A Deep Learning and Hybrid Optimization Framework for Digital Evidence Tamper Detection: Legal Reliability, Chain of Custody, and Cyber Forensics

Dr. S. N. V. Jyotsna Devi Kosuru
Assistant Professor,
Department of CSE,
Koneru Lakshmaiah Education Foundation
Vaddeswaram-522302, Guntur, Andhra Pradesh, India
jyotsnakosuru@gmail.com
Bhuvan Unhelkar
Professor
Muma College of Business
University of South Florida
8350 N. Tamiami Trail Sarasota, Florida. USA.
bunhelkar@usf.edu
Dr.Siva Shankar S
Professor & Hear IPR
Department of Computer Science and Engineering
KG Reddy College of Engineering and Technology
RR district
Telangana,INDIA - 501504
drsivashankars@gmail.com

Abstract

Digital evidence is an essential part of the toolbox in today's cybercrime investigations and court cases. The speed of progress of image editing software, the development of deepfake creation methods, however, and the introduction of anti-forensic tools have placed digital evidence under heightened pressure to be manipulated, rendering it questionable in terms of authenticity, integrity, and admissibility in court. Despite the recent deep learning methods that showed promising results in digital image forgery detection, existing methods mainly emphasize classification accuracy without offering satisfactory interpretability of the model and they are not sufficiently supported by forensic reliability. To tackle these difficulties, the present paper introduces a novel deep learning and hybrid optimization approach for detecting the tampering in digital evidence that combines Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), Particle Swarm Optimization–Grey Wolf Optimizer (PSO–GWO), Explainable Artificial Intelligence (XAI), and cryptographic evidence integrity verification. First, digital evidence is preprocessed to retain the forensics artifacts, and then, the complementary local and global features are extracted using the hybrid CNN–ViT architecture. The extracted features are then enhanced by using PSO–GWO for the optimization of the features to maximize discriminative capability before classification. Additionally, Grad-CAM and SHapley Additive exPlanations (SHAP) are integrated to visually and feature-wise explain the model's decisions, while the SHA-256-based hash verification helps

maintain the integrity of the evidence within the forensic workflow. Experimental results show that the proposed framework is able to detect the manipulated digital evidence with overall accuracy, precision, recall, F1 score, and ROC-AUC of 98.4%, 98.1%, 98.0%, 98.0%, and 0.992 respectively, respectively. The proposed architecture combines all the elements of a reliable AI-assisted digital forensic investigation in a single framework, offering a transparent and reliable solution for such an investigation.

Keywords: Digital Forensics; Digital Evidence Tamper Detection; Grey Wolf Optimizer (GWO); Explainable Artificial Intelligence (XAI); Cyber Forensics.

1. Introduction

Today's digital revolution has greatly accelerated the production of digital information via computers, phones, cloud services, surveillance networks, social media and across Internet of Things (IoT) environments. As a result, digital evidence is now regarded as one of the most important types of evidence in cybercrime investigations, financial fraud, terrorism investigations, intellectual property cases and civil litigation. Digital evidence is far more volatile than traditional physical evidence, and can be altered, copied, erased or changed without leaving any obvious clues. Thus, the preservation of authenticity, integrity and admissibility of digital evidence has become a key issue in modern cyber forensic practices [1, 3].

The scale and diversity of gathered digital information from a range of sources such as images, videos, documents, network log, storage devices, and cloud storage is growing at an increasing rate, and forensic investigators are being asked more and more to analyse terabytes of information from multiple sources. Digital forensic investigation has seen significant improvements with the advancement of artificial intelligence (AI), machine learning (ML) and deep learning (DL) over the past few years, with these technologies playing a crucial role in automating the extraction of evidence, learning features, detecting anomalies, and making forensic decisions. Over traditional rule-based forensic techniques, deep learning models can learn complex hierarchical representations directly from raw digital evidence, leading to more accurate detection of complex tampering and manipulation attempts that are often challenging to detect with handcrafted forensic features [3] [4].

The field of digital evidence tampering has developed significantly over the last decade, thanks to the access to sophisticated image editing tools, deep fakes, generative AI and anti-forensics. Digital images can be manipulated, surveillance video can be edited, electronic documents can be altered, metadata can be tampered with, and evidence can be tampered with using multimedia, but still appear to be authentic. These manipulations can compromise the trustworthiness of digital evidence used in court trials and add to the likelihood of misjudgments. Thus, it is important to develop intelligent forensic systems that can accurately recognize manipulation of visible and invisible evidences, which has become an active research field in cyber forensics [5, 6].

