IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

AI-Driven Original and Tampered Image Detection and Localization Using CNN, Vision Transformers, and LLM-Based Analysis

Dr. K. Ramadevi, J. Jagathish, S. Hari Prasanth, D. Dinesh Kumar
Department of Information Technology
Panimalar Engineering College, Chennai, India
ramadevi_it@panimalar.ac.in, jagathishsmrithi@gmail.com
shariprasanth202@gmail.com, dinesh9786427871@gmail.com

Abstract: *In the current digital age, image manipulation has become rampant with the use of advanced image editing software, thus posing a serious threat in the areas of journalism, forensics, cybersecurity, and social media. This research project presents hybrid framework for a deep learning model that can identify the manipulated areas and differentiate between manipulated and original images. Convolutional Neural Networks (CNNs) are used to extract low-level and spatial features that can distinguish tampering artifacts from original image content, while Vision Transformers (ViTs) are used to introduce global attention mechanisms to identify minute inconsistencies in the spatial context of the image. Large language models (LLMs) are also incorporated to provide text interpretations of image differences, thus facilitating the human-friendly description of results and the semantic analysis of manipulated image areas. The system not only identifies images as original or manipulated but also marks the manipulated areas on the image, thus creating a visual map of the manipulated areas. The approach utilizes Python and deep learning models for efficient training and deployment within a web application setup*

Keywords: Convolutional Neural Networks (CNN), Vision Transformers (ViT), Image Tampering Detection, Copy-Move and Splicing Forgery, Digital Image Forensics, and Large Language Models (LLM).

I. INTRODUCTION

The recent rapid progress in digital image editing software and AI-powered content generation solutions has contributed to the widespread use of image manipulation. Digital images have become essential evidence in journalism, legal inves-tigations, cybersecurity, and national security-related appli-cations. However, advanced manipulation methods such as copy-move forgery, image splicing, inpainting, and AI-driven manipulations can modify semantic information while main-taining perceptual plausibility, making it increasingly difficult to detect using manual analysis. The rising need for effective image forensic solutions to detect both visible and invisible manipulations call for the design of automated, robust, and interpretable image forensic systems.

Conventional image forensic analysis mainly focused on the use of hand-crafted features from compression artifacts, sensor noise characteristics, lighting anomalies, and statistical anomalies. Although these solutions showed promise under controlled settings, their generalization performance is still suboptimal when faced with complex image manipulation, high-resolution images, or post-processing manipulations. The recent advent of deep learning has caused a paradigm shift in digital image forensics by allowing the learning of hierarchical features from data itself.

Convolutional Neural Networks (CNNs) have proven to be highly effective at recognizing local texture irregularities, edge inconsistencies, and image blending artifacts introduced during the tampering process. However, CNN-based

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

models are naturally bound by localized receptive fields, thus limiting their capacity to record global contextual inconsistencies and long-range spatial dependencies typically observed in complex forgeries. To this end, Vision Transformers (ViTs) have recently been incorporated into forensic analysis frameworks because of their self-attention mechanisms capable of capturing global structural dependencies between image patches. Although they possess robust contextual modeling capabilities, transformer- based models generally require extensive amounts of training data and computational resources, and may not perform well when modeling local irregularities independently.

Recent studies have attempted to combine CNN and Transformer architectures to capitalize on the complementary benefits of local and global feature representations. However, current approaches are largely centered on improving detection and localization accuracy while neglecting the need for structured semantic interpretability. In practical forensic analysis, it is important to note that high detection accuracy is not sufficient; rather, decision-makers demand transparent explanations to justify model outputs in a human-comprehensible fashion. Current approaches are largely limited to visual heatmaps such as Grad-CAM or attention maps, which require expert interpretation and lack semantic reasoning capabilities.

This study suggests a hybrid framework to close this gap by combining CNN-based local feature extraction, Vision Transformer-based global contextual representation, and Large Language Model (LLM)-based semantic interpretation into a single end-to-end forensic system. The proposed system not only achieves binary classification between authentic and tampered images but also provides pixel-level localization maps to indicate the manipulated areas. Moreover, the integration of the LLM-based explanation module enables the transformation of quantitative detection results and spatial attention statistics into structured forensic reasoning, thus improving transparency, trustworthiness, and usability in decision-critical settings.

