Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

AI-Assisted Forensic Evidence Recognition From Crime Scene Images: Evaluation of Deep Learning Object Detection Models

Adityasinh Gohil[1], Prof. (Dr.) Pooja Ahuja[2]

[1] Research Scholar, School of Forensic Science, National Forensic Sciences University, Sector-9, Gandhinagar-382007, India, E-mail: aditya.phdfs2330@nfsu.ac.in
[2] Professor, School of Forensic Science, National Forensic Sciences University, Sector-9, Gandhinagar-382007, India, E-mail: pooja.ahuja@nfsu.ac.in

Abstract
Comprehensive evidence recognition is vital to crime scene investigation. Time pressure and scene complexity often cause missed evidence. Missed or misidentified evidence can permanently compromise a case. This study evaluates deep learning object detection for forensic evidence recognition. It forms phase one of C-SAHAI, a human-AI programme. C-SAHAI supports forensic practitioners during on-scene evidence analysis. A domain-specific dataset of 7,028 images was compiled. Images were drawn from simulated scenes and open repositories. Images span 18 forensic evidence classes after taxonomy correction. Classes include firearms, edged weapons, biological traces, and micro-evidence. Two architectures were trained: YOLOv8m and Faster R-CNN. Both models were evaluated at 35 and 100 epochs. Faster R-CNN showed the most balanced detection performance overall. At 100 epochs, its F1-score reached 0.705. Precision was 0.747, and recall was 0.668. YOLOv8m achieved higher mAP@50, reaching 0.453 at 100 epochs. However, its recall remained lower, at only 0.446. This trade-off reflects each architecture's underlying design philosophy. Extended training modestly improved both tested architectures overall. Neither model showed performance degradation across the tested range. Firearms and edged weapons were detected more reliably. Micro-evidence such as hair remained consistently difficult to detect. This reflects a known constraint of current detection architectures. These findings establish a baseline for AI-assisted evidence recognition. The approach supports, rather than replaces, expert forensic judgment.

Keywords: Forensic Evidence Recognition, Crime Scene Analysis, Object Detection, Deep Learning, YOLOv8, Faster R-CNN, AI-Assisted Investigation.

I. Introduction

Crime scene management is foundational to criminal justice work. Investigation reliability depends directly on evidence identification and preservation [1]. Any failure at this stage compromises subsequent forensic conclusions. Evidence overlooked during a walkthrough cannot later be recovered. Poorly preserved material may become inadmissible before laboratory analysis begins.

Physical evidence at crime scenes spans a wide range. This includes firearms, edged weapons, and biological traces [2]. Micro-evidence such as hair and fingerprints also occurs. Each evidence type carries distinct investigative and legal significance. Correct identification shapes collection, preservation, and downstream analysis [3]. A missed firearm and a missed hair strand differ. Both, however, represent a permanent loss once released.

Several recurring challenges complicate on-scene evidence identification. Crime scenes are frequently cluttered and visually complex [3]. Physical accessibility can also be significantly limited. These conditions increase the likelihood of overlooked evidence. Investigators typically operate under considerable time pressure throughout. This pressure, combined with cognitive fatigue, raises oversight risk. Micro-evidence is especially vulnerable to this risk. Hair and fingerprints occupy very small visual areas. They are easily missed against complex, textured backgrounds [2], [3].

Officer training presents a further systemic challenge here. Police personnel are typically first to arrive on-scene. Evidence handling requires strict procedural adherence throughout [2]. Gaps in training can lead to contamination. Contaminated evidence loses much of its scientific value. This reflects a persistent gap between practice and protocol.

Artificial intelligence now performs strongly on visual recognition tasks [4]. Its most productive role supports rather than replaces experts [5]. AI-based detection could assist investigators during active scene work. Deep learning models process complex visual patterns very quickly. A single model can flag multiple evidence types simultaneously. Recent work applied deep learning to bloodstain classification directly [6]. This demonstrates clear relevance to crime scene documentation. Multiple evidence types frequently coexist within a single photograph.

This study evaluates object detection across 18 evidence classes. These span firearms, edged weapons, and micro-evidence types. The dataset reflects realistic, operational crime scene conditions closely. Two architectures were evaluated: YOLOv8m and Faster R-CNN [13], [14]. This work forms the first phase of C-SAHAI. C-SAHAI supports forensic experts during on-scene analysis. The goal is a reliable evidence recognition baseline. Findings identify which evidence types remain difficult to detect. This informs responsible AI deployment in forensic workflows.

II. Related Work

Existing AI research in this area remains narrow. Most studies focus on a single evidence category [5], [25]. Others target security applications like traffic surveillance [7], [8]. Common benchmark datasets also lack forensic relevance here. COCO and ImageNet do not represent crime scene conditions [9], [10]. Broad evaluations across diverse evidence types remain genuinely scarce [11]. This leaves open questions about real-world generalisation.

Much existing work also frames AI as a replacement. Forensic work, however, requires legal accountability and expert judgment [12], [24]. These qualities cannot be replicated by automated systems alone. Any deployed detection system must respect this constraint directly.

Closest to this study is prior region-based detection work. Saikia et al. [23] applied Faster R-CNN to crime scenes. Their system achieved strong average accuracy on indoor classes. Their evaluation, however, excluded micro-evidence categories entirely. It also did not compare architectures directly on shared data. The gap this study addresses is therefore twofold. First, no dataset spans firearms, traces, and micro-evidence together. Second, prior work rarely compares architectures on shared forensic data. This study evaluates 18 evidence classes using both architectures. This forms the first phase of the C-SAHAI programme. The goal remains a reliable, responsibly deployable detection baseline.

