International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 21, Number 2 (2026) pp. 209-223
Research India Publications
https://www.ripublication.com

An Enhanced Image Forgery Detection and Explainability System Using LLM Approach

Dr. K. Kranthi Kumar[1], Mr. P. Mahesh[2], Sudhati Gopi Chand[3], Burri Ganesh[4], Boddupally Srinu[5]

Department of Information Technology, Sreenidhi Institute of Science and Technology, Yamnampet 501301, Telangana, India.
[1]*E-mail: drkkranthikumar@gmail.com* , [2]*maheshkumarpunna@gmail.com*
[3]*maheshkumarpunna@gmail.com* , [4]*ganeshburri1@gmail.com*
[5]*boddupallysrinu979@gmail.com*

Abstract

This paper introduces an improved and explainable image forensics system to determine the uploaded image into three categories: authentic, manipulated, and GAN-generated. The proposed framework incorporates a detector based on the spatial-texture CNN and a detector in the frequency domain of the GAN-based artifact detector, followed by a fusing strategy of the outputs of the detectors by using a confidence aware aggregation strategy along with disagreement check-ups for a reliable decision-making system. In addition to classification, the system produces human-easily-interpretable forensic reasoning results using LLM-based module (with deterministic fall back) real-time outputs for practical application. A production-oriented implementation that includes a FastAPI implementation of the inference backend, a web-based frontend for uploading images and downloading reports, and a modular training pipeline that includes supporting the end to end compile, train, evaluate and deploy of the dataset. Experiments on a 10k samples training setup per task, with 15 epochs, using 2,000 samples split tests, obtain strong results, in which the CNN branch gets 99.85% accuracy for forged vs. authentic detection (F1: 0.9985, ROC-AUC: 0.99996), and the GAN branch gets 97.55% accuracy for synthetic vs. natural detection (F1: 0.9749, ROC-AUC: 0.9988). The results show that hybrid spatial-frequency fusion using explainable reasoning is useful for actual digital image integrity verification.

Keywords— Image forensics, deepfake detection, GAN-generated image detection, image manipulation detection, convolutional neural networks (CNN), frequency-domain analysis, explainable AI (XAI), confidence-aware fusion.

Introduction

The SEM0 modulation of generative models has resulted in image synthesis and manipulation thus rendering it significantly more accessible which has resulted in serious concerns for digital trust in the areas of journalism, social media, law enforcement evidence, and security-sensitive workflows. Modern image forgeries no longer look like such blatant edits; instead, they often maintain realistic structure, texture and lighting, making human analysis unreliable. At the same time, the amount of online visual content has increased to a size that makes manual verifications impractical. These trends have made automated image forensics a critical requirement instead of a research luxury. However, robust forensic systems must accomplish more than generating a label; they must generate consistent estimates of confidence, as well as reasonably interpretive reasoning so that users can gain some insight into why a particular image is deemed suspicious or authentic. Most existing detection pipelines are limited in some way due to their practical application: they only conduct binary classification, rely on a single class of models, fail across distributions, or fail to offer much explanation for decisions. Spatial-domain CNNs tend to be good enough at detection of local inconsistencies in texture and blending artifacts, but tend to overlook signatures that may be present at a frequency level which is common in synthetic generation pipelines. On the other hand, frequency-oriented detectors are able to register periodic artefacts and checkerboard traces, but might be less sensitive to slight localised retouching. Used on their own, the two approaches each have blind spots, which makes them less reliable to be used in the real world. In addition, many research prototypes cease at offline evaluation and fail to involve API integration, inference orchestration, reporting generation that is needed for operational usage. To overcome such limitations, an improved hybrid image forensics framework that models spatial as well as frequency evidence in a combined tri-class decision scenario is proposed in this work. The system is made to classify images as real, manipulated or GAN-generated images, reminiscent of realistic forensic needs as opposed to simple binary schemes. A CNN-based detector catches spatial and texture anomalies related to pixel-level tampering whereas a GAN artifact detector analyzes frequency characteristics that are related to synthetic image generation. Their outputs are fused into a confidence-aware aggregation module which takes into account detector agreement and disagreement pattern prior to deciding with a preliminary decision. This design is explicit about what we consider appropriate types of strengths, in attempt to provide greater robustness for types of manipulations and maintain the clear kinds of decision logic. The architecture implemented is a modularity, production oriented pipeline. An image input layer is used to validate the uploaded content, perform any pre-processing and convert the input into normalized tensors that can be fed into other models. The detection layer is implemented with CNN and GAN branches, which both return score, label, confidence and detector level signals. A layer of aggregation of inference is then used to compute a weighted concatenated score and it uses console checks such as when there is disagreement in detector outputs whereby, consistency penalties are imposed. This makes a successful standard manner of representation to continue with reasoning and reporting. The backend is exposed as endpoints through the FastAPI, including health and metrics interfaces

