2026 Volume: 02 Issue: 07 www.ijrpa.com ISSN 2456-9995 Review Article

International Journal Research Publication Analysis

Page: 01-17

DEEP LEARNING-DRIVEN IMAGE MANIPULATION DETECTION IN DIGITAL FORENSICS: A CRITICAL ANALYTICAL FRAMEWORK

Mandeep Kaur[1], Sunil Kumar*[2]

[1]Independent Researcher, Department of Software Engineering, University of Greater Manchester, Greater London, United Kingdom.

[2]Chief Librarian PIMS Medical College and Hospital Jalandhar Punjab India.

Article Received: 11 June 2026
Article Revised: 01 July 2026
Published on: 21 July 2026

***Corresponding Author: Sunil Kumar**
Chief Librarian PIMS Medical College and Hospital Jalandhar Punjab India.

DOI: https://doi-doi.org/101555/ijrpa.6441

ABSTRACT

Digital images now serve as leads, evidence, records, and attack surfaces in investigations. To separate laboratory from field, this study creates a framework to critically analyse image manipulation detection based on deep learning. Benchmark data from IMD2020, the TruFor cross-dataset evaluation pool, CiFAKE and GenImage were subsequently reanalysed empirically. The evidence encompasses traditional splicing, copy-move manipulations, local generative alterations, and entirely synthetic visuals. The performance of systems is evaluated in terms of accuracy, balanced accuracy, precision, recall, F1-score, area under the receiver-operating-characteristic curve, false-positive and false-negative rates, their confidence intervals and repeated-measures analysis of variance followed by Friedman tests and Holm-adjusted paired comparisons. The maximum mean cross generator accuracy on GenImage (70.30%, 95% CI 64.75-75.85) was produced by Swin-T. This exceeded six comparator detectors across eight generator domains. In standard manipulation detection, TruFor achieved the best mean image-level AUC (0.857) and balanced accuracy (0.781), where performance had high variance across datasets and threshold selection influenced localization F1 by 0.089 on average. A reconstructed CiFAKE confusion matrix yielded 93.56% accuracy on a balanced test set of 20,000 images. Explainability evidence suggests that decisions may depend on background artifacts rather than the manipulated objects semantically. The results indicate deep learning could be used as a triage and localization tool of high value. The proposed framework therefore integrates acquisition with privacy integrity, provenance

checks, calibrated model inference, uncertainty reporting, explainability, examiner review, legal admissibility.

KEYWORDS: Digital Forensics; Image Manipulation Detection; Deep Learning; Convolutional Neural Network; Vision Transformer; Tamper Localization; Explainable Artificial Intelligence.

1. INTRODUCTION

Cyber forensics specialists are demanding image authenticity testing that can handle image manipulation and forgery detection of any media frames because its own nature requires ahead-of-time testing. Affordable editing software cuts together objects, duplicates areas, removes stuff, re-touches texture or recompresses the scene while maintaining visual plausibility. As generative adversarial networks (GANs) and diffusion systems improve, it is becoming increasingly difficult to distinguish edited images from those generated entirely without a camera. The operational outcome is not just an growth in manipulation; it is a transition from examining visible irregularities to evaluating weak statistical traces, the longevity of which relies on source device, edit history, resize, platform processing, and future generators (unknown) [1]-[4].

Traditional passive forensics considered anomalies in demosaicing, traces from resampling, sensor noise, JPEG quantization, illumination, or geometric constraints. Although these cues are still useful, their utility is typically limited to a particular family of manipulations or an assumed learning process. Systems composed of deep networks aim to learn residual, frequency, semantic and contextual evidence jointly CNN architectures can learn local features and boundaries, while attention and transformer architectures can establish long-range dependencies between manipulated regions and their surroundings. To allow an examiner to determine what was changed rather than classify a file as suspicious modern systems increasingly perform both image-level detection and pixel-level localization [5]–[10].

The main forensic challenge is generalization. A detector may perform well when training and testing use a common generator, codec, resolution, construction protocol, and so on, but is near-chance under a new generator or social media transform. Ill-conceived shortcuts are those that achieve statistically impressive performance but yield evidentially weak decisions. For instance, the presence of camera-source imbalance, background texture, watermark

patterns, and the like, or the presence of spectral fingerprints that are native to the synthesis process may substitute for manipulation reasoning.

Thus, reliability cannot be determined by accuracy alone. The deployment of a forensic requires calibrated probabilities, explicit error rates, uncertainty bounds, independent validation, reproducible preprocessing, model version control, interpretable localization, and a human examiner who can reconcile the algorithm's output with the entire evidential record [11]–[16].