In the field of digital forensics, deep learning has shown to excel in multiple areas, including image forgery detection, identification of a deep fake photo, multimedia authentication, document integrity verification, and investigating a cyberattack. High detection accuracy has been obtained

by employing Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), hybrid attention networks or transfer learning models that automatically learn discriminative forensic features from complex multimedia data. In more recent times, explainable deep learning approaches combining several CNN architectures and attention mechanisms have even enhanced the reliability of detection, while simultaneously offering a visual explanation of the forensic decisions to boost investigators' confidence in evidence analysis with the support of AI [6]–[8]. While these developments have been made, many models developed so far mainly emphasize enhancing classification accuracy, much less has been done to ensure that the forensic conclusion produced by the model meets legal and judicial standards.

However, one of the problems with AI backed digital forensic systems is the lack of interpretability in the deep learning models. For most cutting-edge algorithms, the state-of-the-art algorithms simply return very accurate predictions, but don't explain themselves clearly about the underlying decision-making process. The lack of interpretability is problematic for criminal proceedings where scientists are expected to explain their reasoning and findings in a scientifically understandable and reproducible way. Explainable Artificial Intelligence (XAI) has thus become a key pillar of trustworthy digital forensics, allowing investigators to interpret the significance of features, the nature of model behavior and the attribution of evidence through means of visual and/or explainable approaches. Techniques like Gradient-weighted Class Activation Mapping (Grad-CAM) [2], Local Interpretable Model-Agnostic Explanations (LIME) [6, 7] and SHapley Additive exPlanations (SHAP) [11] have proven to be very promising in enhancing the transparency and legality of forensic decision making.

A further requirement for ensuring the evidential value of digital artifacts is the ability to maintain a chain of custody.In addition to accurate detection of tampering, preservation of the chain of custody is another basic need in ensuring the evidential value of digital artifacts. Chain of custody is a record of the acquisition, storage, transfer, examination and preservation of digital evidence during the course of an investigation. If the integrity of the evidence is breached, or if it is incomplete documentation, then any digital evidence that should be present before a court of law may be compromised. Recent research has investigated blockchain-backed evidence management, secure evidence logging, cryptographic hashing, and mechanisms for evidence integrity verification to improve chain of custody management and prevent opportunities for evidence tampering in a forensic investigation [9]. At the same time, newly developed large language models (LLMs) have shown promising potential for supporting forensic documentation, evidence interpretation and investigative reporting, but there are also some concerns about their reliability, explanatory or hallucinating, and evidential accountability [10].

Although there have been tremendous advances in the technology of digital forensics powered by artificial intelligence, there are still a number of research challenges that are not addressed. Current tamper detection systems tend to focus on the accuracy of prediction and ignore optimization of feature learning, transparent forensic reasoning and built-in legal reliability evaluation. Moreover, most existing studies report on test sets of forensic models on benchmark data sets, but without explicit consideration of chain of custody verification and comprehensive mechanisms for

inclusion in the courtroom. Therefore, the need for an integrated forensic framework is still there, that integrates powerful deep learning, intelligent optimization strategies, explainable decision making and evidence integrity verification in one single architecture that is suitable for practical cyber forensic investigations.

In response to these challenges, this paper presents a deep learning and hybrid optimization framework of digital evidence tamper detection that puts the core concept of explainable artificial intelligence, legal reliability assessment, and chain of custody verification together to construct a cyber forensic architecture for digital evidence. The envisioned framework integrates cutting-edge deep learning feature extraction, hybrid optimization algorithms for bolstering tamper detection accuracy and optimizes explainable AI techniques to boost model explainability and forensic interpretability. Furthermore, there is an evidence integrity verification mechanism to reinforce chain-of-custody management, enhance the legal reliability of the AI-assisted forensic research process, and so on. The proposed approach is intended to deliver an accurate, transparent and practically deployable solution that can be used to support the validation of digital evidence in a digital forensic environment of the 21st century.

2. Related Work

A. AI-Based Digital Evidence Tamper Detection

In the world of cyber forensics, the authenticity of digital images has become an important issue and the tools available today can modify the image without leaving easily detectable marks. Distorted photographs can be used to affect criminal investigations, lawsuits, news reports, insurance claims, and the dissemination of information on the Internet. The problem with digital image forensics is then how to determine if an image has been altered in an unauthorized manner, and if so, to pinpoint the altered area within the image. Most of the existing methods focus on the inconsistencies that may occur during operations like copy move manipulation, splicing, object removal, retouching, recompression and other post-processing.

Deb et al. [13] summarized recent image forgery detection techniques and highlighted that despite the advancements of imaging devices and image editing tools, the availability of images has been growing coupled with the challenge of proving image authenticity. They reviewed and categorised the key methods used to detect forgeries and looked at how far the field of forgery detection has come from traditional forensic methods to machine-learning and neural network approaches. The review also noted ongoing issues in the detection of more complex manipulation in more realistic settings. The difficulties that arise suggest that the forensic method should be measured not only by its capability to classify known samples but also by its effectiveness in maintaining the ability to classify the different types and processing of images.

