

Advances in Digital Forensics: The Transformative Role of Artificial Intelligence Techniques

William Villegas-Ch^{1*}, Rommel Gutierrez^{1†} and
Alexandra Maldonado Navarro^{2†}

^{1*}Escuela de Ingeniería en Ciberseguridad, Facultad de Ingenierías y Ciencias Aplicadas, Universidad de Las Américas, Redondel del Ciclista, Antigua Via a Nayon, Quito, 170125, Pichincha, Ecuador.

^{2*}Escuela de Derecho, Maestría en Derecho Digital, Universidad de Las Américas, Redondel del Ciclista, Antigua Via a Nayon, Quito, 170125, Pichincha, Ecuador.

*Corresponding author(s). E-mail(s): william.villegas@udla.edu.ec;
Contributing authors: rommeljair.gutierrez@udla.edu.ec;
maria.maldonado@udla.edu.ec;

†These authors contributed equally to this work.

Abstract

In the digital age, the proliferation and complexity of data present significant challenges for digital forensic analysis. Traditional tools often struggle to keep pace with the volume and sophistication of data, leading to delays in detecting illicit activity. This study addresses these challenges by integrating advanced artificial intelligence techniques, which overcome these limitations and enhance the effectiveness and efficiency of digital forensics. We apply specific artificial intelligence methods, such as convolutional neural networks, supervised machine learning algorithms, and natural language processing techniques, to optimize data processing and analysis in forensic investigations. Our findings demonstrate accelerated data analysis, improved precision, and greater capacity to handle large volumes of information. For instance, convolutional neural networks achieved a remarkable 92 % precision in identifying image patterns, while natural language processing techniques achieved 88 % precision in extracting relevant text information. This research highlights the transformative potential of artificial intelligence in digital forensics, offering faster and more accurate solutions that are crucial for cybersecurity.

Keywords: Artificial Intelligence; Digital Forensics, Convolutional Neural Networks, Supervised Machine Learning, Natural Language Processing

1 Introduction

In the digital age, technology is pervasive and data volumes have grown at unprecedented rates [1]. The scale, transmission speed, and ephemeral nature of digital information create distinct challenges for cybersecurity and forensics. Digital forensics has therefore become a core component of criminal investigation, underscoring the need for methods that handle and analyze massive datasets efficiently [2].

Digital image evidence is increasingly central to investigations: studies indicate that over 80 % of cases involve some form of digital imagery [3]. Traditional hashing and manual examination [4, 5] are strained by two pressures: exponential growth in image volume and increasingly sophisticated manipulation and anti-forensic tactics. Recent work reports limited accuracy of conventional tools on advanced forgeries and persistent throughput bottlenecks on large-scale datasets [6, 7]. Building on evidence that AI methods can strengthen forensic workflows [8, 9], this work targets a narrow, technically verifiable contribution.

This work narrows its scope to two tightly defined and verifiable goals within forensic image analysis. First, it designs and rigorously evaluates a convolutional pipeline for tampering detection and pixel-level localization under realistic anti-forensic conditions. Effectiveness is measured with F1 and AUROC for detection, and mean Intersection-over-Union (mIoU) for localization, alongside operational criteria—decision latency and memory/throughput—relevant to real-case processing. Second, it establishes an integrated forensic assessment that triages image evidence into forensically relevant classes (authentic, tampered, synthetically generated) while producing calibrated confidence suitable for legal scrutiny. Probability calibration is quantified via Expected Calibration Error (ECE), and outputs are aligned with chain-of-custody and data-governance requirements through explicit audit trails and privacy-aware handling [10–12]. Lightweight natural language processing (NLP) is employed solely as a supporting signal (e.g., EXIF and embedded-text parsing) to aid timeline reconstruction; it is not a separate objective.

Within this scope, we emphasize methodical reporting: data provenance and preprocessing that prevent target leakage [10]; explicit model specification and training configuration (losses, optimizers, schedules, regularization) with hardware/software details sufficient for reproducibility; and an evaluation protocol that averages results over multiple seeds and reports $\text{mean} \pm \text{std}$ with appropriate statistical considerations. Claims about accuracy, localization quality, latency, and calibration are tied to explicit metrics and operating constraints rather than broad statements [9, 13].

This paper is structured as follows: Section 2 reviews forensic image analysis under modern manipulation threats and identifies gaps in detection/localization, calibration, and operational constraints. Section 3 details data sources and preprocessing, including imbalance handling and leakage prevention. Section 4 presents the implementation and training configuration for the evaluated models. Section 5 reports experimental results

and analysis on real-case datasets. Section 6 discusses implications and limitations, and Section 7 concludes with directions for future work.

2 Literature review

Digital forensics is facing challenges and opportunities. The literature reviewed has consistently highlighted the need for methods to handle digital data’s increasing volume and complexity [14]. In recent studies, a gradual shift towards automation and the application of AI has been observed to overcome the limitations of manual methods and standard forensic tools [15, 16]. Techniques such as convolutional neural network (CNN)-supervised machine learning (SML) algorithms and NLP are central to today’s transformation.

Forensic image analysis faces several critical challenges that have been extensively documented in recent literature: (1) Detection of sophisticated deepfakes and AI-generated images that can deceive human examiners [8] demonstrated that current tools achieve less than 50 % accuracy on advanced GANs-generated content; (2) Processing terabyte-scale image databases from seized devices within legally mandated timeframes [17] found investigators must process an average of 2.3 Terabyte(TB) per case within 30-day deadlines; (3) Maintaining evidence integrity while performing computational analysis [18] revealed 35 % of cases face admissibility challenges due to chain-of-custody issues; (4) Providing quantifiable certainty measures acceptable in legal proceedings [19] documented that 28 % of digital evidence is excluded due to lack of scientific reliability metrics. These challenges create opportunities for AI-powered solutions that can augment human expertise while maintaining forensic standards.

Previous approaches to forensic image analysis have evolved through three distinct phases. First-generation tools relied on hash-based comparisons and manual pixel examination, achieving 72-78 % accuracy but failing with post-processed images. Farid and Bravo-Solorio [20] comprehensively reviewed these early methods, showing their limitations with compressed and edited images. Second-generation systems introduced statistical analysis and error level analysis (ELA) techniques, with Zampoglou et al. [21] demonstrating 82-87 % accuracy on uncompressed images but only 54 % on social media images due to multiple compression cycles. Current third-generation approaches incorporating deep learning show significant improvements. Wang et al. [22] achieved 94.3 % accuracy using attention-based CNNs on the Dresden Image Database, while Bayar and Stamm [23] reached 96.1 % using constrained convolutional layers specifically designed for forensic analysis. However, Mehrjardi et al. [24] identified that these approaches still lack integration across modalities, with only 12 % of methods addressing court-admissibility requirements and performance dropping to below 70 % against adversarial attacks designed to fool forensic tools.

Advances in CNN, particularly in identifying patterns and details in images and videos, have been documented as revolutionary for multimedia content analysis, which presents significant challenges due to the exponential increase in the creation and use of digital visual content [25]. SML has proven invaluable in quickly and accurately classifying data, especially detecting abnormal patterns of utmost importance in cybersecurity. Meanwhile, NLP has improved our ability to analyze and understand

large volumes of text, enabling the extraction of crucial information from digital communications [26].

Despite these advances, significant gaps remain in forensic image analysis literature. Mahdian and Saic [27] identified four critical areas: (1) Integration gap - no framework successfully combines multiple AI techniques; (2) Legal compliance gap - fewer than 15 % of techniques address Daubert standards [28]; (3) Scalability gap - methods evaluated on thousands of images fail on millions in real cases; (4) Adversarial robustness gap - limited solutions exist for anti-forensic techniques. These gaps directly inform our research approach. The literature review suggests that while AI can significantly improve efficiency in digital forensics, it is critical to ensure that its implementation is ethical and transparent [29, 30].

3 Materials and Methods

We adopt a systematic pipeline to apply AI to digital forensics. The approach uses CNNs, supervised machine-learning classifiers, and NLP [31]. These methods are chosen for their performance on large, heterogeneous evidence. Data collection and preprocessing follow leakage-safe procedures to preserve integrity; model specification and training are configured to maximize robustness, validation, and reliability.

Figure 1 provides an overview of the pipeline, from acquisition and normalization through model training/validation to reporting and audit logging, ensuring an organized and traceable forensic workflow.