III. Materials and Methods

a. Dataset Development

A domain-specific dataset of 7,028 images was compiled. Images span 18 forensic evidence classes overall. These classes represent evidence types commonly encountered during investigations [2], [3]. Two complementary sources contributed images to this dataset. The first consisted of simulated crime scene photography. These photographs were staged under varying indoor lighting conditions. Standard DSLR and mobile forensic imaging equipment was used [3]. This allowed controlled placement of evidence within realistic layouts. The second source was supplementary imagery from Roboflow Universe [26]–[33]. Relevant evidence classes were selectively identified and downloaded there. This broadened dataset coverage beyond what staging alone allowed.

The Roboflow-sourced portion contributes meaningful, distinct scene diversity. These images include outdoor scenes and variable lighting conditions. Motion blur and partial occlusion also appear frequently. Evidence items are sometimes obscured behind other objects. These conditions are common in genuine operational settings. They are, however, difficult to replicate through staged exercises alone. This diversity is treated here as a methodological strength. It is not treated as a dataset limitation. Crime scenes are rarely photographed under clean, ideal conditions. A deployed detection system should be evaluated under comparable variation.

The 18 evidence classes include several firearm subtypes. These include handguns, pistols, revolvers, shotguns, and rifles. Edged weapons such as knives are also included. Biological traces such as blood and bones appear too. Micro-evidence classes include fingerprints, detailed further in Section III.B. Image counts varied naturally across all these classes. Firearms were visually prominent and frequently represented in sources. They therefore appeared most often within the dataset. Micro-evidence classes such as fingerprints were comparatively underrepresented instead. This imbalance was preserved deliberately, without artificial correction. Oversampling or class-weighting was not applied to rare classes. Artificially flattening the distribution would misrepresent real evidence frequencies. It would also risk overstating performance on rare classes.

The dataset was split across three purposes: training, validation, and testing. This produced 5,191 training images (73.9%) for model development. Validation used 1,077 images (15.3%), and testing used 760

images (10.8%). All annotated images were reviewed by the supervising forensic scientist. This is the co-author of this study. This review verified that ground-truth labels met forensic accuracy standards. This reduced the risk of labelling inconsistencies affecting training. Images were resized to 640 × 640 pixels prior to training [13], [14]. This resolution balances computational efficiency against retained spatial detail. This trade-off matters most for small micro-evidence classes. An item there may occupy only a handful of pixels. This is discussed further in Results and Limitations sections.

b. Taxonomy Correction

The dataset was originally compiled using a 22-class taxonomy. A subsequent audit identified several structural inconsistencies present. These inconsistencies arose from merging multiple independently annotated Roboflow sources. Each source had used slightly different labelling conventions. Three classes targeted overlapping human-related annotation regions. These were labelled inconsistently as "Human," "Human-body," and "Victim." They were merged into a single, unified class. Two further classes represented visually identical hand-labelling variants. These were similarly merged into one class. An additional class, "Not Knife," was also identified as problematic. This class was an artifact of one source dataset's convention. It represented negative examples, not genuine forensic evidence. It was therefore removed entirely from the taxonomy. Following these corrections, the dataset used an 18-class taxonomy. All 7,028 label files were re-verified against this correction. This verification occurred prior to any training reported here.

c. Annotation and Augmentation

All images were annotated with bounding boxes using Roboflow. Each evidentiary object was manually labelled according to predefined definitions. These definitions referenced standard forensic evidence terminology throughout [1], [2]. This ensured class boundaries reflected forensically meaningful distinctions. They did not reflect arbitrary or purely visual groupings. Roboflow also handled export of annotated data for training. This streamlined the transition from annotation into model development.

Moderate data augmentation was applied during training [17]. This improved model generalisation across all evidence classes. Techniques included controlled random rotations of up to ±10 degrees. Horizontal and vertical flipping were also applied throughout. Brightness was adjusted within a ±15 percent range. These transformations simulate realistic variation in camera angle. They also simulate variation in ambient lighting conditions [3]. This mirrors conditions encountered during actual crime scene documentation. Augmentation was applied conservatively throughout this study. It was deliberately more restrained than typical general-purpose pipelines. This was due to concern about obscuring fine forensic features. A distorted image could obscure a fingerprint's exact ridge pattern. It could also obscure a firearm's precise silhouette.

d. Ethics Statement

No human participants were involved in this study. All simulated crime scene photography used mannequins exclusively. Staged forensic props were used alongside these mannequins. This avoided ethical concerns tied to identifiable human subjects.

