International Journal of Emerging Engineering and Technology (IJEET)
ISSN (e): 2958-3764
Volume: 4, Number: 2, Pages: 52- 60, Year: 2025

AI in Digital and Mobile Forensics: A Thematic Literature Review

Malik Ahsan Hayat *, and Irshad Ahmed Sumra
Department of Informatics and Systems, School of Science and Technology, University of Management and Technology (UMT), Lahore, 54000, Pakistan
* Corresponding author: Malik Ahsan Hayat (Email: ahsanhayat092@gmail.com)

Received: 12/08/2025, Revised: 22/11/2025, Accepted: 15/12/2025

***Abstract*—The vast amount of information that modern devices produce presents a big challenge for investigators— one that Artificial Intelligence (AI) and Machine Learning (ML) seem perfectly suited to solve. Indeed, AI and ML have been used more and more in digital forensics over the past decade. To better understand this trend, we conducted a thematic literature review of 29 papers published between 2010 and 2025. These papers fell into seven categories: core ML techniques; what practitioners think; image/multimedia classification; explainable AI; robustness/adversarial issues; large language models (LLMs); and governance/legal matters. We also examined three overarching issues: how accurate these tools really are; whether they can be trusted (a question that goes to their legal admissibility); and whether they produce results that humans can understand. Based on our findings, we identified three urgent needs for the field: standardized operating procedures; public benchmarks; and systems that are not only technically sound but also legally admissible and human-friendly.**

***Index Terms*—AI alignment, Digital forensics, Explainable AI, Image classification, large language models, Machine learning, Mobile forensics, Robustness.**

I. Introduction

Digital evidence is a major component in nearly every investigation these days. As such, digital forensics is tasked with managing a huge amount of data and processing it for use in investigations. A decade ago, Garfinkel [1] predicted that digital forensics would have a data overload problem with too much data coming too fast. Later on, Quick and Choo [2] gave numbers showing that indeed there is an ever-increasing flood of data being seized, and that forensics labs are having a hard time keeping up. Phones, laptops, cloud accounts, and IoT devices all create digital data at rates far beyond what any team of humans can handle. No wonder then that DFPulse's practitioner survey [3] found workload and keeping up with new technology as two of the biggest challenges facing digital forensics examiners today a view also backed up by Wickramasekara et al. [4] who see these same things as the main drivers behind calls for more automation in this field.

Increasingly, artificial intelligence and machine learning are being looked at as ways to help deal with these issues. Dunsin et al. [5] talk about this in their paper; they note how machine learning (ML) has begun to be used at all stages of the forensics process– acquisition, recovery, analysis, timeline building, even incident response. Fähndrich et al. [6] make a distinction between two types of AI in digital forensics: AI as a tool (i.e., something used to help investigators do their jobs) versus AI itself being the subject of investigation.

Most of the published work so far focuses on AI tools: developing one particular kind of ML technique to solve one particular problem. What happens if these systems are allowed to make decisions without human involvement? How reliable will they be, from a legal standpoint? What sort of impact could they have on the justice system if used widely? These bigger-picture questions are not getting much attention at present.

A. Scope and Method of This Review

In this paper we do not conduct any new experiments; nor do we provide a simple, descriptive overview of the state-of-the-art in AI applied to digital and mobile forensics. Instead what we offer is a thematic review that differs from existing systematic literature reviews in three important respects. First, whereas systematic reviews typically ask "what tools exist and how accurate are they?", we ask "what does the literature reveal about how AI is actually entering forensic practice? a question that connects technical capability with organisational adoption. Second, rather than evaluating individual techniques in isolation, we synthesise across surveys, tool evaluations, and governance studies to expose structural gaps between what vendors ship, what practitioners need, and what courts require. Third, we foreground the interpretive tension between headline accuracy and real-world utility, a theme that cuts across every technical domain we examine and those existing reviews have tended to treat as secondary. In doing so, we take care not only to provide an up-to-date picture, but one grounded in work that is both rigorously peer-reviewed and has real-world applicability (as opposed to being merely a technical demonstration).

To ensure this sort of quality control we focus mainly on survey articles and systematic literature reviews; tool evaluations; reports on domain-specific standards and best practice guidelines; and research that aims at either methods' general validity, reliability or at their explainability. These sources cluster heavily in the period since 2023, reflecting a field whose output has accelerated sharply.

Our literature search followed a structured, multi-stage protocol designed to maximise coverage while maintaining thematic coherence. We began by querying the following academic databases: IEEE Xplore, ACM Digital Library, Springer Link, ScienceDirect, and Google Scholar. The search string combined variants of the terms "artificial intelligence" OR "machine learning" OR "deep learning" with "digital forensics" OR "mobile forensics" OR "forensic investigation" and was limited to publications from 2010 to 2025. We also screened the



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License, which permits Unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

reference lists of retrieved articles (backward snowballing) and tracked citations of key papers (forward snowballing) to identify additional relevant work.

Inclusion criteria were: (i) peer-reviewed journal articles, conference papers, or established technical reports; (ii) explicit focus on AI or ML techniques in digital or mobile forensics; and (iii) publication between 2010 and 2025. Exclusion criteria were: (i) purely theoretical AI papers with no forensic application; (ii) publications not available in English; and (iii) student theses or non-peer-reviewed blog posts. From an initial pool of approximately 180 candidate papers, 29 were selected for detailed thematic analysis based on their relevance to the seven review themes identified in Section I-B. Fig. 1 summarises the paper selection workflow.

