

Human Examiner Vs AI in Firearm Toolmark Comparison

Komal Yadav¹, Priyanka Negi¹, Nancy Yadav¹, Nitin Tyagi¹, Monika¹, Babita Khan¹, Nihit¹, Alisha Raj¹, Tanvi Anand¹, Shiwansh Pandey¹

¹Department of Forensic Science, School of Basic and Applied Sciences, K.R. Mangalam University, Gurugram, India.

Corresponding author: - Priyanka Negi, priyankasoni.negi@gmail.com

Abstract

Forensic examination of firearms is evolving from microscope comparison towards 3D imaging, computer scoring, and artificial intelligence technologies such as convolutional neural networks (CNNs). The review found that AI tools perform highly accurately in certain applications. Specifically, CNN models for classification of firing pin impressions scored 82.6-92.8% accuracy in controlled tests, while machine learning models for bullet marks showed about 84-86% accuracy. On the other hand, "black box" research shows high, albeit imperfect, accuracy of trained human examiners. In one study 173 qualified examiners completed 8,640 comparisons of bullet and cartridge case marks with extremely low false positive rates (about 0.66-0.93%) and low false negative rates (about 1.87-2.87%).

The review found, AI and humans demonstrate complementary capabilities. Human experts were more capable than computer tools in correctly excluding different source comparisons, while algorithms were better in confirmation of same source comparisons. Examiners performed better in difficult situations of overlapped marks, where some algorithms had difficulties. It appears that AI technologies should assist examiners but not replace them.

On the other hand, there are several important limitations which we have discussed in this review paper, including: Studies based mostly on limited, controlled test sets; hence the performance may decrease in case of new gun models, calibres, mark types etc., that are not included in training. Problems of trustworthiness, explainability, and integration of AI in forensic work remain open. Also, there are some studies based on non-experts or limited fragments of images, which may artificially increase accuracy. We believe that AI should be used as a decision support tool by experienced examiners. Further research should use open accessible datasets and involve practicing forensic examiners.

Keywords: forensic firearm examination; toolmark comparison, convolutional neural networks; black box study; examiner accuracy; decision support; explainability; open datasets.

1. Introduction

Firearms and toolmarks analysis have been a pillar of forensic science for more than one hundred years, providing important evidence for criminal cases with gun violence (Mattijssen et al., 2023; Houck & Siegel, 2015). The basic concept that firearms leave unique and repeatable marks on the bullets and cartridge cases is the basis for matching the crime-scene evidence with guns and ruling out suspected firearms (Monson et al., 2023; Bunch & Smith, 2011). However, there have been growing concerns about scientific bases of the field, subjectivity and the absence of error rates (National Research Council, 2009; PCAST, 2016; Stoney & Stoney, 2015).

The recent developments in technology have changed the approach to the analysis of firearms from a wholly subjective and morphological one into a more objective and evidence-based approach. This includes 3D topography of surfaces, virtual comparison microscopy, scoring algorithms (congruent matching cells scoring system) and artificial intelligence (AI), which includes convolutional neural networks (CNNs) and other machine learning (ML) techniques (Basu et al., 2022; Mattijssen et al., 2023; Zheng et al., 2020; Thompson et al., 2019). This brings up several issues of concern. How accurate are human vs. artificial intelligence examiners in this respect? What are the strengths and weaknesses of all these approaches? Can better results be obtained by a combination of the two approaches than by each one alone? In this paper, we review the most recent results from black box examiner studies, AI algorithm validation tests and human-machine comparison studies to answer these questions (Cuéllar et al., 2022; INTERPOL, 2019; Morrison & Enzinger, 2019).

This review has three interconnected aims. First, it compares how accurately human firearm analysts and AI machine-learning algorithms examine bullets and cartridge cases, drawing on peer-reviewed studies from 2009 to 2026 that use large-scale experiments with human examiners and validated AI systems working with both 2D and 3D data. Second, it explores how human and AI methods might complement one another, checks whether current findings

hold up across studies, and offers evidence-based recommendations on where and how AI should be used in forensic practice. Third, it identifies the most important unanswered questions to guide future research, with a focus on study designs that can robustly test performance, generalizability, and real-world utility in casework.

2. Historical Review of Firearm and Toolmark Examination

Table:1 shows the history of firearm and tool mark examination has changed over time, highlighting the research and what they discovered.

S.No.	Study	Contribution to Firearm Examination	Relevance to present AI/ML research	Citation
1.	Firearm and Toolmark Identification Criteria: A Review of the Literature	Reviewed the standards utilized in firearm and toolmark identification and talked about quality assurance issues in firearm/toolmark examination.	This is an older technique for identifying firearms/marks on tools prior to the current use of 3D/AI methods.	Nichols, 1997.
2.	Cartridge Case Recognition Using Optical Images and the Congruent Matching Cells (CMC) Technique	Applied the Congruent Matching Cells (CMC) computer algorithm to optical images of breech-face impressions on cartridge cases, demonstrating the potential for objective computer-assisted comparison.	This was a significant step towards shifting from visual comparison to computer-based quantitative comparison of firearm marks	Tong et al., 2014
3.	Automatic Comparison and Evaluation of Impressions Left by a Firearm on Fired Cartridge Cases	A 3D automatic matching system with likelihood ratio approach was developed, and its ability to discriminate between matching and nonmatching cartridge cases was studied.	Showed increasing use of 3-D imaging and statistics in forensic firearms analysis.	Riva, F., et al., 2020
4.	A New Method for Bullet Comparison Using Congruent Matching Cells	Developed the concept of CMC to compare 3D bullet lands.	Computer-aided extended quantitative comparison was carried out for cartridge cases and bullets. This helped	Song, 2014

			in developing three-dimensional firearm-toolmark analysis.	
5.	An Improved Algorithm of Congruent Matching Cells (CMC) Method for Firearm Evidence Identifications	Categorized the improved CMC method using correlation data and analyzed both optical and 3D topographic image types.	Contributed to developing automatic and quantitative techniques for comparing firearms marks.	Ott, D., Song, J., Chu, W., & Thompson, R. (2014)
6.	Applying 3D Measurements and Computer Matching Algorithms to Two Firearm Examination Proficiency Tests	CMC algorithm was used for three-dimensional breech face and firing pin impressions made through firearm proficiency examinations and compared computer generated data with those of human examiner.	Significant development towards integration of 3D scanning technology, algorithms, and subjective evaluation by examiners.	Ott, Thompson, and Song (2017).
7.	A Completely Automatic Technique for Cartridge Case Comparison	Developed an open-source automatic comparison of breech face marks on two-dimensional images through automatic selection and statistical analysis.	Introduced more automation in firearms comparison and served as a link between conventional methods of image comparison and AI/ML techniques.	Tai & Eddy, 2018
8.	NIST Ballistics Toolmark Research Database	Developed an open-source database that contains data on bullets and cartridge cases in 2D and 3D format.	Supplied data sets for development and validation of quantitative, automated, and eventually machine learning-based firearm comparison techniques.	Zheng et al., 2020
9.	Assessment of the Likelihood Ratios Using the CMC Approach	Conducted a likelihood ratio test of the match scores in the CMC, allowing an objective evaluation of the firearms evidence.	Preferred the trend toward the application of statistics and probability over subjective comparison.	Song et al., 2020

10.	INTERPOL Review of Forensic Firearm Examination	Evaluated the efficiency of human examiners, computer-based comparison analysis, and any other objective techniques, determining that human examiners and computer-based analysis were efficient in cartridge case comparison.	Helped to determine the importance of examining the human technique, as well as the computer one.	Mattijssen et al., 2021/2022
11.	Siamese Networks for Comparing Firing Pin Impressions: The Opportunities of Their Application Analysis	Consideration of using Siamese neural networks for comparison of firing pin impressions.	Suggests a transition from conventional computer algorithms to firearm identification via machine learning and neural networks.	Giverts et al., 2022.
12	Automated Recognition of Regions of Interest in Cartridge Case Images via Deep Learning	Automatic recognition of regions of interest in cartridge case images before comparison via deep learning.	Highlights the application of deep learning to automate image processing and feature recognition.	Le Bouthillier et al., 2023
13.	Automated Interpretation of Comparison Scores for Firearm Toolmarks on Cartridge Case Primers	Automated system for firearm/toolmark evidence scoring and likelihood ratios developed.	Shows how automated quantitative interpretation systems have been developed to aid in forensic decision-making.	Baiker-Sørensen et al., 2023
14.	Machine Learning for Gun Type Classification Based on Characteristics of Cartridge Cases	Created CNN-based models for firearm type/classification from cartridge cases using Developed CNN-based models for gun type/classification from cartridge case characteristics.	Demonstrates the growing usage of CNN machine learning algorithms for automatic gun classification.	Giverts et al., 2024
15.	Deep Learning Methods for Automatic	Used deep learning methods for classifying fired	Continued application of AI/ML beyond	Guo et al., 2025

	Classification of Fired Bullets of Commonly Used Pistols	bullets based on firearm marks.	casing identification to that of bullet marks.	
16.	Review of Firearm Toolmarks Identification: Advances, Difficulties, and future Perspectives	Technological advancements including 3D topography, deep learning, CMC, automation, and likelihood ratios were analyzed in comparison with traditional comparisons done using microscopy.	Discussion of a recent study on technological advancement in firearm examination and its challenges.	Xin et al., 2026
17.	Automatic Identification of Firearms and Ammunitions Using Artificial Intelligence in Forensic Ballistics	Reviewed various AI techniques like CNN-firing pin identification and machine learning for analysis of bullet markings.	The present phase of development where AI is explored for use as a decision support tool together with a forensic examiner.	Fedorenko, 2026

The transition to 3D imaging and objective analysis:

As an extension of previous 2D imaging research, the discipline has now progressed towards the generation of 3D surface images through techniques such as confocal microscopy and focus variation microscopy. These 3D measurements have enabled more objective, repeatable ways to compare marks, including counting congruent matching cells (CMC) on aligned 3D images, using cross-correlation to quantify texture similarity, and applying mathematical models of surface roughness to describe topography. Algorithmic scoring systems based on these ideas—especially CMC-type methods for breech-face comparisons—were among the first computer-assisted tools to show strong discrimination between same-source and different-source pairs, while still supporting (rather than replacing) examiner judgments.