allowing to monitor performance, and a web frontend allows to upload images, visualize verdicts, display their confidence, the explanations, and to download a report for end users. Beyond classification, the framework focuses on explainability through a reasoning layer which processes the evidence from the detector into human readable forensic stories. The rationale factor suggests an LLM provider ecology to richer contextual clarification and a deterministic fallback ecology to reproducibility in offline or constrained surroundings. This bilateral mode design is a compromise between interpretation and operational accuracy. The system is also stateless, so that it does not keep uploaded images for inference purposes, which helps to meet privacy-conscious deployment requirements. In addition, the loading of models delivers support for local TorchScript artifacts and conditional remote retrieval for comfortable an option of deployment from local machines as well for cloud environments. These engineering choices make the system amenable for experimentation, not only but practical integration situations. Experimental evaluation report has been made using end-to-end 10k pipeline configuration of the N with 15 trainings per detector task using dedicated train/validation/test manifests. On a 2,000 samples CNN test split for forged vs. authentic, the model was able to achieve 99.85% accuracy, 0.998999% precision, 0.998 recall, 0.9985 F1-score and 0.999958 ROC-AUC. On a synthetic versus natural 2,000 sample test data, the model results were 97.55% accuracy, 0.997906 precision, 0.953 recall, 0.9749 F1-score, and 0.998771 ROC-AUC. These results show a very good discriminative performance in both branches, special mention to high precision and AUC. Although the branch of the GAN cannot get through the forensics triage strict accuracy thresholds of 98%, the overall performance is able to perform well in practical regions of forensic triage and explainable decision support. The author has made three critical contributions, which are as follows: a hybrid spatial-frequency detection approach to tri-class image forensics a confidence conscious fusion machine with explicit consistency requirements, and an interpretable end-to-end system that maps model outputs to pupillable forensic reports. In contrast to the detectors of a narrow scope, the system proposed encompasses model training, API inference, explanation generation, as well as report provisioning into a single modular stack. This combination lends to an enhancement in the practical value in forensic AI from both the technical and decision transparency perspectives. Overall the results of the paper prove that merging complementary forensic signals with structured reasoning flow materially improves the workflow of image integrity verification, and also formulates a clear basis for further development toward scenarios such as multimodal metadata analysis, video-level consistency check, and domain shift-robust training strategy.

Literature Survey

Recent advances in image synthesis have been greatly reconstructed by advancements of generative adversarial networks. Karras et al. have enhanced the visual quality of StyleGAN and also examined characteristic artifacts that occur in synthesized images which makes their work very relevant to forensic detection of GAN outputs [1]. Sauer et al. later extended this line through StyleGAN-T, proving that GAN-based text-to-image generation refers to the synthetic creation of large-scale content in a reasonable

amount of time, which in turn raises the demand for effective solutions for their detection [2]. Earlier, Karras et al. proposed progressive growing of GANs which is an important step towards increasing stability and realism of synthesis [3], while Brock et al. have shown that large-scale GAN training can be utilized to produce highly realistic natural images [4]. Together, these works could demonstrate just how fast the development of generative models has progressed, which has made the process of detecting fake images increasingly hard.

Beyond GANs, there was also the synthetic images ecosystem, which was aided by representation learning and multimodal generative modelling. Higgins et al. proposed beta-VAE which gave more interpretable latent representations and had an impact on controllable and image generation research [5]. Radford et al. introduced CLIP, a powerful vision language model that trained a transferable visual features from the natural language supervision [6]. Ramesh and others also made a hard step forward in zero-shot text-to-image generation by demonstrating that synthesis machines can generate convincing images for texts [7]. These developments meant an increased diversity and a realism of artificial imagery, which consequently enhances the importance of forensic systems that are able to make a difference between genuine and generated content.

