forensic sciences

MDPI

Review

Exploring Trends and Gaps in Artificial Intelligence-Assisted Forensic Firearm Evidence Analysis: A Scoping Review

Mohammed Rabia Alkuwari * and Fathelalem Hija

Deanship of Graduate Studies and Scientific Research, Joaan Bin Jassim Academy for Defence Studies, Al Khor P.O. Box 4939, Qatar; fali@jbj.edu.qa
* Correspondence: mr.alkuwari@jbj.edu.qa; Tel.: +974-55518102

Abstract

Background: Artificial intelligence (AI) is increasingly used in forensic firearm evidence analysis to improve pattern recognition, automation, and analytical efficiency. Despite advances in machine learning (ML) and deep learning (DL), limitations in scientific validation, explainability, expert involvement, and judicial readiness constrain operational adoption. This review examines the evolution of computational and AI-assisted approaches and the factors influencing their integration into forensic practice. **Methods:** This PRISMA-ScR scoping review searched Wiley Library, PubMed, IEEE Xplore, ScienceDirect, Google Scholar, and Scopus for literature published between 2000 and December 2025. Of 101 records identified, 3 duplicates were removed, leaving 98 unique records for title and abstract screening. Following screening, 47 records were excluded, and 51 studies met the inclusion criteria and were retained for qualitative synthesis. The findings were interpreted alongside complementary authoritative and foundational literature to support framework development. **Results:** The reviewed literature demonstrates advances in ML, DL, three-dimensional analysis, and automated comparison, particularly in feature extraction, pattern recognition, and similarity assessment. Persistent limitations concern dataset representativeness, external validation, uncertainty quantification, explainability, standardization, expert involvement, governance, and judicial considerations. Integration of these findings with foundational literature informed a Human-Centered Hybrid AI Framework combining scientific validity, computational intelligence, explainability, expert collaboration, governance, and judicial accountability. **Conclusions:** Computational performance alone is insufficient for scientifically reliable and judicially defensible AI-assisted firearm examination. The proposed framework positions AI as an explainable, governed decision-support capability operating with qualified forensic experts rather than as an autonomous decision-maker. It provides an evidence-informed reference architecture for future validation, standardization, system development, and responsible forensic implementation.

Keywords: forensic firearms; firearm toolmarks; artificial intelligence; machine learning; deep learning; explainable artificial intelligence; human-centered AI; Expert-in-the-Loop; scientific validation; judicial admissibility; PRISMA-ScR

Academic Editors: Lina De Paola, Enrico Marinelli and Gabriele Napoletano

Received: 18 July 2026
Revised: 6 September 2026
Accepted: 8 September 2026
Published: 15 September 2026

1. Introduction

Recent advances in artificial intelligence (AI) and computational imaging have substantially transformed forensic firearm examination over the past two decades. Firearm toolmarks, including microscopic impressions produced on bullets and cartridge cases during firearm operation, constitute important physical evidence in criminal investigations and judicial proceedings. Because firearm identification has traditionally relied heavily on examiner expertise and visual comparison, concerns regarding subjectivity, inter-examiner variability, reproducibility, and scientific reliability have stimulated growing interest in computational approaches capable of improving the objectivity and consistency of firearm examinations [1–4].

Machine learning (ML), deep learning (DL), and advanced computational imaging have enabled increasingly sophisticated approaches to firearm toolmark analysis. Techniques including support vector machines (SVMs), convolutional neural networks (CNNs), deep feature-learning methods, surface-topography analysis, three-dimensional imaging, and hybrid computational approaches have demonstrated promising capabilities in automated feature extraction, similarity assessment, pattern recognition, and firearm evidence comparison [5–10]. These developments have expanded the potential of AI-assisted systems to improve analytical efficiency and consistency while supporting, rather than replacing, forensic examiner expertise.

Despite this technological progress, important scientific and operational limitations remain. Existing studies frequently rely on relatively small or laboratory-generated datasets, while validation protocols, uncertainty assessment, external testing, and standardized evaluation remain inconsistent [2,3,7]. Ethical and procedural concerns, including algorithmic bias, limited transparency, and inconsistent evaluation frameworks, further complicate the scientific assessment and responsible deployment of AI-assisted forensic systems [10,11]. Increasing reliance on complex computational models also raises concerns regarding explainability, transparency, expert involvement, and the ability to scrutinize AI-generated findings within forensic and judicial contexts [11–15].

These limitations reveal a broader gap between algorithmic performance and trustworthy forensic decision support. High classification or comparison performance alone does not establish that an AI-assisted method is scientifically valid, interpretable, reproducible, or suitable for evidential use. Research on explainable AI (XAI) and human-centered AI (HCAI) emphasizes the importance of understandable computational outputs, transparency, and meaningful human involvement in high-consequence decision environments [15–18]. In forensic practice, these requirements intersect with established scientific and legal considerations concerning empirical validity, testability, error rates, peer review, expert accountability, procedural fairness, and judicial scrutiny [19–21].

Guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) [22], this study reviews research published between 2000 and December 2025 on AI-assisted firearm evidence analysis. Specifically, the review aims to: (1) map the range and evolution of AI technologies applied to firearm toolmark examination; (2) identify dominant research themes and methodological trends; (3) evaluate technological, scientific, human, and legal challenges; (4) examine issues of explainability, reliability, expert involvement, and judicial admissibility; and (5) synthesize these findings with complementary foundational literature to inform the development of a Human-Centered Hybrid AI Framework and future research directions.

Drawing upon the evidence synthesized from the 51 studies included through the PRISMA-ScR process, together with complementary authoritative and foundational literature used for scientific and theoretical interpretation, this review moves beyond a descriptive mapping of technological developments to examine the conditions required for scientifically reliable, explainable, human-centered, and judicially defensible AI-assisted firearm examination.

https://doi.org/10.3390/forensicsci6030080

2. Materials and Methods

This scoping review was conducted and reported in accordance with PRISMA-ScR [22] to map the extent, characteristics, methodological development, and research gaps of AI-assisted forensic firearm evidence analysis. No review protocol was prospectively registered for this scoping review. Given the heterogeneity of the literature, the review was designed for descriptive and thematic synthesis rather than meta-analysis. The methodological process comprised eligibility definition, literature searching, study selection, data charting, evidence synthesis, and subsequent conceptual framework development.

2.1. Eligibility Criteria

Studies were eligible if they contributed directly to the development, application, validation, or evaluation of computationally assisted firearm evidence analysis. These included studies applying AI, machine learning, deep learning, image analysis, automated comparison, quantitative computational methods, or digital forensic technologies to firearm-related evidence, together with empirical or analytical studies addressing validation, examiner performance, evidential reliability, or related forensic processes relevant to the assessment of AI-assisted firearm examination. Broader methodological and review literature was retained where it contributed directly to understanding the technological, validation, human, or interpretive dimensions examined in the review.

Studies were excluded when they had no substantive relevance to firearm-related forensic evidence, computational or analytical development, validation, or the scientific and operational issues examined in this review. Eligibility was determined according to relevance to the predefined scope of the review and was not based on a formal assessment of methodological quality or risk of bias.

2.2. Information Sources and Search Strategy

A structured literature search was conducted across Wiley Online Library, PubMed, IEEE Xplore, ScienceDirect, Google Scholar, and Scopus. The search covered literature published between 2000 and December 2025. Initial searches were conducted between 16 and 18 January 2025, with a final search update completed in December 2025. The searches identified 101 records: 65 from Wiley Online Library, 7 from PubMed, 8 from IEEE Xplore, 7 from ScienceDirect, 10 from Google Scholar, and 4 from Scopus, as reported in the PRISMA-ScR flow diagram (Figure 1).

A common Boolean search strategy was applied across all six information sources, with minor source-specific adaptations required by the syntax and search functionality of each platform. The core search terms, Boolean structure, search period, and eligibility limits remained consistent across sources. Search terms were organized around two principal concepts: firearm evidence ("forensic firearms" OR "firearm toolmarks" OR "ballistic identification") and artificial intelligence and computational methods ("artificial intelligence" OR "machine learning" OR "deep learning" OR "image analysis" OR "automated firearm comparison"). Terms within each concept were combined using OR, and the two concept groups were combined using AND. The complete common Boolean search expression, together with the search dates, retrieval counts, and source-specific implementation details, is provided in Supplementary Table S3.

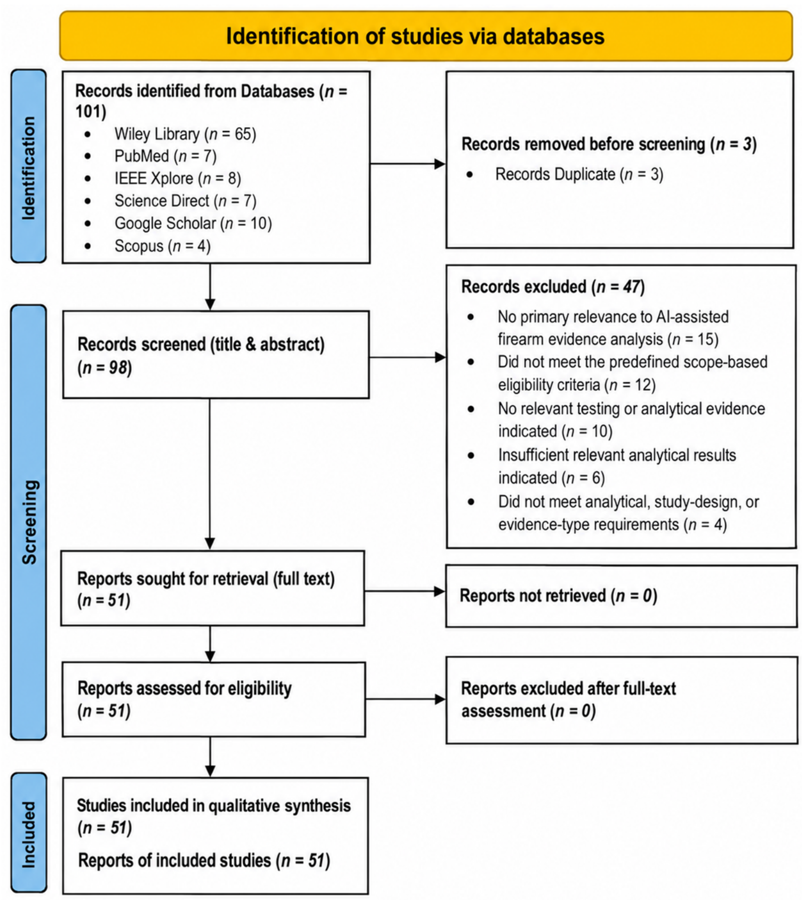

Figure 1. PRISMA-ScR flow diagram of the study identification, screening, eligibility assessment, and inclusion process.

