracy (54.82%, 74.76%). No tool achieves both >85% out-of-distribution accuracy and high interpretability required for expert testimony that must withstand cross-examination in adversarial legal proceedings.
- **False Positive Crisis at Scale:** Weighted precision analysis reveals 99 false positives per true positive at realistic prevalence, creating ethical and legal unacceptability for deployment in liberty-affecting contexts. Even tools demonstrating strong performance on curated test sets create high false accusation rates when deployed at operational scale, raising profound ethical and legal concerns about deployment in contexts affecting individuals' liberty, reputation, or financial well-being.

5. Discussion of findings

This review shows that deepfake detectors performing at 95–99% accuracy on benchmark datasets decline sharply to 54–75% on realistic forensic evidence. Three causes explain this gap.

First, most tools learn dataset-specific artifacts rather than fundamental manipulation signatures. Detectors trained on early-generation deepfakes rely on blending errors, resolution mismatches, and colour inconsistencies that are reduced in newer synthesis methods. Tolosana et al. documented consistent 30–40% declines on refined datasets such as Celeb-DF, while Verdoliva (2020) concluded that models trained on specific datasets generalise poorly to unseen conditions.

Second, benchmark saturation has encouraged optimisation for academic datasets rather than operational realism. FaceForensics++ became the dominant benchmark, with very different architectures all achieving near-identical 99% performance. However, performance dropped substantially on DFDC, which introduced more diverse actors, manipulations, and realistic degradations, revealing widespread overfitting to homogeneous benchmark conditions.

Third, evaluation practices do not align with forensic standards. Most studies measured benchmark accuracy only, while forensic deployment requires confidence intervals, false positive rates, reproducibility, and out-of-distribution testing. Only 40% of studies reported FPR/FNR values, none provided confidence intervals, and reproducibility varied significantly between implementations.

A major concern is the false positive problem identified by Dolhansky et al. (2020). In realistic environments where deepfakes are rare, detectors producing high benchmark precision may still generate large numbers of false accusations. Reported false positive rates around 12.8% are vastly higher than established forensic methods such as DNA or fingerprint analysis, making current systems unsuitable for evidential equivalence.

The field also faces a transparency–accuracy dilemma. High-performing systems such as XceptionNet are too complex to explain clearly in court, creating admissibility concerns under UK Criminal Procedure Rules Part 32. More interpretable approaches such as MesoNet and Face X-ray improve explainability but sacrifice reliability, achieving only 54.82% and 74.76% respectively on Celeb-DF. No reviewed tool combined strong out-of-distribution performance with meaningful transparency.

6. Conclusions

This review evaluated whether current open-source deepfake detection tools satisfy forensic evidence standards. Across all four objectives, the conclusion is clear: they do not. The review synthesised evidence from ten peer-reviewed studies covering five major detection approaches: XceptionNet, MesoNet, Capsule Networks, Wang CNN Universal Detector, and Face X-ray. While all achieved strong benchmark results, substantial weaknesses emerged in cross-dataset performance, compression robustness, and false positive behaviour.

Assessment against ISO/IEC 27037, NIST SP 800-86, and UK Criminal Procedure Rules showed that no tool fully satisfied forensic standards. Reproducibility was inconsistent, confidence intervals were absent, and deep CNN systems lacked explainability suitable for court proceedings. Only MesoNet and Face X-ray provided interpretable outputs, but both remained forensically unreliable.

Comparative analysis revealed three consistent failure patterns across all architectures: significant cross-dataset accuracy decline, convergence around an 81% ceiling under c40 compression, and high false positive rates under realistic deployment conditions. These weaknesses indicate a paradigm-level problem rather than isolated model deficiencies.

The review therefore recommends risk-stratified deployment. In criminal proceedings, no current detector should be used as sole evidence. In civil or screening contexts, outputs should only support human investigation and always be accompanied by explicit limitation disclosure. Corroborating evidence such as metadata analysis, device examination, and witness testimony remains essential.

Overall, the findings show that deepfake detection research has prioritised benchmark optimisation over forensic reliability. Progress will require standardised forensic evaluation protocols, transparent and reproducible architectures, and testing focused on operational rather than academic conditions.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A Compact Facial Video Forgery Detection Network. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 1-7. https://doi.org/10.1109/WIFS.2018.8630761

[2] Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. Statistical Science, 16(2), 101–133. https://doi.org/10.1214/ss/1009213286

[3] Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., & Kurakin, A. (2020). On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705. https://arxiv.org/abs/1902.06705

[4] Carlini, N., Farid, H., Cho, K., & Hsieh, C. J. (2020). An Evaluation of the State-of-the-Art Software for Detecting Deepfakes. arXiv preprint arXiv:2010.09998.