In the more specific area of forensic science, Lyu discussed the challenges of detecting deepfakes in greater detail and the requirement for methods that are general across the types of manipulation and generation pipeline [8]. Wang et al did a survey on the face detection with GAN and pointed out open issues like robustness, transferability and the aging of models [9]. Wang et al. also demonstrated that the images created by CNN were surprisingly easy to detect at the time, as synthetic pipelines would introduce different artefacts which could be detected by classifiers [10]. Zhao et al. enhanced deepfake detection using a multi-attentional framework which proved that spatially focused feature extraction assists to enhance the performance of discriminative [11]. However, Gragnaniello et al. desperately analyzed state-of-the-art GAN image detectors and suggested that sharp benchmark accuracy cannot necessarily pass into reliable the real-world [12].

A related line of work examined the forensic traces that are left by the creators of synthetic images. Yu et al. proposed GAN fingerprint, demonstrating that generated images can have signatures from the source model which can be useful for attribution [13]. Chai et al. investigated what makes fake images potentially detectable, and discovered the existence of low level artifacts that are found to be more generalizable than semantic cues [14]. Liu et al proposed global texture improvement for fake face detection, indicating that texture inconsistencies are a good forensic signal in the unconstrained environments [15]. Zhang et al examined the simulation and detection of GAN artifacts, helping explain the origin of some of the distortions of synthetic images [16]. Frank et al. then showed that frequency analysis is particularly suitable for deepfake recognition and motivated the use of spectral cues in deepfake related forensic pipelines [17].

The study of the manipulation of images has also been affected by benchmark datasets and application-oriented research. Rössler et al. proposed FaceForensics++, which is one of the most influential datasets to learn and test manipulated facial image

detection methods [18]. Marra et al. was focused on spotting GAN produced fake images that were shared over social networks as it has brought to light the problem of degradation due to reposting and compression [19]. Cozzolino et al. later delved into the problem of universal GAN image detection with a goal of creating detectors that are not learned to detect a single known source, but have a larger generalization to multiple generator families [20]. Jeon et al. proposed the T-GD, which is a transferable generative adversarial network-based image detection method designed to achieve better robustness in unknown domains [21]. As a whole, these works highlight the fact that generalization is one of the most significant requirements for practical deployment.

To further increase the robustness, a number of investigators combined complementary types of features. Ju et al fused global and local features for generalized AI synthesized image detection showed that multi scale evidence enhanced detector performance [22]. This idea was further extended in GLFF in which Ju et al. proposed a structured global-local feature fusion framework for AI synthesized image detection [23]. Wang et al. proposed a two-stream CNN containing the PRNU information for fake face detection in the wild using the GAN algorithm (spatial learning) and forensic sensor-level cues [24]. He et al. looked beyond direct analysis of spectra, and instead proposed the use of re-synthesis-based detection that demonstration of reconstruction behavior may reveal synthetic content hard to detect by the simple frequency statistics [25]. Park et al. then compared and visualized different AI-generated-image detection methods and helped provide empirical insight as to performance trade-offs and interpretability [26].

More recent works have also studied multimodal and vision-language for synthetic image detection. Keita et al. suggested a vision-language format called Bi-LORA that hypothesizes that language-directed representations can bolster the detection of synthetics under more general circumstances [27]. In parallel traditional forensic research on source attribution continues to remain relevant. Manisha et al. came up with strong device-specific fingerprints for the identification of source camera (this is done to go further than classical PRNU-based methods to learn resilience) [28]. This was further carried forward towards video-based source camera identification, which demonstrates the development of forensics fingerprinting from images to more complicated medium [29]. The seminal work of Lukas et al. on digital camera identification from sensor pattern noise is still key to this direction, as it helped to establish one of the first and most influential acquisition-trace-based digital forensic methods [30].

Overall, three major trends can be identified from the literature: fast advances in generative image realism, increased interest in the generalized synthetic-image detection problem and comeback of explainable forensic evidence problems. The existing studies warrant the notion that not a single cue is sufficient in all cases. Spatial artifacts, frequency-domain inconsistencies, generator fingerprints, and source-acquisition traces, although they are very useful all provide incomplete evidence. This actually motivates directly a hybrid image forensics framework that uses complementary detectors together and generates interpretable reasoning for final decisions.