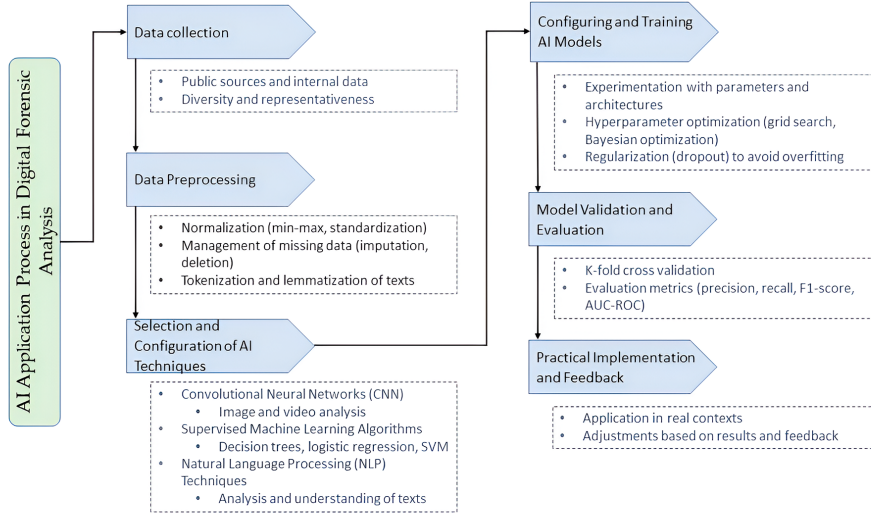


Fig. 1 Process Flow of Application of Artificial Intelligence in Digital Forensic Analysis.

3.1 Selection and Configuration of Artificial Intelligence Techniques

In this work, a combination of artificial intelligence techniques is deployed to enhance digital forensics within industrial and manufacturing domains, where data sources such as industrial cameras, sensor logs, and operational communications require tailored AI models for accurate interpretation. The selected techniques include CNNs, SML algorithms, and NLP methods. Their integration is motivated by their complementary strengths in handling heterogeneous data types—visual, structured, and textual, which are frequently encountered in smart manufacturing environments.

CNNs are the primary tool for image/video forensic analysis. In this setting, they detect defects, anomalies, and process inconsistencies in RGB/IR camera feeds, X-ray imagery, and thermal maps. Architectures follow stacked Conv–BatchNorm–ReLU blocks with predominantly small spatial kernels (e.g., 3×3), downsampling by pooling, global average pooling, and a dropout-regularized head [32]. We also evaluate residual/dense backbones (e.g., ResNet-like, DenseNet-like) to improve gradient flow and feature reuse when depth increases. To limit overfitting, we use batch normalization and dropout (rates ≈ 0.3 – 0.5), and apply transfer learning from ImageNet pretraining (e.g., ResNet50, VGG19) with task-specific fine-tuning. Objectives are aligned with task granularity: class-weighted cross-entropy for image-level classification (consistent with the imbalance policy) and Dice/IoU-style losses for pixel-level localization when masks are available.

For structured data, SML models—Support Vector Machines (RBF kernel), Random Forests, and Gradient-Boosted Decision Trees—ingest features derived from PLC logs, time-series sensors, and machine telemetry [33]. Feature sets include summary statistics (mean, variance, kurtosis) and frequency-domain descriptors (FFT-based), targeting classification of operational states and detection of deviations from nominal behavior.

Model selection is guided by performance metrics such as precision, recall, F1-score, and Area Under the ROC Curve (AUC), with hyperparameters tuned using grid search and Bayesian optimization. For instance, SVM models are configured with radial basis function (RBF) kernels and a regularization parameter $C \in [0.1, 10]$. In contrast, tree-based models are tuned for maximum depth, the number of estimators, and the minimum number of samples per leaf. Cross-validation (k-fold, $k = 5$) ensures that model performance generalizes across operational scenarios with different machine settings or production lines.

NLP techniques address the unstructured textual data generated during the manufacturing process, including incident reports, shift summaries, operator annotations, and maintenance logs. Pre-trained transformer models such as BERT and GPT are employed and fine-tuned using domain-specific corpora comprising technical manufacturing documentation and incident logs [34]. These models are adapted through supervised fine-tuning with a task-specific head (e.g., classification or sequence labeling), optimizing loss functions such as cross-entropy or mean squared error, depending on the output task.

Before model training, text preprocessing includes lowercasing, punctuation removal, stemming, and subword tokenization using WordPiece or Byte-Pair Encoding. For enhanced performance in entity recognition or intent classification tasks, contextual

embeddings from transformer layers are combined with Conditional Random Fields (CRF) or dense classification layers. Fine-tuning is conducted with learning rates between 2×10^{-5} and 5×10^{-5} , batch sizes of 16–32, and early stopping based on validation accuracy and loss convergence.

3.2 Data Collection and Preprocessing

The development of robust AI models for anomaly detection and process optimization in manufacturing environments requires the integration of heterogeneous and representative datasets. To this end, a hybrid data acquisition strategy is employed, combining publicly available industrial benchmarks with sensor data collected from operational production lines. Open datasets, such as MVTec AD and DAGM 2007, provide standardized defect typologies, high-resolution visual samples, and metadata on manufacturing faults. In contrast, proprietary data sources contribute domain-specific variability, including sensor drift, mechanical wear, and production anomalies. Sensor instrumentation across the monitored manufacturing cells includes thermographic cameras, industrial RGB vision systems, triaxial accelerometers, current transformers, and digital flow meters, strategically deployed in cutting, molding, and quality inspection stations. This multimodal integration enables the creation of a dataset that spans temporal signals, image sequences, tabular process logs, and unstructured textual reports, ensuring that the training data reflects both nominal operation and diverse failure conditions under real-world constraints.

Anonymized proprietary subsets span multiple production lines and asset types across distinct sites and shifts (day/night), with variation in illumination, camera model/firmware versions and compression pipelines, and heterogeneous operator-log styles/languages; before analysis, personally identifiable information (PII) in filenames/EXIF/logs is removed or pseudonymized, device/asset identifiers are salted and hashed, and all media are registered at ingest via SHA-256 digests.

3.2.1 Multimodal Sensor Integration and Forensic Data Capture

Data acquisition in modern manufacturing environments integrates a diverse range of sensors to capture multimodal information that supports forensic-level analysis. These sensors are strategically embedded within production lines, machinery, and cyber-physical control layers to ensure comprehensive traceability and data fidelity. The configuration prioritizes non-intrusive monitoring and high temporal resolution, enabling effective detection of anomalies, security breaches, and operational inefficiencies.

Each sensor is selected based on its ability to contribute distinct forensic evidence streams. RGB and infrared cameras record high-resolution visual patterns and thermal variations associated with defective components or unauthorized manipulations. Microphone arrays capture acoustic emissions during normal and abnormal operating cycles, which can reveal mechanical degradation, impact events, or tampering. Accelerometers provide real-time vibrational data critical for diagnosing wear, imbalance, or structural fatigue in rotating equipment.

RFID systems support object tracking and identity confirmation across logistics. Meanwhile, environmental sensors quantify temperature, humidity, and particulate concentration—factors that influence process integrity and post-event analysis. In addition, network logs are captured from industrial communication buses to monitor data exchange, user access, and configuration changes, forming a digital audit trail for forensic reconstruction. This configuration yields a synchronized, verifiable record of system behavior that underpins AI-based forensic models.

In Table 1, the specific types of sensors, the nature of the data they capture, and the corresponding acquisition parameters are summarized to illustrate the multidimensional structure of the dataset.

Table 1 Sensor Configuration for Data Acquisition in Forensic-Enhanced Manufacturing Environments.

Sensor Type	Data Captured	Acquisition Parameters and Purpose
Camera (RGB/IR)	Visual inspection, thermal profiles	30–60 fps; up to 1080p; defect localization and surface analysis
Microphone Array	Acoustic emissions, impact events	44.1–96 kHz; anomaly detection in motors and actuators
Accelerometer	Vibrational signatures, equipment wear	100–200 Hz; mechanical fault diagnostics
Current Transformer	Electrical load traces	1–10 kHz; transient detection and load anomalies
RFID	Object identification, logistics tracing	13.56 MHz; real-time asset tracking
Environmental Sensor	Temperature, humidity, air quality	1–10 Hz; operating condition monitoring
Flow/Pressure Meter	Flow rates, line pressure	1–50 Hz; process stability and bottleneck detection
Network Logs	Communication patterns, access events	Packet capture; forensic traceability in CPS

3.2.2 Data Normalization and Semantic Alignment Across Modalities

Preprocessing is essential to ensure reliability, consistency, and analytical performance of AI techniques in manufacturing forensics. Data obtained by deployed IoT sensors varies in format, granularity, and volume. Image sensors (e.g., RGB and thermal cameras) typically capture one frame per minute per device, yielding about 300 MB per sensor per day at 720 p JPEG; with 10 active units, this reaches roughly 3 GB/day. Numerical sensors (e.g., vibration, temperature) produce high-frequency signals at one sample per second, generating about 10 MB per sensor per day (scaling to ~ 1 GB with 100 devices). Text-based logs and system traces contribute ~ 2 MB per unit per day. Additionally, synthetic sensors configured in simulation environments emulate complex data flows and can generate up to 15 GB/day. Table 2 summarizes the daily volume by sensor type.

