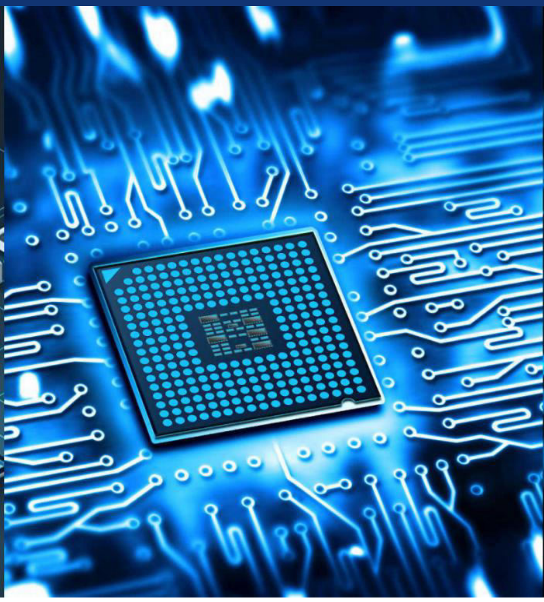
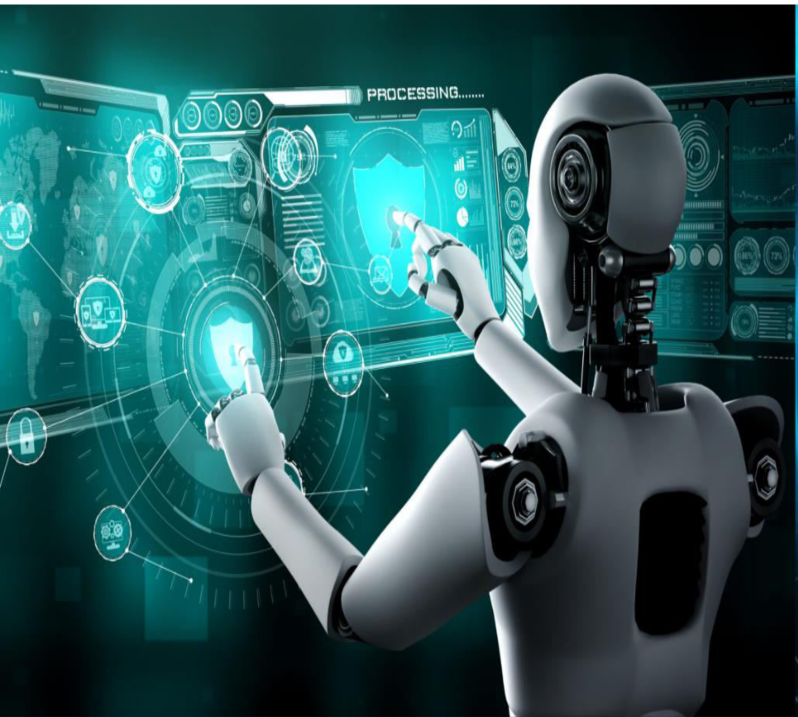
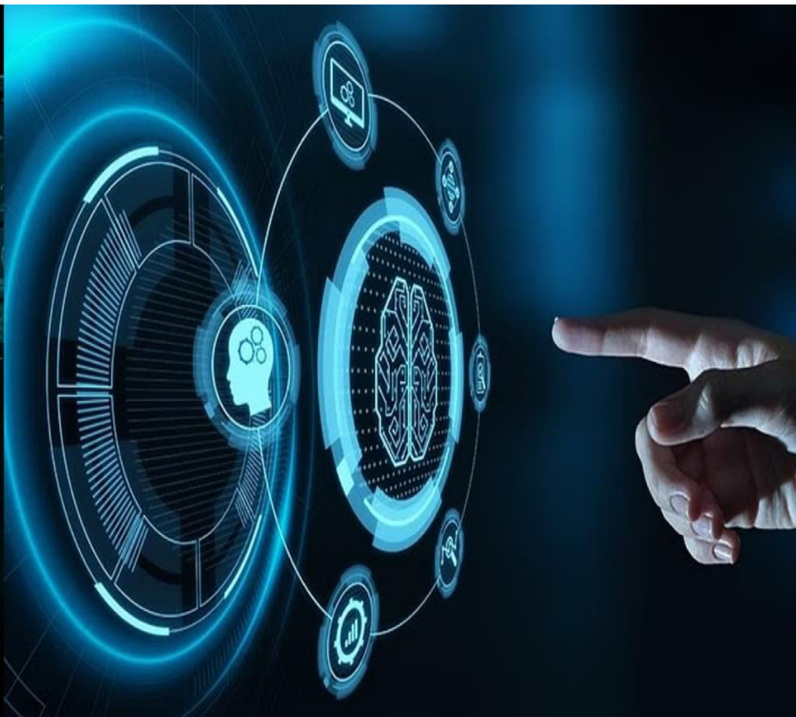
ISSN(O): 2320-9801
ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Impact Factor: 9.927