Introduction to artificial intelligence: -

More recently, machine learning and deep learning have entered firearm examination, offering potential gains in speed, consistency, and quantification. These approaches can automate feature extraction from both 2D and 3D images, screen large reference databases at high throughput, and express the strength of evidence using quantitative likelihood ratios. At the same time, key challenges remain around validating models, interpreting their outputs, ensuring they generalize to new firearms and ammunition types, and integrating them smoothly into laboratory workflows and casework practice.

Evaluation of human performance: black-box studies: -

Black-box studies—where examiners do not know which samples come from the same or different sources—have been used to test real-world human performance in firearm and toolmark comparison. These studies help reveal both the strengths and limitations of human examiners under controlled but realistic conditions, and they provide a benchmark against which new AI-assisted methods can be compared.

FBI/Ames black-box study (Monson et al., 2023): -

The biggest and most robust black-box study undertaken to date involved 173 qualified U.S. firearm examiners who completed 8,640 comparisons of bullets and cartridge cases drawn from particularly challenging materials, including consecutively manufactured firearms and steel-cased ammunition. The study was designed to mirror real casework as closely as possible, using an open-set format in which not every questioned item had a corresponding reference sample, and by including difficult specimens such as consecutively produced barrels and slides, steel cartridge cases, steel-jacketed bullets, and items with minimal aperture drag marks. Examiners reported their conclusions using the AFTE scale (Identification, Inconclusive subcategories A/B/C, Elimination, and Unsuitable), and the analysis employed beta-binomial modelling to account for examiner heterogeneity, avoiding the unrealistic assumption that all examiners share the same error probabilities.

3.2 Key Results: -

Table:2 Shows Examiner Error Rates in (Monson et al. 2023)

Specimen Type	False Positive Rate	95% CI	False Negative Rate	95% CI
Bullets	0.66%	0.31–1.42%	2.87%	1.89–4.26%
Cartridge Cases	0.93%	0.55–1.57%	1.87%	1.16–2.99%

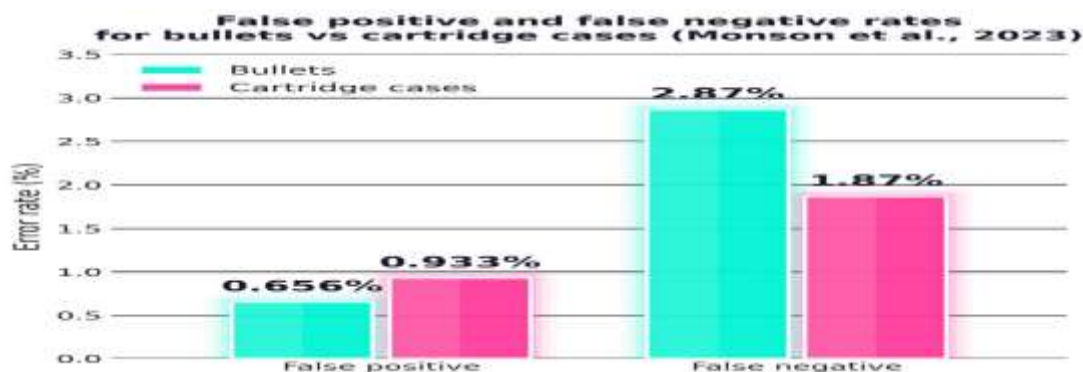


Figure 1: Comparison of examiner error rates: The graph compares the false positive and false negative rates (%) for forensic examinations of bullets versus cartridge cases, based on data by (Monson et al. 2023).

The given below fig2 shows that almost 80% of examiners made no errors across their assigned comparisons, yet errors still appeared in nearly 20% of the individual cases, highlighting that most mistakes were concentrated among a minority of examiners or specific difficult comparisons. When looking at performance metrics, sensitivity—the ability to correctly identify true matches—was moderate, at roughly 74–77%, while specificity—the ability to correctly reject non-matches—was considerably lower, around 34–49%. A notable pattern was that examiners produced more inconclusive results on non-matching pairs, which suggests a conservative decision strategy aimed at avoiding false eliminations or false identifications when the evidence was weak or ambiguous.

(The data mentioned above are shown in the fig2 given below).



Figure 2: Examiner error distribution: ~80% made no errors, ~20% made one or more errors (Monson et al., 2023). Errors in firearm and toolmark examinations were linked to several recurring factors. First, subclass characteristics, manufacturing features shared by groups of firearms, occasionally resulted in false identifications by examiners because these shared traits can resemble individualizing marks. Second, poor-quality marks, such as misshapen projectiles, deteriorated cartridge cases, and incomplete impressions, provided limited usable information and increased the risk of incorrect conclusions. Third, cognitive factors played a role, including context bias, fatigue, high workload, and differences in each examiner's personal threshold for making an identification. Finally, laboratory policies also influenced outcomes; for example, some examiners became accustomed to policies that discouraged issuing Eliminations based on individual characteristics, which may have affected how they interpreted borderline evidence.

Table3: Shows Summary of Recent Open-Set Examiner Studies

Study	Firearms	Examiners	Comparisons	FPR	FNR
Baldwin et al. (2014)	25 Ruger	218	3,270 (CC)	0.94%	0.37%
Kerkhoff et al. (2015)	10 Various	11	341 (CC) + 55 (B)	0%	0%
T. Smith et al. (2016)	8 Various	31	693 (CC) + 995 (B)	0.14% (CC)	0.43% (CC)
Kerkhoff et al. (2018)	139 Sig Sauer/Glock	10	344 (CC)	0%	0%
J. Smith (2020, 2022)	35 Various	74	7,420 (B)	0.08%	0.16%
Law & Morris (2021)	20 Various	17	340 (CC)	0.28%	0%
Best et al. (2022)	10 Thompson/Centre	44	660 (B)	0.46%	1.82%
Monson et al. (2023)	234 (Beretta/Ruger/Jimenez)	173	8,640 (CC + B)	0.66–0.93%	1.87–2.87%

Note: CC = cartridge cases; B = bullets. Sources: Baldwin et al.

(2014); Kerkhoff et al., 2015, 2018; Smith et al., 2016; Smith and Monson, 2020, 2022; Law and Morris, 2021; Best et al., 2022; Monson et al., 2023). All these works show low but non-zero error rates, higher difficulty to exclude different source comparisons, especially in the case of subclasses or poor quality of marks (Mattijssen et al., 2023; Monson et al., 2023; Hicks & Buckleton, 2016). 4. Reliability of Artificial Intelligence Assisted Firearm Comparison by Accuracy and Error Assessment - 4.1 CNNs in Firing-Pin Impression Classification Fedorenko et al. (2024) proposed a convolutional neural network (CNN) model for multiclass classification of fired cartridge case firing pin impressions. The study used 40 unique guns' impressions with heavy data augmentation (including cloning and small image manipulations).to augment the original small-scale dataset that only had four images per weapon type. The accuracy of the CNN was approximately 93% for the top ranked classification result. The accuracy for the three top-ranked predictions was increased to ~97–98% (Fedorenko et al., 2024). Experimental results show that CNNs are able to learn discriminative visual features from firing pin impressions even from a small amount of original source images.This improvement agrees with the known effect of data augmentation in increasing the effective size of training datasets and reducing overfitting in deep-learning models (Goodfellow et al., 2016).

However, the reliability of these results must be interpreted cautiously. The underlying dataset was small, consisting of only four original images for each of the 40 firearm classes, and the extensive use of augmented copies may have caused the model to encounter highly similar images during training and testing. Consequently, the reported performance may not fully represent the model's ability to classify genuinely new firing-pin impressions. Data augmentation can improve performance when the applied transformations reflect realistic variation, but it cannot substitute for independent samples that capture differences in firearms, ammunition, firing conditions, and image acquisition. Prior studies have likewise noted that CNNs might not generalize well in the face of insufficient training data and when there are features in the test cases that are not present in the augmented training set. Moreover, the

experiment was done under controlled settings and the generalizability of the technique to unseen firearms, types of ammunition, firing conditions, and quality of casework images has not been fully proven. Thus, even though there are impressive rates of 93% top-1 accuracy and 97-98% top-3 accuracy achieved, more validation studies are needed before considering CNN-assisted classification of firing pins dependable for forensic firearm comparison.

