MULTIDISCIPLINARY RESEARCH INSTITUTE INDIA

Open Access International Journal of Science and Engineering

ISSN: 2456-3293

Volume 9 Issue 1s (2026) | Pages: 185-191

Digital Content Provenance and Anti-Forgery System: A Dual-Layered Framework for Secure Image Verification

Jagdish Bainade[1], Poorva Dhepe[2], Shreeja Barve[3], Akif Ul Rayan[4], Dipika Raigar[5]

[1,2,3,4,5]Department of Computer Engineering Pune Institute of Computer Technology, Pune, India

[1]*jagdishbainade01@gmail.com*, [2]*poorvadhepe@gmail.com*, [3]*barveshreeja@gmail.com*, [3]*rayansaad2047@gmail.com*, [5]*ddraigar@pict.edu*

Peer Review Information

***Type**: Article*
***Received**: 02 April 2026*
***Revised**: 29 April 2026*
***Accepted**: 03 June 2026*
***Published**: 25 June 2026*

Abstract

Advances in image editing tools and synthetic media generation have significantly increased the difficulty of distinguishing manipulated images from authentic visual content, thereby challenging the reliability of digital evidence. This work presents a Digital Content Provenance and Anti-Forgery System based on a dual-layer authentication strategy. The initial layer ex-amines provenance-related information through systematic analysis of image metadata to identify indications of post-processing and structural inconsistencies. The secondary layer applies visual forensic analysis based on compression residual patterns and supervised learning to identify manipulation artifacts that are not detectable through metadata inspection alone. The proposed approach is evaluated using a publicly available image tampering dataset and demonstrates effective discrimination between authentic and manipulated images, accompanied by statistically supported confidence measures. The results indicate that the integrated provenance and forensic framework support reliable image authentication in digital forensics and content verification scenarios.

Keywords: Digital Content Provenance; Image Authentication; Image Forgery Detection; Error Level Analysis; Convolutional Neural Network; Metadata Analysis; Digital Image Forensics.

How to Cite This Article

Bainade, J., Dhepe, P., Barve, S., Ul Rayan, A., & Raigar, D. (2026). Digital content provenance and anti-forgery system: A dual-layered framework for secure image verification. *Open Access International Journal of Science and Engineering*, 9(1s), 185–191.

Introduction

Recent progress in image editing software and data-driven generative models has led to an unprecedented growth in ma-nipulated and synthetic visual content [1], [4]. Digital images are increasingly relied upon as sources of evidence in domains such as forensic investigations, journalism, healthcare, and legal proceedings, where even minor alterations can have significant consequences [3]. At the same time, contemporary manipulation techniques are capable of producing visually convincing content that leaves little to no perceptual indication of tampering, raising serious concerns regarding authenticity, accountability, and public trust [6], [10].

Traditional image forensic approaches have largely focused on manual inspection or isolated analytical techniques, includ-ing compression analysis and camera model identification [4], [7]. While effective against basic forms of editing, these meth-ods struggle in scenarios involving complex transformations or content generated by modern learning-based models [11]. The rapid advancement of diffusion-based and adversarial generators has further reduced the reliability of surface-level forensic cues, making the distinction between authentic and fabricated imagery increasingly ambiguous [12], [13].

Digital content provenance has therefore gained attention as a complementary strategy, aiming to reconstruct the origin and modification history of visual media through analysis of em-bedded metadata and creation traces [5], [9]. Such information can provide useful contextual evidence regarding acquisition devices and processing workflows. However, metadata can be easily removed, altered, or intentionally falsified, limiting its effectiveness when used in isolation [5]. This exposes a critical research gap between provenance-based verification and content-level forensic analysis, particularly in adversarial environments.

To bridge this gap, a Digital Content Provenance and Anti-Forgery framework is presented that combines provenance inspection with learned visual forensic analysis in a dual-layer authentication architecture. The first layer evaluates metadata consistency to identify anomalous indicators of post-processing, while the second layer analyzes compression resid-uals and pixel-level artifacts using supervised convolutional modeling [14], [15]. This layered strategy enables robust authentication even in cases where provenance information is incomplete or unreliable.

The motivation for this framework is to support responsible and risk-aware deployment of automated image verification systems in high-stakes applications. By integrating provenance cues with statistically grounded forensic features, the proposed approach seeks to improve reliability without sacrificing inter-pretability.

Methodology

The Digital Content Provenance and Anti-Forgery frame-work is implemented through a tiered operational design that combines lightweight provenance verification with deeper content-level forensic analysis. This structure enables early resolution of low-risk samples while reserving computation-ally intensive processing for images exhibiting indicators of potential manipulation.