Volume 14, Issue 6, June 2026

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406050

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Advanced Approaches to Reverse Prompt Engineering in AI-Generated Content

Harish Kumar K, Magna Y, Sankara Narayanan S T.

PG Student, Dept. of Cyber Security, Dr. MGR Educational and Research Institute, Chennai, India

Assistant Professor, Dept. of Cyber Security, Dr. MGR Educational and Research Institute, Chennai, India

Assistant Professor, Dept. of ISDF, Center of Excellence in Digital Forensics, Perungudi, Chennai, India

ABSTRACT: The rapid advancement of generative Artificial Intelligence has significantly increased the creation of AI-generated digital content such as synthetic images and deepfakes. This has introduced major challenges in cyber security, digital forensics, and content authenticity verification. Conventional image analysis techniques often fail to accurately identify AI-generated content because modern generative models produce highly realistic outputs with minimal visible artifacts. This paper attempts to provide an advanced framework for Reverse Prompt Engineering and AI-generated image forensic analysis. The proposed system performs multimodal analysis using metadata inspection, forensic feature evaluation, AI probability estimation, and prompt reconstruction techniques. The framework analyzes uploaded images through integrated forensic modules such as EXIF metadata analysis, Error Level Analysis (ELA), FFT-inspired spectral analysis, and AI-assisted interpretation. The performance of the proposed system is analyzed based on forensic consistency, AI-generation probability, and reconstructed prompt estimation. The proposed system shows effective identification of synthetic media and improves digital forensic investigation capabilities through multimodal forensic analysis.

KEYWORDS: Reverse Prompt Engineering; AI-Generated Content; Digital Forensics; Synthetic Media Detection; Multimodal Analysis; Deepfake Detection; Metadata Analysis; Cyber Security

I. INTRODUCTION

Artificial Intelligence generated content consists of synthetic images, manipulated media, and AI-created visual content produced using advanced generative models such as diffusion models and Generative Adversarial Networks (GANs). AI-generated content can be created rapidly with high visual quality and can be distributed through digital platforms without restriction. The easy accessibility and realistic generation capability of modern AI systems make them highly suitable for media creation, entertainment, advertising, and social networking applications.

AI-generated images have introduced major challenges in cyber security and digital forensics because modern generative models produce highly realistic outputs with minimal visible artifacts. Conventional forensic techniques often fail to accurately identify synthetic media due to the absence of metadata traces and the presence of visually coherent structures. The misuse of AI-generated content in deepfakes, fake news, identity impersonation, and manipulated digital evidence has increased the importance of advanced forensic investigation systems. Researchers and industries are continuously working on techniques to improve synthetic media detection and content authenticity verification. But reverse prompt engineering and multimodal forensic analysis play an important role in AI-generated content investigation because these techniques help identify how the content was generated and whether the media is authentic or manipulated.

The main purpose of reverse prompt engineering and forensic analysis systems is to improve the identification of AI-generated content and reconstruct the probable prompts used for content generation. These systems are not only related to detecting synthetic images but also to analyzing metadata inconsistencies, forensic artifacts, and AI generation patterns to improve digital investigation capabilities. Reverse prompt engineering techniques can be based on multiple forensic metrics: i) AI probability estimation ii) multimodal forensic analysis. The first metric focuses on identifying synthetic media using AI-assisted interpretation and forensic indicators. The second metric focuses on metadata analysis, Error Level Analysis (ELA), FFT-inspired spectral analysis, and prompt reconstruction techniques for improving digital forensic investigation accuracy [1].