Table 1: Literature Review

S.No	Author	Description	Advantages	Disadvantages
1	Karras et al.	Improved StyleGAN and analysed artifacts	High-quality synthesis, useful for forensics	Harder detection due to realism
2	Sauer et al.	StyleGAN-T for text-to-image generation	Scalable image generation	Increases fake content
3	Karras et al.	Progressive GAN training	Improved stability & realism	Complex training
4	Brock et al.	Large-scale GAN training	Highly realistic images	High computation
5	Higgins et al.	Beta-VAE representation learning	Interpretable latent space	Trade-off in reconstruction
6	Radford et al.	CLIP vision-language model	Strong multimodal learning	Can aid fake generation
7	Ramesh et al.	Zero-shot image generation	Flexible synthesis	Hard to detect
8	Lyu	Deepfake detection challenges	Highlights research gaps	No universal solution
9	Wang et al.	GAN face detection survey	Identifies open issues	Model aging problem
10	Wang et al.	CNN-generated image detection	Easy detection (early GANs)	Weak for modern GANs
11	Zhao et al.	Multi-attention deepfake detection	Better feature extraction	Complex architecture
12	Gragnaniello et al.	GAN detection evaluation	Real-world benchmarking	Poor generalization
13	Yu et al.	GAN fingerprinting	Source attribution	Limited to known GANs
14	Chai et al.	Low-level artifact analysis	Generalizable detection	Subtle artifacts vanish
15	Liu et al.	Texture-based detection	Works in real-world	Noise sensitive
16	Zhang et al.	GAN artifact simulation	Explains distortions	Limited application
17	Frank et al.	Frequency-based detection	Strong spectral cues	Can be bypassed

18	Rössler et al.	FaceForensics++ dataset	Benchmark dataset	Dataset bias
19	Marra et al.	GAN detection on social media	Real-world relevance	Compression issues
20	Cozzolino et al.	Universal GAN detection	Better generalization	Lower accuracy
21	Jeon et al.	T-GD transferable detection	Cross-domain robustness	Complex training
22	Ju et al.	Global + local feature fusion	Improved accuracy	High computation
23	Ju et al.	GLFF framework	Structured feature fusion	Complex design
24	Wang et al.	Two-stream CNN with PRNU	Combines spatial + sensor cues	Sensitive to noise
25	He et al.	Re-synthesis detection	Detects subtle patterns	Computationally expensive
26	Park et al.	Method comparison study	Insight into trade-offs	No best model
27	Keita et al.	Bi-LORA vision-language model	Better generalization	Early-stage research
28	Manisha et al. (2022)	Device fingerprinting for camera ID	Strong attribution	Device dependent
29	Manisha et al. (2023)	Image → video-based source identification	Works for video forensics	Compression challenges
30	Lukas et al.	Sensor pattern noise method	Foundational forensic technique	Affected by noise/compression

Proposed Work

The proposed work aims to develop an end-to-end image forensics framework that is reliable to classify the input image as either authentic, manipulated, or generated by a GAN and also provide an understandable, forensic explanation for the human users. The basic motivation is the common failure of single-model detectors in the case of varying manipulation styles, and hence, the system is not built around a single classifier, but rather the complementary sources of evidence. The structure will be applied practically within the context of verification pipelines involving digit counting where detection quality as well as explanability is needed. It is concerned with real deployment constraints such as gain access to API, Latency aware inference flow, and stateless handling of uploaded files.

The first major component of the proposed system is a modular system to infer the architecture that includes the image input layer, dual detector layer, aggregation layer and the reasoning layer. The input image is validated and processed to a normalized representation of the tensor for the compatibility of the model. Two parallel detectors then handle the same input: one branch detecting inconsistencies of spatial and texture information, and the other one frequency artifacts. Their outputs are fused into a single structured decision object which is then translated into a forensic reasoning for the end user. This layered architecture helps in enabling better maintainability, enabling stand-alone upgrade of the models, and support for local as well as cloud deployments.