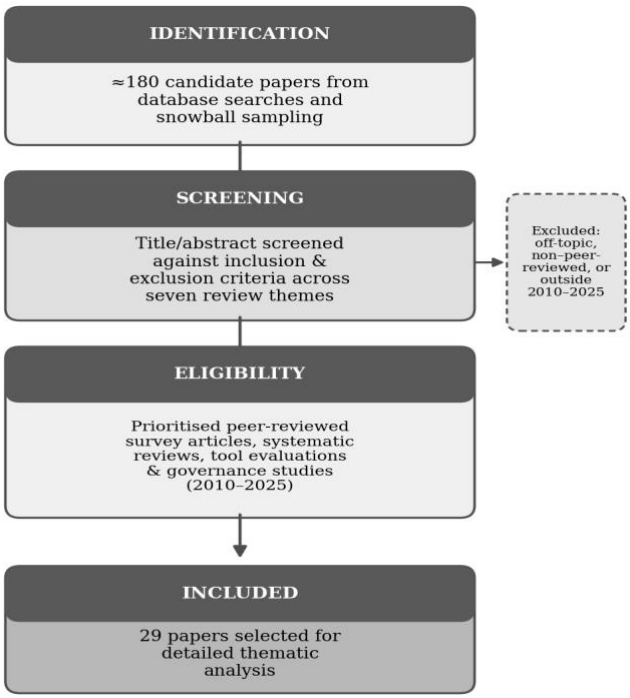


Fig. 1. Paper selection workflow for this thematic review. Starting from an initial pool of approximately 180 candidate papers identified through database searches and snowball sampling, 29 papers were selected for detailed analysis using inclusion and exclusion criteria organised around seven review themes. The multi-stage protocol prioritised peer-reviewed survey articles, systematic reviews, tool evaluations, and governance studies published between 2010 and 2025.

Previously, most studies have looked at one specific tool or survey, but we want to connect different pieces of work and see what they say collectively. By doing this, we hope to identify areas of agreement as well as contradictions that need to be resolved. Our review is not exhaustive— that would require a systematic review with detailed inclusion criteria and database searches— but we do aim for a broad coverage of themes rather than going deeply into just a few.

Our main contribution is to show how AI has become part of digital forensics not because practitioners have deliberately adopted it (in most cases) but because vendors have made it available in products they are buying. This interpretation helps to explain why there seems to be a gap between what some academic surveys say about AI's potential role in digital forensics and what practitioners actually do: The issue isn't so much one of model accuracy as of matching solutions to problems better.

Having said all that, our study is subject to several limitations that readers should keep in mind. For one thing, we only looked at English-language publications; it is possible that we missed relevant work in other languages. Also, because the use of AI in forensics is a relatively new topic, most of the surveys we reviewed were conducted after 2023: That means they may not fully capture earlier trends or provide a historical perspective on this issue.

Furthermore, practitioner surveys remain geographically concentrated, with the majority of respondents in studies such as DFPulse [3] and Vasilaras et al. [7] situated in Europe and North America. Consequently, the generalizability of current findings to other regions remains uncertain. Additionally, the reliance on citation-based snowball sampling as a retrieval strategy introduces a well-documented bias toward highly cited works, potentially excluding relevant but less prominent contributions from the corpus.

While these things do restrict our ability to generalize from this study alone, we believe they also suggest some directions for future research: In particular, follow-up reviews could try to overcome some of the limitations we have identified here by casting a wider net geographically speaking and using more systematic methods overall.

The remainder of the paper is organized by theme rather than by source. Section II surveys the ML techniques used across the forensic pipeline. Section III synthesizes what recent practitioner surveys reveal about adoption and trust. Section IV turns to image and multimedia classification, where the gap between headline accuracy and real performance is starkest. Sections V and VI take up interpretability and robustness, the two properties the literature treats as non-negotiable for courtroom use. Section VII examines the fast-moving frontier of large language models. Section VIII addresses governance, legal admissibility, and AI alignment. Section IX consolidates the research gaps, and Section X concludes.

II. AI and ML Techniques Across the Forensic Pipeline

A useful way to make sense of the literature is to map techniques onto the phases of an investigation, as detailed in Table II. The engine behind most of these techniques is the deep learning revolution that began in earnest when Krizhevsky et al. [9] showed that convolutional neural networks could dramatically outperform prior methods on large-scale image classification; LeCun et al. [10] later consolidated the broader shift toward representation learning that this kindled. Forensics has been a steady beneficiary, borrowing these advances for the visual and textual classification tasks that dominate casework.

In the identification and triage phase, classification and anomaly-detection models help investigators decide which devices and artifacts deserve attention first; Dunsin et al. [5] describe how this kind of automated prioritization is increasingly used to manage scale. During acquisition and recovery, pattern-recognition methods assist with carving and reconstructing fragmented data. The examination and analysis phase is where ML is most visible, covering image and video classification, text and chat analysis, and the clustering of related artifacts. Timeline reconstruction, the task

of assembling the chronological story of an event, is an area Dunsin et al. [5] and others see as especially promising, since correlating events across many sources is exactly the sort of pattern problem ML handles well.

Yet Fähndrich et al. [6] caution that most deployed systems remain narrow and task-specific, and that the field has not yet grappled seriously with what it would mean to trust a more autonomous system. The picture that emerges is of a discipline adopting AI enthusiastically at the level of individual tasks while leaving the harder, system-level question, namely whether the technology is actually aimed at the problems that most constrain a lab, for later.

The methods themselves cluster into a few recognizable families. Convolutional networks dominate the visual tasks, from nudity and weapon detection to document recognition [9]. Sequence and transformer models handle text, chat logs, and the increasingly important problem of language understanding in messaging applications. Clustering and graph-based methods, the latter highlighted by Bokolo and Liu [11] in the social-media context, help investigators surface the communities and relationships hidden in large artifact sets. No single family is a panacea; in practice forensic pipelines stitch several together, which is part of why end-to-end evaluation is so difficult and why the comparative numbers reported in tool benchmarks deserve careful scrutiny.