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. RELATED WORK

In [2] authors used metadata inspection techniques for detecting manipulated and AI-generated images by analyzing camera signatures, software traces, and EXIF inconsistencies present within digital media files. According to their approach missing metadata and editing traces were treated as suspicious forensic indicators for synthetic media detection. Deep learning based synthetic image detection techniques were considered in [3]. Authors used Convolutional Neural Networks and feature extraction mechanisms to identify hidden generation artifacts, texture inconsistencies, and abnormal visual patterns in AI-generated images. Frequency domain analysis and spectral feature extraction were also considered for improving detection accuracy. In [4] multimodal analysis techniques were used for reverse prompt engineering and AI content interpretation. Visual semantics, object relationships, and contextual structures were analyzed to reconstruct probable prompts used for AI image generation. Semantic matching functions were used to improve prompt estimation accuracy. In [5] authors improved AI-generated content analysis by implementing forensic scoring mechanisms into digital investigation systems. Images containing suspicious visual structures and metadata inconsistencies were assigned higher AI probability scores for improving synthetic media identification. In [6] Authors had integrated metadata analysis with AI-assisted interpretation models for improving digital forensic investigations. Only suspicious images were considered for deeper forensic evaluation while remaining images were classified using standard analysis techniques. AI probability scores and forensic consistency values were generated through multimodal investigation methods. In [7] authors considered reverse prompt reconstruction and multimodal forensic analysis for AI-generated content detection, as contextual semantic analysis helped improve prompt estimation and synthetic media identification accuracy. Reverse prompt engineering was performed by analyzing image composition, texture distribution, and visual consistency patterns. After image analysis was completed, the system generated forensic summaries and probable generation prompts. Multimodal analysis techniques were used to improve digital forensic investigation and AI-generated content identification accuracy.

III. PROPOSED ALGORITHM

A. *Design Considerations:*

- Uploaded images are converted into Base64 format for processing.
- Metadata extraction is performed using EXIF analysis techniques.
- AI-generated content probability is calculated using multimodal forensic analysis.
- Considered both single image and batch image analysis modes.
- Reverse prompt reconstruction is performed using AI-assisted interpretation.
- Images containing suspicious forensic indicators are treated as synthetic or manipulated content.
- The generated forensic report is considered as the final investigation output.

B. *Description of the Proposed Algorithm:*

Aim of the proposed system is to improve AI-generated content detection and reverse prompt engineering using multimodal forensic analysis techniques. The proposed system consists of three main steps.

Step 1: Calculating AI Probability Score:

The AI probability score (APSimage) of each uploaded image relative to its forensic indicators is calculated using multimodal forensic analysis techniques.

$$APSimage \propto FIimage$$

$$APSimage = k(FIimage) \qquad \text{eq. (1)}$$

where k is a constant and FIimage represents the forensic indicators extracted from the uploaded image.

Step 2: Selection Criteria:

Images should contain suspicious forensic indicators and metadata inconsistencies to be classified as AI-generated or manipulated content. All uploaded images are analyzed using metadata inspection, AI-assisted interpretation, Error Level Analysis (ELA), and FFT-inspired spectral analysis. Images containing suspicious metadata traces, synthetic visual patterns, or abnormal forensic structures are considered as feasible suspicious samples. Images with higher AI probability scores are assigned higher forensic risk levels. This selection criterion helps improve synthetic media identification and reverse prompt reconstruction accuracy. We tried to avoid incorrect classification of authentic images

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

by combining multiple forensic metrics into a unified investigation process. AI probability score of each uploaded image is calculated according to forensic indicators and the total forensic consistency score (TFCS) for the image is calculated using eq. (2).

$$TFCS = \sum_{i=1}^{m} FI \quad \text{eq. (2)}$$

where m is the number of forensic indicators in the uploaded image and FI represents suspicious forensic characteristics extracted during analysis. The image having maximum forensic consistency score i.e. max (TFCS) will be selected as highly suspicious AI-generated content.

Step 3: Calculating Reverse Prompt Estimation:
After forensic analysis is completed, reverse prompt estimation for each uploaded image is calculated using semantic interpretation and contextual visual analysis techniques. The reconstructed prompt estimation (RPE) is calculated using eq. (3) with parameters AI probability score (APSimage) and forensic indicators (FIimage).