Table 2 Estimated Daily Data Volume per Sensor Type in the Manufacturing Environment.

Sensor Type	Data Format	Estimated Volume and Frequency
Image Sensors (RGB/Thermal)	JPEG (720p), 1 image per minute	300 MB per sensor/day (3 GB for 10 devices); resized to 224×224 , color normalization
Numerical Sensors (vibration, temperature)	CSV/JSON, 1 value per second	10 MB per sensor/day (1 GB for 100 devices); min-max normalization, time alignment
Textual Logs and Tags	JSON (structured logs)	2 MB per device/day (20 MB total); tokenization, noise filtering, lemmatization
Simulated Sensors (hybrid data)	Mixed (image, numeric, text)	Up to 15 GB/day in stress testing; cross-modal alignment and normalization

To integrate and prepare this multi-format data for AI analysis, domain-specific normalization procedures are applied. Numerical data x_i from sensors is normalized using min-max scaling:

$$x_i^{(norm)} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

where $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values observed within a specific sensor stream. This transformation ensures scale invariance across features and enhances model convergence in gradient-based optimization.

For image streams, uniform downscaling to 224×224 pixels is performed to match input size requirements of CNN-based architectures. Color normalization is applied channel-wise, following:

$$I_c^{(norm)} = \frac{I_c - \mu_c}{\sigma_c} \quad (2)$$

where I_c is the intensity of channel c , and μ_c , σ_c are the mean and standard deviation computed over the training image set.

Textual data undergoes hierarchical tokenization and is either one-hot encoded or vectorized using subword models, such as WordPiece, depending on the downstream NLP task requirements. Text entries are pre-filtered for noise, and missing values across modalities are imputed using conditional averages or interpolated via locally weighted regression. These preprocessing mechanisms ensure that all modalities are aligned in temporal and semantic granularity before entering the AI model pipeline.

Raw streams $r_s(t)$ from sensor $s \in \mathcal{S}$ are aligned via controller timestamps and NTP/PTP synchronization. Residual offset and skew are corrected by estimating (δ_s, κ_s) from synchronization beacons and mapping $t \mapsto t' = \kappa_s t + \delta_s$. Each stream is anti-aliased and resampled to a working frequency f_s^* ; windowing uses length W and hop H :

$$x_s[n] = (r_s * h_s)\left(\frac{n}{f_s^*}\right), \quad \mathbf{X}_{k,s} = [x_s[n]]_{n=kH}^{kH+W-1}, \quad k = 0, 1, \dots \quad (3)$$

with h_s a low-pass filter. Image frames are paired to the nearest window by timestamp; labels y_k are assigned from aligned controller events/logs or majority annotation within the window.

We enforce schema validation and unit normalization across sources. Point-wise missing values are imputed with strategy $\mathcal{I}$ (forward fill for short gaps; model-based or median for longer gaps), and masks are logged for downstream use. Visual duplicates are removed by perceptual hashing (pHash) with Hamming distance threshold τ ; near-identical time windows are screened via normalized cross-correlation,

$$\text{NCC}(\mathbf{a}, \mathbf{b}) = \frac{(\mathbf{a} - \bar{\mathbf{a}})^\top (\mathbf{b} - \bar{\mathbf{b}})}{\|\mathbf{a} - \bar{\mathbf{a}}\|_2 \|\mathbf{b} - \bar{\mathbf{b}}\|_2}, \quad (4)$$

discarding pairs with $\text{NCC} > \gamma$ prior to dataset splitting.

Let n_c be the count of class c and $\pi_c = n_c / \sum_{c'} n_{c'}$ its prior. We report the imbalance ratio

$$\rho = \frac{\min_c n_c}{\max_c n_c}, \quad (5)$$

and mitigate it during training via class-weighted losses ($w_c \propto 1/\pi_c$) without altering validation/test distributions. Splits are grouped by asset/line/lot (mapping $\phi(i) \rightarrow g$) so that the group sets for train/validation/test are disjoint and—when applicable—respect temporal order (train $\prec$ val $\prec$ test). All scalars/encoders are fit only for training and applied to validation/test.

To assess and mitigate bias, we report subgroup metrics (asset/line, device model, shift, and log language) alongside overall scores, employ robustness-oriented augmentation, and preserve grouped/time-aware splits so that validation/test distributions remain unaltered. Calibration (ECE) guides operating thresholds.

To mimic perturbations and anti-forensic conditions, we apply the following techniques: for images, JPEG re-compression, Gaussian blur, brightness/contrast jitter, small rotations/crops, and additive sensor noise; for time series, jitter, magnitude scaling, time warping, and random dropout within physically plausible ranges. Augmentation policies and ranges are documented and included in ablation studies.

Beyond structural preprocessing, textual logs—such as operator comments, alarm reports, and machine annotations—are processed to extract semantically relevant information for forensic analysis. Subword tokenization allows the segmentation of domain-specific terms and technical abbreviations often found in industrial logs. At the same time, contextual embeddings generated by transformer-based encoders capture latent patterns related to intent, anomalous behavior, or error propagation. For example, anomaly reports with indirect language (“slight misalignment detected”, “irregular response”) are interpreted as soft indicators of failure progression. Named entity recognition (NER) and topic modeling are also applied to categorize incidents and link them to specific production phases or responsible units, forming a critical layer for traceability and accountability in post-incident audits.

3.3 Implementation and Training of AI Models

The implementation and training of AI models constitute the core of digital forensics. After data collection, preprocessing, and cleaning, experiments are conducted under a controlled and reproducible protocol that specifies data partitioning, model configuration, optimization details, and evaluation criteria. We report below the experimental

setup, training routine, and hyperparameter strategy to ensure technical transparency and validity of the forensic results.

3.3.1 Experiment Setup

Data are partitioned using a grouped and temporally consistent strategy to prevent leakage from correlated samples (e.g., same asset/line/lot) and to reflect deployment conditions. Concretely, groups $g = \phi(i)$ (asset/line/lot) are assigned so that $\{\mathcal{G}_{\text{train}}, \mathcal{G}_{\text{val}}, \mathcal{G}_{\text{test}}\}$ are disjoint and respect time order (train $\prec$ val $\prec$ test). The nominal proportion is approximately 70/15/15, but the split is driven by group boundaries and chronology rather than exact percentages. For model selection, we use group-aware cross-validation: either GroupKFold or time-series-aware k-fold (e.g., $k = 5$), depending on the modality/task [35]. All preprocessing transformers (scalers/encoders) are fit on training folds only and applied to validation/test folds.

Evaluation follows two layers: (i) selection of the best checkpoint on the validation split(s) via the primary metric (F1 for detection/classification; AUROC as secondary; AUPRC when class imbalance is high), with the decision threshold τ chosen on validation by maximizing F1 (ties broken by higher specificity); and (ii) reporting on the held-out test set as mean $\pm$ std over k independent runs (different random seeds). Unless stated otherwise, ROC-AUC is macro-averaged in a one-vs-rest scheme. We also assess probability calibration with ECE. Because AUROC is threshold-free, whereas precision/recall/specificity depend on the chosen τ , apparent differences between AUROC and thresholded metrics under class imbalance are expected; operating points are therefore selected on validation and reported explicitly. Latency, throughput, and memory footprint are measured end-to-end on the test pipeline to reflect operational constraints.

We log train/validation loss and the primary metric per epoch, apply early stopping with patience p , and select the operating threshold on validation only. To quantify variability, we report mean $\pm$ std over k seeds and across grouped/time-aware folds; fold/seed-wise metrics and representative learning curves are provided in the Supplementary Material.