TABLE I
AI/ML Techniques Mapped to Digital Forensic Phases

Forensic phase	**Representative AI/ML techniques**	**Example sources**
Identification & triage	Classification, anomaly detection, automated prioritization	[5], [4]
Acquisition & recovery	Pattern recognition, data carving, fragment reconstruction	[5]
Examination & analysis	Image/video classification, NLP and chat analysis, clustering	[9], [7], [11]
Timeline reconstruction	Event correlation, sequence modelling across sources	[5], [4]
Reporting & presentation	Summarization, report drafting, explanation generation (XAI/LLMs)	[4], [12]

TABLE II
Comparative Overview of AI Techniques in Digital and Mobile Forensics

Technique	**Forensic Domain**	**Key Applications**	**Advantages**	**Limitations**
CNNs / Deep Learning	Image / Multimedia	Nudity detection, weapon identification, document classification, CSAM detection	High accuracy on large datasets; mature tooling; vendor integration	Class imbalance hides poor rare-class recall; adversarial vulnerability; opaque decisions
NLP / Transformers	Text / Chat Analysis	Chat log analysis, language identification, sentiment detection, social-media mining	Handles unstructured text; scalable to large corpora; cross-lingual capability	Context window limits; hallucination in LLMs; bias in training data
Graph Neural Networks	Social Network / Relationship	Community detection, link prediction, fake-profile identification	Captures relational structure; effective for network forensics	Requires large graph data; computationally expensive; limited tool support
Anomaly Detection	Triage / Network	Prioritisation of devices, log anomaly detection, timeline reconstruction	Reduces analyst workload; unsupervised variants need no labels	High false-positive rates; threshold tuning difficult; domain-specific
Large Language Models	General / Reporting	Report drafting, code generation, evidence summarisation, keyword extraction	General-purpose reasoning; rapid prototyping; natural language interfaces	Hallucination; non-determinism; data privacy risks; legal admissibility uncertain
Explainable AI (XAI)	Cross-cutting	Post-hoc explanations, model auditing, courtroom justification	Supports legal admissibility; builds practitioner trust; aids error detection	Explanation fidelity debated; adds computational overhead; standards immature
Multimodal AI	Emerging	Cross-modal evidence correlation, video-audio synchronisation, multi-source triage	Integrates heterogeneous evidence; richer context; aligns with real casework	Limited forensic benchmarks; very new; training data scarce

III. Practitioner Perspectives: What the Surveys Actually Show

Two recent surveys are usually cited together as evidence that AI has arrived in digital forensics. Read side by side, they tell a more awkward story. Vasilaras et al. [7] surveyed 37 mobile-forensics practitioners and reported that every one of them uses at least one tool with AI capabilities, an apparent saturation. DFPulse [3], surveying 122 practitioners more broadly, finds the reverse texture: of the 64 who answered the question, 17 use no AI at all, and those who do use it almost entirely for a single task, the categorisation of media.

The contradiction dissolves once you notice that the two surveys measure different things. Vasilaras measured whether practitioners' tools contain AI; DFPulse measured whether practitioners use AI to do work. Those are not the same question. Every current mobile-forensic suite ships image categorisation built in, so universal adoption in the first sense is close to automatic, reflecting what vendors chose to bundle rather than what investigators chose to deploy. The hundred-percent figure is a fact about procurement, not about practice.

That distinction matters because it relocates the real problem. The familiar reading of these surveys is that AI is useful but not yet accurate enough, and Vasilaras's respondents split along exactly that line, rating the tools useful far more often, around 59.5%, than highly accurate, around 21.6%, as shown in Figs. 2 and 3. But the more telling gap is not the one between usefulness and accuracy; it is the gap between buying software and solving problems.

The tools arrived, boxed into suites that practitioners already licensed. The relief did not. DFPulse makes the unmet need explicit: when respondents rated the pressures on their work, high workload, understaffing, and low budgets ranked as the most impactful by a wide margin (Fig. 4), and reported backlogs were severe, with almost a quarter of laboratories exceeding six months and one exceeding four years. Practitioners are not short of AI features; they are short of time, and the AI sitting in their toolbars is not yet touching the thing that hurts.

Seen this way, the two surveys are not really in tension. Saturation and scarcity describe the same situation from opposite ends: the supply of AI capability is effectively total, while its application to the work that actually constrains a lab, namely triage, data volume, and the backlog, remains marginal. Vasilaras et al.'s unanimous finding that no respondent could point to any standardised procedure governing AI use [7] fits the same pattern, since tooling outran practice and governance never caught up because deliberate adoption never really happened. Table III summarises the studies that recur across this review.

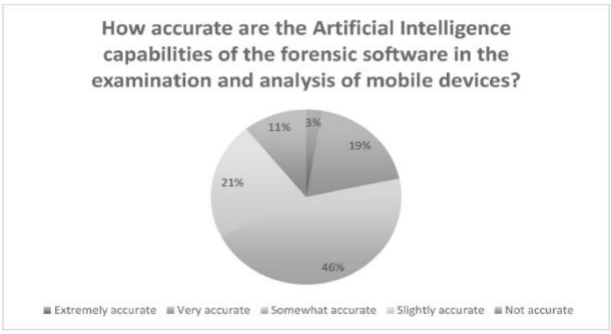


Fig. 2. Survey responses on the perceived accuracy of AI capabilities in mobile forensic software (redrawn from [7]).

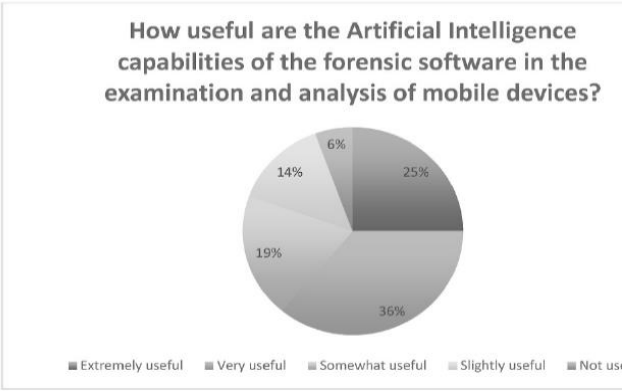


Fig. 3. Survey responses on the perceived usefulness of AI capabilities in mobile forensic software (redrawn from data reported in [7]).