RPE = APSimage + FIimage
eq. (3)

IV. PSEUDO CODE

Step 1: Upload and preprocess all input images.
Step 2: Calculate the APSimage for each uploaded image using eq. (1).
Step 3: Check the below condition for each uploaded image till all forensic analysis operations are completed.
    if (FIimage >= Suspicious Threshold)
        Classify image as AI-generated or manipulated..
    else
        Classify image as authentic content.
    end
Step 4: Calculate the total forensic consistency score for all uploaded images using eq. (2).
Step 5: Select the energy efficient route on the basis of minimum total transmission energy of the route.
Step 6: Calculate the reverse prompt estimation using eq. (3).
Step 7: Generate forensic investigation report.
Step 8: End.

V. SIMULATION RESULTS

The simulation studies involve AI-generated image forensic analysis using uploaded digital images as shown in Fig.1. The proposed reverse prompt engineering and multimodal forensic analysis system is implemented using React, TypeScript, Express.js, and Gemini AI integration. We analyzed different categories of synthetic and authentic images through the proposed forensic framework. Proposed system is compared between two metrics AI Probability Score and Forensic Consistency Score on the basis of synthetic media identification accuracy, reverse prompt estimation, and forensic investigation capability. We considered the forensic analysis time as investigation processing time and processing time is calculated during image analysis and report generation operations. Our results shows that the metric forensic consistency score performs better than AI probability estimation in terms of synthetic media identification, prompt reconstruction, and forensic investigation accuracy..

The forensic analysis system showed in Fig. 1 is able to identify suspicious AI-generated content with higher forensic consistency score and improved reverse prompt estimation accuracy. And the forensic investigation capability is also improved for multimodal forensic analysis. It clearly shows in Fig. 2 that the forensic consistency score provides more accurate synthetic media identification than individual AI probability estimation. As the uploaded images contain different forensic characteristics and metadata structures, the analysis results vary according to image composition and generation patterns. After uploaded images were analyzed the reconstructed prompts and forensic investigation summaries are shown in Fig .3 and forensic probability estimation results are shown in Fig. 4. Our results shows that the proposed multimodal forensic analysis system performs better in terms of AI-generated content detection, reverse prompt estimation, and digital forensic investigation capability.