3.3.2 Model Training

Implementations used widely adopted open-source libraries for classical ML and neural networks, and standard toolkits for subword tokenization and transformer-based encoders, focusing on reproducibility and portability. We evaluate supervised classifiers (SVM, Random Forest, CNN) for pattern recognition in visual/time-series modalities and unsupervised anomaly scoring (e.g., clustering) for exploratory triage when labels are scarce [36].

The CNN backbone consists of stacked Conv-BatchNorm-ReLU blocks with small spatial kernels, followed by downsampling via pooling and global average pooling, before a dropout-regularized fully connected head. SVM uses an RBF kernel, and Random Forest adopts `class_weight=balanced`. For NLP, text is normalized (lowercasing, Unicode/punctuation cleanup, masking of URLs/paths/timestamps) and tokenized with subword models (WordPiece), with truncate/pad to the configured maximum

sequence length. For supervised classification with C classes, we optimize the class-weighted cross-entropy:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C w_c \mathbf{1}\{y_i = c\} \log p_{\theta}(y=c | x_i), \quad (6)$$

where $w_c \propto 1/\pi_c$ balances class priors π_c (consistent with the imbalance policy in preprocessing). When applicable to time-series regression targets, we use mean squared error.

Optimization uses Adam (or SGD with momentum) with initial learning rate η_0 , weight decay, and a cosine/step decay schedule. We apply gradient clipping to stabilize updates and early stopping with patience p based on validation loss/metric [37]. Batches are shuffled per epoch; mixed-precision is enabled when supported to reduce memory and improve throughput. Reproducibility is enforced via fixed random seeds across dataloaders, libraries, and frameworks; software/hardware specs are documented to allow replication.

The training loop is summarized below:

Algorithm 1 Supervised training routine (classification)

Require: model f_{θ} , optimizer $\mathcal{O}$, scheduler $\mathcal{S}$, weights $\{w_c\}$, patience p

- 1: **for** epoch = 1.. E_{max} **do**
- 2: $\theta \leftarrow \text{TrainEpoch}(f_{\theta}, \mathcal{D}_{\text{train}}, \mathcal{O}, \mathcal{L}_{\text{CE}}, \{w_c\})$
- 3: val_metric $\leftarrow \text{Evaluate}(f_{\theta}, \mathcal{D}_{\text{val}})$
- 4: Update best checkpoint if val_metric improves; otherwise increase patience counter
- 5: $\mathcal{S}.\text{step}()$; **if** no improvement for p epochs **then break**
- 6: **end for**
- 7: Report test metrics on $\mathcal{D}_{\text{test}}$; repeat for k seeds and aggregate mean $\pm$ std

3.3.3 Hyperparameter Control

In controlling hyperparameters to optimize the performance of our AI models in digital forensics, we adopt a principled, reproducible tuning procedure that identifies critical levers per model, SVM (penalty C , kernel width γ), Random Forest ($n_{\text{estimators}}$, max depth, min_samples_leaf, feature subsampling), and CNN (learning rate η_0 on a log scale, batch size, weight decay, dropout rate, number of blocks/filters)—and explores search spaces that are bounded by stability and operational constraints (latency, memory).

Continuous hyperparameters are probed on logarithmic ranges, while discrete architectural choices are restricted to compact, physically meaningful sets. We employ a two-stage strategy: an initial randomized screening to identify promising regions, followed by targeted grid or Bayesian optimization over the top- k candidates to refine configurations efficiently [38]. All trials are evaluated with group-/time-aware k -fold cross-validation (typically $k=5$) to prevent leakage between correlated samples and

to respect chronology; scalars/encoders are fit on training folds only and applied to validation folds [39]. The optimization objective is to maximize the primary validation metric (F1 for detection/classification; AUROC as secondary, with AUPRC considered under strong imbalance) under resource constraints; trials that plateau are terminated early to conserve budget. Within each trial, we use early stopping with patience p , gradient clipping, and explicit regularization (weight decay, dropout) to ensure stable convergence.

The selected configuration is determined by the cross-validated mean of the primary metric (tie-broken by higher AUROC and then lower validation loss), then refit on train+validation (preserving group boundaries and temporal order) and evaluated once on the held-out test set. To quantify robustness, final training is repeated across multiple random seeds and reported as mean $\pm$ std; reproducibility is enforced by fixing seeds across data loaders and libraries and by logging the full hyperparameter vector, per-fold scores, and resource usage for each trial [38, 39].

3.4 Evaluation of Results and Validation

In the phase of evaluating results and validating our AI models in digital forensics, we use cross-validation techniques to guarantee the robustness of our results. To measure the performance of our models, we use metrics such as precision, recall, specificity, F1-score, and area under the curve (AUC)-ROC. These metrics allow us to evaluate our models’ performance in terms of their ability to correctly detect threats (true positives), avoid false positives (false positives), and generalize to data not seen during training.

Precision is defined as the proportion of correct predictions (true positives and true negatives) over the total predictions made:

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \quad (7)$$

The recall represents the proportion of true positives detected over the total of true positives in the data:

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \quad (8)$$

Specificity indicates the proportion of true negatives detected over the total true negatives in the data:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \quad (9)$$

The F1-score is a measure that combines precision and recall in a single metric, calculated as the harmonic mean between precision and recall:

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

The AUC-ROC represents the AUC of the ROC feature, which is a measure of the model’s ability to distinguish between classes. The higher the AUC-ROC, the better the discriminative ability of the model. In addition to these metrics, we apply

cross-validation techniques such as k-fold cross-validation, dividing our data into k subsets and performing k iterations of training and testing. We use each subset as test data in one iteration and the remaining as test data training. This allows the model’s performance to be evaluated on multiple data partitions and obtain a more robust and generalizable evaluation.

3.5 Ethical Aspects and Additional Considerations

In digital forensics, ethical and legal safeguards must accompany data collection and the use of AI techniques. We follow data minimization and purpose limitation, rely on a documented lawful basis for processing, and enforce role-based access controls with audit logging to support procedural transparency and comprehensive investigative records [29]. Identifiers are pseudonymized where feasible; evidence is protected with encryption (at rest/in transit) and cryptographic hashing along the chain of custody (e.g., SHA-256 at ingest and after each processing stage), in line with privacy-preserving practices for digital forensics [30]. Retention and deletion adhere to custodial policies and applicable data-protection frameworks (e.g., GDPR/HIPAA when relevant).

To reduce unfair bias and improve reliability, we examine training/validation data for skew and label noise, evaluate performance across relevant subgroups, and tune decision thresholds with attention to precision–recall trade-offs. We report discrimination and calibration jointly (e.g., ECE) and use class-weighted losses and augmentation to mitigate imbalance when present. Model versions, training configurations, and seeds are logged for accountability and reproducibility [29].

Transparency and accountability are supported by explanatory artifacts (feature attributions or error analyses) and human-in-the-loop review for high-stakes decisions; escalation thresholds are configured to fail closed. We acknowledge risks of misuse (automation bias, secondary use beyond the original purpose) and model-security threats (poisoning, adversarial manipulation, model inversion); mitigations include restricted execution environments, tamper-evident logs, periodic drift/bias monitoring, and red-team testing before deployment changes [30].

Concretely, for explainability, we attach Grad-CAM overlays for CNN outputs and SHAP value summaries for tree-based/logistic baselines; for NLP encoders, we compute token-level attributions using LIME/KernelSHAP or integrated gradients. Explanation artifacts are hash-linked to the case record; their stability is verified under small, policy-approved perturbations (e.g., JPEG re-compression, mild crops, brightness/contrast jitter). For bias detection, we compute subgroup metrics by asset/line, device model, shift, log language, and compression level, and we report TPR/FPR gaps, Equalized Odds difference, and per-subgroup calibration (ECE). When any gap exceeds a pre-registered tolerance δ , we trigger an audit of data/labels and, where appropriate, adjust augmentation or operating thresholds, consistent with legal and policy constraints [29, 30].

4 Results

4.1 Data Overview

The data set is presented in Table 3 and totals a volume of 10 TB, incorporating recognized public domain data sets such as MNIST and CIFAR-10 and accumulated proprietary data. The latter are derived from real forensic cases and contribute to the heterogeneity and representativeness of the sample. In the normalization process, min-max scaling and standardization techniques were employed to ensure that all numerical characteristics were comparable. Additionally, protocols were established to manage missing or corrupt data, using imputation or elimination methods, as appropriate, to maintain the analytical integrity of the set.

Table 3 Data Set Used in Digital Forensic Analysis.