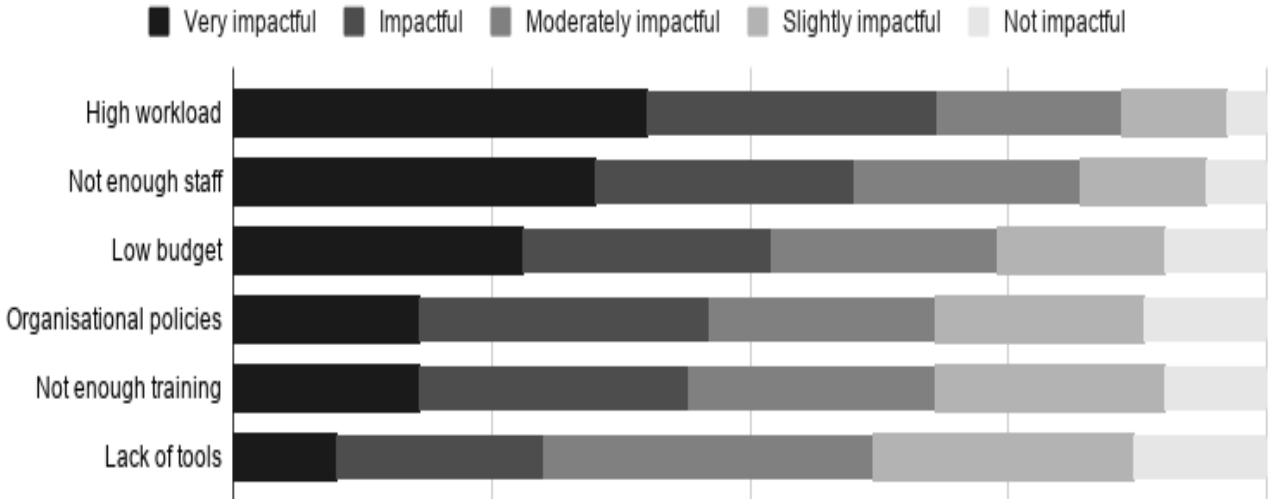


Fig. 4. Practitioner-rated impact of organisational challenges on digital forensic work (redrawn from [3]). High workload, understaffing, and limited budgets were identified as the most significant operational pressures.

TABLE III
Representative Studies on AI in Digital and Mobile Forensics

Study	Focus	Method / Scope	Key takeaway
Vasilaras et al. [7] (2024)	Practitioner survey + image-categorization benchmark	37 institutes/agencies; 3 commercial tools on an Android 12 image	High accuracy but low F1 on imbalanced classes; no standardised procedures
Hargreaves et al. [3] (2024)	Practitioner survey (DFPulse)	122 respondents; environments, techniques, challenges	Workload and research–practice gap dominate; appetite for automation
Dunsin et al. [5] (2024)	Analysis of AI/ML in DF and incident response	Cross-phase: acquisition, recovery, timelines, big data	AI/ML now spans the pipeline; chain-of-custody concerns persist
Fähndrich et al. [6] (2023)	Structured literature review	Strong-AI vs. narrow-ML framing of the field	Most work is narrow; autonomy and legal questions under-explored
Wickramasekara et al. [4] (2025)	Survey of LLMs in DF	176 articles; LLMs across each forensic phase	Promising but constrained by hallucination, bias, and data scarcity
Bokolo & Liu [11] (2024)	Survey of AI in social-media forensics	NLP, graph neural networks, GAN-based content	AI scales analysis but raises integrity and deepfake challenges
Ketsekioulafis et al. [8] (2024)	PRISMA systematic review (forensic science)	Human ID, pathology, and related domains	Gains in automation; validation, bias, and admissibility unresolved
Chernyshev et al. [29] (2026)	Systematic review of LLMs in digital forensics	33 peer-reviewed works mapped to DFRWS phases	Three strategic integration points identified; 85–98% accuracy across tasks; probabilistic–deterministic tension central

IV. Image and Multimedia Classification

Nowhere is the accuracy paradox sharper than in image categorization, which is also one of the most common forensic uses of AI. The benchmark reported by Vasilaras et al. [7] is instructive: across three leading tools, headline accuracy on an Android 12 reference image routinely exceeded 0.98, yet the F1-scores collapsed for several forensically critical categories, with weapons and currency in particular performing very poorly.

This pattern is laid bare in the per-category metrics of Fig. 5, where near-perfect accuracy sits beside F1-scores that fall close to zero for the rarest classes. The culprit is class imbalance. When the vast majority of images are true negatives, a model can score brilliantly on accuracy while missing most of the items an investigator actually cares about. This is not a quirk of one study; it is a structural feature of forensic image sets, and it means accuracy alone is close to meaningless as a measure of forensic usefulness.

The point bites harder still because image and media categorisation is, by some margin, the one task for which practitioners actually report using AI today [3]: the capability they lean on most is precisely the one whose headline numbers are most misleading.