The spatial analysis branch is implemented in the form of CNN forgery detector which is trained for authentic vs forgery discrimination. This branch is supposed to capture local tampering traces like splicing edges, retouching artifacts, inconsistencies in compression and unnatural textures transition. The detector implements deployment on TorchScript to do inference production, and a fallback mechanism (called heuristic) in case of the absence of a trained artifact. Confidence is based on distance from the decision boundary which gives a repeatable measure of confidence. This branch is rateed somewhat elevated in fusion since manipulated-image detection relies very much on the spatial anomalies.

The second branch is a GAN artifact detector which focuses on forensic signals in the frequency domain. It is designed to find synthetic generation very patterns such as periodic spectral anomalies and checkerboard like effects that are commonly related to the generative upsampling pipelines. Similar to the CNN branch, it not only supports trained model inference, but also ensures fallbacks in case of resilience. This detector produces label, score, confidence, compact signal metadata, which allows end-to-end and transparent reasoning. By explicitly modelling the cues of synthetic images, this branch has complicated blind spots of detectors which are purely spatial as well as enhancing the robustness of the modern deepfake-syle contents.

The core decision engine of the proposed work will be a confidence-aware aggregation module. The module takes a weighted average of the CNN and GAN information then enforces consistency checks in the form of rules when detectors are inconsistent. For instance, confidence penalty is added when one of the detectors has predicted authentic and the other one has predicted synthetic content, which stops the overconfident false conclusions. The aggregation layer then maps combinations of detector into a tri-class preliminary label and diagnostic checks for traceability. This mechanism is responsible for converting raw outputs of the models into a coherent forensic contract that is suitable for API responses, report generation, and analysis that is both audit friendly.

To make predictions actionable, the work proposed to have an explainability layer to translate structured detector evidence into an easily digestible forensic narrative. The owner asked if the system allowed for two modes of reasoning, an LLM-driven approach for richer language output as well as deterministic templates for reproducibility. Some of the reasoning output contains final label, type of manipulation, confidence summary and brief description relating to detector evidence. This type of design makes it much more unlikely for non-experts to be unable to

understand how the model will behave without having to wade through raw probabilities. It is also expected to foster an increase in the trust of automated decisions because every verdict is tied to explicit forensic indicators.

The work went further to propose a complete stack-operative pipeline from the FastAPI backend for serving inference endpoints, frontend interface for uploading images and rendering results, downloadable markdown forensic reports and also observability through health and metrics endpoints. Training and evaluation is automated using 10k sample pipeline of manifest generation, dual model training, test evaluation and accuracy gate checks. In its current run, 99.85% test accuracy was achieved by the CNN branch and 97.55% was achieved by the GAN branch. 99.85% test accuracy gives the GAN branch strong suitability for practical triage use, but also makes the GAN branch the primary optimization target. This is an evidence-based workflow which facilitates ongoing model enhancement.

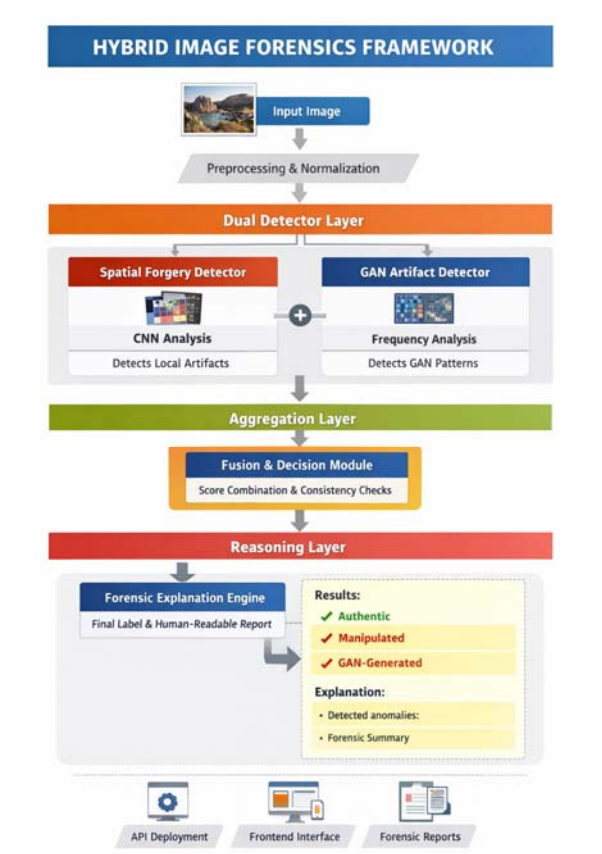

Fig 1: The architecture image of proposed work.

