REVİEW/DERLEME

DOI: 10.17986/blm.1807
Bull Leg Med 2026;31(2):233-242

Artificial Intelligence Applications in Forensic Sciences: A Systematic Review of Methods, Performance, Validation, and Medico-Legal Implications

Melek Şenışık[1]

[1] Istanbul Topkapi University, Plato Vocational School, Department of Medical Services and Techniques, Istanbul, Türkiye

Abstract

Artificial Intelligence Applications in Forensic Sciences: A Systematic Review of Methods, Performance, Validation, and Medico-Legal Implications

Objective: Artificial intelligence (AI) and machine learning (ML) are increasingly used in forensic sciences to analyze large-scale and heterogeneous evidence, providing improved speed, consistency, and decision support. This review aims to categorize current AI/ML application domains in forensic practice, summarize methodological trends, and discuss key validation requirements for operational and court-defensible use. The scope covers forensic image analysis and computer vision (traces, injuries, and crime-scene imagery), biometric identification, AI-assisted interpretation of complex forensic genetic data (including mixtures and low-template DNA), AI-enabled triage and authenticity assessment in digital/cyber forensics, and segmentation/classification tasks in forensic pathology and postmortem imaging. Despite promising performance in research settings, critical barriers remain, including explainability, data representativeness and bias, cross-laboratory reproducibility, robustness against manipulation, privacy-preserving governance, and legal admissibility. The review highlights the need for independent validation, standardized performance reporting, and clearly defined human oversight to ensure responsible and defensible integration of AI into forensic workflows.
Keywords: Forensic science; Artificial intelligence; Machine learning; Forensic genetics; Digital forensics; Forensic pathology.

Özet

Adli Bilimlerde Yapay Zekâ Uygulamaları: Yöntemler, Performans, Doğrulama ve Medikolegal Etkiler Üzerine Sistematik Derleme

Yapay zekâ (YZ) ve makine öğrenmesi (MÖ), adli bilimlerde büyük hacimli ve heterojen delil verilerinin analizinde hız, tutarlılık ve karar desteği sağlayarak giderek daha fazla kullanılmaktadır. Bu derlemenin amacı, adli bilimlerde YZ/MÖ'nün güncel uygulama alanlarını sınıflandırmak, yöntemsel eğilimleri özetlemek ve sahaya/muhakemeye entegrasyon için kritik doğrulama gereksinimlerini tartışmaktır. Kapsam; adli görüntü analizi ve bilgisayarlı görme (iz, yara, olay yeri görüntüsü), biyometrik tanıma, adli genetikte karışım ve düşük miktarlı DNA yorumlama, dijital/siber adli tıpta delil triyajı ve içerik özgünlüğü analizi, adli patoloji ve postmortem görüntülemede segmentasyon ve sınıflandırma uygulamalarını içermektedir. Potansiyel kazanımlara rağmen açıklanabilirlik, veri temsiliyeti ve yanlılık, laboratuvarlar arası tekrarlanabilirlik, manipülasyona dayanıklılık, mahremiyet ve mahkemede kabul edilebilirlik gibi sınırlılıklar sürmektedir. Derleme, YZ çıktılarının insan uzman denetimi altında, standart performans raporlaması ve bağımsız validasyonla desteklenmesi gerektiğini vurgulayarak sorumlu ve savunulabilir kullanım için bir çerçeve sunar.
Anahtar Kelimeler: Adli bilimler, Yapay zekâ, Makine öğrenmesi, Adli genetik, Dijital adli tıp; Adli patoloji

Cite this article: Senisik M. Artificial Intelligence Applications in Forensic Sciences: A Systematic Review of Methods, Performance, Validation, and Medico-Legal Implications. Bull Leg Med. 2026;31(2):233-242. https://doi.org/10.17986/blm.1807

Corresponding Author: Melek Şenışık: Dr., Istanbul Topkapi University, Plato Vocational School, Department of Medical Services and Techniques, Istanbul, Türkiye
Email: meleksenisik@topkapi.edu.tr **ORCID iD:** 0000-0002-6093-3174

Received: 07.02.2026
Accepted: 03.04.2026

INTRODUCTION

Forensic sciences represent an interdisciplinary field in which biological, physical, chemical, and digital evidence is scientifically analyzed to support criminal investigations and judicial processes. Traditional forensic examination procedures rely largely on expert experience and manual interpretation. However, the increasing volume of data, complex evidence structures, and time constraints have challenged the sustainability and efficiency of these conventional approaches. In particular, data-intensive sources such as high-resolution images, genomic datasets, multi-contributor DNA mixtures, and large-scale digital log records may lead to prolonged analysis times and increased interpretative variability when evaluated using classical methods (1,2,3).

Recent advances in artificial intelligence (AI) and machine learning (ML) have introduced a new paradigm in forensic sciences by enabling automated pattern extraction from large and heterogeneous datasets. Deep learning architectures have demonstrated strong performance in image classification, feature extraction, matching, and segmentation tasks relevant to forensic workflows (4,5,6). Consequently, AI systems are increasingly employed in applications such as fingerprint matching, facial recognition, injury pattern analysis, and automated interpretation of crime-scene imagery (4,5,7).

Despite the performance gains observed in biometric systems, challenges such as algorithmic bias and false match rates pose significant risks in forensic contexts. Demographic variations may influence model performance, potentially compromising fairness and equity in judicial decision-making processes (5,8,9). Therefore, AI-based forensic tools must be evaluated not only in terms of technical accuracy but also with regard to ethical and legal implications.