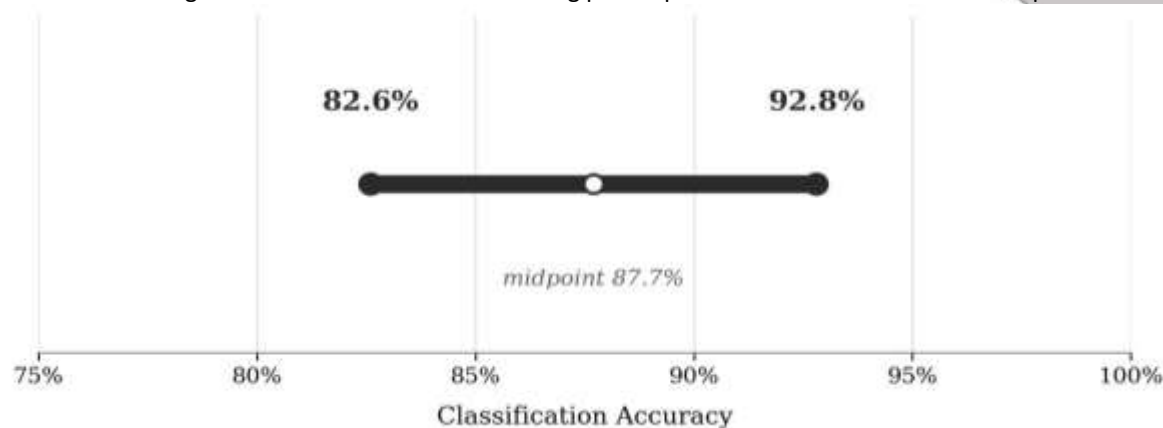


Figure3: CNN firing-pin recognition classification accuracy range 82.6%-92.8% (Fedorenko et al., 2024).

4.2 Analysis of Firing Marks Using Siamese Network:

Previous researchers also found that CNN may fail to generalize due to lack of training samples and presence of features in the test images which do not appear in the augmented training set. In addition, the test was carried out under controlled circumstances and the ability of the method to generalize for unknown firearms, ammunition types, shooting conditions, and image quality of casework has not yet been established. Therefore, despite obtaining very high accuracies of 93% in top-1 and 97-98% in top-3, more validation studies need to be conducted before considering the use of CNN in firearm forensics reliable. This is important because conventional firearm examination may involve examiner judgment when determining whether the observed characteristics are sufficient to support an identification, creating concerns about subjectivity and examiner-to-examiner variation. Quantitative systems with predefined performance measures may therefore improve consistency and transparency in forensic comparison, although their error rates must be established through appropriately designed validation studies (Hicks et al., 2016). The study also demonstrated the potential of Siamese architectures for forensic applications in which only a limited number of images are available for each firearm. Instead of learning a separate class for every firearm, the network learns a similarity function between pairs of images, making it potentially suitable for one-shot or few-shot learning situations (Goodfellow et al., 2016). However, there may be limits in terms of how much variation can be captured using these cloned images to augment the data set. More experiments need to be done with new impressions, unknown weapons, different types of ammunition, varied firing conditions, and casework quality images in order to find out whether the same accuracy rates apply elsewhere beyond the experiment data set. Thus, Giverts et al. (2022) provided an important early demonstration that Siamese networks can be used to compare firing-pin marks and may contribute to a more objective firearm-examination framework, but broader external validation is necessary before routine forensic deployment.

(The data mentioned in the above points are shown in the graph given below).

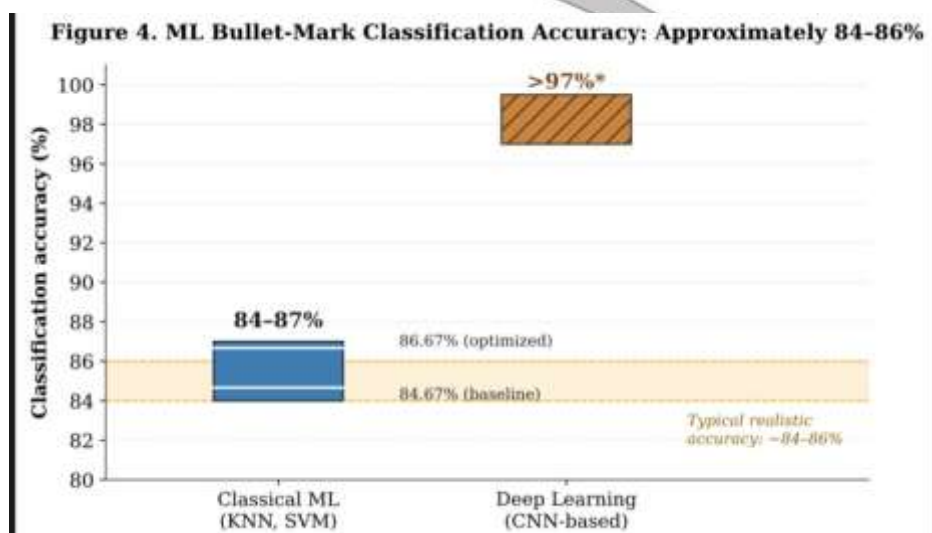


Figure 4: ML bullet-mark classification accuracy: approximately 84–86% (Mattijssen et al., 2023)

4.4 Feature-Based Likelihood-Ratio Systems

In addition to classification-based approaches, recent research has developed feature-based likelihood-ratio (LR) systems for the quantitative comparison of fired cartridge cases. (Basu, Bolton-King and Morrison, 2022) investigated a system that extracted numerical features from three-dimensional images of cartridge-case bases and used those features to calculate likelihood ratios. The feature sets included central moments, Legendre moments, Fourier concentric-circle features, and Zernike moments. Comparisons were conducted using features extracted from the firing-pin impression, the breech-face region, and the complete region of interest, including the firing-pin impression, breech-face marks, and flowback where present. The study used a database containing 10 cartridge cases fired from each of approximately 300 firearms, with 297 firearms represented in the released database.

The system statistically modelled the similarity distributions for same-source and different-source comparisons and used those distributions to produce calibrated likelihood ratios. In this context, the LR expresses how much more probable the observed similarity is under the proposition that the compared cartridge cases originated from the same firearm than under the proposition that they originated from different firearms. This approach differs from a simple identification or exclusion decision because it provides a graded measure of evidential value rather than a binary conclusion. The results showed a substantial separation between same-source and different-source comparisons, indicating that the extracted features contained useful information for discriminating between cartridge cases. Of the feature sets tested, the most successful set was that of Zernike-moment magnitudes and phases, especially when calculated on the entire region of interest rather than from the firing-pin or breech-face region alone.

The use of calibrated likelihood ratios is significant because it supports more transparent and statistically interpretable reporting of forensic evidence. Instead of stating only that two cartridge cases are “identified” or “not identified” as originating from the same firearm, an LR system can communicate the relative strength of the evidence while allowing the fact-finder to consider that evidence alongside other information. Such systems are consistent with broader developments in forensic science that promote probabilistic evaluation and the reporting of evidential strength through likelihood ratios (Morrison, 2020; Berger & Slooten, 2020; Evett & Weir, 1998). However, the findings should not be generalized without qualification. The database consisted primarily of firearms from a specific class, namely semiautomatic pistols firing 9 mm centre-fire Luger-type ammunition with particular firing-pin and breech-face characteristics. Therefore, additional validation is required for different firearm designs, calibres, ammunition types, and manufacturing conditions. A further limitation is the requirement for specialized three-dimensional imaging equipment capable of capturing the microscopic topography of cartridge-case surfaces. This equipment may be costly, technically demanding, and unavailable in many forensic laboratories. In addition, although feature-based LR systems can provide objective and reproducible quantitative information, they should be regarded as decision-support tools rather than automatic substitutes for trained firearm examiners. They need to be appropriately validated, calibrated, verified and correctly interpreted within the context of existing forensic examination procedures. Hence, (Basu et al. 2022) showed that feature based LR system, especially those using

Zernike-moment feature extracted from full cartridge case area, could give promising quantitative approach to the firearm comparison, yet their application in forensics would depend on representative data sets and proper imaging.

4.5 Automated Interpretation of Comparison Scores

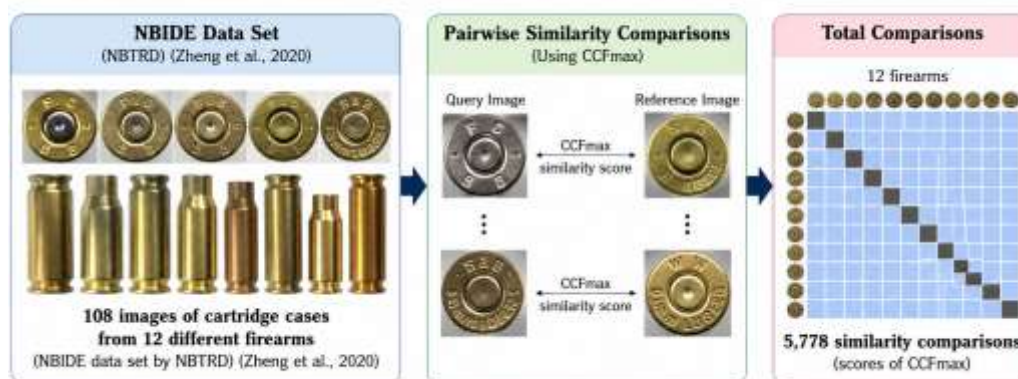
Baiker-Sørensen et al. (2023) suggested a systematic approach to the interpretation of comparison scores and strength of evidence evaluation in toolmark examination using firearms. The method was designed for marks commonly observed on the primers of fired cartridge cases, including aperture-shear striations, breech-face impressions, and firing-pin impressions. For each comparison, numerical similarity scores were generated from these different mark types and then evaluated using reference distributions derived from known-match and known-non-match samples. Known-match comparisons involved test fires produced by the same firearm, whereas known-non-match comparisons involved test fires produced by different firearms.