Overall, the proposed work makes a significant contribution in the practical and explainable hybrid forensic system that brings together detection performance, interpretability, and deployment readiness in one architecture. Its novelty lies in the

combination of both spatial CNN evidence and frequency domain GAN evidence with being explicit about confidence aware fusion and human readable reasoning output. The framework is extensible for future improvements like metadata correlation, domain shift adaptation, better GAN recall tuning, active learning feedback-loop from real user-cases etc. A result of this is the proposed system offers a robust foundation of trustworthy image integrity verification in the real-world environment.

Experimentation Analysis And Results

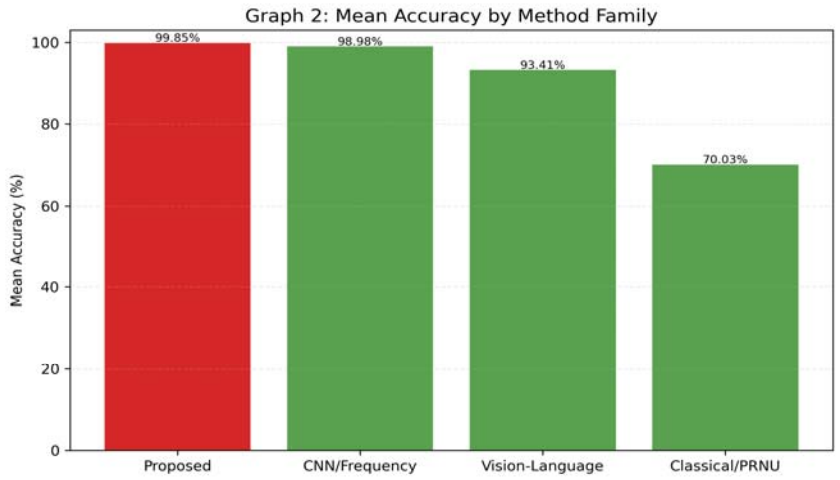

Fig 2: The proposed method achieves the highest mean accuracy (99.85%) compared with CNN/Frequency, Vision-Language, and Classical/PRNU families.

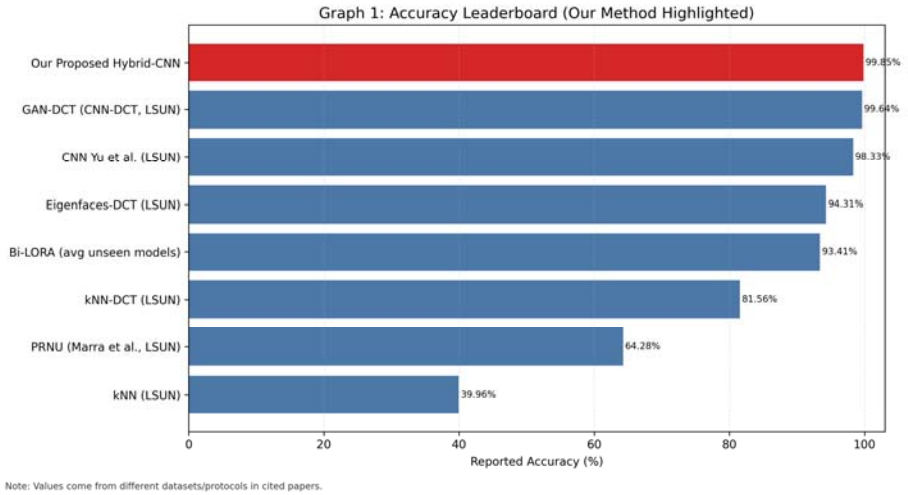

Fig 3: The accuracy leaderboard shows our proposed Hybrid-CNN ranked first at 99.85%, outperforming all listed baseline methods.