In forensic genetics, probabilistic genotyping and ML-based models have substantially improved the interpretation of complex, low-template, and mixed DNA samples. These approaches reduce subjectivity associated with expert judgment and provide statistically more consistent and reproducible outcomes (2,10,11,12). Similarly, in digital forensics, ML techniques play a critical role in automated evidence triage, anomaly detection, and identification of manipulated or synthetic media within large-scale datasets (3,6,13,14,15).

AI applications are also expanding within forensic pathology and postmortem imaging. Deep learning models have been investigated for wound classification, organ segmentation, volumetry, and rib fracture detection, thereby serving as decision-support tools for forensic practitioners (7,16,17,18,19). Nevertheless, issues related to limited explainability, cross-laboratory reproducibility, data privacy, and legal admissibility of AI-generated outputs remain subjects of ongoing debate (20,21,22).

Although the number of publications addressing AI in forensic science has steadily increased, integrative reviews that collectively evaluate all major forensic subdisciplines while simultaneously discussing methodological validation requirements and legal frameworks remain relatively limited. Accordingly, there is a clear need to systematically classify current AI applications and define standards for reliable and defensible forensic integration.

The aim of this review is therefore to evaluate the current state of artificial intelligence applications across forensic image analysis, biometrics, forensic genetics, digital forensics, and forensic pathology; to compare their performance benefits and limitations; and to provide evidence-based recommendations for safe, transparent, and legally robust implementation in forensic practice. A conceptual workflow of AI-supported forensic laboratory processes is illustrated in Figure 1.

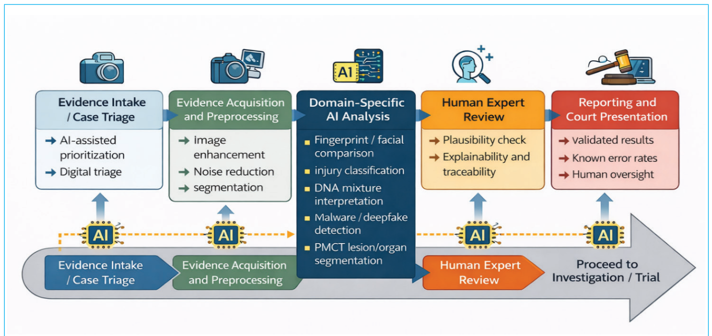

Figure 1. Conceptual workflow of AI-supported forensic laboratory processes.

METHOD

This study was designed as a narrative review informed by a systematic literature search to comprehensively evaluate artificial intelligence (AI) and machine learning (ML) applications in forensic sciences. The literature search covered studies published between 1 January 2010 and 1 February 2026, and the last search was performed on 1 February 2026.

Ethical approval was not required because this study is based exclusively on previously published literature and does not involve human participants, biological samples, animals, or identifiable personal data. The manuscript was prepared in accordance with the principles of the Declaration of Helsinki where applicable.

Electronic databases including PubMed/MEDLINE, Scopus, Web of Science, IEEE Xplore, and Google Scholar were searched. The search strategy combined free-text keywords related to forensic science and artificial intelligence. The main keywords included “forensic science,” “artificial intelligence,” “machine learning,” “deep learning,” “forensic genetics,” "digital forensics," "forensic pathology," "computer vision," "probabilistic genotyping," and "decision support systems," which were used in various Boolean combinations (AND/OR) according to the requirements of each database. An example search string was as follows: ("forensic science" OR "forensic pathology" OR "digital forensics" OR "forensic genetics") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR "computer vision").

A total of 7,463 records were identified through database searching, and 47 additional papers were identified through other sources, including reference mining and expert input, yielding 7,510 records for initial screening. Duplicate records were identified and removed manually by cross-checking titles, author names, publication years, journal details, and DOI information where available. No duplicate records were identified after manual cross-checking. The screening process was performed by a single reviewer. First, titles were screened for relevance, and 7,340 records were excluded at this stage. The remaining 170 records underwent abstract screening, after which 126 records were excluded. A total of 44 full-text publications were assessed for eligibility. Of these, 12 publications were excluded, including 10 due to mismatch with the research question or study design and 2 because the full text was not available. Ultimately, 37 publications were included in the final qualitative synthesis. The literature selection and screening process is summarized in Figure 2.

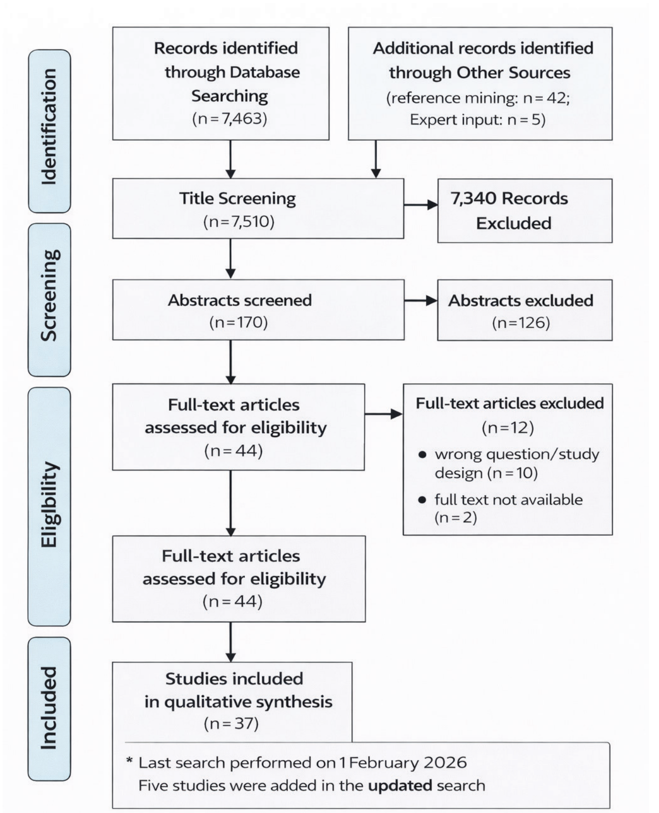

Figure 2. PRISMA-style literature selection flow diagram of the review process.