e. Hardware and Software

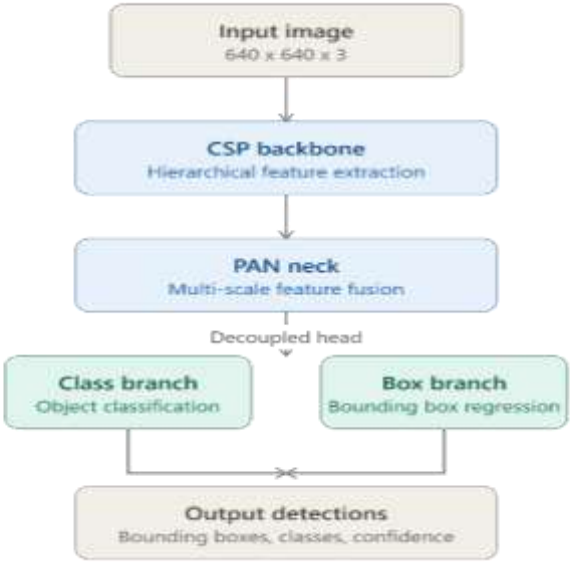
Training and evaluation used an NVIDIA RTX 3070 GPU. This GPU carried 8 GB of VRAM. The system also included 64 GB of system RAM. Python 3.11 was used with the PyTorch framework. This local configuration followed earlier pilot experiments on cloud platforms. Google Colab and Kaggle were both tested initially. Both proved insufficient for sustained training at this scale. This was due to session time limits in some cases. Memory constraints and inconsistent GPU availability also contributed. Local GPU training provided the sustained capacity required. This capacity supported full training runs across all configurations. The RTX 3070's 8 GB VRAM ceiling mattered too. It partially motivated the 640×640 input resolution. Faster R-CNN batching was especially memory-constrained. Larger inputs were not feasible within available memory. Higher input resolutions and tiled inference were therefore reserved for future work.

f. Detection Models

Two architectures were selected for evaluation here. Object detection has evolved substantially across single-stage and two-stage design families [16]. These were YOLOv8m and Faster R-CNN [13], [14]. Both were chosen for complementary operational characteristics. This allowed assessment of a fast, single-stage detector. It also allowed assessment of a slower, two-stage detector. Both were tested against diverse forensic evidence properties. These range from large firearms to small micro-evidence [15], [18].

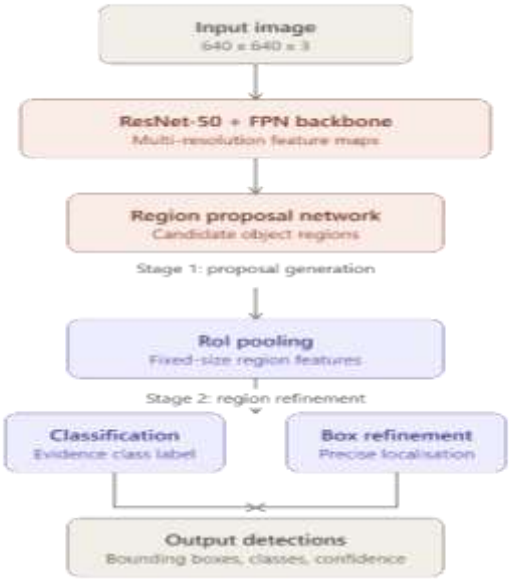
YOLOv8m is an anchor-free, single-stage detection model [13]. It performs localisation and classification within one forward pass. The input image is processed by a CSP-based backbone. This backbone extracts hierarchical feature representations across multiple scales [13], [19]. These features are aggregated through a Path Aggregation Network neck. This neck enables effective fusion of multi-scale information. Both large and small objects benefit from this fusion [13], [18]. A decoupled detection head separates two distinct tasks. These are classification and bounding-box regression, handled independently. Prior work

shows this design improves convergence and accuracy. This holds relative to earlier, coupled-head YOLO variants. The entire image is processed in one pass. This design achieves notably fast inference speeds. YOLOv8m is therefore well suited to rapid preliminary screening. It can screen multiple evidence classes simultaneously across an image. Fig. 1 illustrates the full YOLOv8m pipeline used here.



(**Fig. 1**: YOLOv8m architecture pipeline used for forensic evidence detection.)

Faster R-CNN instead follows a two-stage detection strategy [14]. This strategy prioritises localisation precision over raw inference speed. The architecture combines a ResNet-50 backbone with a Feature Pyramid Network [20]. This combination enables feature extraction across multiple spatial resolutions. It improves sensitivity to small or partially obscured objects [14], [15]. This property is directly relevant to micro-evidence classes. Hair and fingerprints benefit particularly from this sensitivity. In the first stage, a Region Proposal Network scans extracted features. This network generates candidate object regions from objectness scores [14]. In the second stage, each candidate region is further refined. This refinement uses Region of Interest pooling. Classification and bounding-box regression then produce final detections. This two-stage process is considerably more computationally demanding [15], [18]. This holds true compared to YOLOv8m's single-pass design. In exchange, it offers improved localisation reliability in complex scenes. Precise candidate-region refinement can reduce missed detections. It can also reduce false localisations in cluttered environments. Fig. 2 illustrates the Faster R-CNN pipeline used here.



(**Fig. 2**: Faster R-CNN architecture pipeline used for forensic evidence detection)

g. Hyperparameter Configuration

Hyperparameter settings were based on published detection literature [15], [18]. Arbitrary or purely default values were not used. Both YOLOv8m training runs used a matched batch size. This batch size of 16 was consistent across runs. This ensured any performance differences reflected training duration only. Batch composition inconsistency was therefore ruled out as a factor. Epoch counts of 35 and 100 were tested. This applied to both architectures under comparison. This allowed a direct, controlled comparison of training duration effects. This comparison spanned both a fast single-stage detector. It also spanned a slower, two-stage detector.

h. Evaluation Metrics

Model performance was assessed using precision, recall, F1-score, and mAP [21], [22]. These metrics were evaluated at the overall model level. They were also evaluated at the individual evidence-class level. Precision reflects a model's ability to limit false positives. Recall measures the proportion of ground-truth objects correctly identified. This metric carries particular forensic importance in this context. Recall directly indicates the likelihood evidence would go undetected. The F1-score summarises performance as precision and recall's harmonic mean. This provides one balanced measure overall. Neither false positives nor false negatives dominate this measure. mAP was reported at a fixed IoU threshold of 0.50. It was also averaged across the broader 0.50 to 0.95 range [9], [21]. The broader range offers a stricter localisation accuracy measure. It rewards tighter bounding-box alignment with ground truth. An IoU threshold of 0.50 was the primary benchmark throughout. This is consistent with common practice in the literature [21]. Each model-epoch configuration was trained as a single run. Implications of this for result variance appear in Limitations.