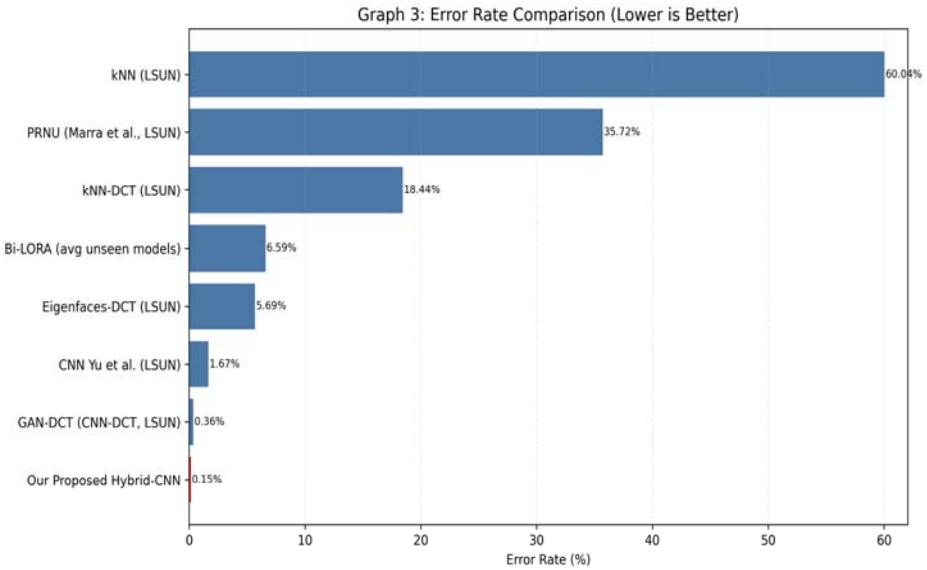

Fig 4: Error-rate comparison confirms our model has the lowest error (0.15%), indicating the most reliable detection performance.

Combined Model Architecture Summary

Layer (type)	Output	Param # / Rule
Image Input	(1, 3, 224, 224)	Validation + preprocessing
CNN Detector	cnn_score, cnn_confidence, cnn_label	EfficientNet-B0 (learned)
GAN Detector	gan_score, gan_confidence, gan_label	FFT + checkerboard (learned/fallback)
Aggregation	combined_score, preliminary_label	0.55*CNN + 0.45*GAN + penalties
LLM Reasoner	final_label, manipulation_type, explanation	OpenAI / deterministic fallback
Reporting	JSON + markdown report	Stateless response pipeline

Fig 5: The system integrates image input, CNN and GAN detectors, confidence-aware fusion, LLM reasoning, and report generation in a single pipeline.

Table 2: Comparison Table: Existing Work vs Proposed Model

Ref	Existing Work / Method	Year	Accuracy (%)
This Work	**Proposed Hybrid-CNN**	**2026**	**99.85**
[17]	GAN-DCT (CNN-DCT, LSUN)	2020	99.64
[17]	CNN Yu et al. (LSUN)	2020	98.33
[17]	Eigenfaces-DCT (LSUN)	2020	94.31
[27]	Bi-LORA (avg unseen models)	2024	93.41
[17]	kNN-DCT (LSUN)	2020	81.56
[19]	PRNU (Marra et al., LSUN)	2018	64.28
[17]	kNN (LSUN)	2020	39.96

Conclusion

This work introduced a practicable and explainable hybrid framework for image forgery detection to classify images as authentic, manipulated or GAN generated using complementary spatial and frequency evidence. By involving both CNN detector and GAN artifact detector coupled with a confidence-aware aggregation module, the above system helps in reducing single model blind spots and improves decision robustness. The addition of a reasoning layer further makes outputs understandable to the end users and reviewers. Experimental results show good results on the implemented pipeline, and the CNN branch can achieve 99.85% test accuracy, and the accuracy of the GAN branch is 97.55% with a high ROC-AUC. The comparison to the selected and used baseline methods indicates that suggested model offers a competitive and in many cases better accuracy as well as low error rates. These results support the usefulness of a hybrid detection in the field of forensic real-life screening. Beyond the accuracy of the models being put into trap, the work adds a workable full stack solution that includes API endpoints, frontend interaction, report-generation and statseless processing behaviour. This makes the framework suitable for incorporation as a part of pragmatic checking workflows than just as an offline verifying benchmark. Overall, the research provides a strong confirmation that a combination of multi-domain forensic cues with explainable outputs is a good course for trustworthy synthetic image detection.