The following is a summary of this paper's main contributions:

1) Hybrid CNN-ViT Architecture: A unified deep learning architecture that integrates local spatial feature extraction and global contextual reasoning for improved tampering detection robustness.

2) Transformer-Guided Localization: An improved spatial mapping approach that combines CNN activation maps and transformer attention weights for precise pixel- level localization of the tampered regions.

3) LLM-Based Semantic Interpretability: A structured post-processing explanation module that translates detection confidence scores and localization results into human-readable forensic reasoning.

4) Comprehensive Empirical Evaluation: Comprehensive benchmarking against existing CNN and transformer-based architectures, including ResNet, DenseNet, Xception, and Vision Transformers, to demonstrate superior classification and localization performance.

II. LITERATURE REVIEW

Deep learning-based digital image forgery detection has improved remarkably. Traditional forensic techniques were based on handcrafted features like JPEG anomalies, noise variance, and illumination inconsistencies. But these methods were not robust enough to handle complex forgeries like copy-move, splicing, inpainting, and AI-based forgeries.Most recently, the focus of research has been on Convolutional Neural Network (CNN)-based models because of their superior ability to extract hierarchical spatial features. Zhang et al. [1] suggested a generative adversarial network that combines CNN and Transformer to identify copy-move source and destination areas. Their model was able to improve localization precision by combining contextual attention mechanisms. Khalil et al. [2] also designed AR-Net, which incorporated adaptive attention and residual refinement modules to improve feature discrimination ability for subtle forgery cues.

Multiscale fusion techniques have been studied to improve robustness against geometric alterations and post-processing attacks. The CMCf-Net, an end-to-end context multiscale cross-fusion network, was proposed by Xiong et al. [3] and demonstrated outstanding generalization ability on test datasets. It has also been demonstrated that transfer learning holds promise.Pre-trained CNN backbones were used by Khalil et al. [4] to improve detection accuracy while lowering computing cost.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, June 2026 Impact Factor: 8.2

Despite their shown effectiveness, CNNs' inherent localization limits their ability to model long-range dependencies. Transformer-based models have been incorporated into forensic analysis frameworks in order to get around this limitation. Transformer auxiliary networks were shown by Shi et al. [5] for operator-aware manipulation localization, demonstrating enhanced contextual comprehension. In [6], long-range spatial dependencies significantly increased the accuracy of copy-move forgery detection in a Vision Transformer (ViT) with attention methods.

Hybrid CNN-Transformer models are a novel paradigm that has been introduced recently. In order to highlight the importance of interpretability in forensic applications, the study in [7] combined explainable AI techniques with CNN and Vision Transformer models. Furthermore, by differentiating between spatial and frequency domain data, dual-path networks in [8] improved localization accuracy.

ManTra-Net in [9] is still a basic anomaly-driven CNN model for manipulation tracing and serves as a benchmark baseline. Moreover, ResTran in [10] represented long-distance dependencies through residual transformer modules, enhancing the integrity of context analysis.

TABLE I: Summary of Related Work

Reference	Architecture	Detection Type	Localizatin	Explainability	Limations
[1]	CNN +Transformer GAN	Copy-Move	Yes	No	Limited generalization
[2]	CNN +Attention t (AR-Net)	Copy-Move	Yes	No	Focused On specific Forgery type
[3]	Multiscale CNN (CMCf-Net)	Copy-Move	Yes	No	High compu-tational cost
[4]	Transfer Learning CNN	General Forgery	Partial	No	Dataset dependency
[5]	Transformer-Auxiliar y Network	Manipulation	Yes	No	Complex architecture
[6]	Vision Transformer	Copy-Move	Yes	No	Requires large training data
[7]	CNN + ViT Hybrid	AI-Generate d Images	Yes	Partial	Limited forensic explanation
[8]	Dual-Path Network	Document Tamperin g	Yes	No	Domain-specific
[9]	ManTra-Net (CNN)	General Manipu-lation	Yes	No	Weak semantic reasoning
[10]	Residual Transform er (ResTran)	Forgery Detection	Partial	No	No structured reporting

III. METHODOLOGY

This section describes the proposed hybrid deep learning architecture for original and tampered image detection and localization. The proposed architecture combines Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and a Large Language Model (LLM) for effective classification, localization, and forensic explanation.