2.3. Study Selection

Study selection followed the PRISMA-ScR process of identification, screening, eligibility assessment, and inclusion [22]. The database and supplementary searches identified 101 records. Records were managed in Zotero (version 7), and 3 duplicate records were removed, leaving 98 unique records for title and abstract screening against the predefined eligibility criteria. Following title and abstract screening, 47 records were excluded because they had no primary relevance to AI-assisted firearm evidence analysis (n = 15), did not meet the predefined scope-based eligibility criteria (n = 12), indicated no relevant testing or analytical evidence (n = 10), indicated insufficient relevant analytical results (n = 6), or did not meet the analytical, study-design, or evidence-type requirements of the review (n = 4). The remaining 51 reports were sought for retrieval, successfully retrieved, and assessed for full-text eligibility. No additional reports were excluded at the full-text stage; consequently, all 51 studies were retained for qualitative synthesis.

Title and abstract screening and full-text eligibility assessment were conducted by the first author using the predefined eligibility criteria, with the second author reviewing uncertain or borderline cases. Any disagreements regarding eligibility were resolved through discussion and consensus.

The complete study-selection process and reasons for exclusion at the title and abstract screening stage are summarized in the PRISMA-ScR flow diagram (Figure 1). The 51 included studies constitute the evidence base retained for qualitative synthesis and are reported individually in Supplementary Table S1. Given the methodological diversity of the included literature, this evidence base encompasses empirical, methodological, and conceptual contributions rather than a uniformly empirical body of evidence.

2.4. Data Extraction and Synthesis

References and study records were managed using Zotero (version 7) for reference management software.

Data extraction was performed by the first author using a structured data-charting approach. The extracted data were subsequently reviewed by the second author for consistency and accuracy, and any discrepancies were resolved through discussion.

Extracted variables included bibliographic information, study objectives, datasets and sample characteristics, AI methods and analytical tools, performance metrics, validation approaches, explainability, reproducibility indicators, limitations, and principal findings.

To account for the methodological heterogeneity of the evidence base, three characteristics—external/cross-laboratory validation (ELV), algorithmic explainability (XAI), and structured human–AI/computational interaction (HAI)—were assessed at the study level using a two-stage applicability-based coding procedure. First, each characteristic was assessed for methodological applicability to the design and analytical approach of each study. Characteristics considered not methodologically relevant were classified as Not Applicable (N/A). Second, among studies for which the characteristic was applicable, it was classified as Reported when explicitly demonstrated or described and as Not Reported when methodologically applicable but not documented. Studies classified as N/A were excluded from the denominator used for characteristic-specific descriptive percentages.

ELV was considered applicable when a study empirically evaluated a specific AI/ML model, automated/computational system, analytical method, or analytical procedure for which independent validation was methodologically meaningful, and was classified as Reported only when the same system, model, method, or procedure was evaluated using genuinely independent external data or samples, an independent institution or laboratory, or a cross-/multi-laboratory validation design. XAI was considered applicable only when a study implemented an AI/ML model and was classified as Reported only when an explicit mechanism for explaining or interpreting model outputs was implemented. HAI was considered applicable when an implemented AI or computational analytical system produced outputs capable of informing an expert analytical task or judgment and was classified as Reported only when the study explicitly described meaningful expert engagement with those outputs or the computational process. The complete study-level assessment is reported in Supplementary Table S5.

The extracted data were organized, and synthesized descriptively and thematically. Given the methodological heterogeneity of the included studies, no pooled quantitative synthesis was undertaken. The synthesis focused on technological trends, methodological and validation limitations, explainability, reproducibility, and requirements related to scientific reliability and judicial admissibility. The complete study-level data charted from the 51 included studies are provided in Supplementary Table S1.

Because the included literature was heterogeneous, studies were additionally characterized according to their primary contribution as AI/ML-based analysis, automated or computational firearm comparison, forensic validation or examiner-performance research, conventional analytical/digital forensic methodology, or review/conceptual evidence. The study-level mapping of all 51 included studies across these five categories is provided in Supplementary Table S4 to ensure transparency and traceability of the classification. This classification was used for descriptive stratification and synthesis and did not alter the original study-selection process or PRISMA counts.

2.5. Data Availability

The study-level data supporting this scoping review are provided in the Supplementary Materials, including the characteristics of the 51 included studies (Table S1), the

https://doi.org/10.3390/forensicsci6030080

Framework Development and Traceability Matrix (Table S2), the study-level methodological stratification of the heterogeneous evidence base (Table S4), and the applicability assessment of external validation, explainability, and human–AI/computational interaction (Table S5).

2.6. Framework Development

Beyond the descriptive synthesis of the scoping review, an iterative qualitative conceptual synthesis was undertaken to develop the proposed Human-Centered Hybrid AI Framework. Framework development integrated two complementary evidence streams: (1) recurring findings, limitations, and research gaps identified across the 51 studies included in the PRISMA-ScR evidence base; and (2) complementary foundational literature in forensic science, explainable AI, human-centered AI, governance, and judicial admissibility. The latter was used to provide established scientific, conceptual, and legal foundations and was not treated as part of the PRISMA-ScR evidence base.

The pathway from evidence to conceptual foundations, design principles, and architectural components is documented in the Framework Development and Traceability Matrix (Supplementary Materials, Table S2).

3. Results

This section presents the findings synthesized from the 51 studies included in the PRISMA-ScR review, with the complete study-level characteristics and extracted data provided in Supplementary Table S1. The results summarize publication trends, technological evolution, methodological characteristics, dataset and validation practices, and recurring research gaps identified across the reviewed literature. Interpretation of these findings, together with their integration with complementary scientific and theoretical literature, is presented in Section 4.

3.1. Publication Trends

The temporal distribution of the reviewed literature demonstrates a clear and sustained increase in research activity on AI-assisted firearm evidence analysis over the past two decades. The number of publications increased gradually during the early years of the study period, followed by a marked acceleration after 2018, coinciding with the widespread adoption of deep learning architectures, advances in high-resolution ballistic imaging, and increased computational capability.

Across the review period, the computational literature shows a transition from conventional image-processing and statistical approaches toward machine learning, deep learning, three-dimensional analysis, and increasingly automated comparison methods. Because the overall evidence base also includes validation, examiner-performance, analytical, and review/conceptual studies, technological trends are interpreted within the relevant computational subset rather than as a classification of all 51 included studies. The methodological composition of the complete evidence base is reported in Supplementary Table S4.

Early studies primarily employed conventional image-processing techniques, handcrafted feature extraction, and statistical approaches, whereas more recent publications increasingly adopted machine learning, convolutional neural networks (CNNs), Siamese networks, three-dimensional imaging, and automated comparison systems. This shift in computational approaches is examined in greater detail in Section 3.2.

3.2. Evolution of AI Methods in the Reviewed Studies

The 51 studies included in this PRISMA-ScR review demonstrate a clear evolution in AI approaches applied to forensic firearm evidence analysis over the review period, reflecting broader advances in computer vision, machine learning (ML), deep learning (DL), and

https://doi.org/10.3390/forensicsci6030080

computational imaging (Figure 2). Collectively, the reviewed studies show a transition from traditional image-processing techniques toward increasingly sophisticated AI architectures capable of supporting automated firearm comparison and forensic decision support.

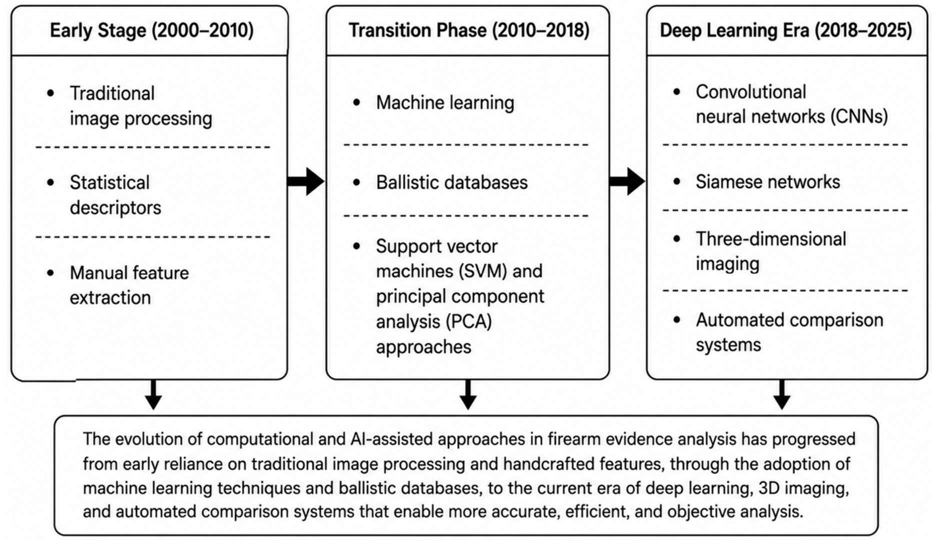

Figure 2. Technological evolution of computational and AI-assisted approaches in firearm evidence analysis. Horizontal arrows indicate the chronological progression across the three technological phases, while downward arrows link each phase to the overall evolutionary synthesis.