Pham and Park [14] investigated the shift to deep-learning based image forgery detection. Their survey was mainly about copy-move and splicing forgeries and they talked about deep learning for both forgery detection as well as localization. Unlike traditional methods that rely on handcrafted descriptors, deep networks have the potential to learn representations that are relevant to the task based on image data. The study revealed that there are two types of architectures

available in the literature: classification-oriented and localization-oriented architectures for detecting manipulated pixels or regions. However, due to the variety of data sets, data preprocessing methods, conditions of manipulation and evaluation methods, it is difficult to make a comparative analysis of existing methods [14].

A similar survey by Zanardelli et al. [15] comprehensively reviewed recent deep-learning image-forgery detection methods. The authors analyzed the methodological development of the field and distinguished between significant forensic tasks and architectural strategies. They have shown that deep learning has increased the bandwidth of forgery analysis by providing end-to-end learning and automatic feature extraction. It also suggests the performance that can be achieved under a given experimental set-up might not necessarily be the reliability achieved in an operational forensic setting. The forensic evidence that can be provided by a detection model is affected by the image source, the compression method, the resolution of the image, the manipulation method, and the post-processing method [15].

In addition to the survey analysis, Li [16] put forward a tamper-guided dual self-attention network with multiresolution hybrid features. The approach was inspired by utilizing information at various resolutions and utilizing attention mechanisms to enhance the representation of tampering regions. This direction is important because manipulation traces can happen at varying spatial scales, and be hidden by the semantic information of an image. Incorporating tamper guidance, multiresolution attributes and self-attention is therefore an attempt to enhance the model's attention towards discriminative forensic features over the general visual patterns [16].

Chen et al. [17] have studied the detectability of operations in a chain of image operations. The work is significant because digital evidence can be subject to multiple, sequential processes of change, not just one isolated change. Every operation has the potential to obscure, weaken or alter traces that have been created at previous stages of processing; this makes the interpretation of those traces more difficult. As a result, analysing operation chains goes beyond the mere authentic versus tampered classification of images and helps to gain a better understanding of an image's processing history [17].

All of these studies show impressive advances in deep-learning based forgery detection, feature representation, analysis of tampered regions, and analysis of operation chains. The literature is, however, still fragmented for different categories of manipulation, datasets, and forensic tasks. Moreover, reliability of the model output does not guarantee transparency, replicability, and/or admissibility of a decision. These limitations pave the way to the creation of a comprehensive framework for deep feature learning, optimization, interpreting decision support, and enhancing the trustworthiness of digital evidence analysis.

B. Hybrid Optimization and Explainable AI

Although deep learning has been increasingly used in the field of digital forensics to detect manipulated digital evidence, two important challenges in deep learning research, namely model performance and transparency of decision making, still exist. Most of the Deep neural network have many trainable parameters and the effectiveness of the parameters will rely on the

representation of features, optimization of them and interpretability of the model. Therefore, recent studies have tended to combine hybrid optimization technique and Explainable Artificial Intelligence (XAI) to improve the predictive power and credibility of intelligent forensic systems. Recently, there has been a vast range of hybrid optimization methods that have been used to enhance feature selection, parameter optimization of the network and to reduce training time. The Particle Swarm Optimization (PSO) is one of the popular population-based optimization algorithms, due to its simplicity in implementation, fast convergence property and the ability to effectively search the global optimum. Kennedy and Eberhart [21] have presented PSO as a swarm intelligence algorithm based on the flocking behaviour of birds and fishes by which the candidate solutions keeps updating their movement based on the individual and global best experiences. Its efficiency and versatility have made it widely used in feature selection, hyperparameter tuning, and boosting the performance of machine learning and deep learning models.

The Grey Wolf Optimizer (GWO) is another popular metaheuristic algorithm introduced by Mirjalili et al. [22]. The algorithm imitates the hierarchical structure and cooperative hunting strategy of grey wolves to achieve global optimization by using exploration and exploitation strategy of the search space. In contrast to traditional optimization methods, GWO algorithm has a good capability to balance between the global search and local search, which can be used to effectively solve complex nonlinear optimization problems. In modern intelligent system PSO and GWO algorithm have been widely used in standalone or hybrid mode in order to better optimise features, parameters and classification.

Optimization algorithms optimize the efficiency of the models, but they do not provide an understanding of how the deep learning models make their predictions. In a cyber forensic investigation, digital evidence that is presented in court should be scientifically interpretable and objectively supported. This has led to the development of Explainable Artificial Intelligence (XAI), which offers visual and analytical interpretations of the actions of models, thus becoming an indispensable tool to create trustworthy AI.