A. Proposed System Architecture

First, the dataset of authentic and tampered images is prepared and preprocessed. The preprocessed images are then fed into a CNN backbone (ResNet or EfficientNet) for extracting local texture-based forensic features. For improved global contextual reasoning, a Vision Transformer is applied to process image patches with self-attention.

The extracted CNN and ViT features are then fused using a weighted approach to obtain:

• Binary classification output (original vs. tampered)

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

• Tampering probability heatmaps

Lastly, a Large Language Model (LLM) is applied to interpret confidence scores and localization maps to provide human-readable forensic explanations. The entire pipeline is implemented using a Python-based web application developed using Flask or Django for real-time user interaction.

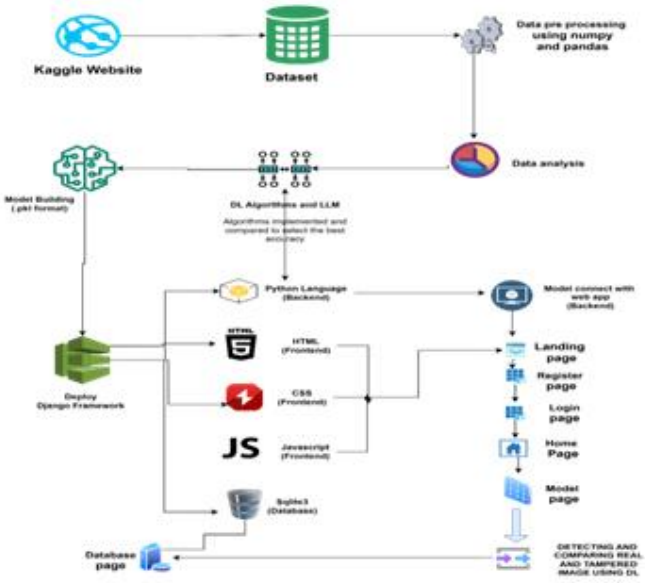

Fig.1: Architecture Diagram

B. Algorithm Selection

Algorithm selection is made based on the complementary learning abilities necessary for digital image forensics.

1) CNN Backbone Selection: The choice of CNN backbones was made based on the following considerations for ResNet and EfficientNet:

• Strong ability to learn local spatial patterns

• Forensic performance

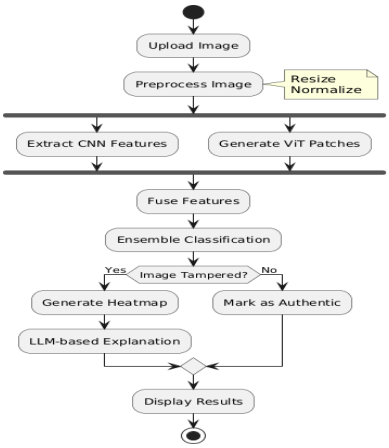

Fig. 2: Activity Diagram

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, June 2026 Impact Factor: 8.2

• Effective gradient backpropagation through residual connections
• Pre-trained ImageNet weights for transfer learning CNNs are very useful in identifying:
• Compression artifacts
• Texture inconsistencies
• Edge discontinuities
• Copy-move boundaries Mathematically, for an input image x:

$$Fcnn = \phi cnn(x)$$

where Fcnn represents hierarchical multi-scale feature maps extracted via convolutional operations.
Vision Transformer Selection:
While CNNs excel at modeling local features, they are limited in capturing long- range dependencies due to fixed receptive fields. Therefore, a Vision Transformer (ViT) is incorporated to model global relationships.