During the early stage (2000–2010), research relied primarily on conventional image-processing methods, statistical descriptors, and engineered features to characterize ballistic signatures. Early automated comparison approaches employed surface-topography measurement, feature extraction, and correlation-based analysis, while remaining dependent on carefully designed feature representations and expert interpretation [4,23].

The transition phase (2010–2018) witnessed increasing efforts to develop quantitative and automated firearm comparison methods, supported by advances in digital imaging, surface measurement, and computational analysis. Striation-based feature extraction, automated consecutive matching striae analysis, correlation methods, and statistical comparison approaches contributed to increasingly quantitative ballistic evidence comparison [5,7,24,25].

Since 2018, rapid advances in DL have transformed firearm evidence analysis. CNN-based architectures, deep feature-learning approaches, surface-topography analysis, three-dimensional imaging, and hybrid computational methods have demonstrated improvements in automated feature extraction, pattern recognition, and ballistic evidence comparison [8,9]. These developments have shifted the field from analytical methods toward increasingly automated AI-assisted comparison systems capable of supporting forensic examiners in complex evidential assessments.

Figure 2 summarizes the principal technological trajectory identified across studies concerned with computational and AI-assisted firearm analysis, highlighting the progression across three developmental stages from engineered image-processing and statistical approaches toward increasingly automated, data-driven AI-assisted firearm comparison. The figure is not intended to classify all 51 included studies as AI/ML model-development studies; rather, it illustrates the technological evolution within the relevant computational literature. The broader methodological diversity of the evidence base is documented in Supplementary Tables S1 and S4 and discussed in Section 3.4.

3.3. Dataset Characteristics and Validation Practices

The included studies showed considerable heterogeneity in dataset size, composition, and validation practices, with reported sample sizes ranging from small laboratory datasets to collections exceeding 20,000 observations. Dataset sources included internally generated laboratory collections, publicly accessible or standardized resources, and case-specific experimental datasets (Supplementary Table S1).

The study-level applicability assessment showed that external/cross-laboratory validation was methodologically applicable to 30 of the 51 included studies. Of these, 2 studies (6.7%) reported external or cross-laboratory validation, whereas 28 (93.3%) did not report such validation. The remaining 21 studies were classified as not applicable and were excluded from the ELV-specific denominator (Supplementary Table S5).

Differences in datasets, experimental settings, validation procedures, and reported performance measures further limited direct cross-study comparison.

Statistical assessment also varied across the reviewed literature. Reported procedures included analysis of variance, Mann–Whitney, Fligner–Killeen, and Kolmogorov–Smirnov tests, whereas explicit reporting of uncertainty measures, confidence intervals, and effect-size estimates was less consistently documented (Supplementary Table S1).

Several studies utilized standardized or publicly accessible resources, including the NIST Ballistics Toolmark Research Database [26].

Overall, the charted evidence demonstrates substantial heterogeneity in dataset composition and validation practices, with external or cross-laboratory validation remaining limited among studies for which such validation was methodologically applicable (Supplementary Tables S1 and S5).

3.4. Methodological Characteristics of the Reviewed Studies

The 51 included studies were retained for qualitative synthesis and constitute a methodologically heterogeneous evidence base rather than a uniformly empirical corpus. As documented in Supplementary Table S1, the studies vary in evidence type, dataset scale, analytical approach, validation strategy, principal findings, and reported limitations. Accordingly, the reviewed literature is characterized here as a qualitatively synthesized methodological landscape rather than as a homogeneous body of experimental AI studies. To make this heterogeneity explicit and traceable at the study level, the 51 included studies were stratified according to their primary contribution into five descriptive categories: AI/ML-based analysis ($n = 9$), automated or computational firearm comparison ($n = 8$), forensic validation or examiner-performance research ($n = 11$), conventional analytical/digital forensic methodology ($n = 10$), and review/conceptual evidence ($n = 13$). The individual study assignments are reported in Supplementary Table S4.

This methodological diversity includes examiner-based firearm comparison studies, reviews of computational and analytical methods, database-oriented studies, and experimental AI-assisted approaches [4,6,7]. Reported limitations likewise varied across studies and included closed-set designs, restricted review scope, data bias and privacy constraints, limited external validation, and unresolved challenges relating to standardization and explainability (Supplementary Table S1).

Performance evaluation commonly employed quantitative measures such as accuracy, precision, recall, F1-score, receiver operating characteristic curves, confusion matrices, and statistical comparison procedures. However, considerable heterogeneity was observed in dataset composition, validation procedures, and reported performance indicators, limiting direct methodological comparison across studies [7,10].

As reported in Section 3.3, external/cross-laboratory validation remained limited among the subset of studies for which this characteristic was methodologically applicable;

the corresponding study-level coding and denominator are provided in Supplementary Table S5.

Methodological attention beyond computational development and performance evaluation was also uneven. Algorithmic explainability was methodologically applicable to 9 studies implementing AI/ML models; none reported an explicit model-explanation mechanism (0/9, 0%). Structured human–AI/computational interaction was applicable to 24 studies, of which 4 (16.7%) explicitly reported meaningful expert engagement with computational systems or their outputs, whereas 20 (83.3%) did not report such interaction. The remaining studies were classified as not applicable to the corresponding characteristic and were excluded from the characteristic-specific denominators (Supplementary Table S5).

Broader issues concerning expert perceptions, workflow and organizational integration, and legal considerations were also less prominently represented in the reviewed literature [10,11].

Overall, the methodological profile of the reviewed studies was predominantly computational and performance-oriented, with substantially less methodological attention devoted to human-centered, organizational, interdisciplinary, and judicial dimensions of AI-assisted firearm evidence analysis. These observed methodological patterns contribute to the research gaps summarized in Section 3.5.

3.5. Major Research Gaps Identified in the Reviewed Studies

The synthesis of the 51 studies included in this PRISMA-ScR review identified recurring gaps across several dimensions of AI-assisted firearm evidence analysis. Dataset-related limitations were prominent, with many studies relying on relatively small, homogeneous, or laboratory-generated datasets and limited access to publicly available benchmarks [6,7,9]. Validation practices were also heterogeneous. Among the 30 studies for which external/cross-laboratory validation was methodologically applicable, only 2 (6.7%) reported such validation, whereas 28 (93.3%) did not (Supplementary Table S5). Reporting of uncertainty, confidence, and error rates also remained inconsistent [2,3].

Beyond these methodological limitations, explainability and structured human–AI/computational interaction remained limited among studies for which these characteristics were methodologically applicable. None of the 9 studies implementing AI/ML models reported an explicit XAI mechanism (0%), while structured human–AI/computational interaction was reported in only 4 of 24 applicable studies (16.7%); 20 of 24 (83.3%) did not report such interaction (Supplementary Table S5). These findings indicate continuing gaps in the systematic integration of algorithmic explainability and meaningful expert engagement within AI-assisted analytical workflows [14–17].

Reproducibility assessment, standardized validation and reporting practices, and comprehensive measures of statistical robustness were inconsistently addressed [2,3,27], while legal and judicial considerations-including transparency, accountability, procedural fairness, and evidential admissibility-received comparatively limited attention [19,20].

Table 1 summarizes the principal research gaps identified across the included studies and their corresponding methodological dimensions.

Collectively, these review findings provide the evidence base for the Discussion. In Section 4, they are interpreted in conjunction with complementary authoritative and foundational literature in forensic science, explainable AI, human-centered AI, and governance to examine their broader scientific, operational, and judicial implications and inform the development of the proposed Human-Centered Hybrid AI Framework.

https://doi.org/10.3390/forensicsci6030080

Table 1. Major research gaps identified in AI-assisted firearm evidence analysis.

Dimension	Main Shortcomings
Data	Small, homogeneous, and laboratory-generated datasets
Validation	Limited external testing; lack of benchmark datasets
Models	Limited interpretability of many complex AI/DL models
Explainability	Limited integration of XAI techniques
Human Factors	Underrepresentation of expert knowledge and human–AI collaboration
Reliability	Limited error rate estimation, uncertainty quantification, and reproducibility assessment
Standards	Absence of standardized validation and reporting protocols
Legal Aspects	Limited attention to judicial admissibility, procedural fairness, and legal accountability

4. Discussion

Section 3 presented the findings synthesized from the 51 studies included in this PRISMA-ScR review, describing publication trends, technological evolution, methodological characteristics, validation practices, and recurring research gaps in AI-assisted firearm evidence analysis. While these findings provide a comprehensive overview of the current state of the field, a descriptive synthesis alone does not fully explain their broader scientific, operational, and judicial significance.

The present discussion therefore extends beyond the review evidence by interpreting these findings in conjunction with complementary authoritative and foundational literature in forensic science, explainable artificial intelligence (XAI), human-centered AI, governance, and judicial admissibility. This complementary body of literature was not part of the PRISMA-ScR evidence synthesis itself; rather, it was incorporated to provide the scientific and theoretical foundations necessary to interpret the observed trends, explain the identified research gaps, and support the development of a coherent conceptual framework.

Hence, the following discussion examines the review findings from successive scientific perspectives, including technological evolution, scientific validation, explainability, human-centered AI, Expert-in-the-Loop collaboration, and judicial governance. Collectively, these complementary perspectives establish the conceptual foundations from which the proposed Human-Centered Hybrid AI Framework is derived.

4.1. Technological Evolution and Performance

The review findings presented in Section 3 demonstrate a clear technological evolution in AI-assisted firearm evidence analysis over the past two decades. The reviewed studies show a progressive transition from traditional image-processing techniques and engineered feature extraction toward machine learning (ML), deep learning (DL), and increasingly sophisticated computational approaches for firearm comparison. This evolution mirrors broader advances in artificial intelligence, computer vision, and digital imaging, which have transformed pattern recognition and decision-support capabilities across forensic science.