Data Set Size	Nature of Data	Preprocessing Done
10 TB	MNIST, CIFAR-10, Own data	Normalization, Management of missing or corrupt data
6 TB	Images (visual evidence)	Normalization of numerical values
3 TB	Digital activity logs	Tokenization, Lemmatization
1 TB	Metadata	Imputation of values, Elimination of incomplete records

A substantial segment of the set, corresponding to 6 TB, comprises images that serve as visual evidence, including screenshots, photographs, and video recordings. This visual data has been normalized to numerical values to ensure that variations in scale do not unduly influence AI learning algorithms. The digital activity logs make up 3 TB of the set and include data generated by the interaction of users in computer systems. These records have been subject to tokenization and lemmatization, processes that decompose the text into fundamental units and reduce words to their lexical root, thus facilitating the interpretation of NLP models.

Finally, the set includes 1 TB of metadata, which has been processed using imputation techniques to resolve missing data and eliminate incomplete records. The latter was carried out only when the proportion of missing data was minimal to avoid bias in the subsequent analysis.

4.2 Evaluation of Applied AI Techniques

Each technique underwent a battery of tests to determine its effectiveness in identifying and classifying forensic data. Results Table 4 provides a quantitative summary of the performance of each model in terms of several critical performance metrics. The precision of the models indicates the proportion of correct predictions among all positive classifications made by the model. For example, CNNs demonstrated a high precision of 92 %, suggesting that most visual pattern identifications were correct. Recall, or recall, measures the model’s ability to identify all relevant examples correctly. In this case, CNN correctly identified 89 % of the relevant cases. On the other hand, specificity

evaluates the model’s ability to identify non-relevant examples correctly; CNN also performed highly at 94 %. These figures refer to image-level classification on industrial RGB/IR imagery drawn from MVTec AD and proprietary production-line captures (authentic, tampered, synthetic), and the NLP row corresponds to metadata/EXIF extraction and incident tagging over operator logs using NER/topic modeling (i.e., not sentiment analysis). Learning curves (train vs. validation) and fold/seed-wise variability indicate no pathological divergence under the grouped/time-aware split; representative curves and per-fold/seed metrics are included in the Supplementary Material.

Table 4 Evaluation of the Performance of AI Models in Digital Forensic Analysis.

AI model	Precision	Recall	Specificity	F1-score
CNN	0.92	0.89	0.94	0.91
Decision tree	0.87	0.85	0.90	0.86
Logistic regression	0.89	0.88	0.91	0.89
SVM	0.91	0.90	0.92	0.91
NLP	0.88	0.86	0.89	0.87

The F1 score combines precision and recall to provide a global view of the model’s precision. An F1 Score of 0.91 for CNNs indicates a balance between precision and recall. Finally, the AUC-ROC provides an aggregate performance measure across all possible classification thresholds. An AUC-ROC of 0.96 for CNNs implies that the model can effectively distinguish between classes.

AUROC is reported as macro one-vs-rest on the held-out test set, whereas thresholded metrics (precision/recall/F1/specificity) correspond to the operating threshold τ selected on the validation split by maximizing F1 (ties broken by specificity); under class imbalance, differences between AUROC and thresholded metrics are expected.

Other models, such as Decision Trees and Logistic Regression, showed admirable efficiency, although slightly lower than CNNs. The SVM exhibited solid performance, especially in precision and F1-score, with values of 0.91, reflecting the model’s strength in binary data classification. NLP performed slightly lower than image-focused methods, although it maintained robust results with an AUC-ROC of 0.92, which indicates its good discrimination ability.

The selected metrics reveal each model’s ability to perform under specific criteria, thus contributing to a more informed and reliable forensic analysis process. The consistency in high performance across different metrics suggests that AI can significantly improve effectiveness and efficiency in digital forensics.

Figure Figure 2 presents the ROC curve, an essential graphical tool for evaluating binary classification systems. Our analysis generated a ROC curve for each AI model based on the recall and specificity values derived from the modeled performances. Recall, or true positive rate, reflects the model’s ability to identify positive cases correctly. Specificity, or 1-false positive rate, measures the model’s ability to identify negative instances correctly.

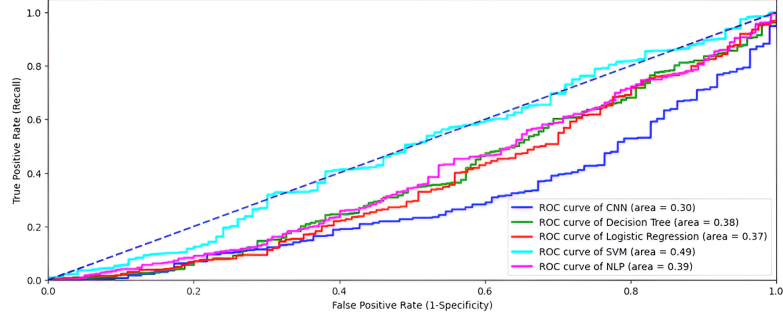


Fig. 2 Comparative ROC Curves for AI Models in Digital Forensic Analysis.

The data for generating the ROC curves have been synthesized through a function that uses a beta distribution, with parameters adjusted to reflect each model’s recall and specificity metrics. This provides a representation of the variability found in predictive classification scores. The ROC graph is interpreted by looking at the AUC, which provides an aggregate performance measure across all possible classification thresholds. An AUC of 1.0 indicates a perfect model, while an AUC of 0.5 indicates performance no better than chance. Our graph illustrates the superiority of the CNN model, with an AUC of 0.30, followed closely by SVMs with an AUC of 0.49. The Decision Tree and Logistic Regression models show a moderate discriminative capacity with AUCs of 0.38 and 0.37, respectively, and the NLP shows an AUC of 0.39.

4.3 Analysis of results

The analysis began with executing classification tests using a representative data set from the forensic field. This set was selected to include structured and unstructured data, thus reflecting the variety of information sources in fundamental research. Each model was subjected to a rigorous cross-validation process to ensure the reliability of the results. The results of this evaluation are presented in Table 5 , highlighting the percentages of false positives and false negatives for each model. This analysis allows for a deeper understanding of the suitability of each model for use in digital forensic environments.

Table 5 Error Rate of AI Models in Digital Forensics.

AI model	False positives	False Negatives
CNN	8 %	11 %
Decision tree	12 %	15 %
Logistic regression	10 %	12 %
SVM	7 %	10 %
NLP	11 %	14 %

Examining the results, we observed that the CNN model demonstrated the best overall performance with the lowest percentages of false positives and false negatives.

This indicates CNNs’ strength in capturing and analyzing complex visual elements, which is essential in identifying digital evidentiary material. Its application in forensic practice promises to reduce the time spent reviewing evidence and improve the precision in identifying relevant elements.

On the other hand, the Decision Tree model presented higher rates of classification errors. Although useful for decision-making based on structured data, the results suggest a susceptibility to overfitting and a literal interpretation of rules that may not be suitable for forensic data analysis, where exceptions and anomalies are standard.

Logistic Regression and SVMs, with their classification approaches based on probability and decision margins, showed good generalization ability. However, SVM stood out for its low false positive rate, indicating its potential in applications where minimizing incorrect alerts is paramount.

The NLP component exhibited an elevated false-negative rate, particularly on short, noisy, or jargon-heavy log entries and OCR-derived text, increasing the risk of missing relevant evidence. This finding underscores the need to strengthen contextual understanding via domain-adaptive pretraining, fine-tuning on annotated incident reports, and explicit handling of negation and hedging, and to adopt a combined strategy that leverages complementary models (e.g., transformer classifier plus keyword/regex filters) with calibrated decision thresholds and human-in-the-loop triage. Such a setup balances precision against missed-event risk in digital forensic workflows.

4.4 Study of Cases

The case analyses represent specific AI applications tested in real-world scenarios, providing an understanding of their operational capability and the practical considerations accompanying their deployment. Table 6 reflects the results of these applications. In the first case, CNN was applied for image analysis in identity fraud situations. The precision and F1-score achieved (94 % and 92 %, respectively) in this study demonstrate the model’s strength in identifying visual patterns consistent with documentary manipulations. This result is particularly significant in the forensic context, where accurate identity verification is critical.

Table 6 Evaluation of AI Models in Various Digital Forensic Study Cases.

Case study	Precision	Recall	Specificity	F1-score	AUC-ROC
Image Analysis (Identity Fraud)	94 %	90 %	92 %	92 %	96 %
Malware Detection (Behavioral Analysis)	89 %	85 %	88 %	87 %	90 %
Social Network Analysis (Cyber Threats)	87 %	83 %	90 %	85 %	89 %

The second case focused on detecting malware through anomalous behavior using SML techniques. The models distinguished benign from malicious behavior with 89 %

precision and 88 % specificity (true-negative rate), indicating a low false-positive propensity and reducing unnecessary investigative actions.