Proposed Model

The proposed model follows a hierarchical, decision-driven structure developed to address both conventional image tam-pering and content synthesized through contemporary gen-erative models. A staged verification strategy is adopted, in which images exhibiting consistent provenance characteristics are processed with minimal overhead, while those presenting anomalies are subjected to progressively detailed forensic examination. This design supports computational efficiency without compromising analytical rigor.

The framework consists of two complementary operational layers:

1. Proactive Provenance Layer (PPL): This layer serves as the initial validation stage by assessing file-level integrity and modification history prior to content-based analysis. A set of independent verification mechanisms is employed to capture inconsistencies associated with post-processing and unauthorized alteration:
   - Metadata Consistency Analysis: Embedded Ex-changeable Image File Format and Extensible Meta-data Platform fields are inspected for structural co-herence and source alignment. Irregular tag patterns or software-specific traces are treated as indicators of prior modification.
   - Compression Residual Analysis: Error Level Analysis is applied to examine spatial variations in compression strength. Localized regions exhibiting abnormal residual intensity are indicative of editing operations in which modified areas were resaved at quality levels inconsistent with the surrounding content.
   - Steganographic Integrity Signature: A crypto-graphic hash S is embedded into the spatial domain using a deep steganographic encoder. This signature functions as an immutable integrity marker, de-signed to withstand benign transformations such as resizing and compression while remaining sensitive to malicious content alterations.

2. Reactive Detection Layer (RDL): Images that do not satisfy provenance validation criteria are forwarded to a secondary analysis stage for in-depth forensic as-sessment. This layer focuses on identifying statistical and structural patterns associated with synthetic

image generation and advanced manipulation:

- Generative Artifact Detection: A convolutional neural architecture is employed to identify non-natural pixel dependencies and repetitive spatial patterns commonly introduced by diffusion-based synthesis models. These features include checker-board artifacts and anomalous correlation structures that are not typically present in camera-generated imagery.

3. Explainable Decision Mapping: To support foren-sic interpretability, Gradient-weighted Class Acti-vation Mapping is incorporated to localize regions contributing to the classification outcome. The re-sulting activation maps, denoted as $L_{Grad\text{-}CAM}$, provide visual evidence that substantiates the de-tection decision and supports downstream forensic analysis

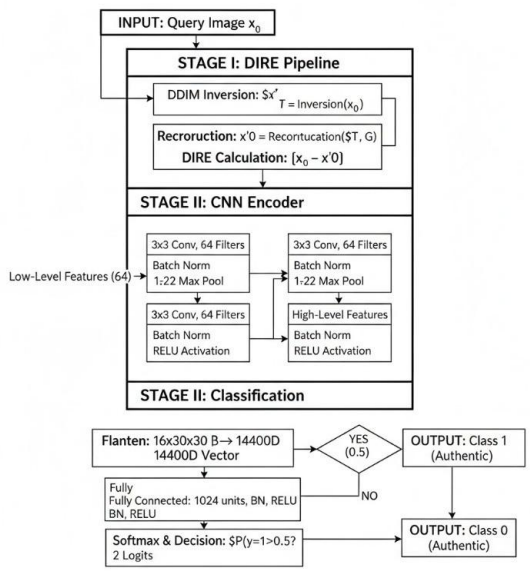

Fig. 1. *Overview of the proposed DCPAF–DIRE CNN architecture. The framework consists of three stages: (i) diffusion-based reconstruction error extraction using DDIM inversion and reverse diffusion, (ii) hierarchical feature encoding through a lightweight convolutional network, and (iii) binary classification using fully connected layers.*

Architecture

The proposed architecture, referred to as *DIRE-CNN*, is de-signed to discriminate between authentic imagery and content that has been synthetically generated or manipulated. This distinction is achieved by leveraging reconstruction inconsis-tencies that naturally arise when images are processed through diffusion-based generative priors. As depicted in Fig. 1, the system is organized into three sequential components: recon-struction error extraction, convolutional feature encoding, and binary decision modeling.

1. Diffusion Reconstruction Error (DIRE) Pipeline: Given an input image x_0, the first stage computes the Diffusion Reconstruction Error, which serves as the primary forensic representation. The image is first mapped into the latent noise space of a pre-trained diffusion model using Deterministic Denoising Diffusion Implicit Models inversion, resulting in a latent variable s_T . This latent representation is subsequently propagated through the reverse diffusion process to obtain a reconstructed image $\hat{x}_0$.