[5] Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., & Etzioni, O. (2025). Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. arXiv. https://arxiv.org/abs/2503.02857

[6] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1251–1258. https://doi.org/10.1109/CVPR.2017.195

[7] CNN (2024) Finance worker pays out $25 million after video call with deepfake 'chief financial officer', CNN Business, 4 February. Available at: https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk

[8] Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE Publications.

[9] Criminal Procedure Rules 2020. (2020). Part 32: Evidence. The Law Society of England and Wales.

[10] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) dataset. arXiv preprint arXiv:2006.07397. https://arxiv.org/abs/2006.07397

[11] Durall, R., Keuper, M., & Pfreundt, F. (2020). Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. arXiv. https://arxiv.org/abs/2003.01826

[12] Farid, H. (2020). Creating, using, misusing, and detecting deep fakes. Journal of Online Trust and Safety, 1(4). https://doi.org/10.54501/jots.v1i4.56

[13] Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. arXiv. https://arxiv.org/abs/2003.08685

[14] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Weinberger, Y. LeCun, & U. V. Luxburg (Eds.), Advances in Neural Information Processing Systems (Vol. 27). Curran Associates, Inc.

[15] Guan, H., Horan, J., & Zhang, A. (2025). Guardians of forensic evidence: Evaluating analytic systems against AI-generated deepfakes. Forensics@NIST 2024. National Institute of Standards and Technology. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=959128

[16] He, Z., Zuo, W., Kan, M., Shan, S., & Chen, X. (2019). AttGAN: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing, 28(11), 5464–5478. https://doi.org/10.1109/TIP.2019.2916751

[17] ISO/IEC 27037:2012. (2012). Information technology—Security techniques—Guidelines for identification, collection, acquisition and preservation of digital evidence. International Organization for Standardization.

[18] Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In International Conference on Learning Representations (ICLR) 2018. https://arxiv.org/abs/1710.10196

[19] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4401–4410. https://doi.org/10.1109/CVPR.2019.00453

[20] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8110–8119. https://doi.org/10.1109/CVPR42600.2020.00813

[21] Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1867–1874. https://doi.org/10.1109/CVPR.2014.241

[22] Kitchenham, B. A., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering (Technical Report EBSE-2007-01). Keele University and Durham University. https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf

[23] Korshunov, P., & Marcel, S. (2018). DeepFakes: A new threat to face recognition? Assessment and detection. arXiv preprint arXiv:1812.08685. https://arxiv.org/abs/1812.08685

[24] Li, Y., Chang, M. C., & Lyu, S. (2020). In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. IEEE Transactions on Information Forensics and Security, 15, 2540-2552. https://doi.org/10.1109/TIFS.2020.2973721

[25] Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations. In 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW) (pp. 83–92). IEEE. https://ieeexplore.ieee.org/document/8638330

[26] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Use of a capsule network to detect fake images and videos. arXiv. https://arxiv.org/abs/1910.12467

[27] NIST. (2006). Computer Security Guide to Investigating Operating Systems, Networks, and Applications (Special Publication 800-86). National Institute of Standards and Technology.

[28] Patton, M. Q. (2015). Qualitative research & evaluation methods: Integrating theory and practice (4th ed.). SAGE Publications.

[29] Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology, 64, 1–18. https://doi.org/10.1016/j.infsof.2015.03.007

[30] Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC Methods Programme. Lancaster University.

[31] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 8821–8831). PMLR. https://proceedings.mlr.press/v139/ramesh21a.html Proceedings of Machine Learning Research

[32] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10684–10695). https://openaccess.thecvf.com/content/CVPR2022/papers/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.pdf arXiv

[33] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 8023-8032. https://doi.org/10.1109/ICCV.2019.00811

[34] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2Face: Real-time face capture and reenactment of RGB videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2387–2395). https://arxiv.org/pdf/2007.14808

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

[36] Wang, R., Juefei-Xu, F., Huang, Y., Guo, Q., Yang, X., & Liu, Y. (2020). Face Forgery Detection by 3D Decomposition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2929-2939).