The inclusion criteria were as follows: studies reporting AI or ML applications in forensic science; peer-reviewed original research articles, review articles, or methodological studies; and publications in English or Turkish. The exclusion criteria were purely theoretical or technical studies without a clear forensic application, duplicated datasets or conference abstracts with limited methodological or outcome data, and editorials or opinion papers.

Selected studies were categorized using thematic content analysis into the following domains: forensic image analysis and biometrics, forensic genetics and DNA interpretation, digital and cyber forensics, forensic pathology and postmortem imaging, and ethical and legal considerations.

For each included study, the following data were extracted and comparatively evaluated: algorithm type, dataset characteristics, forensic application area, performance metrics, validation approach, and reported limitations. This approach aimed to provide a comprehensive overview of current trends, validation practices, and the current level of evidence in the field.

Because of the heterogeneity of study designs, datasets, algorithms, and reported outcome measures, a quantitative meta-analysis was not performed.

Quality Assessment / Risk of Bias

Given the broad methodological diversity of the included studies and the narrative nature of this review, a formal quantitative risk-of-bias tool was not applied. Instead, methodological rigor was considered descriptively during data extraction by taking into account factors such as dataset size, validation strategy, reporting of performance metrics, and stated study limitations.

Statistical Analysis

This review did not perform statistical hypothesis testing or meta-analysis; findings were synthesized descriptively from the included literature.

RESULTS

An overview of AI applications across major forensic subdisciplines and common algorithmic approaches is provided in Table 1. Representative and methodologically relevant studies discussed in this review are summarized in Table 2. An overview of AI application domains in forensic sciences is shown in Figure 3.

Forensic Image Analysis and Biometrics

AI in forensic imaging typically targets three task families: (i) enhancement and preprocessing (noise/illumination/latent ridge quality), (ii) feature learning and matching (fingerprints/traces), and (iii) detection/classification (injury patterns, evidentiary objects). For latent fingerprint matching, classical minutiae-based pipelines are challenged by small print area, distortion, and background noise; therefore, systems that incorporate extended ridge descriptors have been proposed to improve robustness (1). On the deep learning side, convolutional neural network (CNN)-based feature learning and virtual-minutiae concepts have been investigated to strengthen matching under partial and degraded conditions (4). End-to-end latent fingerprint identification frameworks increasingly integrate enhancement, minutiae extraction, and matching into a unified pipeline to automate workflows similar to an automated fingerprint identification system (AFIS) (4).

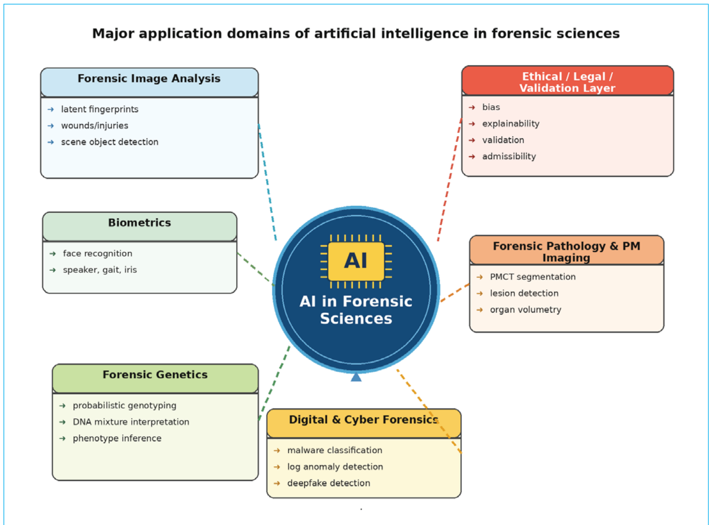

Figure 3. Overview of AI application domains in forensic sciences.

For facial recognition, the most forensic-relevant experimental evidence concerns demographic variability. National Institute of Standards and Technology (NIST) Face Recognition Vendor Test (FRVT) Part 3 reports that false positive rates can differ by one to two orders of magnitude across demographic groups for many contemporary algorithms, underscoring the need for subgroup-aware validation and threshold governance in operational settings (5). Earlier NIST work also documented an other-race effect for face recognition algorithms, reinforcing the need to evaluate subgroup error profiles rather than relying solely on average accuracy (9). Consequently, forensic deployment requires scenario-specific testing, threshold selection, and subgroup-aware performance reporting (5,8,9).

Forensic Genetics and DNA Interpretation

In forensic genetics, AI/ML impact is most visible in probabilistic inference and complex mixture interpretation. Probabilistic genotyping systems such as STRmix, EuroForMix, and MaSTR model stochastic effects, contributor proportions, and likelihood ratios, thereby improving the interpretation of complex DNA mixtures (2,10,11,12).

Table 1. Artificial intelligence applications across forensic subdisciplines and corresponding algorithm types.