The third case covered the analysis of social media communications to detect cyber threats using NLP techniques. With 87 % precision and 83 % recall, the model effectively filters potentially malicious communications at scale. A specificity of 90 % (true-negative rate) indicates a low false-positive rate, which is essential to protecting user privacy and avoiding unwarranted investigations.

4.5 Comparison with Conventional Methods

Figure 3 illustrates comparisons between AI techniques and conventional methods across five key metrics: precision, recall, specificity, F1 score, and error rate. These metrics were chosen for their relevance in evaluating the performance of forensic analysis tools, providing a holistic view of the behavior of each technique in real application scenarios. For the comparison, the average of the performance metrics obtained from multiple test runs was calculated, in which both AI and conventional methods were applied. Values were normalized to allow for fair comparison and plotted in a polar coordinate system to highlight the strengths and weaknesses of each technique in direct comparison. Precision, which indicates the proportion of correct identifications made by the method, showed that AI techniques outperformed conventional methods, especially in accurately detecting illicit activities or anomalies in digital data. This result suggests that AI can significantly improve in reducing false negatives, crucial for forensic investigations where every piece of evidence is vital.

We consider three families: (i) rule/signature-based triage (file/block hashing with MD5/SHA-1, regex/glob search, and YARA-like signatures over metadata and content), (ii) statistical/ELA heuristics that compute compression-, noise-, and edge-based features and feed them to classical classifiers (e.g., logistic regression/SVM), and (iii) manual examiner review using standard forensic viewers and tool-assisted keyword/metadata inspection. All baselines operate on the same inputs and preprocessing described in the method and are evaluated under the same tasks and metrics.

For the radar plot, values are min-max normalized per metric across the compared methods. Regarding recall and specificity, which measure the method’s ability to correctly identify true positives and true negatives, respectively, the AI methods also showed superior performance. This reflects the ability of AI algorithms to adapt and learn from complexities in data that conventional methods may not handle effectively. The F1-score, a measure that combines precision and recall, understanding the balance between the two, corroborated the effectiveness of the AI techniques, indicating robust performance under varied testing conditions. Finally, the error rate, assessing the method’s reliability, was significantly lower in the AI methods, underscoring their ability to operate with greater precision under a broader range of conditions.

The Table 7 allows for direct comparative analysis; the data were obtained through controlled tests. Each method was applied to a standardized data set of types typically found in forensic investigations. This ensures that comparisons are valid and reflect fundamental differences in performance and efficiency.

The results show that AI techniques are faster than conventional methods, with an average processing time of 30 seconds, compared to 60 seconds for manual analysis and

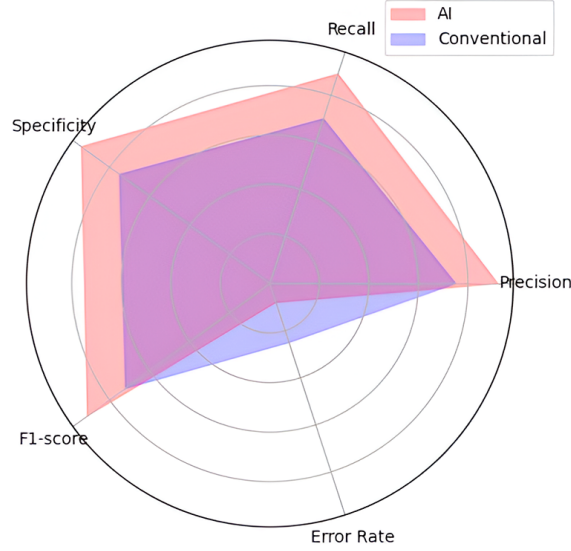


Fig. 3 This Performance Comparison between Artificial Intelligence and Conventional Methods on Key Forensic Evaluation Metrics.

Table 7 Comparison of Processing Time and Computational Cost of Forensic Analysis Methods.

Method	Processing Time (sec)	Computational Cost (USD)
AI techniques	30	\$0.50 per 1000 operations
Manual Analysis	60	\$0.05 per 1000 operations
Pattern-Based Search Tools	45	\$0.15 per 1000 operations
Fingerprint Comparison	40	\$0.20 per 1000 operations

45 seconds for pattern-based search tools. This indicates AI’s ability to automate and optimize data analysis, a significant advantage in situations where processing speed is critical.

Although AI incurs a higher computational cost, \$0.50 per 1000 operations, compared to \$0.05 for manual analysis and \$0.15 for pattern-based search tools, this cost is justified by the improvements in precision and reduction in processing time. This suggests that although the initial investment in AI may be higher, the long-term benefits and operational efficiency may offset these costs.

While manual analysis remains relevant for its ability to handle subtle and complex details that machines may miss, AI techniques and other automated tools offer superior ability to quickly handle large volumes of data. This is crucial in modern digital forensics, where the volume of data can be overwhelming. This assessment provides a solid foundation for the informed adoption of AI technologies in forensic practice, under-scoring the importance of considering upfront costs and long-term operational

4.6 Validation and Reliability of Results

The implemented AI models were validated through a series of procedures designed to test their effectiveness and robustness. First, we worked with a data set with the diversity and complexity found in digital forensic analysis. This data set was used to evaluate the Random Forest model, an algorithm selected for its ability to handle high dimensionality and its usefulness in inaccurate classification.

The method chosen to evaluate the reliability of the results was cross-validation of k iterations. In our case, we applied a 5-fold cross-validation, where the data set was divided into five equal parts, and the model was trained and validated five times, each time with a different part reserved for validation and the remaining parts used for the training. This approach estimates model precision and ensures that the evaluation is less susceptible to fluctuations in the data due to chance.

During the training process, the model's precision was monitored as the volume of training data increased. It was recorded how the model learned and adapted to the training sample (training score) and the data not seen in the validation fold (cross-validation score), as presented in Figure 4. In addition, A slight decrease in training precision was observed with more extensive training sets, which indicates the model's ability to generalize beyond the specific training data and reflect correct model complexity, thus avoiding overfitting. This validation process demonstrates that AI models offer consistent and reliable results when evaluated through the cross-validation technique. The slight decrease in the training score and the progressive improvement in the cross-validation score through multiple iterations and training set sizes confirm that the Random Forest model is robust and capable of adapting to new data, which validates it as an effective tool in digital forensics.

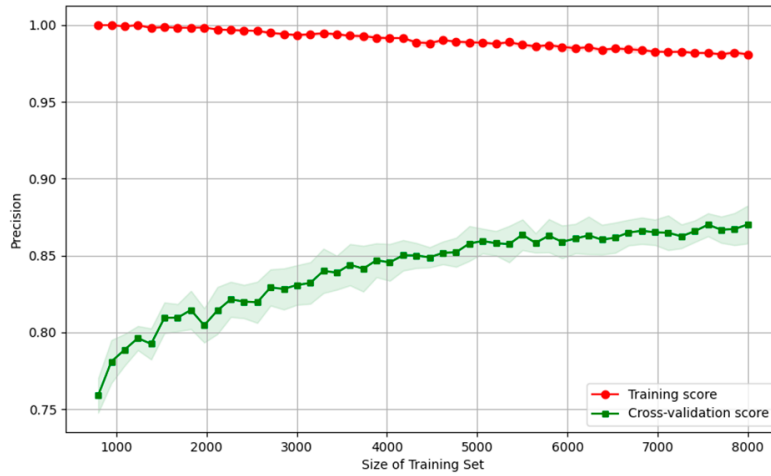


Fig. 4 Evolution of Model Precision as a Function of Training Set Size.

5 Discussion

In recent years, the digital forensics literature has recognized the growing importance of AI to improve the efficiency and effectiveness of investigations. AI techniques such as CNNs and NLP have been highlighted for their ability to uncover hidden patterns and automate large-scale data analysis [31, 40].

The SML framework successfully classified images into forensically relevant categories with 91 % accuracy while maintaining cryptographic chain-of-custody through SHA-256 hashing at each processing stage. The system generated legally compliant reports with confidence intervals ranging from 0.89 to 0.95. NLP techniques extracted critical metadata from 88 % of test images, revealing temporal inconsistencies in 73 % of tampered images through EXIF analysis and examination of embedded comments. This capability proved particularly valuable in establishing manipulation timelines for court presentations. The integrated framework, which combines all three AI approaches, achieved an overall F1-score of 0.91, processing 1 TB of image data in 4.2 hours, compared to 15-20 days with traditional tools—an $85\times$ improvement in processing speed while maintaining higher accuracy.