IV. Results

a. Evidence Recognition Performance

Detection performance varied clearly between the two architectures. Both models were evaluated at 35 and 100 epochs. This comparison reveals distinct trade-offs relevant to forensic deployment.

Faster R-CNN produced the most balanced overall profile. At 35 epochs, its F1-score reached 0.686. Precision stood at 0.699, and recall at 0.674. These three values remained close to each other. This closeness reflects balanced detection behaviour overall. At 100 epochs, all three metrics improved further. F1-score rose to 0.705 at this point. Precision reached 0.747, and recall reached 0.668. This closeness persisted at the higher epoch count too. Forensic implications of this balance are addressed in Section V (b).

YOLOv8m showed a markedly different performance profile. At 35 epochs, precision reached a strong 0.635. Recall, however, remained comparatively low at 0.428. This combination produced an F1-score of 0.512. At 100 epochs, performance improved only modestly. Precision reached 0.628, essentially unchanged from before. Recall improved slightly, reaching 0.446 at this point. F1-score rose modestly to 0.522 overall. YOLOv8m's recall stayed below Faster R-CNN's throughout testing.

A different pattern emerged when examining mAP@50. YOLOv8m consistently outperformed Faster R-CNN on this metric. At 100 epochs, YOLOv8m's mAP@50 reached 0.453. Faster R-CNN's mAP@50 remained lower, at 0.284. This gap reflects an important metric distinction. F1-score depends on one confidence threshold only. mAP@50 integrates performance across every threshold instead. YOLOv8m ranks and localises detections with more confidence. Yet it still misses more ground-truth objects overall. Both patterns matter for real deployment decisions. Extended training benefited both architectures, though only modestly. Neither model showed performance decline with more epochs. This held true across the full tested range. Standard epoch budgets remain reasonable for this dataset. Longer training, however, caused no measurable harm either.

(**Table 1**: Detection performance across models and training epochs)

Model	Epochs	Precision	Recall	F1	mAP@50	mAP@50-95
YOLOv8m	35	0.635	0.428	0.512	0.414	0.323
	100	0.628	0.446	0.522	0.453	0.355
Faster R-CNN	35	0.699	0.674	0.686	0.279	0.197
	100	0.747	0.668	0.705	0.284	0.196

b. Per-Class Evidence Recognition

Detection performance varied substantially across the 18 evidence classes. This variation reflects differences in object size and context.

Firearms achieved the strongest detection performance overall. This pattern held true for both architectures tested. Firearms are large, structurally distinct visual objects. Their visual prominence supports consistently stable localisation. Subclass separation, however, proved more difficult than expected. Pistol, revolver, and gun classes overlapped visually. This overlap caused frequent confusion between these subclasses.

Edged weapons also showed reliable detection performance overall. Knives present clear, well-separated visual boundaries. Both architectures performed comparatively well on this class. Micro-evidence classes showed the weakest performance overall. Hair and fingerprints occupy very small pixel areas. At standard resolution, such items may fall below ten pixels across. This falls below both models' effective receptive fields. These results reflect intrinsic detection difficulty, not model failure. Tiled inference at native resolution may help future work.

Classes with fewer training images also underperformed noticeably. This reflects the natural imbalance preserved in the dataset. This imbalance mirrors realistic evidence distributions at actual scenes.

Confusion matrices for both models appear below. Both training durations are shown, in Fig. 3–6. These provide class-level detail beyond aggregate summary metrics. Firearm subclasses showed visible cross-confusion in both matrices. This confirms the subclass overlap described above. Micro-evidence classes showed high miss rates in both models. Predictions for these classes were often absent entirely. Absence differs meaningfully from outright misclassification here. This distinction matters directly for forensic deployment decisions. A missed detection and a wrong label differ operationally.