Forensic Subdiscipline	Typical Task / Application	Common AI/ML Algorithms	Data Type	Reported Performance (literature range)	Key Advantages	Main Limitations
Forensic Image Analysis (Fingerprints)	Latent fingerprint matching, minutiae extraction, ridge enhancement	CNN, deep CNN, FingerNet, Siamese networks	Latent and rolled fingerprint images	Matching accuracy of approximately 90–97% in controlled datasets	Improved performance in partial or degraded prints; automation of the AFIS pipeline	Sensitive to image quality; false matches remain possible
Facial Recognition / Biometrics	1:1 verification, 1:N identification, watchlist screening	CNN, ArcFace, FaceNet, ResNet	Face images and video frames	High overall accuracy (>95%), although subgroup error disparities have been reported	Rapid large-scale identification	Demographic bias; legal and ethical concerns
Wound and Injury Classification	Differentiation of gunshot, sharp-force, and blunt-force trauma	CNN, ResNet, DenseNet	Forensic photographs	Accuracy of approximately 85–95%	Reduces subjective interpretation	Dataset variability; limited external validation
Crime Scene Object Detection	Detection of weapons, bloodstains, and evidentiary items	YOLO, Faster R-CNN	Scene images and video	Sensitivity >90% in experimental datasets	Rapid automated scanning of large image sets	Performance decreases in cluttered real-world scenes
Forensic Genetics (DNA Mixtures)	Contributor deconvolution, likelihood ratio estimation	Probabilistic genotyping (STRmix, EuroForMix, Bayesian networks, MCMC)	STR and NGS DNA profiles	Higher consistency than manual interpretation; improved likelihood-ratio discrimination	Reduced subjectivity; quantitative output	Parameter sensitivity; requires laboratory-specific validation
Forensic DNA Phenotyping	Ancestry and appearance prediction	Random forest, SVM, neural networks	SNP markers	Variable; population-dependent	Provides investigative leads	Ethical and privacy concerns
Digital Forensics (Malware)	Malware family classification	CNN, deep learning on byte-images	Binary files and logs	Accuracy >90–95% in controlled datasets	Automated triage of large evidence volumes	Adversarial evasion remains possible

Table 1. Artificial intelligence applications across forensic subdisciplines and corresponding algorithm types.

Forensic Subdiscipline	Typical Task / Application	Common AI/ML Algorithms	Data Type	Reported Performance (literature range)	Key Advantages	Main Limitations
Digital Forensics (Anomaly Detection)	Detection of suspicious log and activity patterns	Isolation Forest, autoencoders, LSTM	Log and network data	Marked reduction in manual review time has been reported	Accelerates investigations	False positives are common
Deepfake / Media Forensics	Detection of synthetic or manipulated images and videos	CNN, transformers, frequency-based models	Video, audio, and image data	Accuracy of 90–98% in benchmark datasets	Supports authenticity assessment	Ongoing arms race with generative models
Forensic Pathology (PMCT)	Organ and lesion segmentation, volumetry	U-Net, 3D CNN	Postmortem CT and MRI	Dice scores of approximately 0.85–0.93	Quantitative assessment; improved reproducibility	Small datasets; limited generalizability
Digital Histopathology*	Tissue and lesion classification	CNN, vision transformers	Whole-slide images	>90% accuracy in experimental studies	Automated screening and prioritization	Requires high-quality annotations; forensic applicability remains limited
Cross-domain Decision Support	Case prioritization and risk scoring	Gradient boosting, ensemble ML	Multimodal forensic data	Context-dependent	Supports expert decision-making	Interpretability concerns

Internal and multi-laboratory validation studies indicate that these systems can improve consistency and reduce subjectivity compared with manual interpretation, provided that laboratory-specific validation is performed before routine casework implementation (2,12).

More recent studies have extended probabilistic genotyping to 2–5 person mixtures in MaSTR, to NGS-derived autosomal STR profiles, and to single/few-cell workflows, indicating that these applications are feasible but validation-intensive (10,23,24).

Digital and Cyber Forensics

In digital forensics, ML primarily supports automated classification and triage to cope with evidence scale. One influential study transformed malware binaries into grayscale images and demonstrated accurate family-level classification, illustrating the feasibility of image-based ML pipelines in malware triage (15). More broadly, digital forensic workflows increasingly emphasize standardization, reproducibility, and transparent classification procedures (3,13).

For deepfake forensics, a comprehensive ACM survey stresses that operational deployment requires more than headline accuracy: detectors must meet reliability, protocolized testing, and evidentiary defensibility requirements to be useful in forensic settings (14). This framing highlights robustness, domain shift, and long-term sustainability in an adversarial detection–evasion landscape, while complementary overviews map current method families and evaluation pitfalls across image, video, and audio deepfakes (6,14).

Forensic Pathology and Postmortem Imaging

In postmortem CT (PMCT), segmentation-centric AI approaches enable quantitative forensic assessments. Fehr et al. developed a U-Net-based CNN for volumetric segmentation of multiple thoracic and abdominal structures in PMCT, reporting a best overall Dice score of 0.907 ± 0.029. Their findings illustrate both the promise of quantitative organ volumetry and practical constraints related to organ boundaries and image contrast (17).

Additional forensic pathology studies have extended this line of work to wound classification and rib fracture detection. J. Cheng et al. demonstrated promising deep-learning performance for classifying entrance versus exit gunshot wounds from digital images, whereas Lopez-Melia et al. reported average sensitivity of 70.2% and average precision of 78.1% for automatic rib fracture detection on PMCT data. Collectively, these findings indicate that the most defensible PMCT AI outputs are those tied to measurable imaging endpoints such as segmentation, volumetry, and detection (7,18,19).