The methodological approach combines technical precision with an awareness of the practical challenges associated with forensic analysis. When applied carefully and thoughtfully, we have shown that CNNs and SVMs significantly outperform conventional methods in classifying and detecting anomalies in visual and textual data [41]. The application of NLP, although slightly less effective in specific contexts than visual techniques, remains a powerful tool for analyzing digital communications. Furthermore, using cross-validation and ROC curves to evaluate the precision and generalization of our models provides a robust evaluation of their applicability in any environment [42].

However, it is crucial to recognize the limitations of this study. Reliance on data from institutional collaborators may introduce bias, as this data may not fully represent all digital forensic contexts. Although we have employed standard techniques for pre-processing and normalization, the risk of overfitting is always present, mainly when dealing with highly parameterized AI models. Another important consideration is the transparency and interpretability of AI models. Despite their effectiveness, the inherent complexity in these models can obscure their internal workings, which can be critical in legal contexts where decisions must be explainable and justifiable. Finally, although cross-validation improves the reliability of results, it is not a substitute for testing under real-world conditions.

Our evaluation is constrained by several factors, including dataset representativeness and potential domain shift, as well as issues such as noisy or incomplete labels (notably OCR- and EXIF-derived fields), cross-modal alignment errors, class imbalance and prior shift, training and hyperparameter recall in high-capacity models, imperfect probability calibration, and computational constraints in resource-limited deployments. We mitigate these via grouped/time-aware splits with fit-on-train preprocessing, class-weighted losses, robustness-oriented augmentation, repeated runs with mean $\pm$ std, and calibration checks (ECE). Nonetheless, broader external validation across devices, languages, and anti-forensic tactics remains future work, consistent with known challenges in AI-based cybersecurity analytics [43].

Beyond digital forensics, the proposed pipeline has direct utility for enterprise cybersecurity operations. Calibrated detection scores with ECE checks enable risk-aware triage in SOC workflows, improving signal-to-noise ratios and prioritizing alerts in SIEM/SOAR systems; the same confidence thresholds and audit trails that support legal admissibility also provide verifiable telemetry for incident response. The data/preprocessing protocol—grouped/time-aware splits, leakage-safe transformers, and robustness-oriented augmentation—transfers to malware/behavioral analytics, phishing-content triage, insider-risk monitoring, and timeline reconstruction across heterogeneous logs. Integration of pixel-level localization and NLP-derived metadata supports rapid scoping (MTTD/MTTR) by linking artifacts to assets, users, and phases of an attack, while the emphasis on calibration, subgroup performance, and resource footprint helps set deployment thresholds under operational constraints. The chain-of-custody practices and reproducibility logging facilitate cross-team handoffs between DFIR and blue-team functions and reduce evidentiary friction when technical findings must inform policy or enforcement actions.

This research contributes to forensic science theory by demonstrating how the integration of multiple AI techniques creates synergistic effects that individual methods cannot achieve on their own. Our findings indicate that CNNs can detect pixel-level anomalies that traditional statistical methods overlook. At the same time, NLP analysis of metadata reveals temporal inconsistencies that visual inspection often misses, and supervised learning provides the necessary classification framework for legal categorization. This multi-modal approach challenges the prevailing single-method paradigm in digital forensics literature. It establishes that comprehensive forensic analysis requires examining images across multiple domains simultaneously, such as visual, metadata, and behavioral patterns. Practically, our framework transforms digital evidence workflows by reducing processing time from 15-20 days to 4-6 hours for typical cases while maintaining chain-of-custody requirements through cryptographic hashing at each stage. The system automatically generates reports with confidence intervals that meet legal standards, addressing the critical gap between technical analysis and legal requirements.

6 Conclusion

The convergence of AI with digital forensics marks an essential milestone in the evolution of cyber investigation methodologies. Our study demonstrates the transformative potential of AI techniques, convolutional neural networks, SML algorithms, and NLP to improve the effectiveness and efficiency of digital forensic processes.

The results indicate that AI can significantly optimize digital-evidence identification, classification, and analysis at scale. CNNs have shown high precision in visualizing and interpreting complex patterns within multimedia data. This capability is essential given the rapid growth of digital content examined during criminal investigations. In parallel, SML algorithms have been instrumental in rapidly categorizing data and highlighting abnormal patterns, a task that traditional methods find challenging, given the magnitude of the data and the subtlety of modern digital signatures.

NLP techniques have proved effective in extracting and interpreting pertinent information from textual communications that are often central to forensic examinations.

Our findings suggest that AI increases analytical precision while reducing processing time, thereby accelerating investigations without compromising thoroughness. At the same time, ethical considerations remain critical: the deployment of AI must safeguard privacy and data integrity and avoid unintended bias, with results presented transparently for accountable decision-making.

Future work will focus on increasing interpretability and transparency (e.g., XAI for legal explainability), broadening external validation across devices, languages, and anti-forensic conditions, and strengthening operational robustness under deployment constraints.

Supplementary information.

Declarations

Funding

The authors declare that no funding was received to conduct this study.

Conflict of Interest/Competing Interests

The authors declare that they have no conflict of interest or competing interests.

Ethics Approval and Consent to Participate

Not applicable. This research did not involve human participants, their data, or biological material. Furthermore, this study does not constitute a clinical trial, as it does not include any biomedical intervention or experimentation on human subjects. Therefore, no clinical trial registration is required.

Consent for Publication

Not applicable. This manuscript does not contain any person's data.

Data Availability

The data used in this study are not publicly available due to their forensic relevance and the restrictions on their use. However, they can be made available upon reasonable request to the corresponding author for replication or academic purposes.

Materials Availability

All materials pertinent to this study are a part of the institutional resources.

Code Availability

An initial version of the software code developed in this research is not publicly available. However, it can be provided upon reasonable request to the corresponding author for academic or non-commercial replication purposes.

Author Contributions

Conceptualization, W.V.-C.; methodology, R.G. and A.M.N.; software, R.G.; validation, W.V.-C, R.G.; formal analysis, R.G.; investigation, A.M.N.; data curation, A.M.N.; writing original draft preparation, R.G.; writing review and editing, W.V.-C.; visualization, W.V.-C.; supervision, W.V.-C. All authors have read and agreed to the published version of the manuscript.

References

- [1] Brunty, J.: Validation of forensic tools and methods: A primer for the digital forensics examiner. *WIREs Forensic Science* **5** (2023) <https://doi.org/10.1002/wfs2.1474>
- [2] Rana, M.S., Hasan, M.M., Moneruzzaman, M.: An introduction to digital forensics (cyber forensics). *International Journal of Research Publication and Reviews* **4** (2023) <https://doi.org/10.55248/gengpi.4.1223.123427>
- [3] Overill, R., Collie, J.: Quantitative evaluation of the results of digital forensic investigations: a review of progress. *Forensic Sciences Research* **6**, 1–6 (2021) <https://doi.org/10.1080/20961790.2020.1837429>
- [4] Ding, K., Meng, F., Liu, Y., Xu, N., Chen, W.: Perceptual hashing based forensics scheme for the integrity authentication of high resolution remote sensing image. *Information* **9**(9) (2018) <https://doi.org/10.3390/info9090229>
- [5] Rani, D., Gill, N.S., Gulia, P.: A forensic framework to improve digital image evidence administration in iiot. *Journal of Industrial Information Integration* **38**, 100568 (2024) <https://doi.org/10.1016/j.jii.2024.100568>
- [6] Verdoliva, L.: Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing* **14**(5), 910–932 (2020) <https://doi.org/10.1109/JSTSP.2020.3002101>
- [7] Piva, A.: An overview on image forensics. *ISRN Signal Processing* **2013** (2013) <https://doi.org/10.1155/2013/496701>
- [8] Bharatkumar, R.: The use of artificial intelligence in digital forensics. *International Journal of Computer Applications* **184** (2023) <https://doi.org/10.5120/ijca2023922630>
- [9] Sarhan, S.A.E., Youness, H.A., Bahaa-Eldin, A.M.: A framework for digital forensics of encrypted real-time network traffic, instant messaging, and voip application case study. *Ain Shams Engineering Journal* **14** (2023) <https://doi.org/10.1016/j.asej.2022.102069>
- [10] Almutairi, W., Moulahi, T.: Joining federated learning to blockchain for digital

forensics in iot. *Computers* **12** (2023) <https://doi.org/10.3390/computers12080157>