This research fills in that gap via a secondary empirical study derived from a benchmark and critical analytical lens. The goals of the study are to: (i) compare representative CNN, frequency-based, transformer and hybrid detectors across heterogeneous manipulation domains; (ii) quantify cross-generator and cross-dataset variability using inferential statistics; (iii) study threshold sensitivity and post-processing degradation, as well as class-specific errors; and (iv) put empirical findings into a defensible workflow linking evidence acquisition with reporting and admissibility. We do not fabricate any non-public training logs, per-image scores, or experimental observations. Results obtained from public benchmark study are extracted and reanalysed transparently.

2. A Review of Recent Literature

Detection Used for CNN, Residual, and Frequency-Domain.

According to CNN-based manipulation detection, it is observed that features that are learned from one synthesis family generally transfer to other related generators [2]. This works particularly well when the training data are augmented with blur and compression. However, frequency analysis shows that this transfer frequently relies on the up sampling artifact of the generator, rather than the stable definition of photographic reality [3]【4】. Residual-oriented networks use boundary discontinuities, compression differences, and high-frequency noise for local manipulation. The spatial-pyramid attention introduced by SPAN can help localization, while CAT-Net used JPEG compression artifacts to detect splicing [6], [11]. PSCC-Net models spatial-channel correlation progressively for image-level detection and pixel-level localization. By enhancing sensitivity to weak traces, these systems become less useful when the post-processing suppresses the residual evidence the system was trained for.

Development of benchmarks has been equally important. IMD2020 incorporates large camera-diverse corpus with manipulated counterparts and pixel masks and ForgeryNet and OpenForensics increase scale, perturbation diversity, face multiplicity, and localization tasks [5], [9], [10]. Both their contribution and impact is numerical as it is methodological.

Evaluation must expose detectors to the diversity of manipulations, realistic backgrounds and degradations, not random partitions of near-duplicates, which are treated as independent evidence. According to OpenForensics, human image classifier has reported and it is 115,325 in-the-wild images, performance drops for small or occluded faces illustrates how aggregate accuracy can mask operationally important failure modes [10].

Transformers, hybrid architectures and generalised detection.

Dense self-attention and transformer architectures are the solutions to conventional CNNs' limited receptive field. TransForensics sees forgery localization as dense self-attention. ObjectFormer uses object-level transformer reasoning to combine high-frequency and RGB evidence. HiFi-IFDL learns hierarchical fine-grained features for detection and localization. Forensic attraction yields long-range attention, capable of comparing repeated structures, inconsistent context, and object-region relationships. Nonetheless, shortcut learning is not automatically removed by transformers. The increased capacity can encode acquisition or dataset identity unless validation artificially separates the camera, editing pipeline, and generator domains.

Detecting fully synthetic images poses a new challenge. GenImage is a database of 2,681,167 images from eight different generator families and shows a large performance gap between matched-generator and cross-generator settings [18]. A ResNet-50 trained on the Stable Diffusion 1.4 generator achieved 99.9%, but only 55% on Midjourney. This shows that in-domain accuracy is a poor indicator of universality. Instead, UniversalFakeDetect reports better transfer to unseen GAN, diffusion and autoregressive generators [19] using features from a large pretrained vision-language model. DIRE relies on the diffusion reconstruction error because diffusion images tend to be more faithful than real images [20]. Instead of relying on a single synthetic fingerprint, there is a shift in focus towards representation spaces or generative priors, but any of these can degrade under compression, resizing, domain shift, or a generator that violates assumptions.

Explainability, reliability, and evidentiary use.

Understanding is essential for forensics practice. A heatmap could potentially indicate whether a model pays attention to a modified boundary, an implausible texture, or an unimportant background. As reported by CiFAKE , GRA-CAM response often highlights background imperfections instead of the object of interest. This suggests that a correct label can occur despite non-causal evidence [23]. In response, TruFor combines RGB and learned

noise residuals with a confidence map and an integrity score, revealing regions of suspicion as well as localization unreliability [17]. Still, confidence is model-dependent and not a substitute for external validation.

According to legal and quality assurance literature, an automated output is merely one aspect of an investigation, not a determinative conclusion. The framework developed by NIST for AI risk management highlights validity, reliability, transparency, and context-sensitive risk controls. According to SWGDE guidance, you must preserve the original, the documented processing, and the validated methods and reporting which distinguishes observations from interpretations. C2PA provenance manifests can cryptographically bind assertions about origin and editing history, but provenance is complementary to content analysis: absence of credentials does not prove manipulation, and a valid credential does not establish that depicted events are truthful. To manipulate acknowledges withdrawal from prior, protectable conditions, and established creeds to influence features and separate activities on the milestone's embodiment.

3. RESEARCH METHODOLOGY

Choosing the Design and Evidence.