Ethical, Legal, and Methodological Requirements

The implementation of artificial intelligence in forensic sciences must be evaluated not only in terms of technical performance but also with respect to ethical, legal, and methodological suitability. Unlike clinical or industrial applications, forensic outputs must withstand judicial scrutiny; therefore, scientific validity, reproducibility, and interpretability are essential requirements.

One of the most frequently reported concerns is algorithmic bias. Imbalanced training datasets may lead to substantially different false-positive or false-negative rates across demographic groups. Performance evaluations conducted by NIST and subsequent survey literature have demonstrated that error rates may vary markedly among subpopulations, raising concerns about fairness and due process in legal contexts (5,8,9). Accordingly, performance should be reported not only as overall accuracy but also through subgroup-specific analyses.

Table 2. Representative studies included in this review.

Forensic domain	Dataset / sample size	Algorithm / model	Evaluation metric(s)	Key reported results	Notes	Ref.
Latent fingerprints	NIST latent fingerprint pairs	Minutiae-based latent matching	Rank-1 identification	Provided a strong classical baseline, although performance decreased for low-quality and partial latent prints.	Benchmark classical approach	1
Latent fingerprints	Large latent/reference fingerprint datasets	Deep CNN-based latent matching pipeline	Rank-1 accuracy	Improved automated latent matching under controlled conditions.	Representative deep learning study	4
Face recognition	Millions of mugshot/probe comparisons	Commercial face recognition algorithms	False-positive rate, FNMR	Reported demographic differences reaching one to two orders of magnitude in false-positive rates.	Demonstrated algorithmic bias	5
Face recognition	Controlled face image comparisons	Commercial face recognition algorithms	Verification error rates	Documented an other-race effect in algorithm performance.	Supports subgroup-aware validation	9
Wound classification	Forensic photographs representing seven injury types	Segmentation and classification network	Accuracy, AUC, IoU	Demonstrated the feasibility of automated wound segmentation and classification.	Forensic image classification	7
DNA mixtures	Single-source and mixed STR profiles	STRmix™	Likelihood ratios, concordance	Internal validation supported laboratory-specific implementation.	Routine probabilistic genotyping validation	2
DNA mixtures	Multilaboratory STR datasets	STRmix™	Interlaboratory response, LR behavior	Strengthened evidence for reproducibility across laboratories.	Multilaboratory validation	12
DNA mixtures	STR mixture profiles	EuroForMix	Likelihood-ratio stability	Introduced an open-source continuous probabilistic genotyping model with competitive discriminatory performance.	Open-source probabilistic genotyping framework	11
DNA mixtures	Two- to five-person mixed DNA profiles	MaSTR™	Type I/II error, LR performance	Reported robust internal validation for mixtures with two to five contributors.	MaSTR validation	10
NGS DNA mixtures	Forensic NGS autosomal STR profiles	STRmix™ NGS	Sensitivity, specificity, accuracy, precision	Supported probabilistic genotyping of NGS-derived STR profiles.	Extension of probabilistic genotyping to NGS	23
Single- and few-cell STR analysis	Single- and few-cell subsamples	STRmix™ and EuroForMix	Validation and replicate analysis	Demonstrated feasibility under very low-template conditions.	Single- and few-cell application	24
DNA phenotyping	SNP-based appearance prediction studies	Phenotype prediction models	Classification accuracy	Highlighted potential investigative value, while emphasizing population dependence.	Ethical and privacy concerns noted	28
Malware classification	Malware image dataset (9 families)	Image-based machine learning pipeline	Accuracy	Reported high family-level classification accuracy.	Representative malware triage study	15
Media forensics / deepfakes	Benchmark datasets and review literature	CNN-based forensic detectors	Review of accuracy trends	Reported strong laboratory performance, but emphasized robustness concerns.	Review article	6
Deepfake detection	Multi-source datasets / survey literature	Multiple ML and CNN methods	Survey synthesis	Showed that benchmark success does not necessarily translate into real-world robustness.	Operational caution emphasized	14

Table 2. Representative studies included in this review.

Forensic domain	Dataset / sample size	Algorithm / model	Evaluation metric(s)	Key reported results	Notes	Ref.
Postmortem imaging	Postmortem imaging literature	Deep learning overview	Narrative synthesis	Highlighted the promise of AI in postmortem imaging together with data and expertise bottlenecks.	Review article	16
PMCT organ segmentation	PMCT thoracic and abdominal organ datasets	U-Net CNN	Dice score	Reported a best overall Dice score of 0.907 ± 0.029.	Quantitative organ volumetry	17
Gunshot wound classification	2,418 cropped wound images plus holdout set	ConvNeXt Tiny	Accuracy, precision, recall, F1 score	Reported a holdout accuracy of 87.99%.	Entrance-versus-exit wound classification	18
PMCT rib fractures	50 PMCT scans with comparison to clinical data	nnDetection-based deep learning	Sensitivity, precision	Reported an average sensitivity of 70.2% and an average precision of 78.1%.	Domain shift remained a significant limitation	19

A second critical issue is lack of explainability. The black-box nature of many deep learning models makes it difficult to clearly justify how decisions are produced. However, forensic reporting requires transparent and scientifically interpretable reasoning. Consequently, explainable AI techniques are increasingly recommended for forensic informatics and imaging applications (22).