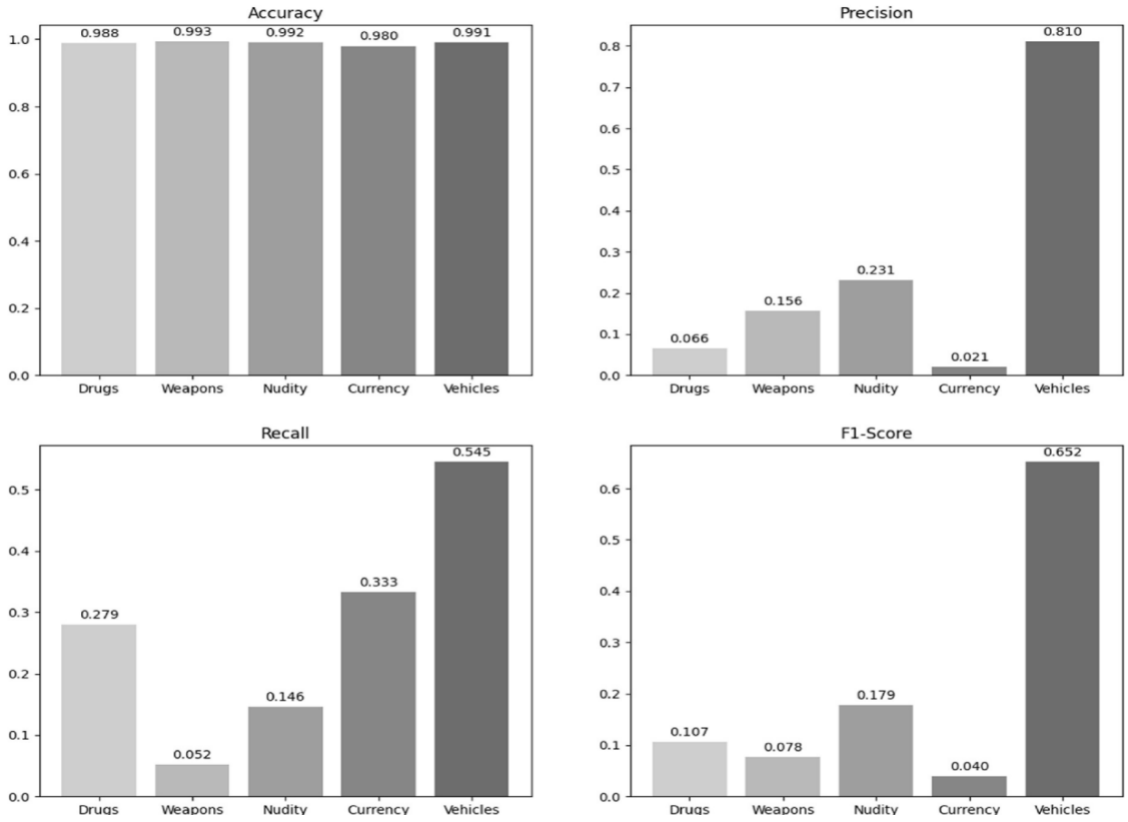


Fig. 5. Per-category evaluation of a leading mobile forensic tool (Magnet Axiom): accuracy, precision, recall, and F1-score, illustrating the accuracy–F1 gap on imbalanced classes (redrawn from data reported in [7]).

The underlying classifiers are typically deep convolutional networks of the kind popularized by Krizhevsky et al. [9], which

excel at the visual recognition that powers categories such as nudity, weapons, and documents. The multimedia challenge is also widening, not narrowing. Bokolo and Liu [11], reviewing AI in social-media forensics, describe a landscape in which natural-language processing, graph neural networks, and generative models are all being pressed into service to detect bullying, extremism, and deceptive profiles. The same generative techniques that aid investigators also produce the deepfakes and synthetic media that forensic tools must now learn to detect, an arms race that puts a premium on robustness and on careful validation against realistic data.

Some applications raise the stakes higher still. The detection of child sexual abuse material is among the most consequential uses of forensic image classification, and it concentrates every tension discussed in this review: the cost of a false negative is enormous, the cost of a false positive is a wrongful accusation, and the imbalance between the rare target class and the overwhelming background is extreme. It is precisely in such settings that the headline-accuracy illusion is most dangerous and that calibrated, recall-aware evaluation matters most [7]. The lesson generalizes: the higher the stakes of a category, the less tolerance there is for a metric that hides the model's true behaviour.

V. Interpretability and Explainable AI

If there is one requirement the literature returns to again and again, it is explainability, for a blunt practical reason: a conclusion that cannot be explained is hard to defend in court. Lipton [13] gives the most cited framework, distinguishing transparency at the level of the whole model, its components, and its training algorithm, and noting that interpretability is less a single property than a family of them. Rudin [14] pushes the argument further, contending that for high-stakes decisions we should prefer inherently interpretable models over post-hoc explanations of black boxes, a position with obvious appeal in a forensic setting even if it collides with the accuracy of deep networks.

Where inherently interpretable models are not practical, the field leans on post-hoc explanation methods. The local surrogate approach of Ribeiro et al. [15], the additive feature-attribution of Lundberg and Lee [16], and the gradient-based saliency maps of Selvaraju et al. [17] have all migrated from mainstream machine learning into forensic tooling, making individual decisions at least partially inspectable. Within forensics specifically, Solanke [12] argues that explainable digital-forensics AI is essential for building trust in AI-assisted investigations and proposes a framework for designing models that confront opacity, bias, and uncertainty directly. The recurring theme is that explainability is not a feature to be added at the end; in a domain where evidence must survive cross-examination, it has to be a design objective from the outset.

It is worth distinguishing the kinds of transparency on offer, since they are not interchangeable. Inspection methods open up a trained model after the fact; training-time methods build interpretability into the learning process; and architectural approaches, favored by Rudin [14], design the model to be interpretable from the start. Each buys clarity at a different price, and the right choice in a forensic setting depends on whether the priority is auditing a tool already in use or commissioning a new one. What unites them is the recognition, increasingly explicit in the literature, that an unexplained output is a liability the moment it enters a courtroom [12].

VI. Robustness, Adversarial Threats, and Reliability

Robustness, the ability of a model to keep performing when conditions change, is the second non-negotiable. Szegedy et al. [18] first revealed that neural networks could be fooled by imperceptible perturbations, and Goodfellow et al. [19] showed how easily such adversarial examples can be generated. In a forensic context where evidence integrity is paramount, that vulnerability is more than academic, since a manipulated image could in principle steer a classifier toward the wrong conclusion without a human noticing. The standard reference text by Goodfellow et al. [20] situates these phenomena within the broader theory of deep learning.

Forensic data are also messy in less adversarial ways: they vary wildly in quality, are often scarce, and frequently look nothing like the data a model was trained on, all of which strain generalization. The inconsistency documented across tools in the Vasilaras benchmark [7], where each tool stumbled on different categories, is itself a robustness signal: these systems do not fail gracefully or predictably. Strategies the literature recommends include adversarial training, diversified and representative datasets, rigorous cross-validation, and continuous auditing. A complementary line of work, rooted in the Bayesian decision framework of Taroni et al. [21], argues for making the differing costs of false positives and false negatives explicit rather than hiding them inside a fixed default threshold, a stance especially well suited to a justice system that must weigh those costs openly.