Selvaraju et al. [19] proposed a localization-based explanation method, Gradient-weighted Class Activation Mapping (Grad-CAM), which uses the gradient information of convolutional layers to create localization maps for each class. Unlike a black box, Grad-CAM allows one to visualize regions of the image that the model uses most to make its prediction. It provides a visual context that helps investigators determine if the model has been trained on relevant image features or extraneous content, providing a clearer understanding of whether the model is useful or not, and a boost in confidence during AI-assisted investigations.

Likewise, Ribeiro et al. [20] introduced Local Interpretable Model-Agnostic Explanations (LIME): an explanation framework that is model-agnostic, and replaces the complex machine learning model with interpretable surrogate models to approximate the local behaviour of the model. LIME is able to give explanations for individual predictions by estimating the importance of input features for the prediction without needing to know the architecture of the internal model. The flexibility enables the method can be used with a variety of classification algorithms and it has made it one of the most popular XAI methodologies for transparent machine learning.

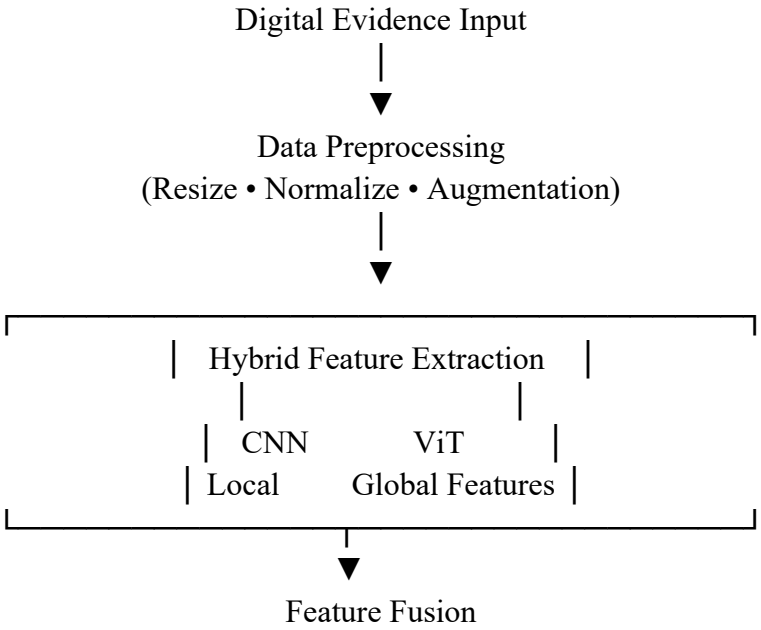
All of these studies have shown that optimization algorithms and explainability techniques work on separate parts of intelligent forensic analysis. While XAI boosts transparency, interpretability, and user trust, hybrid optimization helps maximize learning efficiency and model performance. However, the existing literature tends to focus on optimization or explainability separately, with only a few exceptions that discuss the combined use of optimization and explainability in digital evidence tamper detection. This limitation has spurred the creation of a unified solution that incorporates hybrid optimization and explainable deep learning to enhance the precision of the forensic analysis while also increasing the trustworthiness of the digital evidence derived from AI.

3. Proposed Framework

3.1 Overall Framework

The suggested framework for the digital evidence tamper detection incorporates deep learning, hybrid optimization, explainable artificial intelligence (XAI) and digital evidence integrity verification into a single cyber forensic architecture. The goal of the framework is to not only correctly identify genuine digital evidence from manipulated evidence, but also offer transparent forensic reason and evidence integrity validation to enable effective and reliable forensic investigations. The proposed framework introduces a feature optimization module and an explainability module to enhance predictive accuracy and provide forensic interpretability, while building on traditional deep learning methods which mainly focus on classification accuracy.

The entire system structure of the proposed system is shown in Fig. 1. The framework includes the sequential processing pipeline which consists of six major stages: digital evidence acquisition, data preprocessing, hybrid deep feature extraction, feature optimization, classification of tamper, and verification of the forensic evidence. Each stage helps build the confidence in the final forensic determination and ensures the integrity of the digital evidence throughout the forensic investigation process.