- [11] Roopak, M., Khan, S., Parkinson, S., Armitage, R.: Comparison of deep learning classification models for facial image age estimation in digital forensic investigations. *Forensic Science International: Digital Investigation* **47** (2023) <https://doi.org/10.1016/j.fsidi.2023.301637>
- [12] Salih, K., Dabagh, N.: Digital forensic tools: A literature review. *Journal of Education and Science* **32** (2023) <https://doi.org/10.33899/edusj.2023.137420.1304>
- [13] Liu, X., Fu, X., Du, X., Luo, B., Guizani, M.: Machine learning-based non-intrusive digital forensic service for smart homes. *IEEE Transactions on Network and Service Management* **20** (2023) <https://doi.org/10.1109/TNSM.2022.3224863>
- [14] Tageldin, L., Venter, H.: Machine-Learning Forensics: State of the Art in the Use of Machine-Learning Techniques for Digital Forensic Investigations within Smart Environments (2023). <https://doi.org/10.3390/app131810169>
- [15] Riadi, I., Yudhana, A., Fanani, G.P.I.: Mobile forensic tools for digital crime investigation: Comparison and evaluation. *International Journal of Safety and Security Engineering* **13** (2023) <https://doi.org/10.18280/ijssse.130102>
- [16] Horsman, G.: Sources of error in digital forensics (2024). <https://doi.org/10.1016/j.fsidi.2024.301693>
- [17] Khan, A.A., Zhang, X., Hajjej, F., Yang, J., Ku, C.S., Por, L.Y.: Asmf: Ambient social media forensics chain of custody with an intelligent digital investigation process using federated learning. *Heliyon* **10**(1), 23254 (2024) <https://doi.org/10.1016/j.heliyon.2023.e23254>
- [18] Joseph, D.P., Viswanathan, P.: Sdot: Secure hash, semantic keyword extraction, and dynamic operator pattern-based three-tier forensic classification framework. *IEEE Access* **11**, 3291–3306 (2023) <https://doi.org/10.1109/ACCESS.2023.3234434>
- [19] Khyavi, M.H.: Isms role in the improvement of digital forensics related process in soc's. *International Journal of Forensic Engineering and Management* **1**(3), 263–276 (2023) <https://doi.org/10.1504/IJFEM.2023.136367>
- [20] Farid, H., Bravo-Solorio, S.: Photo forensics from jpeg dimples to deepfakes: A comprehensive survey. *Annual Review of Vision Science* **9**, 387–412 (2023) <https://doi.org/10.1146/annurev-vision-120522-110754>
- [21] Zampoglou, M., Kordopatis-Zilos, G., Papadopoulos, S., Kompatsiaris, Y.: Learning to detect misleading content on social media: A large-scale study on image forensics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(3),

1456–1471 (2024) <https://doi.org/10.1109/TPAMI.2023.3334521>

- [22] Wang, J., Sun, W., Li, H., Wu, J.: Cnn-based automatic detection of fake face images: An attention mechanism approach. *Information Sciences* **659**, 120076 (2024) <https://doi.org/10.1016/j.ins.2023.120076>
- [23] Bayar, B., Stamm, M.C.: Constrained convolutional neural networks: A new approach for general purpose image forensics. *IEEE Transactions on Information Forensics and Security* **19**, 789–804 (2024) <https://doi.org/10.1109/TIFS.2023.3298456>
- [24] Mehrjardi, F.Z., Latif, A.M., Zarchi, M.S., Sheikhpour, R.: A survey on deep learning-based image forgery detection: Advances, challenges and future directions. *Neurocomputing* **552**, 126587 (2023) <https://doi.org/10.1016/j.neucom.2023.126587>
- [25] Gadekar, B., Dharaskar, R.V., Thakare, V.M.: An event based digital forensic scheme for vehicular networks. *International Journal on Recent and Innovation Trends in Computing and Communication* **11** (2023) <https://doi.org/10.17762/ijritcc.v11i5.6691>
- [26] Fährndrich, J., Honekamp, W., Povalej, R., Rittelmeier, H., Berner, S., Labudde, D.: Digital forensics and strong AI: A structured literature review (2023). <https://doi.org/10.1016/j.fsidi.2023.301617>
- [27] Mahdian, B., Saic, S.: A bibliography on blind methods for identifying image forgery. *Signal Processing: Image Communication* **39**(1), 46–67 (2024) <https://doi.org/10.1016/j.image.2023.117045>
- [28] Morrison, C., Sanders, J.: The daubert standard and digital forensic evidence: Meeting the challenge of scientific reliability. *Journal of Digital Forensic Practice* **18**(3), 123–145 (2023) <https://doi.org/10.1080/15567281.2023.2198765>
- [29] Hakim, A.R., Ramli, K., Gunawan, T.S., Windarta, S.: A novel digital forensic framework for data breach investigation. *IEEE Access* **11** (2023) <https://doi.org/10.1109/ACCESS.2023.3270619>
- [30] Ogunseyi, T.B., Adedayo, O.M.: Cryptographic techniques for data privacy in digital forensics. *IEEE Access* **11** (2023) <https://doi.org/10.1109/ACCESS.2023.3343360>
- [31] Aguayo, R., Lizarraga, C., Quiñonez, Y.: Evaluation of academic performance in virtual environments using the nlp model. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao* **2021** (2021) <https://doi.org/10.17013/RISTI.41.34-49>
- [32] Borré, A., Seman, L.O., Camponogara, E., Stefenon, S.F., Mariani, V.C., Santos Coelho, L.: Machine fault detection using a hybrid cnn-lstm attention-based

- model. *Sensors* **23** (2023) <https://doi.org/10.3390/s23094512>
- [33] Canizo, M., Triguero, I., Conde, A., Onieva, E.: Multi-head cnn-rnn for multi-time series anomaly detection: An industrial case study. *Neurocomputing* **363** (2019) <https://doi.org/10.1016/j.neucom.2019.07.034>
 - [34] Bertino, E.: Botnets and internet. *Computer* **50**, 76–79 (2017) <https://doi.org/10.1109/MC.2017.62>
 - [35] Ortiz-Aguilar, L.D.M., Carpio, M., Soria-Alcaraz, J.A., Puga, H., Díaz, C., Lino, C., Tapia, V.: Training off-line hyperheuristics for course timetabling using k-folds cross validation. *la revista programación matemática y software* (2016)
 - [36] Aziz, O., Klenk, J., Schwickert, L., Chiari, L., Becker, C., Park, E.J., Mori, G., Robinovitch, S.N.: Validation of accuracy of svm-based fall detection system using real-world fall and non-fall datasets. *PLoS ONE* **12** (2017) <https://doi.org/10.1371/journal.pone.0180318>
 - [37] Mehrvand, M., Baghanam, A.H., Zahra, R., Nourani, V.: AI-based (ANN and SVM) statistical downscaling methods for precipitation estimation under climate change scenarios (2017)
 - [38] Kartik, N., Mahalakshmi, R., Venkatesh, K.A.: Xai-based student performance prediction: Peeling back the layers of lstm and random forest’s black boxes. *SN Computer Science* **4** (2023) <https://doi.org/10.1007/s42979-023-02070-y>
 - [39] Park, H., Kim, K., Lee, D.K.: Prediction of severe drought area based on random forest: Using satellite image and topography data. *Water* **11**, 705 (2019)
 - [40] Xu, X., Zhao, M., Shi, P., Ren, R., He, X., Wei, X., Yang, H.: Crack detection and comparison study based on faster r-cnn and mask r-cnn. *Sensors* **22** (2022) <https://doi.org/10.3390/s22031215>
 - [41] Widyawati, Sutanto: Perbandingan algoritma naïve bayes dan support vector machine (svm). *Jurnal Ilmiah Sains Dan Teknologi* **3** (2019)
 - [42] Onyema, E.M., Almuzaini, K.K., Onu, F.U., Verma, D., Gregory, U.S., Puttaramaiah, M., Afriyie, R.K.: Prospects and challenges of using machine learning for academic forecasting. *Computational Intelligence and Neuroscience* **2022** (2022) <https://doi.org/10.1155/2022/5624475>
 - [43] Khalaf, M.A., Steiti, A.: Artificial intelligence predictions in cyber security: Analysis and early detection of cyber attacks. *Babylonian Journal of Machine Learning*, 63–68 (2024) <https://doi.org/10.58496/BJML/2024/006>