VII. Large Language Models: A Fast-Moving Frontier

The newest and most volatile strand of the literature concerns large language models. The early, careful assessment by Scanlon et al. [22] of ChatGPT for forensic tasks set the tone, concluding that the model could act as a useful assistant for code generation and artifact understanding but only in the hands of an expert able to catch its mistakes, and that uploading evidence to a third-party service raised obvious problems. Wickramasekara et al. [4], in a survey spanning 176 articles, later mapped LLMs onto every phase of the forensic process, and their related conference work [23] examined where the genuine opportunities lie.

Complementing these broad surveys, Chernyshev et al. [29] conducted a systematic review of 33 peer-reviewed works specifically examining LLM capabilities across the DFRWS process model. Their analysis identifies three strategic integration points where LLMs demonstrate measurable forensic utility: pattern recognition during the examination phase, evidence analysis during the analysis phase, and evidence presentation during the reporting phase.

Performance reported in the reviewed literature is impressive on headline metrics, LLMs have achieved 85–98% accuracy across diverse forensic tasks, including 94.6% precision in automated log anomaly detection and 98% classification accuracy in specialised evidence extraction. The range of applications is also widening rapidly beyond general-purpose text analysis. Purpose-built frameworks now address memory forensic triage (volGPT), mobile device analysis (Thumb), cloud log investigation, password cracking (PassGPT), vehicle infotainment forensics, and threat hunting automation (Thelma) [29]. It should be noted that these figures are reported ranges across heterogeneous benchmarks and task definitions rather than directly comparable metrics.

What unites these diverse applications is the same underlying

tension: the probabilistic nature of LLM output generation, controlled by parameters such as temperature, introduces non-determinism where identical inputs may produce varying outputs across multiple invocations, a characteristic that fundamentally conflicts with the reproducibility requirements essential for forensic method acceptance and legal admissibility [29].

Michelet et al. [24] take the next step, offering recommendations for fine-tuning LLMs to specific forensic tasks and demonstrating the idea on chat summarization. The cautions are consistent across all of these: hallucination, bias inherited from training data, censorship, the cost of infrastructure, and unresolved legal questions surface repeatedly [4], [22]. The broader DFRWS retrospective by Breitinger et al. [25] situates this enthusiasm within a longer arc of forensic research, a reminder that each wave of automation has arrived with similar promise and similar caveats. The sensible reading is that LLMs are a powerful assistant whose output must be verified, not an oracle whose conclusions can be trusted unchecked, an especially important distinction when those conclusions may become evidence.

The concrete applications already being explored give a sense of both the appeal and the danger. LLMs have been used to draft incident narratives, generate keyword lists, write extraction scripts, and translate technical findings into plain language for non-specialist audiences such as judges and juries [22], [4]. Each of these is genuinely useful, and each carries the same risk: a fluent, confident output that is subtly wrong. In a domain where the reader may lack the expertise to detect the error, that fluency is not a neutral feature but a hazard, which is exactly why every source in this strand insists on expert verification and on keeping the model firmly in an assistive rather than a decisive role.

A. Multimodal AI: An Emerging Frontier

A comparatively new direction in forensic AI research involves multimodal models that process and correlate evidence across multiple data types simultaneously. Unlike unimodal systems that analyse images, text, or audio in isolation, multimodal architectures can fuse information from heterogeneous sources such as video footage, audio transcripts, geolocation logs, and messaging records to construct a more comprehensive evidential narrative. Recent advances in vision-language models (VLMs) have demonstrated the ability to answer natural-language questions about image content, a capability with clear forensic relevance for expediting multimedia review.

From a forensic standpoint, the principal attraction of multimodal AI is its alignment with the reality of modern investigations, where evidence rarely arrives in a single format. A typical mobile-device extraction yields photographs, chat messages, application databases, audio recordings, and location history simultaneously. Multimodal systems that can cross-reference these channels, for example, linking a photograph's timestamp to a GPS entry and a contemporaneous text message offer the potential to automate timeline reconstruction in ways that single-modality tools cannot. Early work in this space has explored cross-modal retrieval for video forensics, automated scene understanding in CCTV analysis, and the fusion of facial recognition with voice biometrics for identity verification.

Nevertheless, the challenges are substantial. Benchmark datasets for forensic multimodal tasks are virtually nonexistent, meaning that most published results derive from general-purpose corpora that do not reflect the noise, compression, and fragmentation typical of forensic data. Furthermore, the increased complexity of multimodal architectures amplifies every concern discussed in the preceding sections: interpretability becomes harder when decisions depend on interactions across modalities, robustness testing must cover combinatorial perturbations, and the computational demands of training and inference may exceed the resources available to many forensic laboratories. For the present, multimodal AI in forensics remains a research frontier rather than a deployable capability, but its potential to unify the disparate strands of digital evidence analysis makes it a priority direction for future investigation.

VIII. Governance, Legal Admissibility, and Alignment

Cutting across every theme above is a governance problem that the surveys expose bluntly: there are, as yet, no agreed procedures or legal frameworks for AI in mobile forensics, a finding on which the Vasilaras respondents were unanimous [7]. That vacuum is the predictable consequence of capability that arrived faster than any decision about how to use it.

Regulation has now begun to catch up. The European Union's risk-based AI Act, adopted in 2024 and in force since August of that year [26], is the first binding, comprehensive framework of its kind, and its classification of certain law-enforcement and biometric systems as high-risk applies directly to forensic tools whose outputs can deprive someone of liberty.