Third, within-laboratory and inter-laboratory validation is essential. Probabilistic genotyping tools in forensic genetics are sensitive to parameter settings, thresholds, and instrumentation, necessitating laboratory-specific validation including sensitivity, specificity, and error rate estimation (2,10,11,12,25). Similar requirements apply in digital forensics, where robustness across diverse data types and operational scenarios must be demonstrated.

Furthermore, standardized reporting and quality assurance frameworks are strongly recommended. International forensic networks and laboratory standards emphasize independent testing, transparent performance reporting, traceability, and continued human oversight, positioning AI as a decision-support tool rather than a replacement for expert judgment (20,21,26).

Finally, legal admissibility remains a key consideration. For AI-generated evidence to be acceptable in court, methods must be scientifically recognized, validated, and accompanied by known error rates. Without these criteria, algorithmic outputs may not meet evidentiary standards. Thus, responsible integration of AI into forensic workflows depends primarily on transparency, validation, and accountability rather than solely on technical success (20,21,26). Key validation, ethical, and legal considerations for laboratory and courtroom-defensible use are summarized in Table 3.

Table 3. Validation, ethical, legal, and operational requirements for AI implementation in forensic practice.

Requirement domain	Core requirement	Why it matters in forensic practice	Risk if neglected	Practical implication
Internal validation	Validation under local laboratory conditions	Confirms that the system performs reliably with the laboratory's own workflows, instruments, and case types	Poor performance in routine casework	Each laboratory should validate the system before operational use
External validation	Testing on independent datasets from other centers or sources	Evaluates generalizability beyond the development dataset	Overfitting and limited reproducibility	Multicenter testing should be encouraged whenever possible
Realistic casework conditions	Evaluation using degraded, partial, mixed, noisy, or otherwise challenging forensic material	Reflects real forensic evidence rather than idealized benchmark conditions	Inflated performance estimates	Validation should include realistic case scenarios
Human oversight	Expert review of AI outputs before interpretation and reporting	Forensic conclusions must remain scientifically and legally defensible	Automation-related errors may go unnoticed	AI should support, not replace, expert judgment
Explainability	Transparent and interpretable reasoning for model outputs	Courts and forensic stakeholders may require understandable justification	Black-box evidence may be challenged or rejected	Explainability tools and traceable documentation should accompany AI outputs
Bias assessment	Evaluation of demographic, population, and subgroup performance	Reduces the risk of unfair or systematically skewed outcomes	Discriminatory performance and due-process concerns	Models should be tested across relevant subgroups
Error-rate reporting	Clear reporting of false positives, false negatives, uncertainty, and limitations	Forensic evidence must be presented with known limitations	Overstated evidentiary value	Reports should include known error rates where applicable

Table 3. Validation, ethical, legal, and operational requirements for AI implementation in forensic practice.

Requirement domain	Core requirement	Why it matters in forensic practice	Risk if neglected	Practical implication
Standardized reporting	Consistent reporting language and interpretation frameworks	Supports comparability, transparency, and legal clarity	Inconsistent reporting and interpretive ambiguity	Reporting should align with accepted forensic guidelines
Quality assurance	Ongoing monitoring, version control, and performance review	AI systems may drift over time or behave differently after updates	Undetected model drift and inconsistent outputs	Continuous quality monitoring is required after implementation
Reproducibility	Ability to obtain stable results across time, operators, and settings	Reproducibility is essential for scientific credibility and courtroom defensibility	Poor confidence in findings	Procedures should be documented and repeatable
Data governance	Secure handling of sensitive forensic and biometric data	Forensic datasets often contain highly sensitive personal information	Privacy breaches and regulatory concerns	Data access, storage, and processing should follow institutional and legal safeguards
Legal admissibility	Compatibility with evidentiary standards and judicial expectations	Even technically strong systems may fail if not legally defensible	Risk of evidentiary exclusion	Validation, transparency, and expert interpretation should be documented
Cybersecurity / adversarial robustness	Resistance to manipulation, spoofing, or adversarial interference	Forensic systems may be targeted or deceived intentionally	Vulnerability to tampering or misleading inputs	Robustness testing should be part of implementation
Domain-specific implementation	Adaptation to the requirements of each forensic subdiscipline	Performance expectations differ across fingerprints, DNA, PMCT, and digital forensics	Misapplication across domains	Systems should be validated for their specific forensic context
Ethical governance	Consideration of fairness, proportionality, accountability, and human rights	Forensic AI can affect investigations, charging decisions, and court outcomes	Ethical misuse and erosion of trust	Governance frameworks should accompany technical deployment

Note. This table summarizes validation, ethical, and legal considerations synthesized from the studies cited in the main text, particularly refs 2, 5, 11-12, 17, 20-22.

Overall, current evidence indicates that AI systems can provide substantial benefits as decision-support tools but should not function as autonomous decision-makers without expert supervision.

DISCUSSION

The studies reviewed demonstrate that artificial intelligence and machine learning provide meaningful performance gains across multiple forensic subdisciplines. In data-intensive tasks such as fingerprint matching, face comparison, DNA mixture interpretation, and digital evidence screening, algorithmic approaches consistently offer automation, improved consistency, and greater scalability compared with purely manual evaluation, although performance is often strongest under controlled conditions (4,5,12,15).

However, many of the reported high accuracies originate from controlled or idealized datasets. Numerous investigations rely on small sample sizes, single-center data, and relatively homogeneous conditions, which may limit generalizability to real-world crime scene environments characterized by degraded, partial, or noisy evidence. In postmortem imaging and forensic pathology in particular, the literature remains relatively small, feasibility-oriented, and often single-center, which limits routine operational transferability (16,17,19). Therefore, multicenter and heterogeneous validation studies are essential.