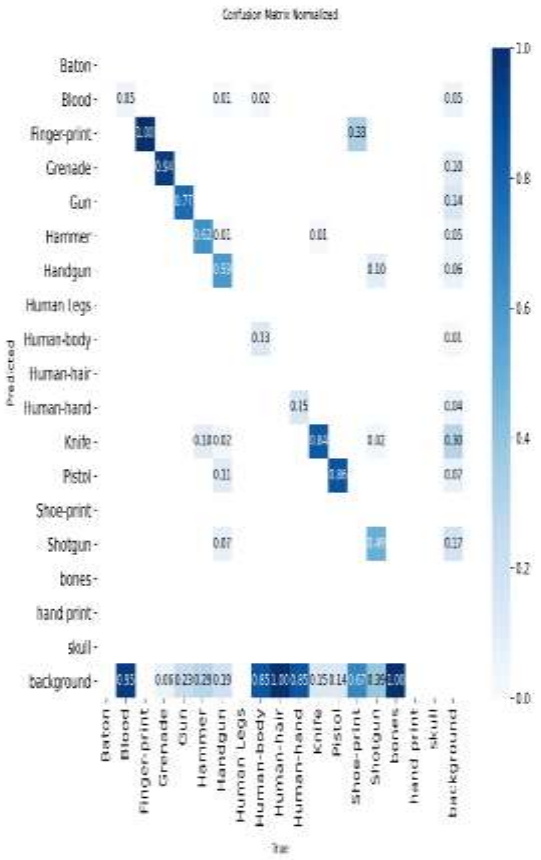


***Fig. 3**: Confusion matrix for YOLOv8m on the test set (35 epochs)*

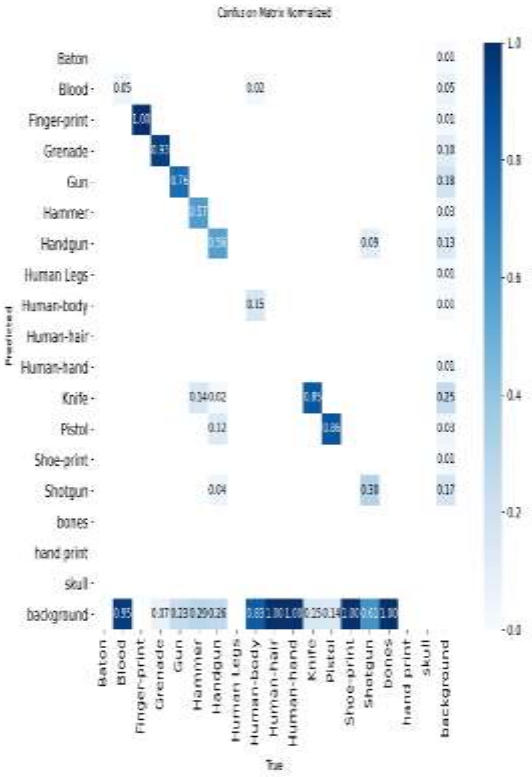


***Fig. 4**: Confusion matrix for YOLOv8m on the test set (100 epochs)*

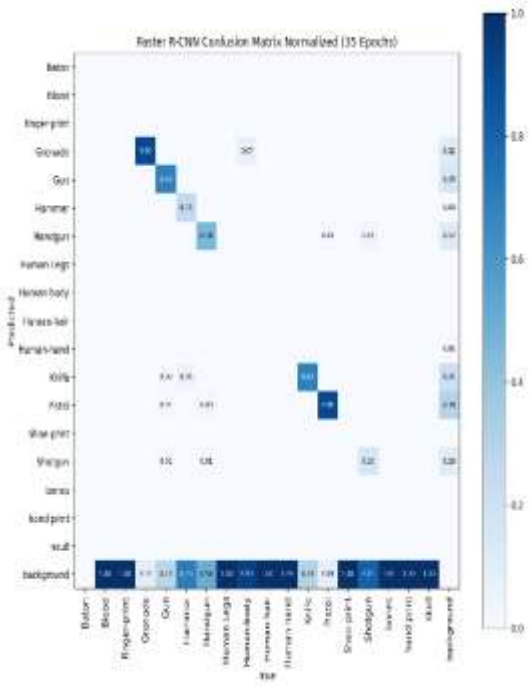


***Fig. 5**: Confusion matrix for FRCNN on the test set (35 epochs)*

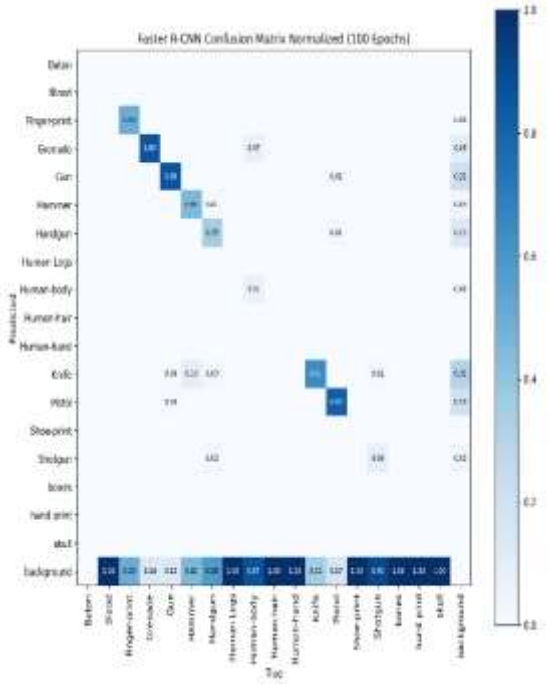


***Fig. 6**: Confusion matrix for FRCNN on the test set (100 epochs)*

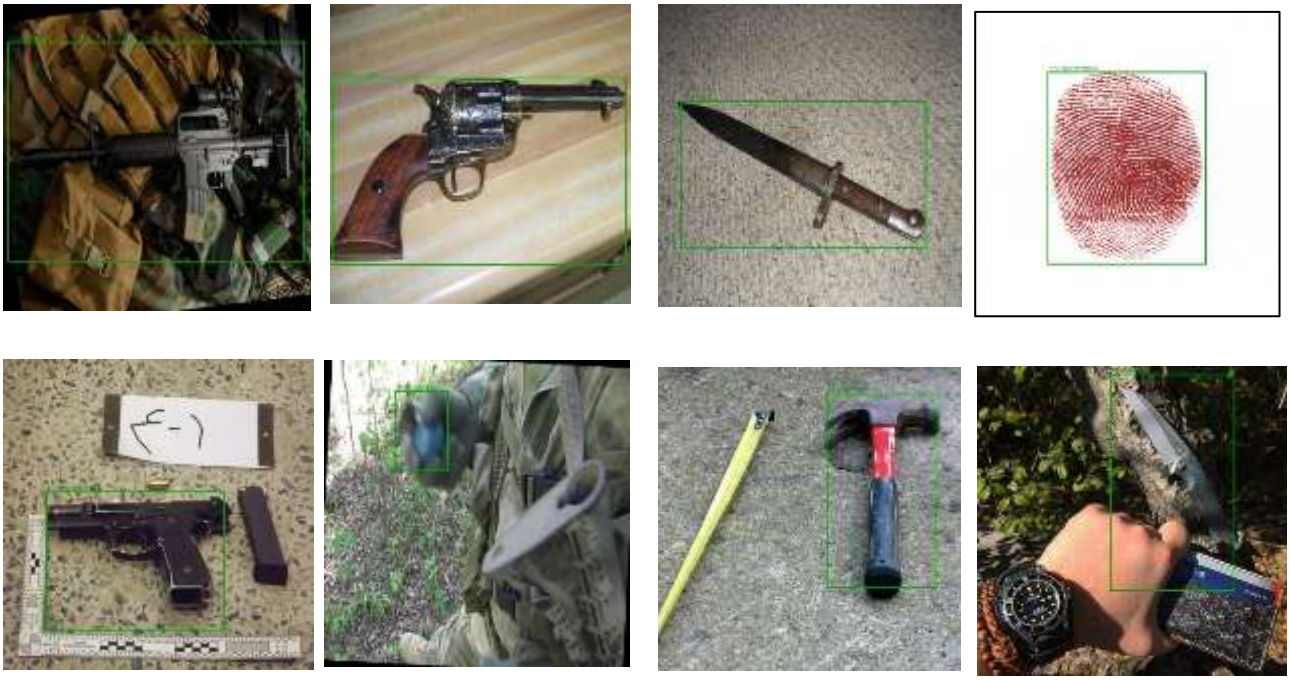
c.

Qualitative Observations

Visual inspection of predictions offered additional useful insight. Fig. 7 shows representative detections from both architectures. These span firearm, edged weapon, and micro-evidence classes.

Firearms were detected with high confidence by both models. Bounding boxes aligned closely with true object boundaries. Faster R-CNN showed slightly tighter localisation in cluttered scenes. This aligns with its Region Proposal Network mechanism. Micro-evidence detection failures were common across both architectures. Hair and fingerprints were frequently missed entirely during inference. When detected, confidence scores remained noticeably lower than firearms. This visual pattern supports the per-class findings above. Firearm subclass confusion also appeared clearly in qualitative samples. A pistol was sometimes labelled as a revolver. This occurred most often in partially occluded views.

YOLOv8m ran at approximately 65 frames per second. Faster R-CNN ran at approximately 12 frames per second. However, crime scene examination is not real-time work. It remains a structured, methodical, deliberate process instead. Thoroughness therefore matters more than raw processing speed. Faster R-CNN's lower speed is not disqualifying here.



(**Fig. 7**: Representative detections from YOLOv8m (top row) and Faster R-CNN (bottom row) across firearm, edged weapon, and micro-evidence classes)

d. Error Analysis

Incorrect predictions followed consistent, identifiable patterns across classes. Over segmentation occurred when one object received multiple boxes. Under-detection was most common for low-contrast items. Reflective surfaces also contributed frequently to under detection. This pattern appeared often for biological traces specifically. Micro evidence classes showed this pattern most consistently overall. Misclassification was most common among firearm subclasses specifically. Error frequency increased noticeably in cluttered scene environments. These patterns remained consistent and systematic, not random. This consistency offers clear direction for future model improvement.