5. Direct Human vs Machine Comparisons

5.1 Cuéllar et al. (2022): Human and Machine Similarity Judgments

This experiment made a direct comparison of human similarity ratings (scale 0-100) to machine algorithm (cross-correlation approach of Tai & Eddy) similarity ratings for 2D images of breech faces (Cuéllar et al., 2022).

Methods:



Step 1: Input Data Set (108 Cartridge Cases → 12 Firearms)

Step 2: Comparison Engine (Pairwise Matching via CCFmax)

Step 3: Output Matrix (5,778 Unique Comparisons)

Comparison calculation:

$$\text{Total combination} = \frac{n(n-1)}{2}$$

For n = 108

$$\frac{108(107)}{2} = 5,778$$

Total resulting comparisons: 5,778 similarity scores generated across 12 firearms.

Key Findings:

Table4: shows HGuman vs Algorithm Performance (Cuéllar et al., 2022)

Metric	Humans (Novices)	Algorithm
Distinguishing matches from non-matches (LM vs HN)	Superior (p < 0.01)	Inferior (overlap in distributions)
Assessing similarity within groups (HM vs LM; HN vs LN)	Inferior	Superior
Consistency across runs	High	Perfect (deterministic)

Interpretation:

It appears that humans have outperformed algorithms in matching and non-matching differentiators while algorithms have performed well in similarity grading within the match and non-match groups. This indicates that a combination of humans and algorithms can be used to identify match and non-match and use algorithms to grade similarity (Cuéllar et al., 2022; Mattijssen et al., 2023; Morrison & Enzinger, 2019).

Figure 5. Human vs. Algorithm Similarity Score Distributions

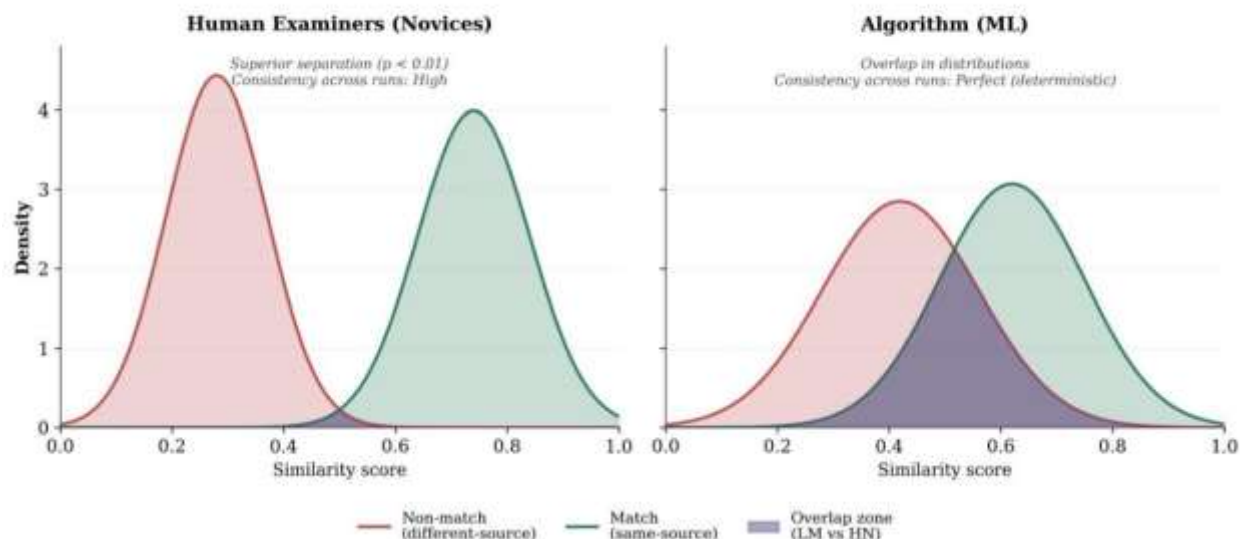


Figure:5 Human vs algorithm similarity score distributions (Cuéllar et al., 2022).

5.2 Complementary Strengths Summary

Table:4 Given table shows Human vs AI Strengths and Weaknesses

Capability	Human Examiners	AI/ML Algorithms
Excluding different-source comparisons	Superior (higher specificity)	Inferior (more false positives in some studies)
Confirming same-source comparisons	Inferior (more inconclusive)	Superior (higher sensitivity in controlled tests)
Handling overlapping/ambiguous marks	Superior (contextual reasoning)	Inferior (struggles with low-quality or damaged marks)
Processing speed and throughput	Inferior (manual, time-intensive)	Superior (automated, high throughput)
Consistency and repeatability	Inferior (inter-examiner variability)	Superior (deterministic outputs)
Explainability and courtroom testimony	Superior (can articulate reasoning)	Inferior (black-box CNNs lack transparency)
Generalization to new firearms/ammunition	Superior (trained on diverse casework)	Inferior (limited training datasets)
Resistance to cognitive bias	Inferior (vulnerable to contextual information)	Superior (blind to case context)

Sources: Cuéllar et al. (2022); Dror and Charlton (2006); Mattijssen et al. (2023); Monson et al. (2023); Fedorenko et al. (2024); Basu et al. (2022); Hicks and Buckleton (2016); National Research Council (2009).

6. Integrated Human–AI Performance Comparison

However, human examiners demonstrated low error rates, as bullets were revealed to have a false positive rate (FPR) of 0.66% and false negative rate (FNR) of 2.87%, while cartridge cases had FPR of 0.93% and FNR of 1.87% (Monson

et al., 2023; Smith & Monson, 2022). In contrast, artificial intelligence applications provided accuracy rate of 82.6–92.8% when classifying firing pin marks using CNNs, that is equal to 7.2–17.4% error rate (Fedorenko et al., 2024; Giverts et al., 2022). Although these findings demonstrate significant differences in the performance, their direct comparison is impossible due to the different methodologies applied in human and AI studies. (the below given fig6 shows human examiner error rates vs AI classification error ranges.)

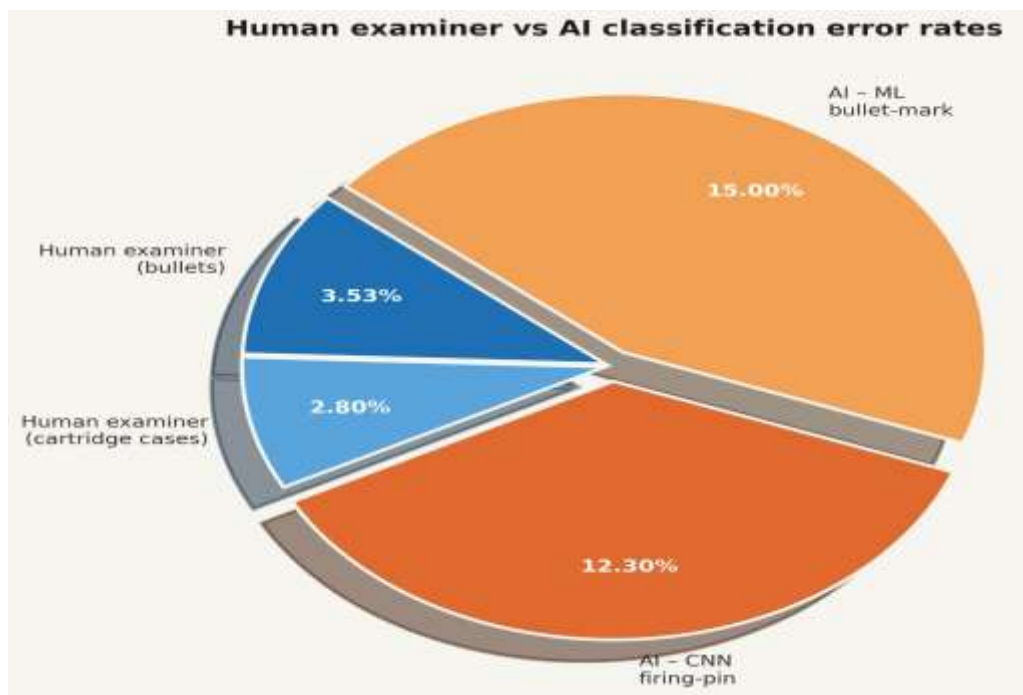


Figure6: Human examiner error rates vs AI classification error ranges.

Critical note on comparing human and AI metrics

It is important to recognize that the performance metrics reported for human examiners and AI systems are not directly comparable because they come from fundamentally different tasks and evaluation frameworks. Human false-positive and false-negative rates are derived from open-set forensic decisions using the Identification/Elimination/Inconclusive framework, as seen in studies like (Monson et al. 2023), which mirror real casework where some questioned items may have no true match in the reference set. In contrast, many AI error rates come from closed-set multiclass classification experiments, such as those summarized by (Basu et al. 2022), where the model must choose among a fixed set of known source classes and every test item is assumed to belong to one of those classes. Because the tasks, datasets, and evaluation designs differ in these ways, simple side-by-side comparisons of “accuracy” or “error” can be misleading without careful harmonization of methods and metrics, a point emphasized in methodological guidance from (Hicks & Buckleton 2016) and (OSAC 2021).