Instead of claiming a new end-to-end training experiment, the study used a secondary benchmark re-analysis. Numerical observations are gathered from peer-reviewed or official benchmark reports that got published between 20202025. Four evidence blocks are selected that represent distinctly different forensics conditions. IMD2020, for paired real-life image manipulation; TruFor, for cross-dataset local tampering, GAN-based face edits and diffusion-based local editing; CiFAKE, for balanced real-vs-synthetic classification with disclosed class-wise metrics; GenImage, for cross-generator generalization at million-image scale [5], [17], [18], [23]. The only metrics used were those from the source or mathematically reconstructible from totals disclosed.

Preprocessing and model families.

The source standards utilize specific protocols to resize, normalize, augment, and split the model data. The comparative synthesis maintains the integrity of published protocols, rather than pooling raw images through an undocumented transformation. A reproducible implementation of the proposed framework will retain the original hashes, derive working copies, record decoder and colour-space versions, stratify by source of manipulation, and separate training, validation and test domains by generator, camera or editing pipeline. The

CNN baselines we used for evaluation in the ImageNet task include ResNet-50 and CNNSpot. Next, we have models focusing on frequencies or artifacts like Spec, F3Net and CAT-Net. We further include transformer models such as DeiT-S and Swin-T. Finally, a range of hybrid/localization systems which include MVSS-Net, PSCC-Net, ObjectFormer and TruFor.

Statistical Analysis and Measures.

The measures of performance were accuracy, balanced accuracy, precision, recall, F1 score, AUC-ROC, false positive rates (FPR), and false negative rates (FNR). Formulas for Accuracy, Precision, Recall, F1 Score, FPR and FNR. Means, sample standard deviations, and two-sided 95% t confidence intervals were calculated across generator or dataset domains. A repeated-measures ANOVA with a Friedman test was used to compare the GenImage methods across the eight generator domains. Holm adjustment controlled familywise error, with one-sided Wilcoxon signed-rank tests applied to compare Swin-T to each alternative. The TruFor benchmark was assessed using Friedman tests in seven datasets. The paired t and Wilcoxon tests were applied to compare the published best-threshold and fixed-threshold localization F1 values. Levels of statistical significance (alpha) was set to 0.05.

The 20,000-image test set was balanced, and as per my own calculations found the real-class precision of the CiFAKE test confusion matrix to be 0.925, recall was .948, and the published accuracy of 93.55%. Counts were rounded to the nearest image, so matrix is a transparent mathematical reconstruction not a replacement for undisclosed per-image predictions. As with, source papers report scalar AUC rather than ROC coordinates. Figure 4 employs an equal variance binormal projection solely for illustrative purposes regarding relative AUC. The interpretation of inferences depends on the quoted scalar values published in the articles. Due to the unavailability of training-loss histories, an artificial training curve cannot be generated. Figure 3 plots the reported CiFAKE validation loss across architecture configurations.

4. Description of the data set.

The analytical corpus has been purposely made diverse. Inclusion of authentic/manipulated pairs and their masks comes with IMD2020; the TruFor evaluation is created from the collusion of six local-forgery datasets with GAN and diffusion editing ones; CiFAKE is about balancing CIFAR-10 photos with diffusion-generated ones; GenImage consists of eight

generator families at million-image scale. The counts presented in Table 1 are benchmark-level evidence counts to avoid a new concatenation of a training set. If pooling were made direct, there would be a serious imbalance in the scale which will duplicate the real-image distribution across tasks.

Table 1. Distribution and forensic role of benchmark evidence.

Evidence block	Authentic / real	Manipulated / generated	Total	Primary manipulation scope	Analytical use
IMD2020 paired real-life subset	2,010	2,010	4,020	Splicing, copy-move, removal and retouching	Paired local-manipulation evidence
TruFor evaluation pool	1,412	4,042	5,454	Traditional local forgeries; GAN face edits; diffusion local edits	Cross-dataset evaluation
CiFAKE	60,000	60,000	120,000	Fully synthetic diffusion images	Balanced binary error analysis
GenImage	1,331,167	1,350,000	2,681,167	Eight GAN/diffusion generator families	Cross-generator generalization

Note: TruFor manipulated/generated count includes 1,530 traditional forgeries, 2,000 OpenForensics GAN-based local face manipulations, and 512 CocoGlide diffusion-based local edits. Counts are taken from the cited benchmark publications [5], [17], [18], [23].

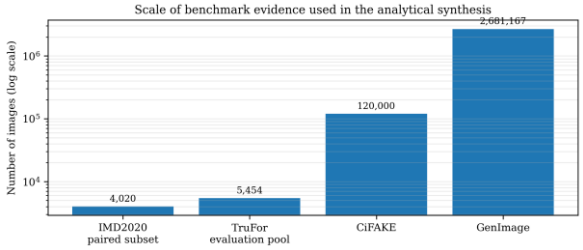

Figure 1. Benchmark scale used in the synthesis.