V. Discussion

a. Evidence Recognition Findings

Automated forensic evidence detection is achievable across most classes. However, performance depends heavily on architecture and evidence type. This distinction matters for how findings get applied.

Saikia et al. [23] applied Faster R-CNN to crime scenes. Their system reported 74.33% average accuracy overall. This aligns reasonably with this study's Faster R-CNN F1. That F1-score reached 0.705 at 100 epochs. Their evaluation, however, covered indoor object classes only. It did not include any micro-evidence categories. Direct comparison should therefore be treated cautiously here. Dataset composition differs meaningfully between the two studies. This study extends coverage to 18 diverse classes. Broader class diversity typically increases overall detection difficulty. The result remains consistent with prior crime-scene-specific findings.

Firearms and edged weapons showed the strongest performance overall. These objects are visually prominent and structurally distinct. They occupy large pixel regions with clear boundaries. This property supports reliable detection even in clutter. Faster R-CNN showed particularly stable localisation for these classes. Its region proposal mechanism likely contributes to this stability. Micro-evidence remained the most persistent challenge throughout testing. Hair and fingerprints occupy minimal pixel space per image. At standard resolution, they fall below detection thresholds. This is not a failure specific to either model. It reflects a known constraint of current architectures. Tiled inference or higher-resolution pipelines may help here. Future work should explore both options directly.

b. Forensic Implications of Detection Performance

These results carry direct implications for responsible deployment decisions. Faster R-CNN achieved balanced performance at 100 training epochs. Its F1 reached 0.705, with recall at 0.668. This balanced profile suits forensic evidence documentation well. In criminal justice contexts, balance matters more than optimisation. Optimising a single metric alone is less defensible here. A missed detection carries more

severe consequences than one false alarm. A human reviewer can quickly dismiss a false alarm. A missed item, however, cannot later be recovered. Once a scene is released, that evidence is gone.