Among the reviewed studies, successive advances in three-dimensional surface analysis and automated comparison methods [4,6], statistical matching algorithms [7], and convolutional neural network (CNN)-based and deep feature-learning approaches [8,9] demonstrated substantial improvements in automated feature extraction, similarity assessment, and ballistic evidence comparison. These developments have enhanced computational capability while reducing reliance on engineered analytical workflows and subjective feature interpretation.

From a broader scientific perspective, this progression represents more than incremental improvements in computational performance. It reflects a transition from algorithm-centered automation toward increasingly intelligent analytical systems capable of supporting forensic examiners through more consistent, objective, and scalable evidence analysis. The progression from statistical matching approaches to modern deep-learning-based methods [7–9] illustrates this broader transformation. This trajectory is also reflected in very recent post-search literature, which continues to identify AI-enabled automation as an emerging direction in firearm and ammunition identification, while emphasizing the continuing importance of forensic expertise, validation, and appropriate integration of computational methods into examination practice [28].

Advances in computational power, digital imaging technologies, and multimodal data processing have further expanded the potential capabilities and applications of AI-assisted firearm comparison systems.

However, the findings also indicate that technological progress has not been matched by equivalent advances in scientific validation, explainability, expert collaboration, and governance. While many studies reported high classification accuracy and comparison performance, comparatively fewer addressed uncertainty quantification, reproducibility, external validation, or operational deployment under realistic forensic conditions. Thus, improvements in computational performance alone cannot establish the scientific reliability or operational readiness of AI-assisted firearm examination.

These observations identify scientific validation as the next critical dimension in assessing the maturity of AI-assisted firearm examination, as examined in Section 4.2. The following sections examine these complementary dimensions and their role in establishing the scientific and theoretical foundations of the proposed Human-Centered Hybrid AI Framework.

4.2. Dataset Quality, Validation, and Scientific Reliability

Despite the technological advances discussed in Section 4.1, the findings of this review indicate that dataset quality and validation practices remain major challenges to the scientific reliability of AI-assisted firearm examination. As identified in Sections 3.3 and 3.5, many of the reviewed studies relied on relatively small, homogeneous, or laboratory-generated datasets that may not adequately represent the variability encountered in operational forensic casework. Variations in firearms, ammunition, surface conditions, imaging systems, and evidence quality further complicate the generalizability of models developed and evaluated under controlled experimental conditions.

The limited availability of publicly accessible benchmark datasets presents an additional challenge. Reliance on institution-specific or internally generated datasets restricts independent replication and makes direct comparison among AI approaches difficult. Consequently, strong performance on a particular dataset cannot necessarily be assumed to generalize across laboratories, jurisdictions, evidence conditions, or operational environments. From a scientific perspective, this highlights the importance of dataset representativeness and independent evaluation when assessing the robustness and external validity of AI-assisted firearm examination systems.

Validation practices also remain inconsistent across the reviewed studies. The study-level applicability assessment reinforces this finding: external/cross-laboratory validation was reported in only 2 of the 30 studies for which it was methodologically applicable (6.7%; Supplementary Table S5).

Although conventional metrics such as accuracy, precision, recall, and F1-score are commonly reported, quantitative approaches to similarity assessment and error-rate estimation have also emerged [29].

Nevertheless, comprehensive uncertainty assessment, independent external validation, and cross-laboratory testing remain comparatively limited [30]. These limitations are particularly significant in forensic science, where analytical conclusions are expected to satisfy established requirements for empirical validity, reproducibility, and scientific scrutiny [2,3].

Accordingly, predictive performance and scientific reliability should be regarded as related but distinct dimensions of AI-assisted firearm examination. Reliable operational adoption requires not only high-performing computational models but also representative datasets, standardized validation protocols, transparent performance and error reporting, uncertainty quantification, reproducibility assessment, and independent testing under realistic conditions.

Together, these requirements establish scientific validity as a distinct conceptual foundation of the framework synthesized in Section 4.7 [2,3,9].

4.3. Explainability and Transparency

The review findings presented in Sections 3.4 and 3.5, supported by the study-level applicability assessment in Supplementary Table S5, show that none of the 9 studies implementing AI/ML models reported an explicit model-explanation mechanism (0/9, 0%).

This empirical gap becomes particularly important when interpreted in light of the broader explainable artificial intelligence (XAI) literature, which identifies explainability as a critical requirement for trustworthy AI in high-stakes domains. Unlike conventional performance metrics, explainability concerns whether AI-generated outputs can be understood and meaningfully interpreted by human users rather than treated as opaque "black-box" predictions [12–15,18].

The increasing use of DL models in firearm examination creates a particular interpretability challenge. Although these architectures can achieve high predictive performance, their internal decision-making processes are often difficult to interpret. This creates a tension between computational performance and the scientific requirements of transparency, reproducibility, and accountability. Foundational work on interpretable and explainable AI emphasizes that this tension becomes particularly important in high-stakes applications where algorithmic outputs may contribute to consequential human decisions [12–14].

In this context, transparency and explainability represent related but distinct properties. Transparency concerns the extent to which the structure, data sources, computational processes, and limitations of an AI system are accessible to scrutiny, whereas explainability concerns the ability to provide meaningful and human-interpretable accounts of particular outputs or predictions. In forensic practice, transparency supports scientific examination of the analytical system, while explainability enables examiners and other stakeholders to understand and critically evaluate the basis of AI-assisted findings.

These principles are particularly relevant to forensic evidence analysis. The broader XAI literature characterizes explainability as a means of making algorithmic outputs understandable to human users and thereby supporting informed evaluation of AI-generated results [15]. Similarly, the NIST principles for explainable AI emphasize that systems should provide explanations, ensure that those explanations are meaningful and accurate, and operate within their knowledge limits [18]. Applied to firearm examination, these principles suggest that an AI system should communicate not only its analytical output but also the relevant evidential features, reasoning process, confidence or uncertainty, and limitations associated with that output.

The importance of explainability becomes even greater when forensic conclusions may be subjected to scientific scrutiny, adversarial examination, or judicial review. An AI-assisted conclusion must therefore be not only accurate but also sufficiently interpretable to permit meaningful scientific and forensic scrutiny. Rudin [14], for example, argues for the use of inherently interpretable models rather than relying exclusively on post hoc explanations of black-box systems in high-stakes decision contexts. For AI-assisted firearm examination, this reinforces the need to integrate explainability into the analytical architecture rather than treating it solely as an additional visualization or post-processing function.

Explainability therefore emerges as a distinct conceptual foundation for trustworthy AI-assisted firearm examination. Its practical value, however, depends on knowledgeable human users who can interpret, challenge, and contextualize AI-generated explanations, providing the transition to human-centered AI examined in Section 4.4.

4.4. Human-Centered AI

The applicability-based assessment likewise showed limited implementation of structured human–AI/computational interaction. Among the 24 studies for which such interaction was methodologically applicable, only 4 (16.7%) explicitly documented meaningful expert engagement with computational outputs or processes, whereas 20 (83.3%) did not (Supplementary Table S5).

The growing emphasis on explainability raises the broader question of how human expertise should be positioned within AI-assisted forensic decision-making. The practical value of explainability depends on the capacity of qualified forensic practitioners to interpret, evaluate, and contextualize AI-generated outputs. This requirement leads to the broader paradigm of human-centered AI (HCAI), in which AI systems are designed to augment rather than replace human expertise.

The review findings presented in Sections 3.4 and 3.5, together with the study-level assessment in Supplementary Table S5, indicate that structured human–AI/computational interaction remains unevenly represented within applicable studies. When considered alongside the increasing computational sophistication identified in Section 3.2, this imbalance highlights an important distinction between improving algorithmic capability and developing AI systems that can effectively support professional forensic decision-making.

HCAI provides a broader conceptual perspective for addressing this distinction by combining computational capability with meaningful human involvement throughout the decision-making process. Rather than pursuing full analytical autonomy, HCAI emphasizes systems that preserve meaningful human control over consequential decisions while ensuring that algorithmic outputs remain understandable, verifiable, and subject to professional oversight. This perspective is particularly relevant to forensic science, where evidential conclusions involve not only computational analysis but also contextual interpretation, professional judgment, and accountability.

Foundational HCAI literature reinforces this interpretation. Holzinger [16] emphasizes the importance of effective human–AI collaboration in domains characterized by uncertainty, complexity, and consequential decision-making, where computational outputs should support rather than replace expert reasoning. Similarly, Shneiderman [17] argues for reliable, safe, and trustworthy AI systems designed to enhance human capabilities rather than pursue complete autonomy. Together, these perspectives establish meaningful human involvement as a defining principle of trustworthy AI in consequential decision environments [16,17]. Similar concerns have been identified in other high-consequence professional domains, where the integration of AI requires balancing technological capability with professional judgment, responsibility, and appropriate human oversight [31].

Applied to firearm examination, HCAI supports the integration of computational analysis with the domain knowledge, contextual understanding, and professional judgment of qualified firearm examiners. AI can therefore function as a transparent decision-support capability assisting with feature analysis, evidence comparison, uncertainty assessment, and interpretation, while the forensic examiner retains responsibility for evaluating the evidence and formulating the forensic conclusion.

Human-centered AI thus emerges as another conceptual foundation for trustworthy AI-assisted firearm examination. It extends explainability beyond making computational outputs understandable by establishing the human role through which those outputs are interpreted, evaluated, and incorporated into forensic decision-making. This human-centered perspective provides the conceptual basis for the more operational Expert-in-the-Loop paradigm examined in the following section.

4.5. Expert-in-the-Loop Paradigm

Building upon the principles of human-centered AI discussed in the preceding section, the Expert-in-the-Loop (EITL) paradigm provides an operational mechanism for integrating professional expertise into AI-assisted forensic decision-making. Rather than positioning AI systems as autonomous decision-makers, EITL architectures treat computational intelligence as a decision-support capability operating in structured interaction with qualified experts. Furthermore, the findings presented in Sections 3.4 and 3.5 indicate that comparatively limited attention has been devoted to systematically integrating human expertise into AI-assisted firearm examination workflows, despite substantial advances in computational capability.