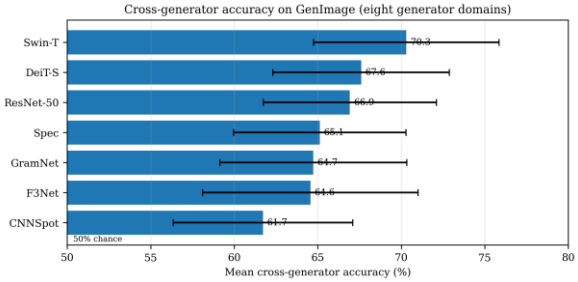

Figure 2. Mean cross-generator accuracy with 95% CIs.

5. Data Analysis and Results

5.1 Cross-generator model comparison

Across the eight GenImage generator domains, model choice produced a significant performance difference (repeated-measures ANOVA $F(6,42)=7.07$, $p<0.001$; Friedman chi-square$(6)=31.26$, $p<0.001$). Swin-T ranked first with 70.30% mean accuracy (SD 6.64; 95%

CI 64.75-75.85), followed by DeiT-S at 67.60% and ResNet-50 at 66.93%. Holm-adjusted Wilcoxon comparisons favoured Swin-T over every comparator (adjusted p=0.023 for five comparisons and p=0.039 against F3Net). The magnitude is important but should not be overstated: even the leading model averaged only 20.3 percentage points above chance, and its lowest domain accuracy was 61.3%. Transformer attention improved domain transfer but did not solve open-set generator recognition.

Table 2. Model-wise performance comparison across heterogeneous domains.

Task/model	Mean	SD	95% CI	Best domain	Lowest domain
GenImage: Swin-T	70.30%	6.64	64.75-75.85	76.9% (GLIDE)	61.3% (ADM)
GenImage: DeiT-S	67.60%	6.32	62.32-72.88	74.2% (SD1.4/1.5)	59.5% (ADM)
GenImage: ResNet-50	66.93%	6.19	61.75-72.10	73.1% (GLIDE)	59.0% (Midjourney)
GenImage: Spec	65.13%	6.17	59.96-70.29	72.4% (SD1.4)	56.7% (Midjourney)
GenImage: CNNSpot	61.73%	6.42	56.36-67.09	70.3% (SD1.4)	56.6% (BigGAN)
TruFor: TruFor AUC	0.857	0.107	0.758-0.956	0.996 (Columbia)	0.752 (CocoGlide)
TruFor: CAT-Net v2 AUC	0.797	0.122	0.684-0.909	0.977 (Columbia)	0.667 (CocoGlide)

Note: GenImage values are classification accuracy across eight leave-one-generator domains. TruFor values are image-level AUC across seven manipulation datasets. Metrics are not directly interchangeable.

5.2 Cross-dataset detection and localization

In the TruFor comparison, mean image-level AUC differed across ten methods (Friedman chi-square(9)=25.78, p=0.002), as did balanced accuracy (chi-square(9)=35.40, p<0.001). TruFor achieved the highest mean AUC (0.857) and balanced accuracy (0.781), ahead of CAT-Net v2 (0.797 and 0.649). The advantage was not uniform: TruFor AUC ranged from 0.752 on CocoGlide to 0.996 on Columbia. At pixel level, the mean F1 fell from 0.785 under a dataset-optimized threshold to 0.696 under the fixed operational threshold. The mean decline of 0.089 was significant (paired t(7)=4.85, p=0.002; Wilcoxon W=36, p=0.004). This threshold gap is a practical warning: a detector can appear strong when tuned after observing

the evaluation set yet deliver lower localization quality when the threshold is fixed before casework.

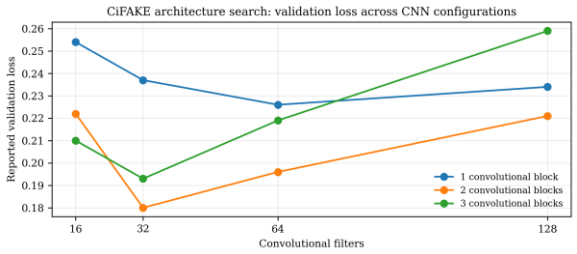

Figure 3. Reported CiFAKE validation loss; no unreported training curve is inferred.