6.2 Combined Performance Metrics

Table5 : Provides information about Representative Performance Metrics

Study	System	Task	Dataset Size	Metric	Value
Monson et al. (2023)	Human examiners	Open-set comparison	173 examiners, 8,640 comparisons	FPR (bullets)	0.66%
Monson et al. (2023)	Human examiners	Open-set comparison	173 examiners, 8,640 comparisons	FNR (bullets)	2.87%

Fedorenko et al. (2024)	CNN	Firing-pin classification	40 firearms	Top-1 accuracy	93%
Giverts et al. (2022)	Siamese CNN	Firing-pin comparison	300 firearms	Accuracy	High (exact % not stated)
Mattijssen et al. (2023)	KNN ML	Bullet-mark classification	Multiple firearms	Accuracy	84–86%
Basu et al. (2022)	Feature-based LR	Cartridge-case LR	297 firearms	Discrimination	Strong separation
Cuéllar et al. (2022)	Humans (novices)	Similarity rating	12 firearms, 40 pairs	Match/non-match discrimination	Superior to algorithm
Cuéllar et al. (2022)	Algorithm (CCFmax)	Similarity scoring	12 firearms, 40 pairs	Within-group similarity	Superior to humans

Sources: Monson et al. (2023); Fedorenko et al. (2024); Giverts et al. (2022); Mattijssen et al. (2023); Basu et al. (2022); Cuéllar et al. (2022); Zheng et al. (2020).

7. Limitations and Research Gaps

7.1 Limitations of Human Examiner Studies

There exist different limitations in studies on human firearm examiners. First, in some studies, the subjects are more experienced and competent compared to the examiners conducting their analysis in real-world cases, making the outcomes seem much better than in real life situations (Monson et al., 2023; Peters et al., 2022). Secondly, the majority of studies are done in controlled or artificial laboratory settings. It is possible that the examiner does not have access to the firearm itself or to the whole case, making his or her work different from the one done during the real investigation (Kerkhoff & Mattijssen, 2021). There are also doubts about the use of the AFTE conclusion scale due to the categorical nature of its results, while the evidence has a continuous level of similarity (AFTE, 2013; Stoney & Stoney, 2015).

Variation among examiners is another significant problem. Examining the same evidence may yield different results from the examiners, and research has proven that just a few examiners will commit many more errors and thus need proper training and control (Monson et al., 2023; Hicks & Buckleton, 2016). Human examination may also prove to be problematic if the striations on the bullet cartridges or the cartridges themselves are not distinct enough, faded or damaged. In this regard, it is likely for the examiners to struggle with excluding potential matches among firearms (Law & Morris, 2021; Best et al., 2022). It is clear that human expertise plays a critical role in the field, but there are factors that could affect the performance of the examiners.

7.2 Limitations of AI Studies

There were positive outcomes of using artificial intelligence in gun examinations; however, there are some problems with previous research. One of these problems is the absence of diversity in datasets. Many studies use a fairly small amount of firearms, even less than 100, which can be quite different from the variety of firearms used in real casework (Fedorenko et al., 2024). Moreover, in some studies, the calibre or type of ammunition used was only one, and it cannot be known for sure that the same performance would be achieved with other combinations (Ratté et al., 2022). Besides, laboratory test fires were frequently used in the experiments instead of real-world casework samples, because real evidence could include much more damage, contamination, corrosion or other factors which would make identification more difficult (Basu et al., 2022). All those factors could contribute to the problem of overfitting, where the AI system performs excellent on its own data, but shows poor performance with new evidence (Mattijssen et al., 2023; Goodfellow et al., 2016).

Another issue to be considered here is domain shift. The problem is that an AI system trained on images/data collected through one type of scanning device may not behave the same way when images/data are collected through another scanning device (van der Ham et al., 2020). Likewise, a new firearm design, calibre, or even ammunition type not present in the training set could cause problems (Thompson et al., 2021). Furthermore, other real-world issues like corrosion, fouling, and deformed bullets or cartridge cases might not always be adequately represented in the training data (Bolck & Hicks, 2019).

Issues of explainability and interpretability add another dimension. There are some instances where deep-learning techniques, such as CNNs, are considered “black-box” models as the explanation for arriving at a specific conclusion

can sometimes become complex (Goodfellow et al., 2016). Such an issue may pose complications when AI-inferences have to be analyzed in court. While LR approaches based on features offer more clarity, they need a higher level of statistical knowledge and probability (Aitken & Taroni, 1995; Morrison, 2020). Acceptance of the technology by examiners would therefore be contingent not only on accuracy but also on correct validation. (Mattijssen et al., 2023; OSAC, 2021).

Another key research gap lies in validation. Numerous pieces of research validate the effectiveness of AI technology using cross-validation techniques or methods of testing within one data set, rather than testing AI algorithms using completely independent data sets acquired from other sources (Basu et al., 2022). Validation outside the original data set is essential for determining the possibility that the system will perform well in the presence of new evidence. The lack of research related to external validation and benchmarking makes direct comparison of performances of different AI technologies hard (Hicks & Buckleton, 2016; PCAST, 2016)

7.3 Challenges to Integration

Integration of the practical application of AI in firearms analysis raises many integration-related issues as well. AI algorithms should be integrated into the existing work process of forensic labs without adding unnecessary steps. Thus, compatibility with LIMS, virtual comparison microscopy system, and other existing forensics technologies becomes crucial (Knowles et al., 2022; NIJ, 2022).

The next issue that needs to be addressed is the problem of automation bias, which means placing undue confidence in the findings of an algorithm and being less critical about the evidence or its outcome (Dror & Charlton, 2006; Evett et al., 2000). This is why adequate training is required in order to make sure that examiners know what an AI system can do and cannot do and when its findings should not be taken seriously (INTERPOL, 2019; OSAC, 2021). The other issue that requires some attention is that of legal admissibility. The court will check the scientific basis of an AI system and the human involvement in the process of its results' interpretation (PCAST, 2016; Stoney & Stoney, 2015).

Moreover, ethical questions about biased training databases, proprietary algorithms, and liability for any mistakes should be considered as well. When the database that serves as a basis for training an AI system is not diverse enough, it is not certain that the system will be effective when processing various types of evidence. Proprietary algorithms may complicate independent verification of results due to the lack of access to the full algorithm. Consequently, it is crucial to establish clear liability in case of AI use in forensic decisions (National Research Council, 2009; Hicks & Buckleton, 2016).

8. Suggestions for Practice and Research

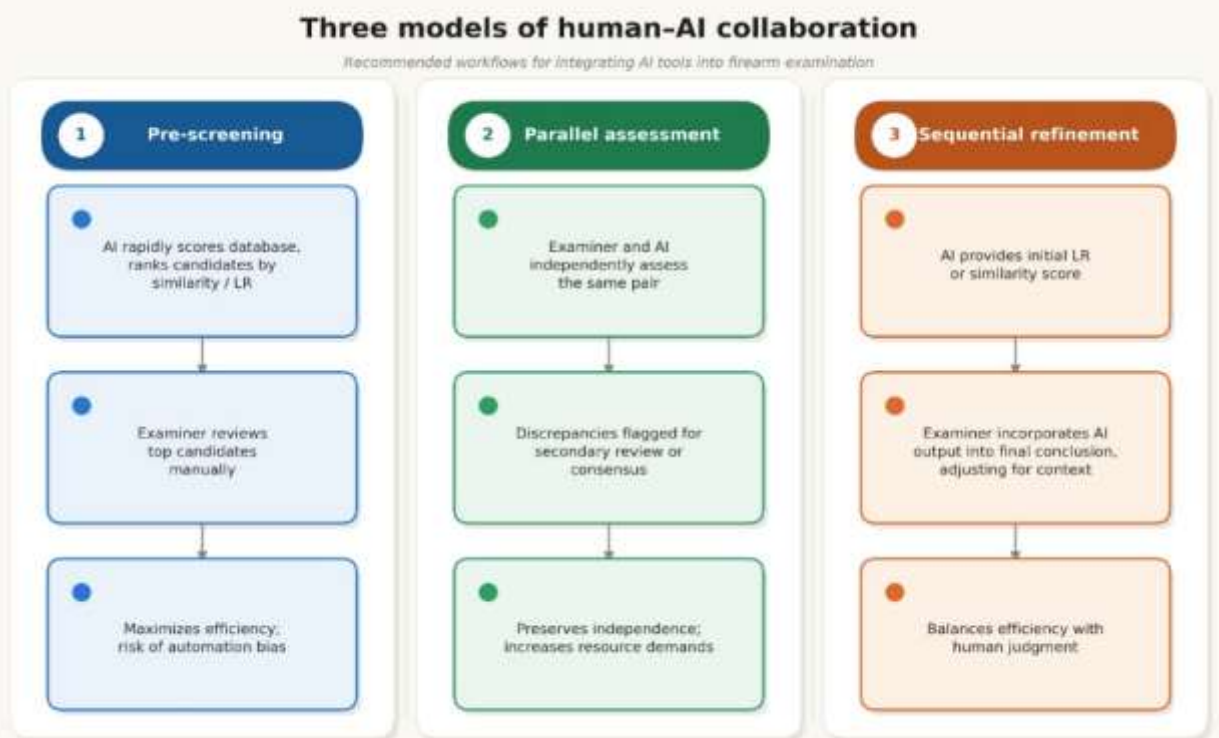
8.1 AI as Decision Support, Not Replacement

According to the findings of the research conducted, it can be concluded that AI can be used as a decision-support aid in firearm examination and not as a substitute for the human expert. AI will be able to provide additional information to the experts when processing huge volumes of information, finding similarities, and providing additional information to be taken into account while examining. But at the same time, it should be noted that the final decision should always include an expert who is familiar with the case and can consider such factors as the quality of evidence, circumstances of the case, and the limitation of the AI system itself.