YOLOv8m showed a meaningfully different operational trade-off instead. Its mAP@50 of 0.453 exceeded Faster R-CNN's clearly. This reflects stronger overall ranking and localisation confidence. However, its recall of 0.446 remained substantially lower. At standard confidence thresholds, YOLOv8m misses more evidence items. This trade-off shapes how each model should deploy. YOLOv8m may suit rapid, preliminary scene screening. Faster R-CNN may better suit thorough, primary evidence review.

Extended training modestly benefited both tested architectures overall. Neither model showed performance decline with more epochs. This suggests stable optimisation across the tested training range. Standard epoch budgets remain reasonable given these findings. Prolonged training, however, caused no measurable harm either.

c. Implications for the C-SAHAI Model

Faster R-CNN is identified as the stronger C-SAHAI candidate. This reflects its balanced recall and precision profile. It also aligns closely with C-SAHAI's core philosophy. C-SAHAI does not position AI as autonomous decision-maker. Detected evidence is instead shown to attending forensic experts. Experts then perform final manual verification themselves. This reduces manual workload while preserving human oversight fully. The automated layer reduces cognitive load under time pressure. The human layer provides judgment and legal accountability instead. Together, these layers reduce both algorithmic and human error.

d. Practical Implications for Forensic Practitioners

AI-based detection is now mature enough for practical support. These tools can flag items otherwise easily overlooked. Time pressure, scene complexity, and fatigue all contribute here.

Responsible deployment, however, requires careful attention to data security. Crime scene imagery constitutes highly sensitive material throughout. It must never reach unsecured external systems or servers. Local, institutionally controlled infrastructure remains necessary for real deployment. Ethical and institutional approvals must precede any operational use. Continuous evaluation also remains important going forward. Evidence types vary meaningfully across different case categories. Performance should therefore be retested against new data regularly. This ensures reliability holds across evolving operational conditions.

VI. Limitations

a. Single Training Run per Configuration

Each model-epoch configuration was trained only once. No repeated runs were conducted to estimate variance. Reported metrics therefore reflect a single training outcome. Random initialisation could influence results to some degree. Future work should include repeated runs with variance reporting. Mean and standard deviation values would confirm result stability.

b. Interclass Confusion Among Firearm Subclasses

Pistol, revolver, and gun classes shared visual overlap. This overlap reduced per-class classification accuracy within firearms. Category-level firearm detection, however, remained reliable throughout testing. Future dataset development should consider consolidating visually similar subclasses. This consolidation may reduce ambiguity between closely related evidence types.

c. Micro-Evidence Detection at Standard Resolution

Hair and fingerprint classes showed persistent detection difficulty throughout. This reflects limited pixel representation at standard resolution. This constraint affects current object detection architectures generally. It is not a dataset-specific failure or shortfall. Tiled inference or higher-resolution pipelines may address this. Future work should test both approaches directly against baselines.

d. Dataset Scope

The dataset combines simulated exercises with open-source repository imagery. Certain rare, real-world conditions remain underrepresented despite this diversity. Extreme weather conditions fall into this underrepresented category. Heavily decomposed biological evidence also remains underrepresented currently. Continued dataset expansion would improve coverage of these scenarios. This expansion is planned as part of future work.

VII. Conclusion

This study evaluated deep learning object detection for forensic evidence. It forms the first phase of the C-SAHAI programme. Evaluation spanned 18 evidence classes across a domain-specific dataset. YOLOv8m and Faster R-CNN were compared directly.

Faster R-CNN achieved the more balanced detection profile overall. At 100 epochs, its F1-score reached 0.705. Precision reached 0.747, while recall reached 0.668. YOLOv8m instead achieved a higher mAP@50

of 0.453. However, its recall remained lower, at only 0.446. This trade-off reflects each architecture's underlying design philosophy. Two-stage detection favours thorough, balanced recognition overall. Single-stage detection favours fast, confident localisation instead.

Extended training modestly improved both tested architectures. This indicates stable optimisation across the tested epoch range. Neither model showed overfitting or degraded generalisation at 100 epochs. Firearms and edged weapons were detected reliably by both models. Micro-evidence such as hair remained consistently difficult to detect. This reflects a known limitation of current detection systems. It is not a dataset-specific shortfall or failure.

Faster R-CNN is proposed as the stronger C-SAHAI candidate. This reflects its balanced recall and precision profile, discussed in Section V.(c). Responsible deployment requires secure, local infrastructure throughout. Ethical and institutional approvals must precede any operational use. These systems should support expert forensic judgment, never replace it.

Future work will expand the dataset across underrepresented classes. Development will continue on C-SAHAI's SOP guidance component. Reliable evidence recognition remains essential to forensic investigation broadly. This study offers a corrected, data-audited baseline toward that goal.

References

[1] National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington (DC): National Academies Press; 2009. https://doi.org/10.17226/12589

[2] Saferstein R. Criminalistics: An Introduction to Forensic Science. 13th ed. New York: Pearson; 2020.