The EITL paradigm recognizes that forensic conclusions cannot be reduced solely to algorithmic outputs. While AI can identify complex patterns and process large volumes of ballistic data, experienced examiners contribute contextual reasoning, uncertainty assessment, evidential interpretation, and professional judgment that may be difficult to formalize computationally. Human expertise can therefore contribute at multiple points in the analytical lifecycle, including evaluation of relevant features, interpretation of model outputs, assessment of ambiguous findings, and formulation of the final evidential conclusion. Interactive machine-learning research similarly demonstrates the value of incorporating human knowledge and feedback into computational processes [32].

Broader human-in-the-loop research similarly supports structured expert participation as a mechanism for strengthening the interpretability, evaluation, and accountability of AI-assisted decision-making [30]. In firearm examination, such participation is particularly relevant when evidence is ambiguous, or model confidence is limited.

Importantly, EITL should be distinguished from simple post hoc human oversight. The expert is not merely positioned at the end of an automated analytical pipeline to approve or reject a computational result. Rather, EITL establishes structured and potentially bidirectional interaction in which computational analysis informs expert reasoning, while expert evaluation and feedback can guide interpretation, additional analysis, and system assessment. This collaborative relationship is consistent with the interactive machine-learning perspective of Amershi et al. [32] and the broader human-in-the-loop approaches reviewed by Mosqueira-Rey et al. [33], both of which emphasize meaningful human participation in AI-supported learning and analytical processes.

Expert-in-the-Loop collaboration therefore represents the operational expression of the human-centered principle established in Section 4.4, embedding expert knowledge, judgment, and authority within the analytical workflow while preserving examiner responsibility for the final evidential conclusion.

The integration of expert authority within AI-assisted workflows also raises broader questions concerning professional responsibility, traceability, accountability, and the defensibility of AI-supported conclusions [34]. These considerations extend beyond human–AI interaction itself and lead directly to the governance and judicial implications examined in Section 4.6.

https://doi.org/10.3390/forensicsci6030080

4.6. Judicial Reliability and Admissibility

The review findings presented in Sections 3.4 and 3.5 indicate that legal admissibility, procedural accountability, and judicial considerations have received comparatively limited attention in the reviewed literature. This gap becomes particularly significant when considered against the scientific and legal requirements governing the reliability and admissibility of forensic evidence [35].

Longstanding critiques of forensic pattern-comparison disciplines provide an important scientific context for interpreting this finding. Saks and Koehler [1,27] argued for a transition from predominantly experience-based identification practices toward empirically grounded methods supported by measurable uncertainty and probabilistic reasoning. Similarly, the National Research Council [2] emphasized the need for objective validation, standardized methodologies, and rigorous quality assurance across forensic disciplines. These concerns were reinforced by the President's Council of Advisors on Science and Technology [3], which emphasized empirical demonstration of foundational validity and scientifically established estimates of error rates for feature-comparison methods. Collectively, these foundational sources establish scientific validity, reproducibility, error characterization, and transparent evaluation as central considerations in assessing the reliability of forensic evidence.

More recent empirical studies of firearm examiner performance have contributed quantitative evidence concerning comparison accuracy, error rates, repeatability, reproducibility, and examiner variability [28,29,36,37]. Such research reinforces the importance of evaluating uncertainty and reproducibility when considering how AI-generated outputs may contribute to firearm examination and evidential interpretation.

From a legal perspective, the *Daubert* decision [19] provides an influential framework within the United States for assessing the reliability and admissibility of expert scientific evidence, including considerations such as empirical testability, peer review and publication, known or potential error rates, standards controlling the technique, and general acceptance within the relevant scientific community. These considerations present particular challenges for AI-assisted forensic systems when their analytical processes are opaque, insufficiently validated, or difficult to independently scrutinize. In this context, explainability and transparency acquire significance beyond their technical functions by supporting scientific scrutiny, expert interpretation, and the communication of AI-assisted findings within judicial proceedings. Recent scholarship has likewise highlighted the challenges that AI introduces for the reliability and legal validity of scientific evidence, particularly in relation to explainability, algorithmic opacity, admissibility, and the need for appropriate procedural safeguards [34]. The human-centered and Expert-in-the-Loop principles discussed in Sections 4.4 and 4.5 provide an additional dimension of judicial reliability by preserving identifiable human authority and professional accountability over forensic conclusions. In this context, HCAI and EITL contribute to judicial reliability by preserving identifiable human authority and professional accountability over AI-assisted forensic conclusions [20].

More broadly, recent work on technologically mediated forensic evidence similarly emphasizes that scientific reliability depends on transparent methodology, verifiability, repeatability, traceability, and appropriate judicial scrutiny when advanced digital technologies are introduced into criminal proceedings [35]. Governance and judicial accountability therefore emerge as complementary requirements for trustworthy AI-assisted firearm examination. Scientific validity establishes whether an analytical method is empirically supportable; explainability enables its outputs to be understood and scrutinized; expert involvement preserves professional judgment and responsibility; and governance provides the mechanisms through which validation, traceability, documentation, quality assurance, accountability, and legal requirements can be maintained across the analytical lifecycle. Together, these dimensions provide the institutional and judicial context within which AI-assisted firearm evidence can be responsibly evaluated and used.

Taken together, the review findings and complementary foundational literature examined in Sections 4.1–4.6 converge on a set of interdependent requirements for trustworthy AI-assisted firearm examination. Their integration and architectural translation into the proposed Human-Centered Hybrid AI Framework are presented in Section 4.7.

4.7. Conceptual Synthesis and Development of the Human-Centered Hybrid AI Framework

The interpretation developed across Sections 4.1–4.6 reveals an imbalance between technological advances in AI-assisted firearm examination and the scientific, human, governance, and judicial requirements necessary for trustworthy forensic practice. Taken together, these findings indicate the need to integrate computational capability with scientific validity, explainability, human expertise, and governance with judicial accountability.

The proposed Human-Centered Hybrid AI Framework was derived from two complementary evidence streams. The first comprised the review findings synthesized from the 51 studies included in the PRISMA-ScR review, which identified technological developments, methodological characteristics, validation practices, and recurring research gaps. The second comprised complementary authoritative and foundational literature in forensic science, explainable AI, human-centered AI, human–AI collaboration, governance, and judicial admissibility, which provided the scientific and theoretical principles required to interpret those findings. Through iterative qualitative synthesis, these evidence streams were integrated into a set of conceptual foundations and design principles that informed the architecture of the proposed framework.

The framework development process is documented in the Framework Development and Traceability Matrix provided in Supplementary Materials (Table S2). The matrix establishes an explicit pathway from the review findings and complementary foundational literature to the synthesized conceptual foundations, associated design principles, and their subsequent architectural realization. Figure 3 translates this conceptual synthesis into an integrated socio-technical architecture for trustworthy AI-assisted firearm toolmark examination.

Figure 3 represents the architectural realization of the conceptual synthesis developed through this review. Rather than treating AI as an isolated or autonomous computational component, the framework integrates the identified scientific, computational, human, and governance requirements within a unified socio-technical decision-support architecture. AI-generated outputs can therefore be scientifically evaluated, professionally interpreted, transparently documented, and appropriately communicated for forensic and judicial use.

The framework is organized into four complementary components. The first defines the core scientific, technical, human, and legal requirements of trustworthy forensic AI; the second identifies the enabling tools and resources supporting these requirements; and the third translates them into operational forensic activities, including model development, explainable analysis, expert review, validation, documentation, and quality assurance. Their integration culminates in the fourth component: scientifically reliable and judicially trustworthy AI-assisted firearm evidence. This structure distinguishes the requirements that trustworthy forensic AI should satisfy from the mechanisms through which they can be implemented in practice.

Importantly, the framework positions AI as a decision-support capability rather than an autonomous decision-maker, with qualified forensic experts retaining responsibility for interpretation and evidential conclusions. It should therefore be understood as an evidence-informed reference architecture rather than a validated operational system. The framework provides a structured foundation for future empirical validation, system development, cross-laboratory evaluation, standardization, and operational implementation.