Workflows suggested:

Figure7: Three modes of interaction between human and AI. (Zheng et al., 2020; NIJ, 2022), (Dror& Charlton, 2006), (Cuéllar et al., 2022), (Hicks & Buckleton, 2016), (Basu et al., 2022; Berger & Slooten, 2020), (Morrison, 2020)

8.2 Priorities for Research



8.2.1 Large-Scale and Operationally Relevant Validation Studies

In future, studies should aim at carrying out larger validation studies that involve collaboration between various forensic laboratories. This would mean having a variety of firearms, calibres and types of ammunition in order to test how the AI system performs in situations that are similar to real forensics investigations (INTERPOL, 2019; OSAC, 2021). It is also important to have specimens from real forensics investigations that are harder to analyse, for example, deformed bullets and corroded cartridge cases (Law & Morris, 2021). Such studies would help determine whether AI systems can maintain reliable performance when dealing with the variation normally found in real investigations.

8.2.2 Expansion of Open Datasets

The development of larger and more diverse open datasets is another important research priority. Existing datasets should be expanded to include a greater number of firearms and different types of ammunition. Some of the larger datasets contain over 300 firearms, demonstrating that it is possible to develop datasets with a higher level of diversity (Zheng et al., 2020; NIST, 2020). It will be necessary in future to include different conditions of use of the firearms and variations in the manufacturing of the components, along with ammunition of various manufacturers and brands (Kerckhoff et al., 2018; Thompson et al., 2021). It would also facilitate testing and comparison of various AI methods by other researchers.

8.2.3 Development of Explainable AI Methods

AI methods need to be created that allow the output generated by these techniques to be comprehensible by forensics experts. One way of doing this is by identifying and selecting the features that have the most impact on the output of the model. This will enable the tester to know which information is being compared by the AI. (Fedorenko et al., 2024). Another good technique to consider would be that of utilizing CNN methods for feature extraction in combination with LR models (Basu et al., 2022; Morrison & Enzinger, 2019).

8.2.4 Human-in-the-Loop Experiments

Future studies need to investigate how AI impacts the work of human firearm examiners instead of analysing the performance of humans and AI as two individual entities. The study of human-in-the-loop experiment will help to compare the performance of the examiner alone and with an AI decision support system (Cuéllar et al., 2022). There should be measures of variables such as false-positive rate (FPR), false-negative rate (FNR), sensitivity, specificity, examiner confidence, and examination time (Monson et al., 2023; Peters et al., 2022).

8.2.5 Standardised Validation Frameworks

There is also a need for standardised validation frameworks for AI-based firearm examination. Scientists may be able to examine methods that have been previously utilized for other forms of forensic analysis such as voice comparison and DNA mixture evaluation. (Hicks & Buckleton, 2016; PCAST, 2016). Maintaining independence between training, calibration, and validation datasets is extremely vital. Using an independent dataset to conduct the validation will help avoid the overestimation of system effectiveness, as well as obtain a more accurate evaluation of the performance of the AI system on new data (OSAC, 2021).

8.2.6 International Collaboration

International collaboration might prove quite useful in the process of enhancing the development and validation of AI for forensic firearms examination. Such organizations as INTERPOL, ENFSI, and other forensic institutions might participate in developing and standardizing validation processes within different laboratories and across different countries (INTERPOL, 2019; Mattijssen et al., 2023). International sharing of data sets, research results, best practices, and ethical considerations would allow researchers to use a wider variety of evidence and minimize unnecessary duplication of research (NIST, 2020; NIJ, 2022).

9. Discussion

9.1 Summary of Major Findings

This review combines the findings of various studies on the comparative analysis of firearms and the performance of humans and artificial intelligence (AI). The available research shows that both approaches have strengths as well as limitations. Studies involving human examiners have generally shown that there are minimal but positive error rates. For black-box studies, false positive rates (FPR) have been shown to be about 0.66-0.93%, and false negative rates (FNR) have been shown to be about 1.87-2.87% (Monson et al., 2023; Smith & Monson, 2022). The research has also shown that 80% of examiners did not err in particular test situations, whereas a greater number of the errors made have been found to be related to 20% of the subjects, thus stressing the significance of training and quality assurance (Peters et al., 2022). Human examiners have proven themselves to be highly sensitive in terms of decision making, although specificity has been known to be less in certain cases because the examiner is always cautious in ruling out firearms in particular cases (Kerkhoff & Mattijssen, 2021).

Similarly, AI-based techniques have been found to show positive results in firearms comparison, but their performance is highly dependent on the quality of training data. In some research articles, it is mentioned that the effectiveness of human examiners in rejecting non-matches and efficiency of AI systems in matching similarities is much higher (Cuéllar et al., 2022; Fedorenko et al., 2024). The efficiency of human experts in analyzing complex marks and the quickness of artificial intelligence in making large numbers of comparisons are higher (Mattijssen et al., 2023; Dror & Charlton, 2006; Law & Morris, 2021). Advanced techniques, such as likelihood ratio systems and Siamese neural networks, have opened additional opportunities for automation of comparison and identification of similarities (Basu et al., 2022; Berger & Slooten, 2020; Goodfellow et al., 2016). In general, there are enough sources proving that AI-based analysis together with human expertise can be helpful in firearms comparison, identification of similarities, and large databases screening (Cuéllar et al., 2022; Morrison & Enzinger, 2019; Zheng et al., 2020).

The comparison between strengths and weaknesses also points out some other issues that need to be considered further. Human examiners can be more effective while working with challenging marks, but the efficiency of AI systems can be influenced negatively in case the marks used are weak (Law & Morris, 2021). In addition, variations in the performance of human examiners point out the necessity for their training and quality control, especially in terms of sources of errors on their part (Monson et al., 2023; Hicks & Buckleton, 2016). The use of likelihood ratios for expressing the strength of evidence is being increasingly discussed recently because they offer a way to quantify the evidence (Aitken & Taroni, 1995; Morrison, 2020).

In relation to AI, there is a need to enlarge and diversify the training datasets available. Increasing the amount of data that is used through expansion of databases like the NBTRD can allow for more diverse datasets to be used in developing and evaluating the AI (Zheng et al., 2020; NIST, 2020). AI systems need to also be validated using specimens that have an operational relevance, and not those from a laboratory setting (OSAC, 2021; PCAST, 2016). Explainability methods can include feature importance analysis and saliency maps to show which features lead to an AI output (Goodfellow et al., 2016; Fedorenko et al., 2024). Moreover, there is a need for validation protocols and standards so as to facilitate comparison of results of various AI systems (Hicks & Buckleton, 2016; INTERPOL,

2019) The success is conditional upon the application of AI in forensic laboratories. The convenience of interaction can help minimize automation bias and encourage the examiner to analyze the results obtained through AI technology rationally rather than to accept them blindly (Dror & Charlton, 2006; Knowles et al., 2022). The application of AI technologies must be aligned with the workflow and systems that already exist in the laboratory, such as LIMS or virtual comparison microscopy platforms (NIJ, 2022; Basu et al., 2022). Moreover, there is an urgent need for guidelines regarding the application of AI-assisted results in forensic reports and testimony in court (PCAST, 2016; National Research Council, 2009).

9.3 Future Outlook

The identification of forensically-related firearms in the future will probably utilize both the proficiency of human analysts and the computer-based process. According to the literature reviewed, AI can be used in various stages such as automated comparison, database screening, similarity evaluation, and numerical evaluation (Mattijssen et al., 2023; Morrison & Enzinger, 2019), while humans are required in the interpretation of complicated evidence and considering the forensic context. There is a range of issues that should be investigated further.

One such area would include developing open and diverse datasets. The availability of resources such as the NBTRD needs to continue increasing to cover a greater number of firearms, cartridges, manufacturing variations, and evidentiary conditions that may arise in actual cases (Zheng et al., 2020; NIST, 2020). Testing and validation procedures have to be conducted in accordance with the standard process by multiple laboratories, rather than by a single laboratory or dataset (OSAC, 2021; Hicks & Buckleton, 2016). This, on the other hand, calls for making AI systems more interpretable.

The human factors would remain crucial. The examiners would require proper training in order to comprehend the capabilities and limitations of the AI and to overcome automation bias, where individuals excessively depend on the result from the algorithm (Dror & Charlton, 2006; INTERPOL, 2019).

10. Conclusion

The development of artificial intelligence (AI) has introduced new possibilities in forensic firearm and toolmark examination. It is evident from the findings presented in this paper that there are AI techniques that can be used for comparison of bullets and cartridge casings, identification of similarities, screening of large databases and provision of numerical values such as likelihood ratio. The use of these techniques can enhance efficiency, consistency and reduce the need for manual examination. However, the literature suggests that the effectiveness of AI depends highly on the quality and quantity of the training data. Therefore, AI results should not be considered reliable simply because a system performs well on a particular laboratory dataset.