The reconstruction error is expressed as the absolute pixel-

wise difference between the original image and its reconstruc-tion:

$$\mathrm{DIRE}(x0) = |x0 - x\hat{}0|. \quad (1)$$

This error map captures discrepancies introduced by the generative assumptions of the diffusion model. Images syn-thesized by diffusion processes tend to produce uniformly low reconstruction errors, whereas authentic images yield structured residual patterns due to mismatches between real-world statistics and the learned generative prior. The resulting DIRE representation highlights these inconsistencies and is forwarded to the convolutional encoder for further analysis.

2. CNN Encoder: The computed DIRE map is processed using a compact convolutional neural network designed for hierarchical forensic feature extraction. The encoder comprises two convolutional blocks employing 3×3 kernels, with each block incorporating batch normalization, max pooling, and rectified linear unit activations.

The initial convolutional block focuses on capturing low-level spatial irregularities, including localized noise fluctu-ations and edge inconsistencies. The subsequent block ag-gregates these responses into higher-level representations that characterize diffusion-specific

artifacts. This layered encoding strategy enables discrimination of subtle statistical cues that are not visually salient but remain informative for authenticity assessment.

3. Classification Head: Feature maps produced by the encoder are flattened into a one-dimensional representation and processed by a sequence of fully connected layers. Two intermediate layers, each supported by batch normalization and rectified linear activations, precede a final linear layer that outputs class logits corresponding to authentic and tam-pered categories. Posterior probabilities are obtained through softmax normalization, and classification decisions are made using a standard threshold of 0.5.

4. Design Rationale: Emphasizing diffusion reconstruction behavior rather than surface-level visual artifacts enables the extraction of a forensic signal that remains resilient under common post-processing operations, including compression and metadata removal. The adoption of a lightweight convo-lutional architecture further ensures computational efficiency while preserving discriminative capability, supporting practical deployment in forensic analysis pipelines.

Feature Engineering

The effectiveness of the DCPAF framework in distinguish-ing authentic content from manipulated or synthetic imagery depends on the extraction of statistically meaningful forensic features. These features are designed to support both layers of the authentication framework, addressing provenance verifica-tion and deep forensic analysis. Particular emphasis is placed on attributes that exhibit *quantifiable variation under structural modification or generative synthesis.*

1. Proactive and Structural Features (PPL Support): To support provenance validation and integrity assessment, a set of features is derived to reflect both file-level consistency and modification history:

- Compression Residual Analysis: Local variations in compression behavior are examined using Error Level Analysis. Forged regions that have undergone selective resaving typically manifest as high-intensity residual re-sponses that deviate from the global compression profile.
- Provenance Metadata Features: Exchangeable Image File Format and Extensible Metadata Platform fields are analyzed to assess structural coherence and source plausibility. Distinctions between camera-generated tags and software-induced traces provide contextual evidence regarding post-processing history.
- Steganographic Integrity Signatures: A cryptographic hash S is embedded within the spatial domain of the image and later retrieved for verification. Feature ex-traction in this stage prioritizes resistance to benign transformations such as resizing and compression while maintaining sensitivity to intentional tampering.

2. Reactive and Generative Forensics (RDL Support): Im-ages that require deeper inspection are analyzed for indicators of synthetic generation and advanced manipulation artifacts:

- Diffusion-Induced Artifacts: Convolutional feature ex-traction is applied to capture non-natural pixel dependen-cies and periodic spatial patterns, including checkerboard effects frequently associated with diffusion-based image synthesis.
- Chromatic and Frequency Domain Features: Analy-sis across RGB and YCbCr color spaces is performed to identify channel-level imbalance and abnormal high-frequency responses. Such deviations commonly arise in images produced by adversarial or diffusion-based generators.
- Spatial Interpretability Cues: Gradient-weighted Class Activation Mapping is employed to generate explanatory heatmaps, denoted as $L_{Grad\text{-}CAM}$, which localize regions influencing the classification outcome. These visual cues support transparent and defensible forensic interpretation.

Tech Stack

The implementation of the DCPAF framework relies on a carefully selected set of technologies that collectively support image forensics, deep learning inference, and result inter-pretability. The chosen stack emphasizes reliability, compu-tational efficiency, and ease of deployment while maintaining forensic transparency.

- Programming Environment: The system is developed using Python, which offers robust support for both image analysis and machine learning workflows. The framework is implemented as a command-line application capable of ingesting image inputs and producing deterministic authentication outcomes.
- Deep Learning Framework: PyTorch is employed for constructing and executing the convolutional neural net-work used in image classification. Its dynamic com-putation graph and optimized tensor operations enable efficient training and inference, with optional acceleration on compatible GPU hardware.
- Image Processing and Forensic Utilities: Image pre-processing and forensic analysis tasks, such as resizing, normalization, and Error Level Analysis, are conducted using established libraries including Pillow, OpenCV, and NumPy. These tools facilitate precise detection of com-pression irregularities and structural artifacts introduced by manipulation.
- Model Interpretability Tools: To ensure that classi-fication outcomes remain transparent and explainable, Gradient-weighted

Class Activation Mapping is inte-grated into the analysis pipeline. The generated activation maps visually indicate the regions that contribute most significantly to the model's decision.