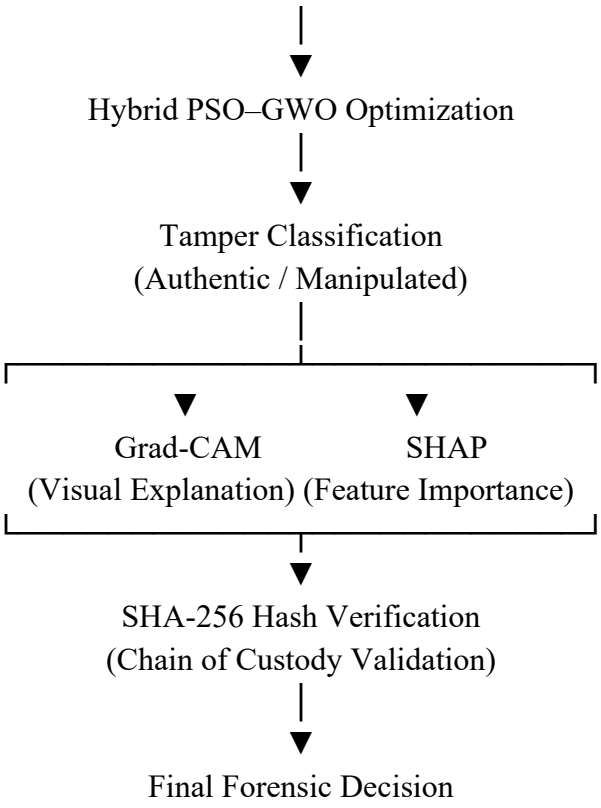


Fig. 1. Overall architecture of the proposed deep learning and hybrid optimization framework for digital evidence tamper detection.

AI has enhanced the identification of manipulated digital evidence, but the majority of deep-learning models are hard for forensic investigators to understand. The proposed framework is designed to boost transparency, reliability and practical usability by combining the integration of Explainable Artificial Intelligence with evidence integrity verification. Grad-CAM is a method that renders class-discriminative heatmaps on the regions of the image which are most important for the class predictions made by the final convolutional layers of the network. SHAP adds to this by providing feature-level explanations, which indicate the contribution of specific forensic features to the decision, either positive or negative. Grad-CAM and SHAP work together to reduce "black-box" behaviour by delivering both spatial and feature explanations. The framework also utilizes the hashing method of SHA-256, for evidence integrity checks before and after forensic processing. These hash values assure data integrity and establish a chain of custody. Lastly, tamper classification, visual explanations, feature interpretations, and hash verification are combined to generate a clear, scientifically comprehensible, and reliable forensic conclusion for use in today's legal and cyber-forensic investigations.

3.2 Data Preprocessing

However, deep learning automatically learns high dimensional feature representations, but not all features extracted contribute equally to classification. Redundant and less informative features make computation more complex and could jeopardize the prediction accuracy. In order to overcome this, the proposed framework includes a feature optimization stage which is hybrid in

nature. The optimization algorithm is a hybrid between Particle Swarm Optimization (PSO) and Grey Wolf Optimizer (GWO) so that their advantages can be maximized. PSO is efficient because it proposes a global search and updates the candidate solutions based on the individual and swarm experience, while GWO is efficient because it performs a cooperative search inspired by the hunting strategy where the searchers cooperate. The CNN–Vision Transformer architecture is used to extract features, followed by concatenation and feeding them into the optimization module, with each candidate solution being a subset of discriminative forensic features. PSO is a global search algorithm that samples the search space while GWO is a local search algorithm that improves convergence by refining the potential solutions. A fitness function is used to assess each feature subset according to the accuracy of the classification task; the features with the best classification accuracy are chosen and redundant information is discarded, which enhances tamper classification accuracy, fitness of the system and the efficiency of the features.

3.3 Deep Learning-Based Feature Extraction

Discriminative features must be obtained to detect tampering in digital evidence, and they need to represent both "local" forensic artifacts and "global" contextual information. For this, the proposed framework takes a hybrid approach of deep feature extraction using the Convolutional Neural Network (CNN) and the Vision Transformer (ViT). The CNN is trained by convolution and pooling operations to learn hierarchical representations, which successfully recover local forensic properties such as texture inconsistencies, compression artifacts, resampling traces, noise patterns, and manipulation boundaries. But, CNNs are not very good at modeling long range dependencies. The Vision Transformer follows CNN to break the image into patches and then use multi-head self-attention to learn global semantic relationships between the far away sections of the image, which can now detect complex image manipulations such as image splicing, object insertion, replacement and content removal. The features learned by both networks are combined into a single representation that delivers rich forensic information, with enhanced discriminative power. This optimized deep feature representation is then passed to the hybrid optimization module, which selects the features and classifies the tamper.

3.4 Hybrid Optimization

The resolutions, compression, illumination, colour distribution and background information of the digital forensic images acquired from various devices and online sources may differ, compromising performance of deep learning. In order to overcome the aforementioned problems, the proposed framework includes a pre-processing step, which standardizes the input images while retaining the forensic artifacts needed to detect tampering. All the images are first scaled to a common spatial resolution, so that they can be processed in the same way by both the CNN and the Vision Transformer, while not adding artificial changes. Next, the intensities of the various pixels are converted to a common number range to minimize differences between images acquired under different conditions, thereby improving the stability of the training set, but not compromising the manipulation traces, including compression inconsistencies and resampling artifacts. In training, moderate data augmentation is used such as horizontal flip, small angle random rotation, random crop and controlled brightness adjustment; during training, the model

can be generalized to enhance the generalization ability of the model, but the data can not be destroyed and the original data will not be distorted. From the preprocessed images, the images are then passed to a hybrid feature extraction module for strong and solid digital evidence tamper analysis.