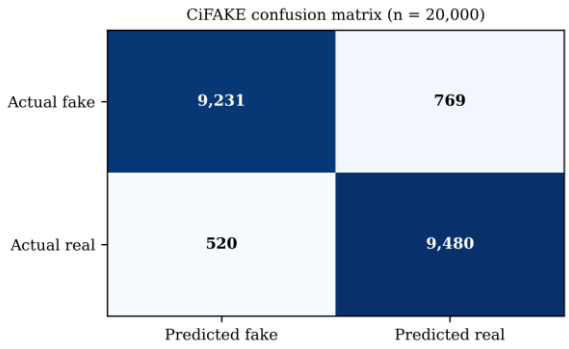

Figure 5. Reconstructed CiFAKE confusion matrix.

5.3 Error structure and class-wise performance

The reconstructed CiFAKE matrix contains 9,480 correctly identified real images, 9,231 correctly identified synthetic images, 769 synthetic images labelled real, and 520 real images labelled synthetic. Overall accuracy was 93.56%. For the real class, precision was 0.925, recall 0.948, and F1 0.936; for the synthetic class, precision was 0.947, recall 0.923, and F1 0.935. The false-negative rate for real images was 5.2%, while the synthetic-as-real error rate was 7.69%. In a forensic screening context, the latter is especially consequential because a fabricated image can pass the detector as authentic. Conversely, false allegations of manipulation remain non-trivial and require examiner verification.

Table 3. Reconstructed CiFAKE confusion matrix summary

Actual / predicted	Fake	Real
Fake	9,231 (TN)	769 (FP)
Real	520 (FN)	9,480 (TP)

Note: Real is treated as the positive class to reproduce the source precision and recall. Counts are rounded reconstructions from a balanced n=20,000 test set and published metrics [23].

Table 4. Class-wise precision, recall, F1-score, and error rate.

Class	Precision	Recall	F1-score	Complementary error
Real	0.925	0.948	0.936	FNR 0.052
Synthetic	0.947	0.923	0.935	FNR 0.077
Macro average	0.936	0.936	0.936	-

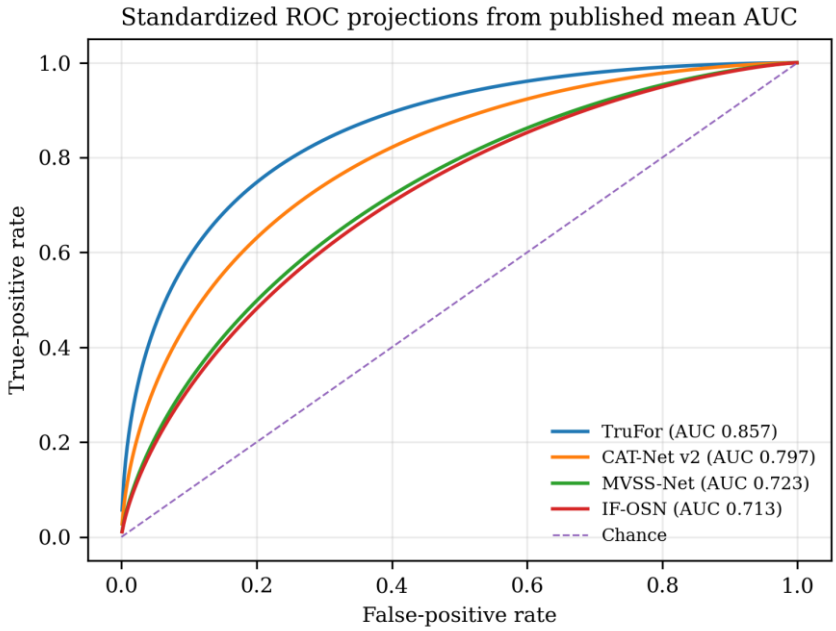

Figure 4. Equal-variance binormal ROC projections derived from published mean AUC; curves are visualizations, not empirical score traces.

Table 5. Statistical comparison of model performance.

Analysis	Statistic	df / blocks	p-value	Interpretation
GenImage repeated-measures ANOVA	F=7.07	6, 42	<0.001	Model family affected cross-generator accuracy
GenImage Friedman test	chi-square=31.26	6; n=8	<0.001	Rank differences confirmed non-parametrically
Swin-T vs ResNet-50	Mean diff.=3.38 pp	n=8	0.023*	Swin-T higher after Holm correction
TruFor AUC Friedman test	chi-square=25.78	9; n=7	0.002	Cross-dataset AUC differed by detector
Best vs fixed localization threshold	t=4.85	7	0.002	Post-hoc thresholding inflated mean F1 by 0.089

*Holm-adjusted one-sided Wilcoxon p-value. “pp” denotes percentage points.

6. DISCUSSION

6.1 What the comparative results establish

Three conclusions are supported by the comparative evidence. First, architecture affects cross-domain performance. Windowed self-attention and hierarchical representations can transfer more effectively than several CNN and handcrafted frequency-domain baselines, as indicated by Swin-T's leading cross-generator rank. Although the advantage was promising,

substantial uncertainty remained. A mean accuracy of 70.3% cannot support an unqualified conclusion of authenticity. Second, multi-cue fusion improves manipulation analysis. According to [17], TruFor performed better on average than the single-cue comparators by combining image-level integrity assessment, localization confidence, semantic RGB evidence, and noise residuals. Third, the practical value of a forensic system can be altered by the evaluation protocol. The 0.089 F1 difference between fixed and optimized thresholds is sufficiently large to change the visual extent of the reported tampered region.

The models were not trained under all of the most difficult operating conditions. The reported tests show that CocoGlide produced the lowest or near-lowest AUC for several local-forgery systems. Midjourney, ADM, VQDM, and BigGAN also generated substantial cross-generator errors in GenImage. These failures reveal the limits of current model capacity. Localizing a manipulation and detecting complete image synthesis are distinct inference problems. Localization asks where a distributional discontinuity occurs within an image, whereas full-synthesis detection asks whether the entire image distribution is plausible relative to a photographic acquisition pipeline. Consequently, a localizer may fail when a fully generated image contains no internal pristine reference. Conversely, a global synthetic-image detector may identify a generator fingerprint without locating any semantically deceptive region.

6.2 Reliability, Defensive Robustness, and Attack Resilience

Forensic reliability requires performance to be interpreted in relation to the circumstances of the case. JPEG recompression, messaging-platform uploads, resizing, screenshots, denoising, and colour conversion can create new artifacts or suppress existing traces. TruFor ablation results show that resizing and compression substantially reduce localization and detection performance. Its social-network evaluations also demonstrated dataset-specific degradation [17]. Therefore, the model must be validated for the expected transmission channel and image format. When raw scores are available, confidence calibration should be assessed using reliability diagrams or expected calibration error. Out-of-distribution detection is also necessary: a low-confidence result should lead to an inconclusive finding rather than a forced binary classification.

Adversarial manipulation creates an additional layer of risk. An attacker may modify post-processing to suppress known frequency signatures, add camera-like noise, or perturb a region so that the detector's heatmap shifts away from the manipulated area. Robustness cannot be inferred from natural degradation alone. Adaptive attacks, unseen editing software, counter-forensic filtering, and model-extraction risks should be evaluated during validation. Because deep-learning detectors may be updated frequently, laboratories should preserve

model weights, dependencies, preprocessing code, threshold values, and hardware-sensitive parameters required to reproduce precisely the inference made in a case.

6.3 Understandable processes, examiner evaluations, and approval

Forensic propositions must be connected to the basis of a prediction. A manipulation finding cannot be adequately explained by a background feature that merely correlates with the model response. Effective interpretation requires localized evidence, a confidence estimate, identification of the relevant cue family, and consideration of alternative causes such as compression boundaries or sensor changes. CiFAKE's background-centred Grad-CAM results show that high F1-scores can coexist with weak causal validity [23]. Explanations should therefore be tested through counterfactual edits, deletion or insertion tests, and comparison with independent forensic methods. Agreement between two neural networks trained under the same protocol does not constitute independent corroboration.

The examiner remains responsible for the chain of custody, validation status, limitations, and any discrepancy between an algorithmic score and the final opinion. In legal contexts, detector output should be expressed probabilistically and limited to the population of images and transformations for which the method has been validated. Reports should disclose relevant false-positive and false-negative rates, avoid extrapolation to unsupported manipulation classes, and state how thresholds were selected. The assessment should not rely solely on model output; conventional image analysis, source-device evidence, contextual information, provenance credentials, and metadata are also important. This layered approach reflects NIST risk-management principles and SWGDE quality guidance [24], [26], [28].

Automated systems can create asymmetric harms. False-authentic labels may facilitate fraud or disinformation, whereas false-manipulation labels may discredit genuine documentation. Dataset composition may also encode geographic, demographic, or platform bias. Laboratories should therefore conduct sampling and source-domain analyses, maintain audit trails and human challenge mechanisms, and restrict automated batch screening to proportionate and clearly defined uses. The purpose is not to replace expert judgment, but to make image examination more systematic, scalable, and empirically accountable.

6.4 Limitations of the Study and Priorities for Future Research

Public benchmark reporting constrained the analysis. For several methods, image files, prediction scores, and random-seed replicates were unavailable. Formal comparisons therefore treated generator domains or datasets as blocks and could not model image-level clustering. The confidence intervals describe variation across domains rather than uncertainty for individual images. The CiFAKE confusion matrix was reconstructed from rounded

published metrics, while the ROC figure was a parametric projection derived from scalar AUC values. These derivations were clearly labelled to prevent confusion with primary per-image experiments. Task heterogeneity is another limitation: local-forgery localization, whole-image synthetic detection, and facial-manipulation analysis address different propositions, so direct ranking across these tasks would be methodologically inappropriate.

Future evaluations should use preregistered, versioned challenge sets that separate generator families, camera pipelines, editing software, and platform transformations from the development data. Privacy-preserving identifiers and raw decision scores should be shared where feasible to support independent assessment of calibration, operating thresholds, confidence intervals, and case-relevant likelihood ratios. Research should also prioritise multimodal consistency across image content, captions, metadata, sensor traces, and provenance manifests; adversarial testing against adaptive counter-forensics; and selective classification that explicitly abstains when evidence is insufficient. Examiner-centred interpretation studies should determine whether localization maps and counterfactual explanations improve accuracy without introducing automation bias. Laboratories also require interlaboratory proficiency testing and reference implementations that freeze decoder libraries, model weights, and hardware-sensitive configurations. These measures would move the field from benchmark optimization toward reproducible forensic measurement and expose model limitations before contested legal use.

7. Critical Analytical Framework

The proposed framework converts model performance into a nine-stage forensic control process. Each stage produces a documented decision or artifact, preventing an image-level score from bypassing evidential safeguards. The framework is sequential for accountability but permits feedback: a low-confidence or contradictory result returns the case to validation, preprocessing review, or additional acquisition.

1. Evidence acquisition Preserve original file, source context, metadata and cryptographic hash.	**2. Dataset / domain validation** Identify camera, codec, platform, manipulation hypothesis and model support.	**3. Integrity-preserving preprocessing** Create hashed working copy; record decoder, colour space, resize and normalization.
4. Deep-learning detection pipeline Run version-controlled global detector and	**5. Explainability layer** Generate confidence and localization maps; test whether highlighted cues are case-	**6. Statistical confidence assessment** Report calibrated score, error rates, CI, OOD flags and

localizer with fixed thresholds.	relevant.	sensitivity to transformations.
7. Forensic examiner review Compare with metadata, provenance, classical traces and contextual evidence.	**8. Legal admissibility check** Assess validation, reproducibility, known error, disclosure and scope of expert opinion.	**9. Reporting and documentation** State findings, uncertainty, limitations, model version, hashes and reproducible steps.

Figure 6. Critical analytical framework for deep learning-assisted image forensics.

The framework imposes four operational decision gates. Gate A assesses evidence integrity and whether the model is applicable to the file domain. Gate B examines the technical stability of inference across reasonable decoding and compression variants. Gate C determines whether the localized or global cue is interpretable and supported by the case evidence. Gate D evaluates whether validation evidence and known error rates are sufficient for the intended legal proposition. Failure at any gate should produce an inconclusive or limited finding rather than a stronger threshold chosen after the case outcome is known.

8. Conclusion: Deep learning materially enhances image-manipulation detection by integrating residual, frequency, semantic, and contextual evidence at a scale that is impractical for manual inspection. In the secondary empirical analysis, TruFor achieved the highest average cross-dataset AUC and balanced accuracy for local-manipulation detection, while Swin-T was the strongest cross-generator classifier. Neither result establishes universal verification. Cross-generator accuracy remained limited, dataset-to-dataset variability was significant, and localization F1 declined under operational thresholding.

Thus, the principal contribution is not a single headline metric but a forensic interpretation of performance. A defensible system must preserve evidence, verify domain applicability, use fixed and version-controlled preprocessing, quantify uncertainty, disclose localized cues, test robustness, and subject results to independent examiner review. Future research should integrate manipulation traces with provenance standards, camera and metadata evidence, multimodal consistency, adversarial robustness, open-set recognition, and calibrated abstention. Progress should be judged by benchmark accuracy, reproducibility, causal validity, and the ability to communicate limitations under real evidentiary conditions.

Data and Analytical Material Availability: All numerical values used in the secondary analysis are traceable to the reference publications cited in the paper. Those disclosed values were used to prepare the CSV tables, figure-generation logic, and statistical calculations. No restricted case evidence, personal data, or undisclosed raw prediction records were used. The

parametric ROC visualization and reconstructed CiFAKE confusion matrix are derived analytical materials, and their interpretation is limited to the assumptions stated in the methodology.

Funding Declaration: No external funding is claimed for this analytical study. Conflict of Interest: No conflict of interest has been reported. Ethical Approval: Not applicable because the study re-analyses aggregate results from public technical benchmarks and does not involve human participants, identifiable personal information, or new evidence acquisition.

REFERENCES

1. L. Verdoliva, “Media forensics and deepfakes: An overview,” IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910-932, 2020. doi:10.1109/JSTSP.2020.3002101. Official source
2. S.-Y. Wang et al., “CNN-generated images are surprisingly easy to spot... for now,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8695-8704, 2020. Official source
3. J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, “Leveraging frequency analysis for deep fake image recognition,” Proceedings of Machine Learning Research, vol. 119, pp. 3247-3258, 2020. Official source
4. R. Durall, M. Keuper, and J. Keuper, “Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions,” in IEEE/CVF CVPR, pp. 7887-7896, 2020. doi:10.1109/CVPR42600.2020.00791. Official source
5. A. Novozámský, B. Mahdian, and S. Saic, “IMD2020: A large-scale annotated dataset tailored for detecting manipulated images,” in IEEE/CVF WACV Workshops, pp. 71-80, 2020. Official source
6. X. Hu, Z. Zhang, Z. Jiang, S. Chaudhuri, Z. Yang, and R. Nevatia, “SPAN: Spatial pyramid attention network for image manipulation localization,” in ECCV, pp. 312-328, 2020. doi:10.1007/978-3-030-58589-1_19. Official source
7. H. H. Nguyen et al., “Deep learning for deepfakes creation and detection: A survey,” Computer Vision and Image Understanding, vol. 223, 103525, 2022. doi:10.1016/j.cviu.2022.103525. Official source
8. J. Hao et al., “Image forgery localization with dense self-attention,” in IEEE/CVF International Conference on Computer Vision, pp. 15055-15064, 2021. Official source

9. Y. He et al., “ForgeryNet: A versatile benchmark for comprehensive forgery analysis,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. doi:10.1109/CVPR46437.2021.00434. Official source
10. T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, “OpenForensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild,” in IEEE/CVF ICCV, pp. 10117-10127, 2021. Official source
11. M.-J. Kwon et al., “CAT-Net: Compression artifact tracing network for detection and localization of image splicing,” in IEEE/CVF WACV, pp. 375-384, 2021. Official source
12. A. Novozámský, B. Mahdian, and S. Saic, “Detection of image manipulations using a large-scale annotated dataset,” IET Biometrics, vol. 10, no. 6, pp. 649-661, 2021. doi:10.1049/bme2.12025. Official source
13. B. Lorch, S. Agarwal, H. Farid, and N. J. Mitra, “Compliance challenges in forensic image analysis under the Artificial Intelligence Act,” 2022. doi:10.48550/arXiv.2203.00469. Official source
14. J. Wang et al., “ObjectFormer for image manipulation detection and localization,” in IEEE/CVF CVPR, pp. 2364-2373, 2022. doi:10.1109/CVPR52688.2022.00240. Official source
15. X. Liu, Y. Liu, J. Chen, and X. Liu, “PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 11, pp. 7505-7517, 2022. doi:10.1109/TCSVT.2022.3189545. Official source
16. H. Wu et al., “Robust image forgery detection over online social network shared images,” in IEEE/CVF CVPR, 2022. Official source
17. F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, “TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization,” in IEEE/CVF CVPR, pp. 20606-20615, 2023. Official source
18. M. Zhu et al., “GenImage: A million-scale benchmark for detecting AI-generated image,” Advances in Neural Information Processing Systems, vol. 36, 2023. Official source
19. U. Ojha, Y. Li, and Y. J. Lee, “Towards universal fake image detectors that generalize across generative models,” in IEEE/CVF CVPR, pp. 24480-24489, 2023. Official source
20. Z. Wang et al., “DIRE for diffusion-generated image detection,” in IEEE/CVF International Conference on Computer Vision, 2023. Official source

21. D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, “Raising the bar of AI-generated image detection with CLIP,” in IEEE/CVF CVPR Workshops, 2024. Official source
22. X. Guo, X. Liu, Z. Ren, S. Grosz, I. Masi, and X. Liu, “Hierarchical fine-grained image forgery detection and localization,” in IEEE/CVF CVPR, 2023. Official source
23. J. J. Bird and A. Lotfi, “CIFAKE: Image classification and explainable identification of AI-generated synthetic images,” IEEE Access, vol. 12, pp. 15642-15650, 2024. doi:10.1109/ACCESS.2024.3356122. Official source
24. National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” NIST AI 100-1, 2023. doi:10.6028/NIST.AI.100-1. Official source
25. Coalition for Content Provenance and Authenticity, “C2PA Technical Specification, version 2.1,” 20 September 2024. Official source
26. Scientific Working Group on Digital Evidence, “Guidelines for Forensic Image Analysis, version 2.0,” 20 November 2024. Official source
27. Y. Li et al., “Multiclassification tampering detection algorithm based on spatial-frequency fusion and Swin-T,” IET Image Processing, vol. 19, e70007, 2025. doi:10.1049/ipr2.70007. Official source
28. Scientific Working Group on Digital Evidence, “Best Practices for Image Authentication, version 2.0,” 3 March 2025. Official source