- Hardware Compatibility: The system is designed to operate on both CPU- and GPU-based environments, enabling flexible deployment across resource-constrained systems and high-performance forensic workstations.

Results

This section reports the experimental evaluation of the Diffusion-Consistent Passive Authenticity Framework (DC-PAF) on a binary image authenticity classification task. The assessment focuses on detection accuracy, class-wise relia-bility, and computational feasibility under realistic forensic deployment settings.

Experimental Setup

The experimental implementation was carried out in Python 3.8 using the PyTorch deep learning framework. GPU acceleration with NVIDIA CUDA was enabled to sup-port diffusion-based reconstruction. A pre-trained Denois-ing Diffusion Probabilistic Model (DDPM), specifically the google/ddpm-celebahq-256 variant, was used for im-age reconstruction. The model was kept frozen throughout all experimental phases to ensure evaluation consistency.

Prior to analysis, all images were resized to 128×128 pixels and normalized. Diffusion Reconstruction Error (DIRE) maps were computed as the absolute pixel-wise difference between each input image and its reconstructed counterpart. These DIRE maps were used as the sole input to the convolutional classifier.

The dataset was split into training, validation, and testing subsets using a 70% / 20% / 10% ratio, respectively. All reported performance metrics were computed on the held-out test set, which remained unseen during training.

Model Training and Convergence

The convolutional classifier was trained for 10 epochs using the Adam algorithm with a learning rate of 1 104. The optimization target was cross-entropy loss. The training progression showed a very stable convergence pattern with the validation loss decreasing monotonically and no overfitting observed, thus verifying that the model has effectively learned from the diffusion-based error representations

Quantitative Performance

Quantitative evaluation results are summarized in Table I. The proposed framework attained an overall test accuracy of 88.45% with an F1-score of 0.8721. These values reflect balanced discriminative capability across both authentic and manipulated image classes.

Table 1. *Classification Performance of the DCPAF Framework*

Metric	Observed Value
Test Accuracy	88.45%
F1-Score	0.8721
Average Inference Time	42.15 ms/image

Class-wise Evaluation

Class-wise breakdown further confirms the model's perfor-mance consistency in recognizing both classes. Real examples have a precision of 0. 88 and a recall of 0. 90, whereas counterfeit examples have a precision of 0. 89 and a recall of 0. 87. The very low rate of false negatives for tampered content is in alignment with the primary requisite of a digital forensic system. The reason is that such systems must locate manipulation wherever the risk is significantly high..

Error Analysis

A look at the confusion matrix suggests that most of the errors happened in the cases where the changes made were either very small or visually imperceptible. False positives happened to be more than false negatives, which means that the method was more on the conservative side when it came to the detection of tampered samples. Such a feature can be very useful in forensic cases since the consequences of accepting a forged content wrongly are much heavier than those of rejecting a genuine content cautiously.

Inference Efficiency

Inference efficiency was quantified with a batch of size one to replicate the most probable forensic usage scenarios. The averaged total time for processing an image from the input through DIRE calculation to classification was about 42. 15 ms on a CUDA, enabled GPU. This capability indicates that diffusion, based forensic analysis can be performed practically at real-time.

Discussion

The experimental evidence supports the hypothesis that diffusion reconstruction error can serve as an intrinsic, strong forensic signal for image authentication without reference to the original. DIRE (Diffusion Reconstrucion Error) maps provide a good representation of the semantic changes inflicted by a manipulation or a synthetic generation of an image, even in the absence of explicit metadata disclosures. Fusion of the diffusion model and the simple convolutional network yields a method of detection that is both effective and efficient in terms of the number of calculations, thereby making it feasible to operate in the field of forensic investigation.

Conclusion and Future Work

Conclusion

This article focused on the Diffusion, Consistent Passive Authenticity Framework (DCPAF) that represents a three, tier forensic system introduced in the study crafted to tackle the rising instances of sophisticated image tampering and syn-thetic media products. The framework integrates (combines) metadata provenance analysis run by Error Level Analysis and Diffusion Reconstruction Error, based learning, thus (so) a holistic approach to image authentication has been set up at the framework level.