$$x \rightarrow \{p1, p2, \ldots, pn\}$$

$$Attention(Q, K, V) = Softmax\ (QKT)\ V \surd dk$$

The resulting global representation is:
Fvit = ϕvit(x) ViTs enhance detection of:
• Contextual misalignment
• Inpainting artifacts
• Subtle region blending
• Semantic inconsistencies

2) Hybrid Fusion Strategy: To exploit complementary strengths of CNN and ViT representations, a weighted feature fusion strategy is adopted:
where α is a learnable parameter optimized during training. $M(x) = \{$ 0 otherwise
These fusion balances:
• Local forensic cues (CNN)
• Global semantic anomalies (ViT)
The fused feature vector is passed to fully connected layers for binary classification.

3) LLM Selection for Interpretability: A Large Language Model (LLM) is incorporated to improve interpretability. The LLM is chosen because:
• Heatmaps are not semantically interpretable on their own
• Forensic analysis needs to be explained in text form
• Enhances trust and usability of the system The LLM takes structured inputs such as:
• Classification confidence score
• Localization mask coordinates
• Attention heatmap summaries
The LLM produces human-understandable forensic text explanations of probable manipulated areas and discovered inconsistencies. Notably, the LLM is designed as a post-processing component and does not affect classification acuracy.

C. Dataset Preparation and Training Protocol
Let the dataset be defined as:

$$D = \{(x_i, y_i)\}_{i=1}^{N}$$

where:
• xi = input image

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, June 2026 Impact Factor: 8.2

• yi ∈ {0, 1} (0 = authentic, 1 = tampered) The dataset is divided into:
• 80% training data
• 20% testing data
Preprocessing Steps
• Resizing to 224 × 224
• Pixel normalization
• Data augmentation (horizontal flip, rotation, brightness scaling)
Loss Function
Binary Cross-Entropy loss is used for classification:

$$L_{cls} = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right]$$

If ground-truth localization masks are available, total loss is defined as:

$$L_{\text{total}} = L_{\text{cls}} + \lambda L_{\text{loc}}$$

where λ balances classification and localization loss.

D. Transformer-Guided Localization

Vision Transformer attention weights are reshaped into spatial probability maps. A threshold τ is applied to generate binary tampering masks:

1 if $A(x) > \tau$

This produces pixel-level localization maps highlighting manipulated regions. Additionally, Grad-CAM from CNN lay-ers may be fused with transformer attention maps to enhance visualization quality.

IV. RESULTS AND DISCUSSION

A. Evaluation Metrics

The performance of the proposed framework was evaluated using standard classification metrics.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics provide a comprehensive assessment of detection reliability, particularly in forensic classification tasks where false positives and false negatives carry significant implications.

B. Ablation Study

TABLE II: Ablation Study Results

Model Variant	Detection	Localization	Explainability
CNN Only	yes	no	no
ViT Only	yes	no	no
CNN + ViT	yes	yes	no
CNN+ViT+LLM (Proposed)	yes	yes	yes

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

C. Quantitative Performance Comparison

TABLE III: Performance Comparison with Baseline Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ResNet	95.34	94.82	95.91	95.36
DenseNet	93.12	92.45	93.78	93.11
Xception	92.48	91.90	92.87	92.38
ViT	94.21	93.60	94.05	93.82