The EU AI Act is not the only regulatory development relevant to forensic AI. In the United States, the National Institute of Standards and Technology (NIST) has published the AI Risk Management Framework, which, although non-binding, provides detailed guidance on assessing and mitigating risks in AI systems used in high-stakes contexts, including law enforcement. Several U.S. states have introduced legislation requiring algorithmic impact assessments for AI tools used in criminal proceedings, though no federal statute yet matches the comprehensiveness of the EU approach. The United Kingdom has adopted a principles-based, sector-led regulatory framework that delegates oversight to existing regulators rather than creating a dedicated AI agency; for forensic applications, this means that the Forensic Science Regulator and the Information Commissioner's Office share responsibility, a division that practitioners may find difficult to navigate. In the Asia-Pacific region, Singapore's Model AI Governance Framework emphasises transparency and human oversight, while Australia's Attorney-General's Department has issued discussion papers on AI in the justice system that signal forthcoming regulatory attention. China's algorithmic recommendation regulations and draft AI law focus on content moderation and social control, with less direct relevance to forensic science but potentially significant implications for cross-border evidence sharing. These divergent approaches create a complex compliance landscape for forensic tool vendors and laboratories that operate internationally, and they highlight the need for harmonised standards that preserve both innovation and due process.

Underneath the regulatory question sits a deeper one about alignment. Russell [27] frames the core challenge as ensuring that capable systems act in accordance with human values, and Gabriel

[28] separates the technical problem of encoding values from the normative problem of choosing them. In forensics these abstractions become concrete: a misaligned or opaque model does not merely make an error, it can distort the evidentiary record. The Bayesian, cost-aware decision framing of Taroni et al. [21] offers one practical bridge, making explicit the trade-offs that a justice system has a legitimate interest in scrutinizing. Aligning tools, procedures, and competencies with this body of research is, in the end, what will let the field adopt AI responsibly rather than recklessly.

There is also a quieter, practical dimension to governance that the literature increasingly stresses: data protection and the integrity of the evidentiary chain. Feeding case data to a third-party model, as Scanlon et al. [22] note, may itself breach the confidentiality that forensic work demands, and any AI step inserted into the pipeline must preserve a documented, auditable chain of custody. Regulation such as the AI Act [26], data-protection law, and established forensic standards will have to be read together rather than in isolation, and the sooner the forensic community helps shape that synthesis, the better the outcome is likely to be for both investigators and the people their work affects.

Chernyshev et al. [29] sharpen these governance concerns into a taxonomy of seven interrelated challenge categories. Technical constraints include the fundamental conflict between probabilistic LLM outputs and deterministic forensic requirements, alongside context window limitations and hallucination risks. Methodological gaps encompass the lack of standardised evaluation frameworks and the predominant reliance on synthetic datasets. Forensic integrity concerns centre on black-box decision-making and reproducibility failures, while data governance and security risks arise from cloud-based processing of sensitive evidence. Operational deployment barriers include prompt engineering skill gaps and substantial computational demands. Domain-specific limitations range from knowledge cutoff dates to multilingual performance degradation, and systemic bias and fairness issues may disproportionately affect certain demographic groups [29].

Their analysis highlights that chain-of-custody maintenance in LLM-assisted investigations requires specialised extensions to traditional protocols, encompassing not only evidence artefacts but also the specific LLM invocations, prompts, parameters, and outputs that contributed to analytical conclusions. Bias concerns are particularly acute: language-specific model training significantly impacts forensic analysis performance, with accuracy degrading to 58.7% when processing mixed-language content, potentially introducing systematic biases that disproportionately affect certain demographic groups [29].

IX. Research Gaps and Future Directions

Reading these strands together, a consistent set of gaps emerges. They are worth stating plainly:

Open, shared benchmarks. Evaluations remain fragmented and tool-specific; the field lacks standardized, openly available forensic datasets and metrics that would make results comparable [4], [7]. Metrics that respect imbalance. Accuracy continues to flatter performance; F1-score, recall on rare classes, and cost-sensitive measures should be reported as standard [7], [21]. Explainability for admissibility. Interpretable-by-design models and forensic-grade explanation methods are needed so that AI-assisted conclusions can withstand cross-examination [14], [12]. Adversarial robustness. Tools must be hardened against manipulated inputs and validated on realistic, diverse data rather than curated benchmarks [18], [19].

Standardised procedures and legal frameworks. The unanimous absence of governing procedures reported in practitioner surveys is an urgent gap for standards bodies and policymakers [3], [7].

Reliable, auditable LLM use. As LLMs enter casework, the field needs verification protocols, provenance tracking, and clear limits on their evidentiary weight [4], [22].

Forensic-ready LLM architectures. General-purpose models retrofitted for forensic tasks through prompt engineering alone are insufficient. The field needs purpose-built architectures that embed audit capabilities, transparent reasoning mechanisms, and minimally probabilistic operation modes supporting reproducible outputs across multiple invocations of identical inputs [29][30]. Standardised validation frameworks and real-world benchmarking. Current evaluations rely predominantly on synthetic datasets and inconsistent metrics that preclude cross-study comparison. Forensic-specific performance metrics, comprehensive benchmark datasets capturing the complexity of real investigations, and reliability assessment methodologies quantifying consistency across multiple invocations are needed [29][31].

X. Conclusion

Taken together, the studies in this review point to a single diagnosis. Artificial intelligence has spread across digital and mobile forensics less through deliberate adoption than through procurement: the capability arrived bundled in the tools that practitioners already license, while the problems that actually constrain a forensic laboratory, namely the volume of evidence, the backlog, and the shortage of time and staff, remain largely untouched by it. The field, in short, bought AI before it decided what problem it was buying it to solve. The work ahead is therefore less about better models than about that decision: identifying the tasks where automation genuinely relieves the burden on practitioners, and then demanding the benchmarks, the explanations, the robustness, and the governance that those particular tasks require. Forensic AI will earn its place not when it is merely capable, but when it is aimed, deliberately, at the problems that matter. For practitioners, we offer the following practical recommendations. First, treat vendor accuracy claims with skepticism, demand per-class metrics and validation on data that matches your caseload, not headline figures from balanced benchmarks. Second, insist on explainability before procurement: any AI tool you consider should provide human-interpretable outputs that can withstand courtroom scrutiny. Third, establish internal protocols for AI use now, even if external standards do not yet exist: document which tools were used, how their outputs were verified, and what thresholds triggered human review. Fourth, avoid uploading sensitive evidence to third-party cloud services until the legal and data-protection implications are fully understood. Fifth, invest in professional development: the skills required to supervise AI tools effectively are different from those needed for traditional forensic analysis, and the gap will only widen as the technology advances. These recommendations are interim measures; the ultimate goal is a forensic ecosystem in which AI tools are not merely available but genuinely fit for the

purpose of justice.