[3] Gouse S, Karnam S, Girish HC, Murgod S. Forensic photography: Prospect through the lens. J Forensic Dent Sci 2018;10(1):2–4. https://doi.org/10.4103/jfo.jfds_2_16

[4] Castillo Camacho I, Wang K. A comprehensive review of deep-learning-based methods for image forensics. J Imaging 2021;7(4):69. https://doi.org/10.3390/jimaging7040069

[5] Ketsekioulafis I, Filandrianos G, Katsos K, et al. Artificial intelligence in forensic sciences: A systematic review of past and current applications and future perspectives. Cureus 2024;16(9):e70363. https://doi.org/10.7759/cureus.70363

[6] Bergman T, Klöden M, Dreßler J, et al. Automatic classification of bloodstains with deep learning methods. Künstl Intell 2022;36:135–41. https://doi.org/10.1007/s13218-022-00760-y

[7] Ahmed I, Das R. Comparative analysis of YOLO and Faster R-CNN models for detecting traffic objects. Int J Adv Comput Sci Appl 2025;16(3). https://doi.org/10.14569/IJACSA.2025.0160342

[8] Zhao D, Shao F, Zhang S, et al. Advanced object detection in low-light conditions: Enhancements to YOLOv7 framework. Remote Sens 2024;16:4493. https://doi.org/10.3390/rs16234493

[9] Lin TY, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context. Lecture Notes Comput Sci 2014;8693:740–55. https://doi.org/10.1007/978-3-319-10602-1_48

[10] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. Proc IEEE Conf Comput Vis Pattern Recognit 2009:248–55. https://doi.org/10.1109/CVPR.2009.5206848

[11] Peserico G, Morato A. Performance evaluation of YOLOv5 and YOLOv8 object detection algorithms on resource-constrained embedded hardware. Proc IEEE ETFA 2024:1–7. https://doi.org/10.1109/ETFA61755.2024.10710707

[12] Kafadar K, Carriquiry AL. Challenges in modeling, interpreting, and drawing conclusions from images as forensic evidence. Stat Data Sci Imaging 2024;1(1). https://doi.org/10.1080/29979676.2024.2401758

[13] Jocher G, Chaurasia A, Qiu J. YOLO by Ultralytics. 2023. https://github.com/ultralytics/ultralytics (accessed 15 July 2026).

[14] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 2015;28:91–9. https://papers.nips.cc/paper_files/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html

[15] Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data 2019;6(1):1–54. https://doi.org/10.1186/s40537-019-0192-5

[16] Zou Z, Shi Z, Guo Y, Ye J. Object detection in 20 years: A survey. Proc IEEE 2023;111(3):257–312. https://doi.org/10.1109/JPROC.2023.3241644

[17] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019;6:60. https://doi.org/10.1186/s40537-019-0197-0

[18] Huang J, Rathod V, Sun C, et al. Speed/accuracy trade-offs for modern convolutional object detectors. Proc IEEE Conf Comput Vis Pattern Recognit 2017:7310–9. https://doi.org/10.1109/CVPR.2017.351

[19] Tripathi M. Analysis of convolutional neural network based image classification techniques. J Innov Image Process 2021;3:100–17. https://doi.org/10.36548/jiip.2021.2.003

[20] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE Conf Comput Vis Pattern Recognit 2016:770–8. https://doi.org/10.1109/CVPR.2016.90
[21] Everingham M, Van Gool L, Williams CKI, et al. The Pascal Visual Object Classes (VOC) challenge. Int J Comput Vis 2010;88:303–38. https://doi.org/10.1007/s11263-009-0275-4
[22] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432
[23] Saikia S, Fidalgo E, Alegre E, Fernández-Robles L. Object detection for crime scene evidence analysis using deep learning. In: Image Analysis and Processing – ICIAP 2017. LNCS vol 10485. Cham: Springer; 2017. p. 13–24. https://doi.org/10.1007/978-3-319-68548-9_2
[24] Taroni F, Bozza S, Biedermann A, Aitken C. Data Analysis in Forensic Science: A Bayesian Decision-Making Approach. Hoboken (NJ): Wiley; 2020.
[25] Bharath Kumar R, Baplawat A, Prabakaran P, et al. Artificial intelligence in forensic sciences: Bridging systematic challenges with next-generation applications. J Neonatal Surg 2025;14(5S):639–50. https://doi.org/10.52783/jns.v14.2105
[26] CrimeSceneObjectsDetection. Crime scene Dataset. Roboflow Universe; 2023. https://universe.roboflow.com/crimesceneobjectsdetection/crime-scene-oma5u
[27] maheshchhetri. weapon-detection Dataset. Roboflow Universe; 2023. https://universe.roboflow.com/maheshchhetri/weapon-detection-e6otc
[28] Perception07. Grenade Dataset. Roboflow Universe; 2024. https://universe.roboflow.com/perception07-xuigk/grenade-1jsgf
[29] set2. Crime Scene Dataset. Roboflow Universe; 2025. https://universe.roboflow.com/set2-v3223/crime-scene-ujkww
[30] E. Evidence Detection Dataset. Roboflow Universe; 2025. https://universe.roboflow.com/e-oujox/evidence-detection-iq6a3
[31] anas. hammer Dataset. Roboflow Universe; 2023. https://universe.roboflow.com/anas-loscp/hammer-9e6jt
[32] PhD. pistol, fire, and gun Dataset. Roboflow Universe; 2023. https://universe.roboflow.com/phd-xqusx/pistol-fire-and-gun
[33] Rapidev. Gun-Knife-Pistol-Grenade Dataset. Roboflow Universe; 2022. https://universe.roboflow.com/rapidev-pmcmr/gun-knife-pistol-grenade