Digital histopathology may also be considered a methodologically adjacent field for forensic pathology. Although current evidence is derived mainly from clinical computational pathology, whole-slide AI systems demonstrate how image-based decision support can assist tissue interpretation and workflow prioritization. Their forensic applicability remains promising but unproven, and future use will require forensic-specific datasets, external validation, and careful attention to explainability and legal defensibility (22,27).

Demographic bias observed in biometric applications represents one of the most critical methodological challenges. Documented disparities in error rates across population subgroups raise serious concerns regarding fairness and due process (5,8,9). Given the legal implications of forensic decisions, reliance solely on average accuracy metrics is insufficient. Subgroup-specific performance and error distributions must be considered before operational deployment.

In forensic genetics, probabilistic genotyping has introduced a quantitative and statistically robust framework that reduces subjective interpretation. Nevertheless, sensitivity to parameter settings, differences among software implementations, and the need for laboratory-specific validation remain central issues (2,10,11,12). Similarly, in digital forensics, deepfake detection tools remain promising but operationally fragile, as real-world robustness, distribution shift, and adversarial adaptation continue to limit deployment in forensic settings (6,14).

From ethical and legal perspectives, AI in forensic science should be positioned as a decision-support technology rather than a fully autonomous system. The prevailing consensus is that AI enhances expert performance but cannot replace expert judgment (20,21,22,26). This hybrid human–AI model appears more sustainable both scientifically and legally.

Future research should focus on

- large-scale external validation using multicenter datasets,
- development of explainable AI approaches,
- standardized performance reporting frameworks,
- integration of ENFSI/ISO-compliant quality assurance protocols, and
- systematic ethical and legal impact assessments.

Such efforts will facilitate the safe and reliable adoption of AI systems in forensic practice.

In summary, while AI offers substantial potential benefits for forensic sciences, its true value can only be realized through rigorous validation, transparency, and continuous expert oversight.

CONCLUSION

This review demonstrates that artificial intelligence and machine learning are increasingly adopted across multiple subdisciplines of forensic sciences and provide significant operational and methodological advantages, particularly in data-intensive analytical tasks. In forensic image analysis, biometric identification, DNA mixture interpretation, digital evidence triage, and postmortem imaging, algorithmic approaches contribute to improved speed, reproducibility, and quantitative evaluation, effectively complementing traditional methods.

Nevertheless, the current body of evidence indicates that AI systems should be positioned as decision-support tools rather than independent decision-makers. Algorithmic bias, limited explainability, dependence on data quality, and insufficient external validation remain critical limitations that must be carefully addressed in forensic contexts. Accordingly, AI implementations in forensic laboratories should be supported by comprehensive internal and multicenter validation, standardized performance reporting, and robust quality assurance frameworks.

Future research should prioritize larger and more heterogeneous datasets, development of explainable AI techniques, and clarification of ethical and legal standards to ensure safe and sustainable integration of AI into forensic practice. In conclusion, while AI holds substantial promise for forensic sciences, its effective adoption depends fundamentally on transparency, validation, and continuous expert oversight.

LIMITATIONS OF THE STUDY

- This manuscript synthesizes a heterogeneous literature base; study designs, datasets, and evaluation metrics varied substantially, limiting direct comparability across domains.

- Although a systematic search strategy was used, this is not a meta-analysis; quantitative pooling of performance was not feasible due to methodological heterogeneity and incomplete reporting in primary studies.

- Many included studies rely on controlled, single-center or benchmark datasets; real-world operational performance (including cross-laboratory reproducibility and robustness to manipulation) may differ.

- Publication bias and rapid technological change may affect the completeness and durability of findings; ongoing updates and external validation are required.

ACKNOWLEDGEMENT

Peer-Review

Double blind internally peer reviewed

Conflict of Interest

The author declares that she has no conflict of interest regarding the content of this article.

Financial Support

The author reports no financial support regarding the content of this article.

Ethical Declaration

Ethical approval was not required because this study is based exclusively on previously published literature and does not involve human participants, biological samples, animals, or identifiable personal data. The manuscript was prepared in accordance with the principles of the Declaration of Helsinki where applicable.

Authorship Contributions

Concept: MŞ, Design: MŞ, Supervising: MŞ, Financing and equipment: MŞ, Data collection and entry: MŞ, Analysis and interpretation: MŞ, Literature search: MŞ, Writing: MŞ, Critical review: MŞ

REFERENCES

1. Jain AK, Feng J. Latent fingerprint matching. IEEE Trans Pattern Anal Mach Intell. 2011;33(1):88-100. https://doi.org/10.1109/TPAMI.2010.59

2. Moretti TR, Just RS, Kehl SC, Willis KD, Buckleton JS, Bright JA, et al. Internal validation of STRmix™ for the interpretation of single source and mixed DNA profiles. Forensic Sci Int Genet. 2017;29:126-144. https://doi.org/10.1016/j.fsigen.2017.04.004

3. Garfinkel SL. Digital forensics research: the next 10 years. Digit Investig. 2010;7 Suppl:S64-S73. https://doi.org/10.1016/j.diin.2010.05.009

4. Cao K, Jain AK. Automated latent fingerprint recognition. IEEE Trans Pattern Anal Mach Intell. 2019;41(4):788-800. https://doi.org/10.1109/TPAMI.2018.2818162