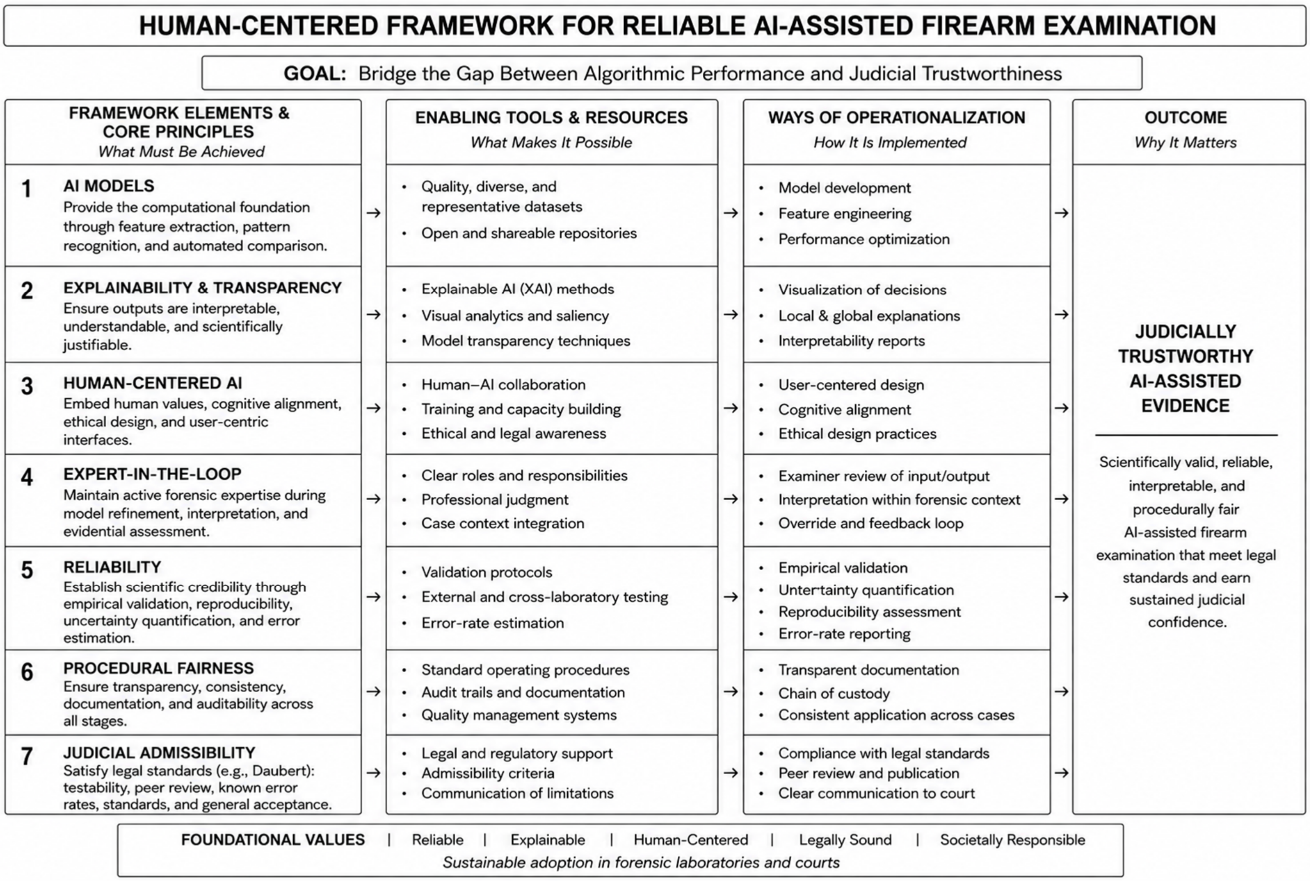

Figure 3. Human-Centered Hybrid AI Framework for trustworthy AI-assisted firearm toolmark examination. Numbers identify the seven framework elements; horizontal arrows indicate their translation from core principles through enabling resources and operationalization toward the intended outcome.

4.8. Future Research Directions

The findings of this review and the conceptual synthesis developed in Section 4.7 identify several priorities for future research. First, larger, more diverse, and publicly accessible benchmark datasets are needed to support independent and cross-laboratory validation, improve model generalizability, and enable meaningful comparison among AI-assisted firearm examination approaches.

Future studies should also strengthen methodological validation through standardized evaluation protocols, uncertainty quantification, confidence estimation, error-rate reporting, reproducibility assessment, and testing under representative operational conditions. Greater attention should be devoted to explainable AI (XAI), human-centered AI (HCAI), and Expert-in-the-Loop (EITL) approaches, including empirical evaluation of how explanations and structured expert interaction influence analytical performance, interpretation, decision consistency, and professional accountability.

A further priority is the empirical implementation and validation of the proposed Human-Centered Hybrid AI Framework. Future research should examine its individual architectural components and their interactions using representative firearm evidence, multimodal benchmark datasets, and realistic forensic workflows. Cross-laboratory studies will be particularly important for evaluating the framework's robustness, scalability, interoperability, and applicability across different institutional and operational environments.

The governance and judicial dimensions of forensic AI require greater interdisciplinary investigation involving forensic scientists, computer scientists, legal scholars, standards organizations, and policymakers. Such collaboration can support the development of common validation frameworks, reporting standards, governance mechanisms, ethical guidelines, and regulatory approaches for the responsible use of AI-assisted forensic systems. Collectively, these research directions provide a pathway from the evidence-informed framework proposed in this review toward its empirical validation, standardization, operational implementation, and evaluation within judicial contexts.

This scoping review has several methodological limitations. The restriction to English-language publications and the heterogeneity of study designs, datasets, analytical methods, and reporting practices may have limited the breadth and comparability of the evidence synthesized. In addition, no formal critical appraisal or risk-of-bias assessment was undertaken, consistent with the mapping purpose of the scoping-review design. Methodological characteristics and reported limitations were descriptively charted where relevant but were not used as a formal quality-rating or risk-of-bias framework. Accordingly, the findings should be interpreted as a structured synthesis of the available evidence and research gaps rather than as an assessment of comparative effectiveness, methodological quality, or evidential certainty.

5. Conclusions

This scoping review examined the evolution of AI-assisted forensic firearm evidence analysis through a qualitative synthesis of 51 studies published between 2000 and December 2025. While demonstrating substantial progress in machine learning (ML), deep learning (DL), three-dimensional analysis, and automated comparison technologies, the review also identified persistent challenges related to dataset representativeness, external validation, uncertainty assessment, explainability, standardization, expert involvement, governance, and judicial readiness.

A central finding is that current research remains predominantly focused on algorithmic and computational performance, with comparatively less attention devoted to explainability, expert knowledge integration, scientific reliability, procedural accountability, and legal considerations. When interpreted in conjunction with complementary foundational literature, these findings demonstrate that trustworthy AI-assisted firearm examination requires more than predictive accuracy; it requires the integration of scientific validity, computational capability, transparent reasoning, meaningful human involvement, and appropriate governance and judicial accountability.

Building upon this synthesis, the review proposes a Human-Centered Hybrid AI Framework derived through the integration of the synthesized PRISMA-ScR evidence base with complementary authoritative and foundational literature in forensic science, explainable AI, human-centered AI, human–AI collaboration, and governance.

The framework development process is documented through a traceability matrix linking the synthesized review evidence and complementary literature to the conceptual foundations, design principles, and their architectural realization. The resulting framework therefore represents an evidence-informed reference architecture that integrates scientific validity, computational intelligence, explainability, expert collaboration, and governance within a unified socio-technical approach to trustworthy AI-assisted firearm toolmark examination.

Rather than positioning AI as a replacement for forensic expertise, the proposed framework positions computational intelligence as an explainable and governed decision-support capability operating in collaboration with qualified forensic examiners. As an evidence-informed framework rather than a fully validated operational system, it provides a structured foundation for future empirical validation, cross-laboratory evaluation, standardization, technological development, and policy consideration. In this way, the study shifts the emphasis from algorithmic performance alone toward scientifically validated, explainable, human-centered, and accountable AI-assisted firearm examination capable of supporting reliable forensic decision-making and judicially defensible use.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/forensicsci6030080/s1, Table S1: Characteristics of the 51 studies included in the scoping review. Table S2: Framework Development and Traceability Matrix. Linking the Evidence Synthesis, Foundational Literature, Design Principles, and Architectural Components of the Proposed Human-Centered Hybrid AI Framework. Table S3: Common Search Strategy and Source-Specific Search Details. Table S4: Study-Level Methodological Stratification of the Included Evidence Base. Table S5: Study-Level Methodological Classification and Applicability Assessment of External Validation, Explainability, and Human–AI Interaction. The completed PRISMA-ScR checklist is provided separately as non-published Supplementary Materials [36–86].

Author Contributions: Conceptualization, M.R.A. and F.H.; methodology, M.R.A. and F.H.; investigation, including literature searching, M.R.A.; study screening and eligibility assessment, M.R.A., with review of uncertain or borderline cases, F.H.; data curation, including data extraction and organization of the PRISMA-ScR evidence corpus, M.R.A.; validation of extracted data, F.H.; formal analysis, M.R.A. and F.H.; writing—original draft preparation, M.R.A.; writing—review and editing, M.R.A. and F.H.; visualization, M.R.A.; supervision, F.H.; project administration, F.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study did not involve human or animal participants and therefore did not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: The materials supporting this scoping review, including the study-level evidence table, framework traceability matrix, methodological stratification of the evidence base, and applicability assessment of external validation, explainability, and human–AI/computational interaction, are available in the Supplementary Materials.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
PRISMA-ScR	Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network
SVM	Support Vector Machine
XAI	Explainable Artificial Intelligence
HCAI	Human-Centered Artificial Intelligence
EITL	Expert-in-the-Loop
3D	Three-Dimensional
CMC	Congruent Matching Cells
NBTRD	NIST Ballistics Toolmark Research Database