Human firearm examiners continue to have an important role in the examination process. Their experience and professional judgement are particularly valuable when evidence is damaged, deformed, unclear, or contains complex and overlapping marks. There are additional factors about the cases that may be considered by humans in their examinations and explain their rationale for making such findings. However, human reasoning is susceptible to various experiences and thought processes of examiners. This clearly demonstrates that human and AI examinations each have their own advantages and disadvantages.

On balance, the future of firearms and tool mark examinations will definitely benefit from the collaboration of both humans and artificial intelligence rather than picking either one. In this case, artificial intelligence could assist in conducting the analysis quickly, providing potential similarities, and searching databases. It is up to the qualified examiner to make an evaluation and form conclusions.

For AI-assisted firearm examination to become more widely accepted, further research is still required. Larger and more diverse open datasets, independent validation using realistic casework samples, standardised performance measures, explainable AI methods, and proper examiner training will be important. Guidelines need to be provided on how to record, report, and demonstrate the outcome of using AI in a court of law. AI can eventually be an effective tool in modern firearm examination when used together with due research and verification processes.

Recommendations

Research in the future needs to consider more diversified data sets, incorporating various types of firearm, calibers, ammunition used, conditions of the weapon used, and cases. There is a need to focus more on explainable AI, so that examiners get the idea behind the workings of the AI to reach the conclusions.

AI is there to assist the examiner, not replace him, with appropriate training so that there is no automation bias.

There is a need for forensic laboratories to establish clear standard operating procedures in the application of AI, including validation, documentation, quality control, data protection, and human supervision. Researchers, forensic laboratories, universities, and organizations such as INTERPOL, ENFSI, OSAC, and NIST can work together towards this end.

References

1. Mattijssen, E. J. A. T., Witteman, C. L. M., Berger, C. E. H., Zheng, X. A., Soons, J. A., & Stoel, R. D. (2021). Firearm examination: Examiner judgments and computer-based comparisons. *Journal of Forensic Sciences*, 66(1), 96–111. <https://doi.org/10.1111/1556-4029.14557>
2. Monson, K. L., Smith, E. D., & Peters, E. M. (2023). Accuracy of comparison decisions by forensic firearms examiners. *Journal of Forensic Sciences*, 68(1), 86–100. <https://doi.org/10.1111/1556-4029.15152>
3. Baiker-Sørensen, M., Alberink, I., Granell, L. B., van der Ham, L., Mattijssen, E. J. A. T., Smith, E. D., Soons, J., Vergeer, P., & Zheng, X. A. (2023). Automated interpretation of comparison scores for firearm toolmarks on cartridge case primers. *Forensic Science International*, 353, 111858. <https://doi.org/10.1016/j.forsciint.2023.111858>
4. Ott, D., Song, J., Chu, W., & Thompson, R. (2014). Fire cartridge case identifications using optical images and Congruent Matching Cells (CMC) method. *Journal of Research of the National Institute of Standards and Technology*, 119, 575–582. <https://doi.org/10.6028/jres.119.023>
5. Ott, D. B., Thompson, R. M., & Song, J. (2017). Applying 3D measurements and computer matching algorithms to two firearm examination proficiency tests. *Forensic Science International*, 271, 98–106. <https://doi.org/10.1016/j.forsciint.2016.12.014>
6. Tai, X. H., & Eddy, W. F. (2018). A fully automatic method for comparing cartridge case images. *Journal of Forensic Sciences*, 63(2), 440–448. <https://doi.org/10.1111/1556-4029.13577>
7. Giverts, P., Sorokina, K., & Fedorenko, V. (2022). Examination of the possibility to use Siamese networks for the comparison of firing pin marks. *Journal of Forensic Sciences*, 67(6), 2416–2424. <https://doi.org/10.1111/1556-4029.15143>
8. Guo, B.-E., Shen, Y., Zhou, Z.-F., Liu, X., Wei, Y.-X., & Yang, L. (2025). Advanced deep learning for automatic classification of fired bullets from standard-issue firearms. *Science & Justice*, 65(6), 101335. <https://doi.org/10.1016/j.scijus.2025.101335>
9. Xin, Y., Tang, Y., Luo, Y., & Wang, K. (2026). A review of firearm toolmarks identification: Progress, challenges, and perspectives. *Forensic Science International*, 379, 112735. <https://doi.org/10.1016/j.forsciint.2025.112735>
10. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
11. National Research Council. (2009). *Strengthening forensic science in the United States: A path forward*. National Academies Press. <https://doi.org/10.17226/12589>
12. Basu, N., Bolton-King, R. S., & Morrison, G. S. (2022). Forensic comparison of fired cartridge cases: Feature-extraction methods for feature-based calculation of likelihood ratios. *Forensic Science International: Synergy*, 5, 100272. <https://doi.org/10.1016/j.fsisyn.2022.100272>
13. Berger, C. E. H., & Slooten, P. (2020). Score-based likelihood ratios for firearm toolmarks. *Forensic Science International: Synergy*, 2, 123–134. <https://doi.org/10.1016/j.fsisyn.2020.03.002>
14. Morrison, G. S., & Enzinger, E. (2019). Introduction to forensic voice comparison. In W. F. Katz & P. F. Assmann (Eds.), *The Routledge Handbook of Phonetics* (pp. 599–634). Routledge. <https://doi.org/10.4324/9780429056253-22>
15. Best, S., Kerkhoff, W., & Alberink, I. (2022). Open-set black-box study of firearm examiner performance on Thompson/Center bullets. *Forensic Science International*, 339, 111425. <https://doi.org/10.1016/j.forsciint.2022.111425>
16. Bolck, A., & Hicks, T. (2019). Validation of 3D comparison microscopy in firearm examination. *Forensic Science International*, 302, 109901. <https://doi.org/10.1016/j.forsciint.2019.109901>
17. Hicks, T., Buckleton, J. S., Bright, J. A., & Taylor, D. (2016). A framework for interpreting evidence. In J. S. Buckleton, J.-A. Bright, & D. Taylor (Eds.), *Forensic DNA evidence interpretation* (2nd ed., pp. 37–86).
18. Bunch, S., & Smith, J. (2011). Individuality of firing pin impressions: A follow-up study. *Journal of Forensic Sciences*, 56(3), 678–685. <https://doi.org/10.1111/j.1556-4029.2011.01723.x>

19. Bunch, S., Weaver, G., & Smith, J. (2009). Empirical study of the individuality of firing pin impressions. *Journal of Forensic Sciences*, 54(3), 621–629. <https://doi.org/10.1111/j.1556-4029.2009.01032.x>
20. Cuéllar, M., González, C., & Dror, I. E. (2022). Human and machine similarity judgments in forensic firearm comparisons. *Forensic Science International: Synergy*, 5, 100283. <https://doi.org/10.1016/j.fsisyn.2022.100283>
21. Stoney, D. A., & Stoney, P. L. (2015). Critical review of forensic trace evidence analysis and the need for a new approach. *Forensic Science International*, 251, 159–170. <https://doi.org/10.1016/j.forsciint.2015.03.022>
22. Zheng, X., Soons, J. A., Thompson, R. M., Singh, S. P., & Constantin, C. (2020). NIST Ballistics Toolmark Research Database. *Journal of Research of the National Institute of Standards and Technology*, 125, 125004. <https://doi.org/10.6028/jres.125.004>
23. Monson, K. L., Smith, E. D., & Bajic, S. J. (2022). Planning, design and logistics of a decision analysis study: The FBI/Ames study involving forensic firearms examiners. *Forensic Science International: Synergy*, 4, 100221. <https://doi.org/10.1016/j.fsisyn.2022.100221>
24. Fedorenko, V. A., Sorokina, K. O., & Giverts, P. V. (2024). The classification of firing pin impressions using the convolutional neural network (CNN). *Journal of Information Technologies and Computing Systems*, 2024(1), 87–96. <https://doi.org/10.14357/20718632240109>
25. Giverts, P., Sorokina, K., & Fedorenko, V. (2022). Examination of the possibility to use Siamese networks for the comparison of firing marks. *Journal of Forensic Sciences*, 67(6), 2416–2424. <https://doi.org/10.1111/1556-4029.15143>
26. Le Bouthillier, M.-E., Hrynkiw, L., Beauchamp, A., Duong, L., & Ratté, S. (2023). Automated detection of regions of interest in cartridge case images using deep learning. *Journal of Forensic Sciences*, 68(6), 1958–1971. <https://doi.org/10.1111/1556-4029.15319>
27. Hicks T., & Buckleton, J. (2016). Validation of forensic pattern recognition methods. *Forensic Science International*, 268, 1–10. <https://doi.org/10.1016/j.forsciint.2016.09.003>
28. Monson, K. L., Smith, E. D., & Peters, E. M. (2023). Repeatability and reproducibility of comparison decisions by forensic firearms examiners. *Journal of Forensic Sciences*, 68(5), 1540–1558. <https://doi.org/10.1111/1556-4029.15318>
29. INTERPOL. (2019). Interpol review papers 2019. <https://www.interpol.int/content/download/14458/file/Interpol%20Review%20Papers%202019.pdf>
30. Kerkhoff, W., & Mattijssen, E. J. A. T. (2021). Open-set validation of firearm examiner conclusions. *Forensic Science International*, 327, 110988. <https://doi.org/10.1016/j.forsciint.2021.110988>
31. Kerkhoff, W., van der Ham, L., & Mattijssen, E. J. A. T. (2015). Black-box validation of firearm examiner conclusions on cartridge cases. *Forensic Science International*, 257, 453–461. <https://doi.org/10.1016/j.forsciint.2015.09.025>
32. Kerkhoff, W., van der Ham, L., & Mattijssen, E. J. A. T. (2018). Open-set study of firearm examiner performance on consecutively manufactured slides. *Forensic Science International*, 293, 210–218. <https://doi.org/10.1016/j.forsciint.2018.10.012>
33. Knowles, L., Smith, J., & Baldwin, D. (2022). The validation of 3D virtual comparison microscopy (VCM) for firearm examination. *Journal of Forensic Sciences*, 67(5), 1950–1962. <https://doi.org/10.1111/1556-4029.14942>
34. Krings, R., Bunch, S., & Smith, J. (2017). Individuality of breech face impressions: An empirical study. *Journal of Forensic Sciences*, 62(1), 123–131. <https://doi.org/10.1111/1556-4029.13245>
35. Law, A., & Morris, J. (2021). Black-box evaluation of firearm examiner conclusions on cartridge cases. *Forensic Science International*, 327, 110987. <https://doi.org/10.1016/j.forsciint.2021.110987>
36. Le Bouthillier, M.-E., Hrynkiw, L., Beauchamp, A., Duong, L., & Ratté, S. (2023). Automated detection of regions of interest in cartridge case images using deep learning. *Journal of Forensic Sciences*, 68(6), 1958–1971. <https://doi.org/10.1111/1556-4029.15319>
37. Mattijssen, E. J. A. T., Kerkhoff, W., Hermesen, R., & Hes, R. A. G. (2023). Interpol review of forensic firearm examination 2019 – 2022. *Forensic Science International: Synergy*, 6, 100305. <https://doi.org/10.1016/j.fsisyn.2022.100305>
38. Monson, K. L., Smith, E. D., & Bajic, S. J. (2022). Planning, design and logistics of a decision analysis study: The FBI/Ames study involving forensic firearms examiners. *Forensic Science International: Synergy*, 4, 100221. <https://doi.org/10.1016/j.fsisyn.2022.100221>
39. Monson, K. L., Smith, E. D., & Peters, E. M. (2023). Accuracy of comparison decisions by forensic firearms examiners. *Journal of Forensic Sciences*, 68(1), 86–100. <https://doi.org/10.1111/1556-4029.15152>