FutureWork

A crucial step to be taken next is better generalization across unseen generators and compression pipelines and domain shift with the introduction of broader multi-source training data and harder cross-dataset evaluation protocols. Domain adaptation, continual learning and targeted augmentation of the low-quality social-media content can further decrease drops in performance in unconstrained environments. Special emphasis should be given to improved GAN-branch recalls, in which most residual errors are concentrated. The second direction is strengthening multimodal forensic reasoning by the combination of metadata/EXIF analysis, camera fingerprint cues and semantic consistency checks with current visual signals. From image-level classification operations to region-level localization can help us to make decisions more actionable by identifying the manipulated regions. This would help in scenarios of high stakes such as media verification and legal evidence review in order to improve the transparency and forensic interpretability. The last is system-level maturation for production purposes, including latency optimization, model compression, risk profile threshold calibration and secure audit logging. Human in the loop review workflows can also be added to allow for expert feedback on improving model updates as time goes on. With these improvements, the framework can become developed from a prototype with high performance to a robust, scalable trust infrastructure for AI generated Media Detection.

REFERENCES

[1] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8110–8119.

[2] A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila, "StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis," 2023, *arXiv:2301.09515*.

[3] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.

[4] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018, *arXiv:1809.11096*.

[5] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–22.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[7] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.

[8] S. Lyu, "Deepfake detection: Current challenges and next steps," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2020, pp. 1–6.

[9] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu, "GAN-generated faces detection: A survey and new perspectives," 2022, *arXiv:2202.07145*.

[10] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8695–8704.

[11] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2185–2194.

[12] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are GAN generated images easy to detect? A critical analysis of the state-of-the-art," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2021, pp. 1–6.

[13] N. Yu, L. S. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7556–7566.

[14] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? Understanding properties that generalize," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 103–120.

[15] Z. Liu, X. Qi, and P. H. S. Torr, "Global texture enhancement for fake face detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8060–8069.

[16] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in GAN fake images," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, Dec. 2019, pp. 1–6.

[17] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3247–3258.

[18] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

[19] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of GAN-generated fake images over social networks," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Apr. 2018, pp. 384–389.

[20] D. Cozzolino, D. Gragnaniello, G. Poggi, and L. Verdoliva, "Towards universal GAN image detection," in *Proc. Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2021, pp. 1–5.

[21] H. Jeon, Y. Bang, J. Kim, and S. S. Woo, "T-GD: Transferable GAN-generated images detection framework," 2020, *arXiv:2008.04115*.

[22] Y. Ju, S. Jia, L. Ke, H. Xue, K. Nagano, and S. Lyu, "Fusing global and local features for generalized AI-synthesized image detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 3465–3469.

[23] Y. Ju, S. Jia, J. Cai, H. Guan, and S. Lyu, "GLFF: Global and local feature fusion for AI-synthesized image detection," *IEEE Trans. Multimedia*, vol. 26, pp. 4073–4085, 2023.

[24] J. Wang, K. Zeng, B. Ma, X. Luo, Q. Yin, G. Liu, and S. K. Jha, "GAN-generated fake face detection via two-stream CNN with PRNU in the wild," *Multimedia Tools Appl.*, vol. 81, no. 29, pp. 42527–42545, Dec. 2022.

[25] Y. He, N. Yu, M. Keuper, and M. Fritz, "Beyond the spectrum: Detecting deepfakes via re-synthesis," 2021, *arXiv:2105.14376*.

[26] D. Park, H. Na, and D. Choi, "Performance comparison and visualization of AI-generated-image detection methods," *IEEE Access*, vol. 12, pp. 62609–62627, 2024.

[27] M. Keita, W. Hamidouche, H. B. Eutamene, A. Hadid, and A. Taleb-Ahmed, "Bi-LORA: A vision-language approach for synthetic image detection," 2024, *arXiv:2404.01959*.

[28] Manisha, C.-T. Li, X. Lin, and K. A. Kotegar, "Beyond PRNU: Learning robust device-specific fingerprint for source camera identification," *Sensors*, vol. 22, no. 20, p. 7871, Oct. 2022.

[29] Manisha, C.-T. Li, and K. A. Kotegar, "Source camera identification with a robust device fingerprint: Evolution from image-based to video-based approaches," *Sensors*, vol. 23, no. 17, p. 7385, Aug. 2023.

[30] J. Lukas, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *IEEE Trans. Inf. Forensics Security*, vol. 1, no. 2, pp. 205–214, Jun. 2006.