References

1. Saks, M.J.; Koehler, J.J. The coming paradigm shift in forensic identification science. *Science* **2005**, *309*, 892–895. [CrossRef] [PubMed]
2. National Research Council. *Strengthening Forensic Science in the United States: A Path Forward*; National Academies Press: Washington, DC, USA, 2009. [CrossRef] [PubMed]
3. President's Council of Advisors on Science and Technology. *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*; Executive Office of the President of the United States: Washington, DC, USA, 2016.
4. Banno, A.; Masuda, T.; Ikeuchi, K. Three dimensional visualization and comparison of impressions on fired bullets. *Forensic Sci. Int.* **2004**, *140*, 233–240. [CrossRef] [PubMed]
5. Soons, J.A.; Bramble, S.K.; Vorburger, T.V.; Zheng, A.; Thompson, R.M. Automated bullet signature identification based on topography measurements and correlations. *Forensic Sci. Int.* **2014**, *238*, 68–73.
6. Tai, X.H.; Eddy, W.F. A fully automatic method for comparing cartridge case images. *J. Forensic Sci.* **2018**, *63*, 440–448. [CrossRef] [PubMed]
7. Hare, E.; Hofmann, H.; Carriquiry, A. Automatic matching of bullet land impressions. *Ann. Appl. Stat.* **2017**, *11*, 2332–2356. [CrossRef]
8. Pisantanaroj, P.; Tanpisuth, P.; Sinchavanwat, P.; Phasuk, S.; Phienphanich, P.; Jangtawee, P.; Tantibundhit, C. Automated firearm classification from bullet markings using deep learning. *IEEE Access* **2020**, *8*, 78236–78251. [CrossRef]
9. Vargas, R.; Mosavi, A.; Ruiz, R. Deep Learning: A Review. *Preprints* **2018**, 2018100218. [CrossRef]
10. Bollé, T.; Casey, E.; Jacquet, M. The role of evaluations in reaching decisions using automated systems supporting forensic analysis. *Forensic Sci. Int. Digit. Investig.* **2020**, *34*, 301016. [CrossRef]
11. Solanke, A.A. Explainable digital forensics AI: Towards mitigating distrust in AI-based digital forensics analysis using interpretable models. *Forensic Sci. Int. Digit. Investig.* **2022**, *42*, 301403. [CrossRef]
12. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, arXiv:1702.08608.
13. Lipton, Z.C. The mythos of model interpretability. *Commun. ACM* **2018**, *61*, 36–43. [CrossRef]
14. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
15. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Herrera, F. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
16. Holzinger, A. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Inform.* **2016**, *3*, 119–131. [CrossRef] [PubMed]
17. Shneiderman, B. Human-centered artificial intelligence: Reliable, safe and trustworthy. *Int. J. Hum. Comput. Interact.* **2020**, *36*, 495–504. [CrossRef]
18. Phillips, P.J.; Hahn, C.A.; Fontana, P.C.; Yates, A.N.; Greene, K.K.; Broniatowski, D.A.; Przybocki, M.A. *Four Principles of Explainable Artificial Intelligence*; NISTIR 8312; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2021. [CrossRef]
19. *Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579*; JUSTIA: Mountain View, CA, USA, 1993.
20. Zalnieriute, M.; Moses, L.B.; Williams, G. The rule of law and automation of government decision-making. *Mod. Law Rev.* **2019**, *82*, 425–455. [CrossRef]
21. Socol de la Osa, D.U.; Remolina, N. Artificial intelligence at the bench: Legal and ethical challenges of informing—Or misinforming—Judicial decision-making through generative AI. *Data Policy* **2024**, *6*, e59. [CrossRef]
22. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Straus, S.E. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [CrossRef] [PubMed]
23. Chu, W.; Song, J.F.; Vorburger, T.V.; Yen, J.H.; Ballou, S.M.; Bachrach, B. Pilot study of automated bullet signature identification based on topography measurements and correlations. *J. Forensic Sci.* **2010**, *55*, 341–347. [CrossRef] [PubMed]
24. Chu, W.; Song, J.F.; Vorburger, T.V.; Thompson, R.M.; Silver, R.M. Selecting valid correlation areas for automated bullet identification systems based on striation detection. *J. Res. Natl. Inst. Stand. Technol.* **2011**, *116*, 647–653. [CrossRef] [PubMed]
25. Chu, W.; Thompson, R.M.; Song, J.F.; Vorburger, T.V. Automatic identification of bullet signatures based on consecutive matching striae criteria. *Forensic Sci. Int.* **2013**, *231*, 137–141. [CrossRef] [PubMed]
26. Zheng, X.; Soons, J.; Thompson, R.; Singh, S.; Constantin, C. NIST Ballistics Toolmark Research Database. *J. Res. Natl. Inst. Stand. Technol.* **2020**, *125*, 125004. [CrossRef] [PubMed]
27. Saks, M.J.; Koehler, J.J. The individualization fallacy in forensic science evidence. *Vanderbilt Law Rev.* **2008**, *61*, 199–219.
28. Herke, C.; Fedorenko, V.A. Automated firearm and ammunition identification: The role of artificial intelligence in forensic ballistics. *Forensic Sci.* **2026**, *6*, 69. [CrossRef]
29. Song, J.; Vorburger, T.V.; Chu, W.; Yen, J.; Soons, J.A.; Ott, D.; Zhang, N.F. Estimating error rates for firearm evidence identifications in forensic science. *Forensic Sci. Int.* **2018**, *284*, 15–32. [CrossRef] [PubMed]

30. Baiker-Sørensen, M.; Alberink, I.; Granell, L.B.; van der Ham, L.; Mattijssen, E.J.; Smith, E.D.; Zheng, X.A. Automated interpretation of comparison scores for firearm toolmarks on cartridge case primers. *Forensic Sci. Int.* **2023**, *353*, 111858. [CrossRef] [PubMed]
31. Marinelli, S.; De Paola, L.; Stark, M.; Montanari Vergallo, G. Artificial intelligence in the service of medicine: Current solutions and future perspectives, opportunities, and challenges. *Clin. Ter.* **2025**, *176*, 77–82. [CrossRef] [PubMed]
32. Amershi, S.; Cakmak, M.; Knox, W.B.; Kulesza, T. Power to the people: The role of humans in interactive machine learning. *AI Mag.* **2014**, *35*, 105–120. [CrossRef]
33. Mosqueira-Rey, E.; Hernández-Pereira, E.; Alonso-Ríos, D.; Bobes-Bascarán, J.; Fernández-Leal, Á. Human-in-the-loop machine learning: A state of the art. *Artif. Intell. Rev.* **2023**, *56*, 3005–3054. [CrossRef]
34. Ottaviani, M.; Spadazzi, F.; Bubbico, T.; De Paola, L.; La Russa, R. The evolution of scientific evidence between artificial intelligence and legal validity. *Leg. Med.* **2026**, *83*, 102852. [CrossRef] [PubMed]
35. Di Santo, M.; Fais, P.; De Paola, L. Virtual reconstruction as scientific evidence in criminal proceedings. *Forensic Sci.* **2026**, *6*, 60. [CrossRef]
36. Law, E.F.; Morris, K.B. Evaluating Firearm Examiner Conclusion Variability Using Cartridge Case Reproductions. *J. Forensic Sci.* **2021**, *66*, 1704–1720. [CrossRef] [PubMed]
37. Monson, K.L.; Smith, E.D.; Peters, E.M. Repeatability and reproducibility of comparison decisions by firearms examiners. *J. Forensic Sci.* **2023**, *68*, 1721–1740. [CrossRef] [PubMed]
38. Sengupta, S.; Basak, S.; Saikia, P.; Paul, S.; Tsalavoutis, V.; Atiah, F.D.; Ravi, V.; Peters, R.A., II. A Review of Deep Learning with Special Emphasis on Architectures, Applications and Recent Trends. *arXiv* **2019**, arXiv:1905.13294v3.
39. Mattijssen, E.J.A.T.; Witteman, C.L.M.; Berger, C.E.H.; Zheng, X.A.; Soons, J.A.; Stoel, R.D. Firearm Examination: Examiner Judgments and Computer-Based Comparisons. *J. Forensic Sci.* **2020**, *66*, 96–111. [CrossRef] [PubMed]
40. Uysal, S.; Uzun, L.; Armutcu, C.; Mercan, M. A Spectroscopic Approach for Rapid and Simple Serial Number Restoration on Polyamide 6 Parts of Firearms: The Use of Video Spectral Comparator 8000. *J. Forensic Sci.* **2021**, *66*, 2381–2386. [CrossRef] [PubMed]
41. Xin, L.; Zhou, J.; Rong, G. A Cartridge Identification System for Firearm Authentication. In *Proceedings of the IEEE International Conference on Signal Processing (ICSP'00), Beijing, China, 21–25 August 2000*; IEEE: Piscataway, NJ, USA, 2000; pp. 1405–1408. [CrossRef]
42. Li, D. A New Approach for Firearm Identification with Hierarchical Neural Networks Based on Cartridge Case Images. In *Proceedings of the 5th IEEE International Conference on Cognitive Informatics (ICCI'06), Beijing, China, 17–19 July 2006*; IEEE: Piscataway, NJ, USA, 2006; pp. 923–928.
43. Tong, M.; Song, J.; Chu, W. An Improved Algorithm of Congruent Matching Cells (CMC) Method for Firearm Evidence Identifications. *J. Res. Natl. Inst. Stand. Technol.* **2015**, *120*, 102–109. [CrossRef] [PubMed]
44. Emmert-Streib, F.; Yang, Z.; Feng, H.; Tripathi, S.; Dehmer, M. An Introductory Review of Deep Learning for Prediction Models with Big Data. *Front. Artif. Intell.* **2020**, *3*, 4. [CrossRef] [PubMed]
45. Abada, R.; Abubakar, A.M.; Bilal, M.T. An Overview on Deep Learning Application of Big Data. *Mesop. J. Big Data* **2022**, *1*, 31–35. [CrossRef] [PubMed]
46. Hamby, J.E.; Warren, E.M.; Brundage, D.J.; Petraco, N.D.K.; Thorpe, J.W. An Update on a World Wide Study of Bullets Fired from 10 Consecutively Rifled 9mm Ruger Pistol Barrels: Analysis of Examiner Error Rate Using Bayesian Statistics. *Arab J. Forensic Sci. Forensic Med.* **2024**, *6*, 102–114. [CrossRef]
47. Riva, F.; Champod, C. Automatic Comparison and Evaluation of Impressions Left by a Firearm on Fired Cartridge Cases. *J. Forensic Sci.* **2014**, *59*, 637–647. [CrossRef] [PubMed]
48. Li, D. Ballistics Image Processing and Analysis for Firearm Identification. In *Image Processing*; Chen, Y.-S., Ed.; IntechOpen: Rijeka, Croatia, 2009; pp. 142–154. [CrossRef] [PubMed]
49. Ott, D.; Thompson, R.; Song, J. Applying 3D Measurements and Computer Matching Algorithms to Two Firearm Examination Proficiency Tests. *Forensic Sci. Int.* **2017**, *271*, 98–106. [CrossRef] [PubMed]
50. Guarnera, L.; Giudice, O.; Livatino, S.; Paratore, A.B.; Salici, A.; Battiato, S. Assessing Forensic Ballistics Three-Dimensionally through Graphical Reconstruction and Immersive VR Observation. *Multimed. Tools Appl.* **2023**, *82*, 20655–20681. [CrossRef]
51. Dalby, O.; Butler, D.; Birkett, J.W. Analysis of Gunshot Residue and Associated Materials—A Review. *J. Forensic Sci.* **2010**, *55*, 924–943. [CrossRef] [PubMed]
52. Hartwig, S.; Tsokos, M.; Byard, R.W. Black Powder Handgun Deaths Remain an Uncommon Event. *Am. J. Forensic Med. Pathol.* **2009**, *30*, 350–353. [CrossRef] [PubMed]
53. Fernandez-Basso, C.; Gutierrez-Batista, K.; Gómez-Romero, J.; Ruiz, M.D.; Martin-Bautista, M.J. An AI Knowledge-Based System for Police Assistance in Crime Investigation. *Expert Syst.* **2025**, *42*, e13524. [CrossRef]
54. Zhou, Z.; Wu, J.; Li, Y.; Jia, L.; Shen, Y.; Zhang, G.; Cui, B. Firearm Image Recognition System Based on Multi-Task Cascaded Residual Network. *Comput. Eng.* **2022**, *48*, 214–219. [CrossRef]