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406050

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

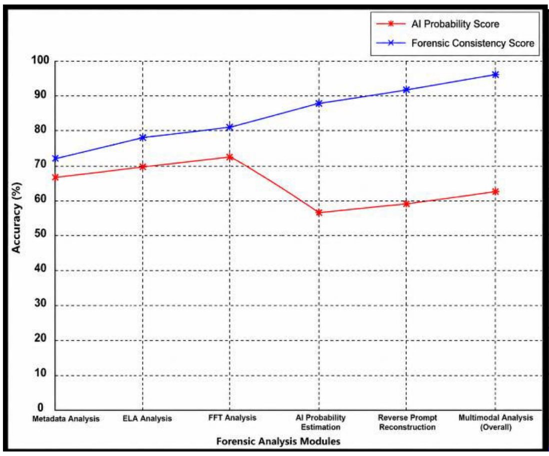


Fig.1. Forensic Analysis Structure

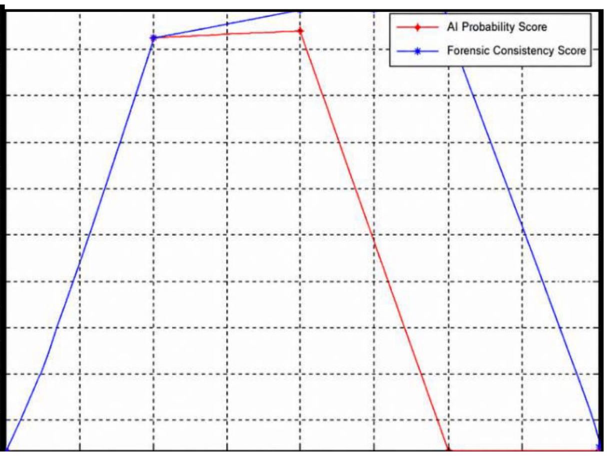


Fig. 2. Performance Comparison Results

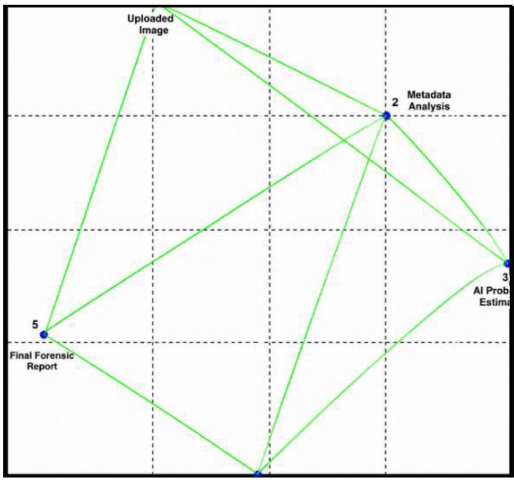


Fig. 3. Reverse Prompt Estimation Results

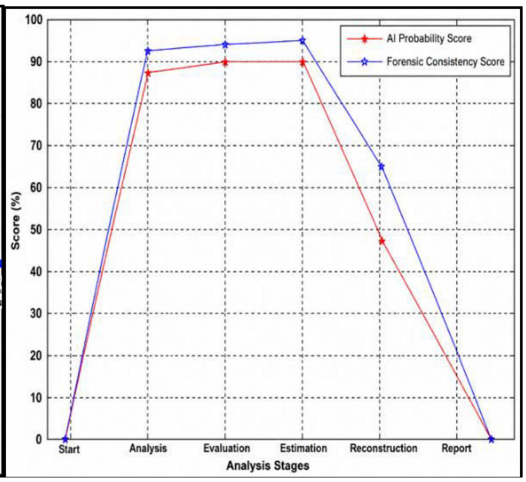


Fig 4. Multimodal Analysis Accuracy

VI. CONCLUSION AND FUTURE WORK

The simulation results showed that the proposed reverse prompt engineering system performs better with the forensic consistency score metric than the AI probability score metric. The proposed system provides improved AI-generated content detection and enhances digital forensic investigation capability through multimodal analysis techniques. As the performance of the proposed system is analyzed between two forensic metrics, in future with some modifications in forensic design considerations the performance of the proposed system can be compared with other AI-generated content detection and reverse prompt engineering techniques. We have used a limited number of forensic analysis modules, as the number of forensic parameters increases the complexity will increase. We can increase the number of forensic investigation modules and analyze the performance.

REFERENCES

1. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio, "Generative Adversarial Networks", Advances in Neural Information Processing Systems (NIPS), pp. 2672-2680, 2014.
2. Jonathan Ho, Ajay Jain and Pieter Abbeel, "Denoising Diffusion Probabilistic Models", Advances in Neural Information Processing Systems (NeurIPS), Vol. 33, pp. 6840-6851, 2020.

3. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684-10695, 2022.
4. Tero Karras, Samuli Laine and Timo Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4401-4410, 2019.
5. Hany Farid, "Image Forgery Detection", IEEE Signal Processing Magazine, Vol. 26, Issue 2, pp. 16-25, 2009.
6. Matthew Stamm, Min Wu and K. J. Ray Liu, "Information Forensics: An Overview of the First Decade", IEEE Access, Vol. 1, pp. 167-200, 2013.
7. David Cozzolino, Giovanni Poggi and Luisa Verdoliva, "Recasting Residual-based Local Descriptors as Convolutional Neural Networks: An Application to Image Forgery Detection", ACM Workshop on Information Hiding and Multimedia Security, pp. 159-164, 2017.
8. Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies and Matthias Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images", IEEE International Conference on Computer Vision (ICCV), pp. 1-11, 2019.
9. Yue Wu, Wael Abd-Almageed and Prem Natarajan, "ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9543-9552, 2019.
10. Sarah Kreps, Miles McCain and Miles Brundage, "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation", Journal of Experimental Political Science, Vol. 9, Issue 1, pp. 104-117, 2022.
11. Yisroel Mirsky and Wenke Lee, "The Creation and Detection of Deepfakes: A Survey", ACM Computing Surveys, Vol. 54, Issue 1, pp. 1-41, 2021.
12. Ning Yu, Larry Davis and Mario Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints", IEEE International Conference on Computer Vision (ICCV), pp. 7556-7566, 2019.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING