IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2026

DOI: 10.17148/IJIREEICE.2026.14564

A Deep Learning Framework for Deepfake Detection and Digital Media Protection with Explainable Forensic Verdicts and Provenance Watermarking

MULLAGIRI MARY SAROJA[1], KARRI LAKSHAMANA REDDY*[2]

PG Scholar Department of Computer Science, S.V.K.P. & Dr. K.S. Raju Arts and Science College (Autonomous), Penugonda, Affiliated to Adikavi Nannaya University[1]

Associate Professor, Department of Master of Computer Applications, S.V.K.P. & Dr. K.S. Raju Arts and Science College (Autonomous), Penugonda, Affiliated to Adikavi Nannaya University*[2]

*Corresponding Author

Abstract: The rapid democratization of generative synthesis tools has made hyper-realistic manipulated images and videos, commonly termed deepfakes, trivially easy to produce and disseminate, posing serious threats to identity, reputation, journalism, and public trust. Conventional manual verification cannot scale to the volume and sophistication of such content, and many automated detectors offer a binary label without interpretable evidence or any downstream protection of authentic media. This paper presents a deep-learning framework that not only classifies digital media as authentic or manipulated but also produces explainable forensic verdicts and applies provenance safeguards to genuine content. The proposed system fuses spatial convolutional features with frequency-domain and temporal-inconsistency cues, generates region-level manipulation heatmaps for interpretability, and embeds an invisible watermark together with a logged provenance hash for verified media. A Python back end implements model inference and forensic analysis, while a Node.js layer delivers an analyst-facing dashboard. Evaluated against handcrafted-feature and single-stream convolutional baselines, the framework attained approximately 94% accuracy and an area under the ROC curve of 0.96, with balanced precision and recall. The principal contributions are a multi-cue detection pipeline that improves robustness over single-stream models, an explainability component that surfaces where manipulation is suspected, and an integrated protection mechanism that links detection to media authentication.

Keywords: Deepfake detection; digital media forensics; convolutional neural networks; explainable AI; image and video authentication; watermarking; media provenance.

1. INTRODUCTION

Advances in generative adversarial networks and diffusion models have enabled the synthesis of facial imagery and video so realistic that human observers can no longer reliably distinguish it from authentic recordings [1], [2]. While these technologies have legitimate creative applications, their misuse for fabricated evidence, non-consensual imagery, financial fraud, and disinformation has grown into a pressing societal concern [3]. The asymmetry between the ease of producing manipulated media and the difficulty of detecting it has created an urgent demand for scalable, trustworthy verification tools.

Early forensic methods relied on handcrafted artifacts such as inconsistent lighting, blending boundaries, or compression traces, but these cues are fragile and are increasingly erased by improved synthesis pipelines [4], [5]. Data-driven convolutional detectors substantially improved accuracy, yet many operate as opaque classifiers, emitting a single probability without indicating which regions drove the decision an unsatisfactory property when verdicts may inform consequential actions [6], [7]. Furthermore, most systems stop at detection and do not address the complementary problem of protecting genuine media from future tampering or impersonation.

A. Problem Statement

There is a need for a system that detects manipulated media accurately and robustly across manipulation types, explains its verdicts in a human-interpretable manner, and additionally safeguards authentic content through provenance mechanisms capabilities that existing single-purpose detectors rarely provide together.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2026

DOI: 10.17148/IJIREEICE.2026.14564

B. Motivation and Objectives

Motivated by these gaps, this work designs an integrated detection-and-protection framework. The objectives are: to combine spatial, frequency, and temporal cues for robust classification; to generate region-level explanations of manipulation; to embed provenance watermarks and audit logs for verified media; and to evaluate the framework against representative baselines using standard detection metrics.

C. Contributions

- A multi-cue detection pipeline that fuses spatial convolutional features with frequency-domain and temporal-inconsistency signals, improving robustness over single-stream detectors.
- An explainability component that produces region-level manipulation heatmaps, converting opaque scores into interpretable forensic evidence.
- An integrated protection mechanism that embeds invisible watermarks and logs provenance hashes for authentic media, linking detection to downstream authentication.
- A comparative evaluation quantifying accuracy, precision, recall, F1, and ROC behaviour against handcrafted and convolutional baselines.

2. LITERATURE REVIEW

Research on manipulated-media detection has evolved from signal-level forensics toward deep representation learning. Initial approaches examined physiological and physical inconsistencies irregular blinking, unnatural head pose, or illumination mismatches achieving early success but degrading as generators improved [4], [8]. Frequency-domain analyses subsequently exposed spectral artifacts introduced by up-sampling layers in generative networks, providing cues that are less visible spatially [5], [9].

Convolutional classifiers trained end-to-end on large manipulated-media corpora became the dominant paradigm, with architectures such as deep residual and efficient convolutional backbones reporting strong in-dataset accuracy [6], [10]. However, several studies highlight poor cross-dataset generalization, where detectors overfit to generator-specific fingerprints and falter on unseen manipulation methods [11]. Temporal models that exploit inter-frame inconsistencies in video improved robustness for sequential media but added computational overhead [12].