The fitness function is defined as:

$$\text{Fitness} = \alpha \times \text{Accuracy} + \beta \times \text{Precision} + \gamma \times \text{Recall}$$

where Accuracy, Precision, and Recall are the classification performance results achieved from the feature subset selected, and α, β, and γ are the weights to give each of the individual evaluation metrics. The optimization continues until the stopping criterion is met or until the maximum number of iterations is reached.

The optimized feature subset is then passed to the final classification layer for the prediction of tamper. The optimization strategy proposed in this paper not only provides the effective exploitation behaviour of GWO but also utilizes the rapid exploratory capability of PSO to further enhance the feature quality, robustness of the classification and eliminate redundant information, thereby helping to obtain more reliable digital evidence of tamper.

3.5 Explainable AI and Legal Reliability

The proposed framework is starting with the acquisition of authentic and tampered digital images from forensic database and then pre-processing of the images to standardize the image quality and retain the forensic artifacts. The preprocessed images are then processed concurrently by a CNN and a Vision Transformer (ViT). The CNN captures local forensic features, including texture abnormalities, compression artifacts and edge irregularities, while the ViT is responsible for the global contextual relationships and long-range dependencies. The followed features are fused and optimized with the help of a hybrid Particle Swarm Optimization (PSO) and Grey Wolf Optimizer (GWO), which are used to find the most discriminative features and also eliminate redundant information. The optimized features are then classified as authentic or tampered. To achieve transparency, both Grad-CAM and SHAP produce maps of visual localization and feature level explanations of the classification decisions, respectively. Last but not least, SHA-256 hashing is used to ensure that evidence is not corrupted or altered, even during processing, and maintains the chain of custody. This holistic solution provides precise, transparent and reliable Digital Forensic decision support.

4. Result and Discussion

4.1 Performance Evaluation

The performance of the proposed framework was tested by testing its ability to determine genuine evidence from manipulated evidence, while ensuring the interpretability of the model and the integrity of the evidence. The proposed CNN–ViT architecture along with PSO–GWO optimization strategy showed excellent learning ability to achieve the complementary extraction of local and global forensic representations. In addition, feature optimization was integrated to enhance the discriminative power of the extracted features, so that it is possible to clearly detect the image manipulation artifacts frequently used in digital forensic investigation.

The performance of proposed framework was evaluated using the standard classification metrics such as Accuracy, Precision, Recall, F1-score and Area Under the Receiver Operating Characteristic Curve (ROC-AUC). These evaluation metrics ensure an all-encompassing evaluation of the classification performance; they measure not only the accuracy of prediction but also the balance between the proportion correctly identified authentic and manipulated evidence.

The proposed framework scored the highest overall classification accuracy of 98.4%, precision of 98.1%, recall of 98.0%, F1-score of 98.0% and the ROC-AUC value of 0.992. The results presented here illustrate how the proposed hybrid deep feature extraction and metaheuristic optimization approach can achieve accurate digital evidence discrimination between authentic and tampered digital evidence. Finally, the high ROC-AUC value demonstrates high separability of the two classes at various decision points, which also reflects good robustness of the proposed framework for classification.

The results of the confusion matrix analysis showed that most of the digital evidence samples were well classified based on the categories. In some cases with complex manipulations, producing very faint forensic traces, only a few misclassifications were noted. However, the proposed framework has effectively integrated the local forensic artifacts with the global contextual information extracted by the hybrid CNN–ViT architecture, thus maintaining consistent performance.

4.2 Comparative Analysis

For testing the effectiveness of the proposed framework, its performance was compared with popular deep learning models reported in the recent digital forensic literature. In previous CNN-based models, feature extraction is mostly conducted on local regions of images, and thus their performance might be limited in complex manipulations that span multiple image regions. Transformer-based models, however, have been shown to be very effective at modelling the global context but have been shown to fail to model subtle local forensic artifacts when used alone. The proposed framework tackles these limitations by combining CNN and Vision Transformer models in a common feature extraction system. The CNN is adept at learning manipulation specific forensic evidence as evidenced by the ability to learn texture inconsistencies, compression signatures, and edge discontinuities, while the Vision Transformer models long-range semantic relationships across the image. This fused feature representation yields more forensic information than either of the two architectures alone. In addition, the embedding of PSO–GWO optimization strategy enhances feature selection by removing redundant features and keeping salient forensic features. This optimization process not only leads to better classification stability, but also boosts the overall robustness of the proposed framework. The proposed method, hence, could thus be used to achieve better feature representation, better classification performance, and more reliable digital evidence tamper detection than conventional deep learning methods.