55. Ramachandran, G.; Kannan, S. Artificial Intelligence and Deep Learning Applications: A Review. *J. Artif. Intell. Mach. Learn. Neural Netw.* **2021**, *1*, 10–13. [CrossRef]
56. Le Bouthillier, M.-E.; Hrynkiw, L.; Beauchamp, A.; Duong, L.; Ratté, S. Automated Detection of Regions of Interest in Cartridge Case Images Using Deep Learning. *J. Forensic Sci.* **2023**, *68*, 1958–1971. [CrossRef] [PubMed]
57. Neuman, M.; Hundl, C.; Grimaldi, A.; Eudaley, D.; Stein, D.; Stout, P. Blind Testing in Firearms: Preliminary Results from a Blind Quality Control Program. *J. Forensic Sci.* **2022**, *67*, 964–974. [CrossRef] [PubMed]
58. Redouté Minzière, V.; Werner, D.; Schneider, D.; Manganelli, M.; Jung, B.; Weyermann, C.; Gassner, A.L. Combined Collection and Analysis of Inorganic and Organic Gunshot Residues. *J. Forensic Sci.* **2020**, *65*, 1102–1113. [CrossRef] [PubMed]
59. Costa, S.T.; Freire, A.R.; Matoso, R.I.; Daruge Júnior, E.; Rossi, A.C.; Prado, F.B. Computational Approach to Identify Different Injuries by Firearms. *J. Forensic Sci.* **2017**, *62*, 361–367. [CrossRef] [PubMed]
60. Mookiah, M.R.K.; Puch-Solis, R.; Nic Daeid, N. Identification of bullets fired from air guns using machine and deep learning methods. *Forensic Sci. Int.* **2023**, *349*, 111734. [CrossRef] [PubMed]
61. Egger, J.; Pepe, A.; Gsaxner, C.; Jin, Y.; Li, J.; Kern, R. Deep Learning—A First Meta-Survey of Selected Reviews across Scientific Disciplines. *PeerJ Comput. Sci.* **2021**, *7*, e773. [CrossRef] [PubMed]
62. O'Keeffe, C.; Champion, S.; Allsop, D. Demonstrating the Effect of Forensic Firearm Countermeasures: Bullet Characteristics Generated Due to Barrel Modifications. *Forensic Sci. Int.* **2015**, *257*, 13–19. [CrossRef] [PubMed]
63. Gagliano-Candela, R.; Colucci, A.P.; Napoli, S. Determination of Firing Distance: Lead Analysis on the Target by Atomic Absorption Spectroscopy (AAS). *J. Forensic Sci.* **2008**, *53*, 321–324. [CrossRef] [PubMed]
64. Law, E.F.; Morris, K.B.; Jelsema, C.M. Determining the Number of Test Fires Needed to Represent the Variability Present within 9 mm Luger Firearms. *Forensic Sci. Int.* **2017**, *276*, 126–133. [CrossRef] [PubMed]
65. Duez, P.; Weller, T.; Brubaker, M.; Hockensmith, R.E.; Lilien, R. Development and Validation of a Virtual Examination Tool for Firearm Forensics. *J. Forensic Sci.* **2018**, *63*, 1069–1078. [CrossRef] [PubMed]
66. Kara, I. Investigation of ballistic evidence through an automatic image analysis and identification system. *J. Forensic Sci.* **2016**, *61*, 775–781. [CrossRef] [PubMed]
67. Gambino, C.; McLaughlin, P.; Kuo, L.; Kammerman, F.; Shenkin, P.; Diaczuk, P.; Petraco, N.D. Forensic surface metrology: Tool mark evidence. *Scanning* **2011**, *33*, 272–278. [CrossRef] [PubMed]
68. Tong, M.; Song, J.; Chu, W.; Thompson, R.M. Fired cartridge case identification using optical images and the congruent matching cells (CMC) method. *J. Res. Natl. Inst. Stand. Technol.* **2014**, *119*, 575–582. [CrossRef] [PubMed]
69. Kamaruddin, S.B.A.; Ghani, N.A.M.; Liong, C.Y.; Jemain, A.A. Firearm recognition based on whole firing pin impression image via backpropagation neural network. In *Proceedings of the International Conference on Pattern Analysis and Intelligent Robotics, Putrajaya, Malaysia, 28–29 June 2011*; IEEE: Piscataway, NJ, USA, 2011; pp. 177–182. [CrossRef]
70. Carriquiry, A.; Hofmann, H.; Tai, X.H.; VanderPlas, S. Machine learning in forensic applications. *Significance* **2019**, *16*, 29–35. [CrossRef]
71. Li, S.Y.; Turner, J.; Golightly, S.; Zelbst, P.; Yu, J. Potential impacts of 3D modeling and 3D printing in firearm toolmark examinations. *J. Forensic Sci.* **2021**, *66*, 2201–2207. [CrossRef] [PubMed]
72. Aggadi, N.; Zeller, K.; Busey, T. Quantifying the strength of firearms comparisons based on error rate studies. *J. Forensic Sci.* **2025**, *70*, 84–97. [CrossRef] [PubMed]
73. Alhijaja, J.A.; Khudeyera, R.S. Techniques and Applications for Deep Learning: A Review. *J. Al-Qadisiyah Comput. Sci. Math.* **2023**, *15*, 114–126. [CrossRef]
74. Gutierrez, R.E.; Prokesch, E.J. The false promise of firearms examination validation studies: Lay controls, simplistic comparisons, and the failure to soundly measure misidentification rates. *J. Forensic Sci.* **2024**, *69*, 1334–1349. [CrossRef] [PubMed]
75. Minzière, V.R.; Gassner, A.-L.; Gallidabino, M.; Roux, C.; Weyermann, C. The relevance of gunshot residues in forensic science. *WIREs Forensic Sci.* **2023**, *5*, e1472. [CrossRef]
76. Giverts, P.; Sorokina, K.; Barash, M.; Fedorenko, V. The use of machine learning for the determination of a type/model of firearms by the characteristics on cartridge cases. *Forensic Sci. Int.* **2024**, *358*, 112021. [CrossRef] [PubMed]
77. Senin, N.; Groppetti, R.; Garofano, L.; Fratini, P.; Pierni, M. Three-Dimensional Surface Topography Acquisition and Analysis for Firearm Identification. *J. Forensic Sci.* **2006**, *51*, 282–295. [CrossRef] [PubMed]
78. Mihna, F.K.H.; Habeeb, M.A.; Khaleel, Y.L.; Ali, Y.H.; Al-saeedi, L.A.E. Using Information Technology for Comprehensive Analysis and Prediction in Forensic Evidence. *Mesop. J. Cybersecur.* **2024**, *2024*, 4–16. [CrossRef]
79. Dorfman, A.H.; Valliant, R. Inconclusives, errors, and error rates in forensic firearms analysis: Three statistical perspectives. *Forensic Sci. Int. Synerg.* **2022**, *5*, 100273. [CrossRef] [PubMed]
80. Murdock, J.E.; Petraco, N.D.K.; Thornton, J.I.; Neel, M.T.; Weller, T.J.; Thompson, R.M.; Collins, E.R. The Development and Application of Random Match Probabilities to Firearm and Toolmark Identification. *J. Forensic Sci.* **2017**, *62*, 619–625. [CrossRef] [PubMed]

81. Monson, K.L.; Smith, E.D.; Bajic, S.J. Planning, design and logistics of a decision analysis study: The FBI/Ames study involving forensic firearms examiners. *Forensic Sci. Int. Synerg.* **2022**, *4*, 100221. [CrossRef] [PubMed]
82. Bird, D.K.; Agg, K.M.; Barnett, N.W.; Smith, T.A. Time-resolved fluorescence microscopy of gunshot residue: An application to forensic science. *J. Microsc.* **2007**, *226*, 18–25. [CrossRef] [PubMed]
83. Izraeli, E.S.; Tsach, T.; Levin, N. Optimizing FEG-SEM combined with an SDD EDX system for automated GSR analysis. *X-Ray Spectrom.* **2013**, *43*, 29–37. [CrossRef]
84. Giverts, P.; Sorokina, K.; Fedorenko, V. Examination of the possibility to use Siamese networks for the comparison of firing pin marks. *J. Forensic Sci.* **2022**, *67*, 2416–2424. [CrossRef] [PubMed]
85. Baruah, S.; Suresh, R.; Govindarajulu, R.B.; Kumar, C.J.; Chanda, B.; Dugar, L.; Saikia, M.J. Feature Extraction and Comparative Analysis of Firing Pin, Breech Face, and Annulus Impressions from Ballistic Cartridge Images. *Forensic Sci.* **2025**, *5*, 62. [CrossRef]
86. Monson, K.L. Accuracy of comparison decisions by forensic firearms examiners. *J. Forensic Sci.* **2023**, *68*, 86–100. [CrossRef] [PubMed]