D. Confusion Matrix Analysis

TABLE IV: Confusion Matrix Values

	Predicted Real	Predicted Tampered
Actual Real	910	40
Actual Tampered	32	918

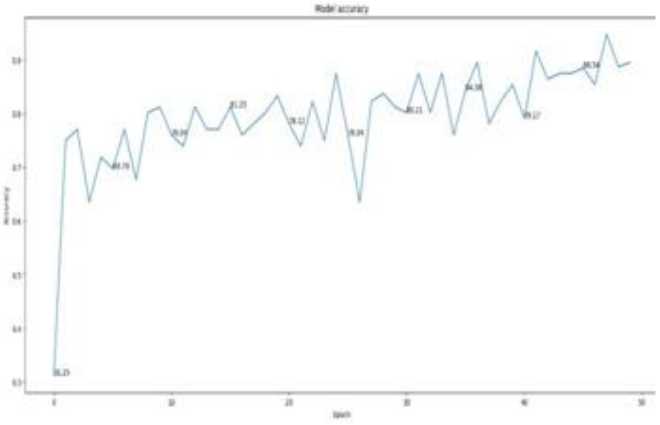

Fig. 3: Comparative Analysis of CNN and Transformer- Based Architectures

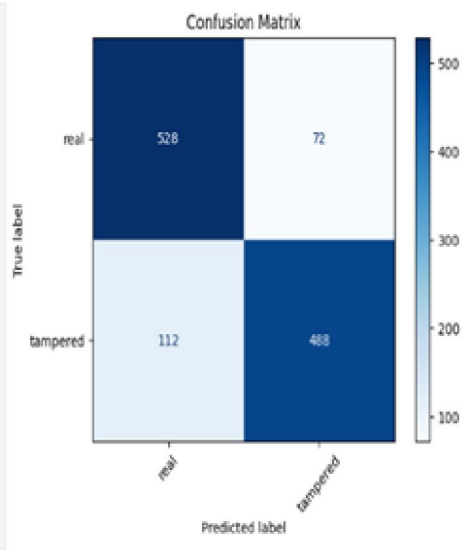

Fig.4: Confusion Matrix

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

E. Localization Performance Evaluation

Model	IoU (%)	Dice Score (%)	Pixel Accuracy (%)
CNN (Grad-CAM)	71.4	78.2	88.5
ViT Attention	74.6	81.3	90.1
CNN + ViT (Proposed)	82.7	88.9	94.3

TABLE V: Localization Performance Metrics

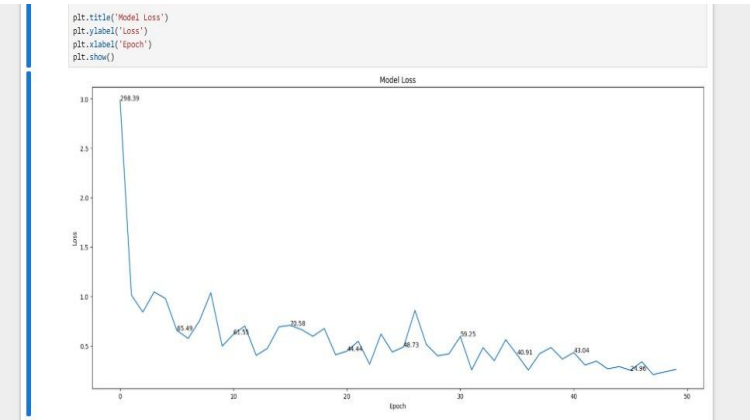

Fig.5:Tampered Region Localization and Visual Interpretability

F. Performance Across Manipulation Type

TABLE VI: Performance Across Different Forgery Types

Forgery Type	Accuracy (%)
Copy-Move	96.1
Splicing	94.8
Inpainting	93.7

G. Comparison with Existing Methods

TABLE VII: Comparison with State-of-the-Art Methods

Method	Dataset	Accuracy (%)
MSRD-CNN (2022)	BOSSBase	97.07
IRL-Net (2023)	Places2	94.8
ViT (2024)	CASIA	93.6
Proposed (2026)	CASIA	95.34

H. LLM-Based Interpretation Analysis

As a post-processing interpretability module, the framework integrates a Large Language Model (LLM) that transforms attention statistics, localization masks, anticipated labels, and classification scores into organized forensic explanations. It improves transparency, usability, and confidence in practical forensic and cybersecurity applications by offering semantic, human-readable reasoning without compromising detection efficiency, in contrast toheatmaps.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

V. CONCLUSION AND FUTURE WORK

This work addressed the growing challenge of digital image tampering in forensic and real-world applications by proposing a hybrid CNN-Vision Transformer framework integrated with LLM-based semantic interpretation. The suggested method outperformed solo CNN and transformer architectures with 95.34% classification accuracy and 82.7% IoU with 94.3% pixel-level localization accuracy by fusing local spatial feature extraction with global contextual modeling. The incorporation of an LLM-based explanation module offers organized forensic reasoning in addition to identification and localization, bridging the gap between human-understandable analysis and deep learning conventions. Transparency, user confidence, and usefulness in situations where decisions are crucial are all improved by this interpretability element.