Funding Statement

The authors received no specific funding for this study.

Conflicts of Interest

The authors declare no conflicts of interest to report regarding the present study.

Author Contributions

All authors contributed to the conception, literature review, drafting, and critical revision of this manuscript and approved the final version for submission.

Data Availability Statement

Data is available on reasonable request.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

References

[1] S. L. Garfinkel, "Digital forensics research: The next 10 years," Digit. Invest., vol. 7, pp. S64–S73, 2010.

[2] D. Quick and K.-K. R. Choo, "Impacts of increasing volume of digital forensic data: A survey and future research challenges," Digit. Invest., vol. 11, no. 4, pp. 273–294, 2014.

[3] C. Hargreaves, F. Breitinger, L. Dowthwaite, H. Webb, and M. Scanlon, "DFPulse: The 2024 digital forensic practitioner survey," Forensic Sci. Int.: Digit. Invest., vol. 51, art. 301844, 2024.

[4] A. Wickramasekara, F. Breitinger, and M. Scanlon, "Exploring the potential of large language models for improving digital forensic investigation efficiency," Forensic Sci. Int.: Digit. Invest., vol. 52, art. 301859, 2025.

[5] D. Dunsin, M. C. Ghanem, K. Ouazzane, and V. Vassilev, "A comprehensive analysis of the role of artificial intelligence and machine learning in modern digital forensics and incident response," Forensic Sci. Int.: Digit. Invest., vol. 48, art. 301675, 2024.

[6] J. Fähndrich, W. Honekamp, R. Povalej, H. Rittelmeier, S. Berner, and D. Labudde, "Digital forensics and strong AI: A structured literature review," Forensic Sci. Int.: Digit. Invest., vol. 45, art. 301617, 2023.

[7] A. Vasilaras, N. Papadoudis, and P. Rizomiliotis, "Artificial intelligence in mobile forensics: A survey of current status, a use case analysis and AI alignment objectives," Forensic Sci. Int.: Digit. Invest., vol. 49, art. 301737, 2024.

[8] I. Ketsekioulafis, G. Filandrianos, K. Katsos, K. Thomas, C. Spiliopoulou, and E. I. Sakelliadis, "Artificial intelligence in forensic sciences: A systematic review of past and current applications and future perspectives," Cureus, vol. 16, no. 9, art. e70363, 2024.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2012, pp. 1097–1105.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[11] B. G. Bokolo and Q. Liu, "Artificial intelligence in social media forensics: A comprehensive survey and analysis," Electronics, vol. 13, no. 9, art. 1671, 2024.

[12] A. A. Solanke, "Explainable digital forensics AI: Towards mitigating distrust in AI-based digital forensics analysis using interpretable models," Forensic Sci. Int.: Digit. Invest., vol. 42, art. 301403, 2022.

[13] Z. C. Lipton, "The mythos of model interpretability," Commun. ACM, vol. 61, no. 10, pp. 36–43, 2018.

[14] C. Rudin, "Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, no. 5, pp. 206–215, 2019.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. ACM SIGKDD, 2016, pp. 1135–1144.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4765–4774.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.

[18] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in Proc. Int. Conf. Learn. Represent. (ICLR), 2014.

[19] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[20] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[21] F. Taroni, A. Biedermann, S. Bozza, P. Garbolino, and C. Aitken, Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science. Chichester, U.K.: Wiley, 2014.

[22] M. Scanlon, F. Breitinger, C. Hargreaves, J.-N. Hilgert, and J. Sheppard, "ChatGPT for digital forensic investigation: The good, the bad, and the unknown," Forensic Sci. Int.: Digit. Invest., vol. 46, art. 301609, 2023.

[23] A. Wickramasekara, F. Breitinger, and M. Scanlon, "Where is the potential for large language models in digital forensic investigations?" in Proc. Digit. Forensics Res. Workshop Europe (DFRWS EU), 2024.

[24] G. Michelet, H. Henseler, H. van Beek, M. Scanlon, and F. Breitinger, "Fine-tuning large language models for digital forensics: Case study and general recommendations," Digit. Threats: Res. Pract., 2025.

[25] F. Breitinger, J.-N. Hilgert, C. Hargreaves, J. Sheppard, R. Overdorf, and M. Scanlon, "DFRWS EU 10-year review and future directions in digital forensic research," Forensic Sci. Int.: Digit. Invest., vol. 48, art. 301685, 2024.

[26] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, L series, 12 July 2024.

[27] S. Russell, Human Compatible: Artificial Intelligence and the Problem of Control. New York, NY, USA: Viking, 2019.

[28] I. Gabriel, "Artificial intelligence, values, and alignment," Minds Mach., vol. 30, no. 4, pp. 411–437, 2020.

[29] M. Chernyshev, Z. Baig, N. Syed, R. Doss, and M. Shore, "Large language models in digital forensics: Capabilities, challenges and future directions," Forensic Sci. Int.: Digit. Invest., vol. 56, art. 302043, 2026.

[30] Sumra,I.A, I. Ahmad, H. Hasbullah and J. -l. bin Ab Manan, "Behavior of attacker and some new possible attacks in Vehicular Ad hoc Network (VANET)," *2011 3rd International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, Budapest, Hungary, 2011, pp. 1-8.

[31] Sumra, I.A., Hasbullah, H.B., AbManan, Jl.B. (2015). Attacks on Security Goals (Confidentiality, Integrity, Availability) in VANET: A Survey. In: Laouiti, A., Qayyum, A., Mohamad Saad, M. (eds) Vehicular Ad-hoc Networks for Smart Cities. Advances in Intelligent Systems and Computing, vol 306. Springer.