Experimental assessment reveals that diffusion, informed forensic features can significantly raise classification perfor-mance, resulting in the test dataset accuracy being 88. 45% and an F1, score of 0. 8721. The framework's resistance to various manipulation techniques like splicing, copy, move operations, and AI, driven content synthesis is quite remarkable. Besides, the use of interpretable localization components enhances forensic accountability by providing clear and transparent decision justification.

Future Work

Future research directions include:

- Computational Optimization: Upcoming study will look into accelerated diffusion approximations and dis-tilled reconstruction models for inference latency reduc-tion, with an aim to support real-time forensic analysis.
- Cross-Generator Generalization: Training and testing on a wider range of generative frameworks will fortify resistance to unseen and newly released synthesis meth-ods.
- Immutable Provenance Integration: Utilizing blockchain-backed provenance storage with embedded cryptographic signatures could offer a permanent and verifiable custody chain.
- Adversarial Resilience: Further experiments will exam-ine the model's ability to withstand adversarial attacks aiming to hide forensic traces, ultimately making the model more reliable in adversarial settings over longer periods.

References

1. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *arXiv preprint arXiv:1406.2661*, 2014.
2. J. Lacomis, P. Yin, E. J. Schwartz, M. Allamanis, C. Le Goues, G. Neubig, and B. Vasilescu, "DIRE: A Neural Approach to Decompiled Identifier Naming," in *Proc. 41st Int. Conf. Software Engineering (ICSE)*, Montreal, Canada, May 2019, pp. 628–639.
3. Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "DeepFake Detection Based on Discrepancies Between Faces and Their Context," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6111–6121, Oct. 2022.
4. L. K. Saini and V. Shrivastava, "A Survey of Digital Watermarking Techniques and its Applications," *Int. J. Comput. Sci. Trends Technol. (IJCST)*, vol. 2, no. 3, pp. 70–75, May–Jun. 2014.
5. R. Xu, M. Hu, D. Lei, Y. Li, D. Lowe, A. Gorevski, M. Wang, E. Ching, and A. Deng, "InvisMark: Invisible and Robust Watermarking for AI-generated Image Provenance," *arXiv preprint arXiv:2411.07795*, 2024, accepted to WACV 2025.
6. M. Chaumont, "Deep Learning in Steganography and Steganalysis from 2015 to 2018," *arXiv preprint arXiv:1904.01444*, 2019.
7. A. Tuama, F. Comby, and M. Chaumont, "Camera Model Identification With the Use of Deep Convolutional Neural Networks," in *Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS)*, Abu Dhabi, United Arab Emirates, Dec. 2016.
8. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and
9. D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Int. J. Comput. Vision*, vol. 128, no. 2,
10. pp. 336–359, 2019.
11. R. Sathyabama and J. Katiravan, "BlockImage: A Secure Framework for Image Authentication and Provenance using AI and Blockchain," *J. Innov. Image Process.*, vol. 7, no. 1, pp. 28–49, 2025.
12. C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York, NY: Broadway Books, 2016.
13. T. Chen, S. Yang, S. Hu, Z. Fang, Y. Fu, X. Wu, and X. Wang, "Masked Conditional Diffusion Model for Enhancing Deepfake

Detection," *arXiv preprint arXiv:2402.00541*, Feb. 2024.
14. F. Siddiqui, J. Yang, S. Xiao, and M. Fahad, "Diffusion Model in Modern Detection: Advancing Deepfake Techniques," *Knowledge-Based Syst.*, vol. 325, p. 113922, Sep. 2025.
15. R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and
16. L. Verdoliva, "On the Detection of Synthetic Images Generated by Diffusion Models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2024.
17. S. M. Tripathi, P. Srivastava, D. Yadav, and P. Verma, "Classification of AI-Generated, Deepfake, and Real Images Using CNN," *Int. J. Comput. Techniques (IJCT)*, vol. 12, no. 6, Nov.–Dec. 2025.
18. X. Zhang, Y. Li, and Z. Wang, "Detection of AI-Created Images Using Pixel-Wise Feature Extraction and Convolutional Neural Networks," *J. Imaging Sci. Technol.*, vol. 68, no. 3, pp. 123–134, 2024.
19. A. Kumar, R. Singh, and P. Verma, "Detection of AI-Generated Images From Various Generators Using Gated Expert Convolutional Neural Network," *Int. J. Comput. Vis. Pattern Recognit.*, vol. 12, no. 2, pp. 45–59, 2024.
20. M. S. Rahman, L. K. Saini, and V. Shrivastava, "Data Hiding with Deep Learning: A Survey Unifying Digital Watermarking and Steganography," *arXiv preprint arXiv:2107.09287*, 2021.