5. Grother PJ, Ngan M, Hanaoka K. Face recognition vendor test (FRVT) part 3: demographic effects. NIST Interagency Report 8280. 2019. https://doi.org/10.6028/NIST.IR.8280

6. Verdoliva L. Media forensics and deepfakes: an overview. IEEE J Sel Top Signal Process. 2020;14(5):910-932. https://doi.org/10.1109/JSTSP.2020.3002101

7. Zimmermann N, Sieberth T, Dobay A. Automated wound segmentation and classification of seven common injuries in forensic medicine. Forensic Sci Med Pathol. 2024;20(2):443-451. https://doi.org/10.1007/s12024-023-00668-5

8. Drozdowski P, Rathgeb C, Dantcheva A, Damer N, Busch C. Demographic bias in biometrics: a survey on an emerging challenge. IEEE Trans Technol Soc. 2020;1(2):89-103. https://doi.org/10.1109/TTS.2020.2992344

9. Phillips PJ, Jiang F, Narvekar A, Ayyad J, O'Toole AJ. An other-race effect for face recognition algorithms. NIST Interagency Report 7666. 2010. https://doi.org/10.6028/NIST.IR.7666

10. Adamowicz MS, Rambo TN, Clarke JL. Internal validation of MaSTR™ probabilistic genotyping software for the interpretation of 2-5 person mixed DNA profiles. Genes (Basel). 2022;13(8):1429. https://doi.org/10.3390/genes13081429

11. Bleka Ø, Storvik G, Gill P. EuroForMix: an open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts. Forensic Sci Int Genet. 2016;21:35-44. https://doi.org/10.1016/j.fsigen.2015.11.008

12. Bright JA, Richards R, Kruijver M, Kelly H, McGovern C, Magee A, et al. Internal validation of STRmix™: a multi-laboratory response to PCAST. Forensic Sci Int Genet. 2018;34:11-24. https://doi.org/10.1016/j.fsigen.2018.01.003

13. Casey E, Nelson A, Hyde J. Standardization of file recovery classification and authentication. Digit Investig. 2019;31:100873. https://doi.org/10.1016/j.diin.2019.06.004

14. Mirsky Y, Lee W. The creation and detection of deepfakes: a survey. ACM Comput Surv. 2021;54(1):1-41. https://doi.org/10.1145/3425780

15. Nataraj L, Karthikeyan S, Jacob G, Manjunath BS. Malware images: visualization and automatic classification. In: Proceedings of the 8th International Symposium on Visualization for Cyber Security; 2011. p. 4:1-4:7. https://doi.org/10.1145/2016904.2016908

16. Dobay A, Ford J, Decker SJ, Ampanozi G, Franckenberg S, Affolter R, et al. Potential use of deep learning techniques for postmortem imaging. Forensic Sci Med Pathol. 2020;16(4):671-679. https://doi.org/10.1007/s12024-020-00307-3

17. Fehr J, Kurmann A, Handschin P, Garyfalia A, Franckenberg S, Golomingi R, et al. Automated volumetric estimation of six basic thoracic and abdominal organs in postmortem computed tomography data using deep learning techniques. Forensic Imaging. 2025;42:200642. https://doi.org/10.1016/j.fri.2025.200642

18. Cheng J, Schmidt C, Wilson A, Wang Z, Hao W, Pantanowitz J, et al. Artificial intelligence for human gunshot wound classification. J Pathol Inform. 2023;15:100361. https://doi.org/10.1016/j.jpi.2023.100361

19. Lopez-Melia M, Magnin V, Schranz S, Andrearczyk V, Depeursinge A, Marchand-Maillet S, et al. Automatic rib fracture detection on postmortem CT data using deep learning. Int J Legal Med. 2026;140(2):857-866. https://doi.org/10.1007/s00414-025-03669-x

20. European Network of Forensic Science Institutes (ENFSI). Guideline for evaluative reporting in forensic science: strengthening the evaluation of forensic results across Europe (STEOFRAE). Wiesbaden: ENFSI; 2015.

21. Morrison GS. What should a forensic practitioner's likelihood ratio be? II. Sci Justice. 2017;57(6):472-476. https://doi.org/10.1016/j.scijus.2017.08.004

22. Samek W, Wiegand T, Müller KR. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv. 2017. https://doi.org/10.48550/arXiv.1708.08296

23. Cheng K, Bright JA, Kelly H, Liu YY, Lin MH, Kruijver M, et al. Developmental validation of STRmix™ NGS, a probabilistic genotyping tool for the interpretation of autosomal STRs from forensic profiles generated using NGS. Forensic Sci Int Genet. 2023;62:102804. https://doi.org/10.1016/j.fsigen.2022.102804

24. Huffman K, Ballantyne J. Validation of probabilistic genotyping software for single cell STR analysis. Genes (Basel). 2023;14(3):674. https://doi.org/10.3390/genes14030674

25. Meuwly D, Ramos D, Haraksim R. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. Forensic Sci Int. 2017;276:142-153. https://doi.org/10.1016/j.forsciint.2016.03.048

26. International Organization for Standardization. ISO/IEC 17025:2017: general requirements for the competence of testing and calibration laboratories. Geneva: ISO; 2017.

27. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole-slide images. Nat Med. 2019;25(8):1301-1309. https://doi.org/10.1038/s41591-019-0508-1

28. Kayser M. Forensic DNA phenotyping: predicting human appearance from crime scene material for investigative purposes. Forensic Sci Int Genet. 2015;18:33-48. https://doi.org/10.1016/j.fsigen.2015.02.003