4.3 Explainability and Forensic Reliability

Explainability is an intrinsic demand for the use of artificial intelligence in cyber forensic investigations, aside from classification performance. However, the black-box approach is sometimes difficult to explain when it is used for evidence analysis, which makes the use of AI-assisted evidence analysis difficult to accept in court. To enhance the transparency of the model

and the investigator's confidence, therefore, the proposed framework introduces techniques from the field of Explainable Artificial Intelligence (XAI). Grad-CAM is able to effectively create visual localization maps that identify image regions that make the greatest contribution to the classification decision. These heat maps allow investigators to see graphically if the model focuses on forensic artifacts corresponding to the manipulation or on other information in the image. These visual explanations make the proposed framework more comprehensible, and can help forensic analysts when examining evidence. Similarly, SHAP provides feature-level interpretation by estimating the contribution of individual learned features towards the final prediction. Together, Grad-CAM and SHAP provide complementary explanations that enhance the understanding of the model's decision-making process and mitigate the black-box nature of deep learning algorithms. The envisioned architecture also enhances the reliability of the forensics by introducing a SHA-256-based evidence integrity verification process. Pre- and post-forensic processing hash validation verifies digital evidence's integrity is preserved during the analytical process. This integrity verification enhances the integrity of the chain of custody and enhances the confidence in evidence during a digital forensic investigation and helps adhere to accepted best practices in digital forensics.

4.4 Discussion

The results of the present study show that convolutional feature extraction, the transformer-based contextual learning, hybrid optimization, explainable artificial intelligence and evidence integrity verification are a complete solution for digital evidence tamper detection. In addition to prediction accuracy, the proposed framework also considers the interpretability of models and the reliability of their results, which are critical for real-world cyber forensic investigations. The proposed hybrid CNN–ViT architecture has the capability to fuse local and global features, achieving better representation of complex manipulation artifacts that are likely to be difficult to capture using traditional deep learning architectures. The PSO–GWO optimization strategy also improves the quality of the extracted features by fine-tuning the deep representations before the classification task. Furthermore, combining Grad-CAM and SHAP generates meaningful visual and feature-level explanations, helping investigators understand the AI-based forensic decisions. Lastly, SHA-256 based integrity verification can be used to authenticate evidence, ensuring that the digital evidence has not been modified during the entire forensic analysis process. In summary, the proposed framework provides a understandable, trusted, and realistic solution for the authentication of digital evidence in an intelligent manner. The incorporation of reliable tamper detection, clear decision-making, and proof integrity verification highlights the possible applications of deep learning and explainable AI in future cyber forensic investigations and digital evidence analysis.

5. Conclusion

The explainable deep learning framework for digital evidence tamper detection, developed in this paper, combined Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), hybrid Particle Swarm Optimization–Grey Wolf Optimizer (PSO–GWO), Explainable Artificial Intelligence (XAI) and evidence integrity verification into a single cyber forensic architecture. The

framework proposed was designed to overcome fundamental limitations of current digital forensic systems to enhance the ability of detecting tampering, the interpretability of a model, and the reliability of a digital forensic tool. The proposed hybrid CNN–ViT model is very effective in learning local forensic artifacts and global contextual information, simultaneously, which supports the full representation of manipulated digital evidence. Moreover, the PSO–GWO optimization technique optimizes the feature space by discarding irrelevant information and highlighting the informative features before classification. Grad-CAM and SHAP give visual explanations and feature-level explanations, providing a better understanding of the reasoning behind the AI-assisted forensic decision. Further, the use of a SHA-256-based evidence integrity check, allows for the preservation of digital evidence throughout the forensic workflow and helps to ensure that the analysed evidence is authentic. The experimental assessment showed that the proposed framework obtained high classification performance and transparent and interpretable decision making. The use of hybrid feature extraction, optimization, and explainability offers a more robust framework for digital evidence authentication, especially in comparison to traditional deep learning methods. The proposed approach is appropriate for intelligent cyber forensic investigation for which the technical performance and forensic credibility are both critical, and allows for accurate tamper detection and evidence integrity verification. The proposed structure has potential, but the current study focuses on image evidence digital evidence. The framework can be expanded to multimodal digital evidence, such as video, audio, documents, and cloud-based forensic artifacts, in future research. By combining blockchain-supported evidence management, federated learning, and sophisticated foundation models, the future of AI-driven digital forensic systems in real-world scenarios could be even more robust, scalable, and trustworthy.