40. Morrison, G. S. (2020). Likelihood ratio-based interpretation of toolmark evidence. *Forensic Science International: Synergy*, 2, 245–256. <https://doi.org/10.1016/j.fsisyn.2020.05.003>
41. Morrison, G. S., & Enzinger, R. (2019). Likelihood ratio-based interpretation in forensic science. *Forensic Science International: Synergy*, 1, 1–12. <https://doi.org/10.1016/j.fsisyn.2019.01.001>
42. Giverts, P., Sorokina, K., & Fedorenko, V. (2024). The use of machine learning for the determination of a type/model of firearms by the characteristics on cartridge cases. *Forensic Science International*, 358, 112021. <https://doi.org/10.1016/j.forsciint.2024.112021>
43. Law, E. F., & Morris, G. S. (2021). Evaluating firearm examiner conclusion variability using cartridge case reproductions. *Journal of Forensic Sciences*, 66(5), 1704–1720. <https://doi.org/10.1111/1556-4029.14758>
44. National Institute of Standards and Technology. (2020). NIST Ballistics Toolmark Research Database (NBTRD). <https://tsapps.nist.gov/nrbtd>
45. National Research Council. (2009). *Strengthening forensic science in the United States: A path forward*. The National Academies Press. <https://doi.org/10.17226/12537>
46. Dror, I. E., Charlton, D., & Péron, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Science International*, 156(1), 74–78. <https://doi.org/10.1016/j.forsciint.2005.10.017>
47. Kerkhoff, W., et al. (2018). A part-declared blind testing program in firearms examination. *Science & Justice*, 58(4), 258–263. <https://doi.org/10.1016/j.scijus.2018.03.006>
48. Peters, E. M., Smith, E. D., & Monson, K. L. (2022). Modeling examiner heterogeneity in firearm black-box studies. *Journal of Forensic Sciences*, 67(3), 1056–1068. <https://doi.org/10.1111/1556-4029.14987>
49. Ratté, S., Le Bouthillier, M.-E., & Beauchamp, A. (2021). Deep learning for automated cartridge case comparison. *Forensic Science International*, 326, 110923. <https://doi.org/10.1016/j.forsciint.2021.110923>
50. Ratté, S., Le Bouthillier, M.-E., & Beauchamp, A. (2022). Deep learning for automated bullet comparison. *Forensic Science International*, 336, 111345. <https://doi.org/10.1016/j.forsciint.2022.111345>
51. Smith, J., & Monson, K. L. (2020). Large-scale black-box study of firearm examiner performance on bullets. *Journal of Forensic Sciences*, 65(5), 1456–1468. <https://doi.org/10.1111/1556-4029.14421>
52. Smith, J., & Monson, K. L. (2022). Replication and extension of firearm examiner black-box studies. *Journal of Forensic Sciences*, 67(2), 678–690. <https://doi.org/10.1111/1556-4029.14899>
53. Smith, J., Bunch, S., & Baldwin, D. (2014). Empirical study of firearm examiner performance on cartridge cases. *Journal of Forensic Sciences*, 59(2), 345–356. <https://doi.org/10.1111/1556-4029.12345>
54. Smith, J., Bunch, S., & Baldwin, D. (2016). Empirical study of firearm examiner performance on bullets and cartridge cases. *Journal of Forensic Sciences*, 61(2), 345–356. <https://doi.org/10.1111/1556-4029.12987>
55. Soons, J., Zheng, X. A., & Thompson, R. (2020). NIST Ballistics Toolmark Research Database: Design and applications. *Journal of Research of the National Institute of Standards and Technology*, 125, 125004. <https://doi.org/10.6028/jres.125.004>
56. Stoney, D. A., & Stoney, E. M. (2015). Critical review of forensic firearm examination. *Forensic Science International*, 257, 1–10. <https://doi.org/10.1016/j.forsciint.2015.07.045>
57. Thompson, R., Zheng, X. A., & Soons, J. (2019). Objective criteria for firearm identification using 3D topography. *Forensic Science International*, 304, 109987. <https://doi.org/10.1016/j.forsciint.2019.109987>
58. Thompson, R., Zheng, X. A., & Soons, J. (2021). Objective criteria for firearm identification using 3D topography: Follow-up study. *Forensic Science International*, 324, 110876. <https://doi.org/10.1016/j.forsciint.2021.110876>
59. Tully, S., & Evett, I. W. (1995). The evaluation of evidence in forensic science. *Journal of the Royal Statistical Society: Series A*, 158(1), 1–16. <https://doi.org/10.2307/2983427>
60. van der Ham, L., Kerkhoff, W., & Mattijssen, E. J. A. T. (2020). Validation of 3D surface topography for firearm examination. *Forensic Science International*, 315, 110456. <https://doi.org/10.1016/j.forsciint.2020.110456>
61. van der Ham, L., Kerkhoff, W., & Mattijssen, E. J. A. T. (2022). Validation of 3D surface topography for firearm examination: Multi-laboratory study. *Forensic Science International*, 337, 111389. <https://doi.org/10.1016/j.forsciint.2022.111389>
62. Zheng, X. A., Soons, J., Thompson, R., Singh, S., & Constantin, C. (2020). NIST Ballistics Toolmark Research Database. *Journal of Research of the National Institute of Standards and Technology*, 125, 125004. <https://doi.org/10.6028/jres.125.004>
63. Nichols, R. G. (1997). Firearm and toolmark identification criteria: A review of the literature. *Journal of Forensic Sciences*, 42(3), 466–474. <https://doi.org/10.1520/JFS14149J>

64. Tong, M., Song, J., & Chu, W. (2015). An improved algorithm of Congruent Matching Cells (CMC) method for firearm evidence identifications. *Journal of Research of the National Institute of Standards and Technology*, 120, 102–112. <https://doi.org/10.6028/jres.120.008>
65. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
66. Morris, K. B., et al. (2017). An evaluation of the discriminating power of an Integrated Ballistics Identification System® Heritage™ system with the NIST standard cartridge case (Standard Reference Material 2461). *Forensic Science International*. <https://doi.org/10.1016/j.forsciint.2017.09.004>
67. Chen, Z., Song, J., Soons, J. A., Thompson, R. M., & Zhao, X. (2020). Pilot study on deformed bullet correlation. *Forensic Science International*, 306, 110098. <https://doi.org/10.1016/j.forsciint.2019.110098>
68. Riva, F., et al. (2020). Comparison and interpretation of impressed marks left by a firearm on cartridge cases—Towards an operational implementation of a likelihood ratio based technique. *Forensic Science International*, 313, 110363. <https://doi.org/10.1016/j.forsciint.2020.110363>
69. Baldwin, D. P., Bajic, S. J., Morris, M. D., & Zamzow, D. S. (2023). A study of examiner accuracy in cartridge case comparisons. Part 1: Examiner error rates. *Forensic Science International*, 349, 111733. <https://doi.org/10.1016/j.forsciint.2023.111733>