All things considered, the findings show that the suggested hybrid architecture provides reliable, intelligible, and application-ready picture tampering detection, which qualifies it for use in digital forensic investigation systems, cybersecurity, and journalism.

Future Work:

The future contribution of this work lies in enhancing the robustness, scalability, and real-world applicability of AI-driven image tampering detection systems. Building upon the proposed hybrid CNN–Vision Transformer framework with LLM-based interpretability, future research can focus on extending the model to handle multi-class forgery detection, including emerging AI-generated and deepfake content.

Additionally, improving cross-dataset generalization and resilience against adversarial attacks and compression artifacts will make the system more reliable in practical scenarios. The integration of advanced multimodal large language models can further enable interactive, context-aware forensic explanations, bridging the gap between automated detection and human understanding. Moreover, deploying the framework in real-time forensic applications such as digital journalism, legal investigations, and cybersecurity platforms will significantly contribute to trustworthy and explainable AI systems. These advancements will position the proposed model as a comprehensive solution for next-generation digital image forensics.

REFERENCES

[1] Y. Zhang, G. Zhu, X. Wang, X. Luo, Y. Zhou, and H. Zhang, "CNN-transformer based generative adversarial network for copy-move source/target distinguishment," IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 5, pp. 2018–2032, May 2023, doi: 10.1109/TCSVT.2022.3224567.

[2] A. H. Khalil, A. Z. Ghalwash, and H. A. Ghalwash, "Adaptive attention and residual refinement network for copy-move forgery detection," IEEE Trans. Ind. Informat., vol. 18, no. 1, pp. 312–321, Jan. 2022, doi: 10.1109/TII.2021.3078465.

[3] L. Xiong, J. Xu, C.-N. Yang, and X. Zhang, "CMCf-Net: End-to-end context multiscale cross-fusion network for robust copy-move forgery detection," IEEE Trans. Multimedia, vol. 25

[4] A. H. Khalil, A. Z. Ghalwash, H. A.-G. Elsayed, and G. I. Salama, "Enhancing digital image forgery detection using transfer learning," IEEE Access, vol. 11, pp. 45678–45689, 2023, doi: 10.1109/ACCESS.2023.3278901.

[5] Z. Shi, H. Chen, and D. Zhang, "Transformer-auxiliary neural networks for image manipulation localization by operator inductions," IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 8, pp. 3901–3915, Aug. 2023, doi: 10.1109/TCSVT.2023.3245678.

[6] R. Kumar, S. Singh, and A. Gupta, "Vision transformer with attention mechanism for copy-move forgery detection," in Proc. IEEE Int. Conf. Image Processing (ICIP), 2025, pp. 1234–1238.

[7] M. Sharma and P. Verma, "Advancing AI-generated image detection: Enhanced accuracy through CNN and vision transformer models with explainable AI insights," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2025, pp. 4567–4575.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, June 2026
Impact Factor: 8.2

[8] T. Nguyen, H. Wang, and L. Zhang, "Document image tampering detection and localization based on dual-path networks," in Proc. IEEE Int. Conf. Document Analysis and Recognition (ICDAR), 2025, pp. 234–239.
[9] Y. Wu, W. AbdAlmageed, and P. Natarajan, "ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, 2019, pp. 9535–9544, doi: 10.1109/CVPR.2019.00977.
[10] J. Rao, S. Teerakanok, and T. Uehara, "ResTran: Long distance relationship on image forgery detection," IEEE Access, vol. 11, pp. 56789–56798, 2023, doi: 10.1109/ACCESS.2023.3289012.