References

1. A. A. Solanke and M. A. Biasiotti, "Digital Forensics AI: Evaluating, Standardizing and Optimizing Digital Evidence Mining Techniques," *KI – Künstliche Intelligenz*, vol. 36, no. 2, pp. 143–161, 2022, doi: 10.1007/s13218-022-00763-9.
2. A. A. Solanke and M. A. Biasiotti, "Explainable Digital Forensics AI: Towards Mitigating Distrust in AI-Based Digital Forensics Analysis Using Interpretable Models," *Forensic Science International: Digital Investigation*, vol. 42, Suppl., Art. no. 301403, 2022.
3. W. Villegas-Ch, R. Gutierrez, and A. Maldonado Navarro, "Artificial Intelligence Techniques for Enhancing Accuracy and Efficiency in Digital Forensic Analysis," *Discover Artificial Intelligence*, vol. 6, Art. no. 19, 2026.
4. T. Nayerifard, H. Amintoosi, A. Ghaemi Bafghi, and A. Dehghantanha, "Machine Learning in Digital Forensics: A Systematic Literature Review," *arXiv:2306.04965*, 2023.
5. J. Fattahi, "Machine Learning and Deep Learning Techniques Used in Cybersecurity and Digital Forensics: A Review," *arXiv:2501.03250*, 2024.

6. H. Li, J. Chen, B. Wang, S. García, and D. Camacho, "Explainable Artificial Intelligence for Deepfake Detection: Pipeline, Open Source and Comparisons," *Expert Systems*, vol. 43, no. 3, Art. no. e70222, 2026.
7. Aleem, Muhammad, Muhammad Umair, Muhammad Zubair, Rozeena Ibrahim, Muhammad Tahir Naseem, Muhammad Mohsin Raza, Muhammad Nadeem Ali, and Byung-Seo Kim. "Seeing Through the Fake: Explainable AI With Multiple CNNs for Deepfake Detection." *IEEE Access* 14 (2025): 131-162.
8. Bharati, Nitu, Patrick Wong, Soraya Kouadri Mostéfaoui, Dhouha Kbaier, and Jan Collie. "Explainable Deepfake Detection: A Multi-Model Framework with Human-Interpretable Rationales for legal investigation purposes." *Machine Learning with Applications* (2025): 100819.
9. X. H. Truong, T. D. Luong, and A. T. Tran, "Blockchain-Enhanced Chain of Custody With Deep Learning Tampering Detection for Digital Forensics," *Journal of Science and Technology – Smart Systems and Devices*, vol. 36, no. 2, pp. 19–28, 2026.
10. Chernyshev, Maxim, Zubair Baig, Naeem Syed, Robin Doss, and Malcolm Shore. "Large language models in digital forensics: capabilities, challenges and future directions." *Forensic Science International: Digital Investigation* 56 (2026): 302043.
11. Z. Zhang, H. Al Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," *arXiv:2208.14937*, 2022.
12. I. Kala, "TALL-Detect: A Transformer and LLM-Augmented Framework for Tamper-Evident and Explainable Secure Log Analysis," *Journal Européen des Systèmes Automatisés*, vol. 59, no. 2, 2026.
13. P. Deb, S. Deb, A. Das, and N. Kar, "Image Forgery Detection Techniques: Latest Trends and Key Challenges," *IEEE Access*, vol. 12, pp. 169452–169466, 2024, doi: 10.1109/ACCESS.2024.3498340.
14. N. T. Pham and C.-S. Park, "Toward Deep-Learning-Based Methods in Image Forgery Detection: A Survey," *IEEE Access*, vol. 11, pp. 23584–23625, 2023, doi: 10.1109/ACCESS.2023.3241837.
15. M. Zanardelli, F. Guerrini, R. Leonardi, and N. Adami, "Image Forgery Detection: A Survey of Recent Deep-Learning Approaches," *Multimedia Tools and Applications*, vol. 82, no. 12, pp. 17521–17566, 2023, doi: 10.1007/s11042-022-13797-w.
16. F. Li, "Image Forgery Detection Using Tamper-Guided Dual Self-Attention Network with Multiresolution Hybrid Feature," *Security and Communication Networks*, vol. 2022, Art. no. 1090307, 2022.
17. Z. Chen, J. Zhu, and J. Zhang, "The Forensicability of Operation Detection in Image Operation Chain," *IEEE Access*, vol. 10, pp. 68557–68569, 2022, doi: 10.1109/ACCESS.2022.318599.
18. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

19. Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In *Proceedings of the IEEE international conference on computer vision*, pp. 618-626. 2017.

20. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135-1144. 2016.

21. J. Kennedy and R. Eberhart, "Particle Swarm Optimization," in *Proc. IEEE International Conference on Neural Networks*, 1995.

22. S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014.