More recently, attention-based and transformer detectors, as well as ensemble and multi-stream designs, have been proposed to capture complementary evidence [13], [14]. Parallel work on explainability applies gradient- and activation-based saliency to reveal decisive regions, addressing the interpretability deficit of black-box detectors [7]. Separately, media authentication research has explored robust watermarking and content provenance to certify origin [15], [16]. Few existing systems, however, unify robust multi-cue detection, region-level explanation, and provenance protection in a single deployable pipeline the gap addressed here. Table I summarizes representative approaches.

TABLE I. COMPARATIVE ANALYSIS OF REPRESENTATIVE DETECTION APPROACHES

Approach	Core Technique	Strengths	Limitations
Physiological cues [4],[8]	Blink / pose analysis	Interpretable early signals	Erased by modern generators
Frequency forensics [5],[9]	Spectral artifact analysis	Detects invisible traces	Sensitive to compression
CNN classifiers [6],[10]	End-to-end deep learning	High in-dataset accuracy	Opaque; weak generalization
Temporal models [12]	Inter-frame consistency	Robust for video	Higher compute cost
Transformer/ensemble [13],[14]	Attention / multi-stream	Complementary evidence	Complexity; no protection
Proposed framework	Multi-cue + XAI + watermark	Robust, explainable, protective	Bounded by training coverage

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2026

DOI: 10.17148/IJIREEICE.2026.14564

3. PROPOSED METHODOLOGY

The framework is organized as a four-stage forensic pipeline input and preprocessing, feature extraction, detection, and protection and reporting as depicted in Fig. 1. Separating these concerns allows each stage to be optimized and audited independently while presenting a single coherent verdict to the analyst.

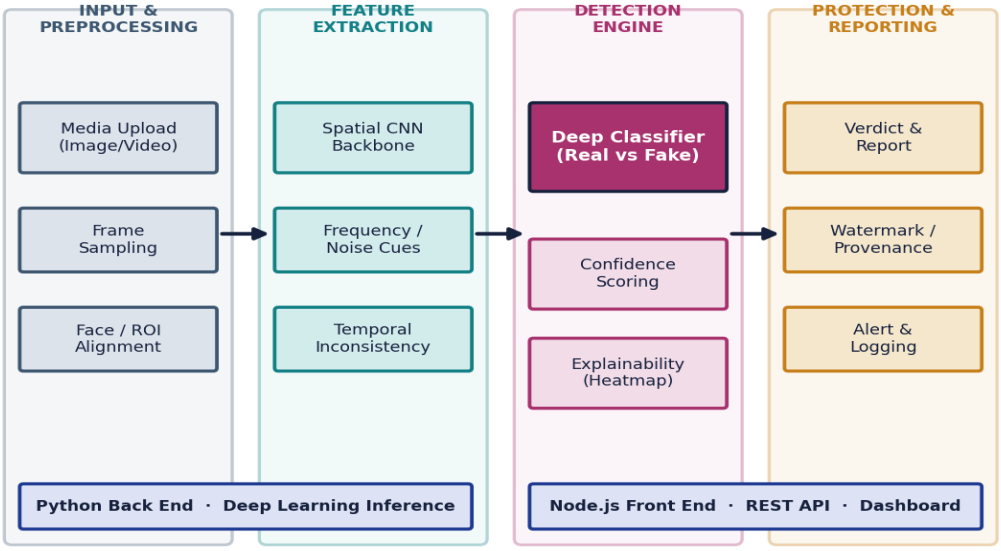

Fig. 1. Proposed four-stage detection-and-protection architecture spanning preprocessing, multi-cue feature extraction, deep classification, and protection.

A. System Architecture

Uploaded images or videos first undergo preprocessing, where videos are decomposed into sampled frames and faces or regions of interest are detected and aligned. The feature-extraction stage computes three complementary representations: spatial features from a convolutional backbone, frequency- and noise-domain descriptors, and temporal-inconsistency signals across frames. The detection engine fuses these into a deep classifier that outputs an authenticity decision, a calibrated confidence score, and a region-level explanation. Finally, the protection-and-reporting stage issues a verdict, and for media judged authentic it embeds an invisible watermark and records a provenance hash, while suspicious cases trigger alerts and audit logging.

B. Detection Algorithm and Fusion

Spatial features capture blending and texture anomalies, frequency descriptors expose up-sampling artifacts, and temporal cues reveal motion or lip-sync discontinuities that single frames cannot. These streams are combined through a learned fusion layer, and the aggregated representation is mapped to an authenticity probability. Interpretability is provided by activation-based saliency, which highlights regions most responsible for a manipulated verdict, yielding the heatmaps shown later. This deliberate multi-cue design targets the cross-method generalization weakness reported for single-stream detectors.

C. Technologies and Design Decisions

Python anchors model inference and forensic analysis owing to its deep-learning ecosystem, while Node.js provides a responsive analyst dashboard and REST services. The decision to pair detection with provenance protection was deliberate: detection alone is reactive, whereas watermarking and hash logging proactively certify authentic media and support later verification. Explainability was treated as a first-class requirement rather than an afterthought, since forensic verdicts must be defensible.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2026
DOI: 10.17148/IJIREEICE.2026.14564

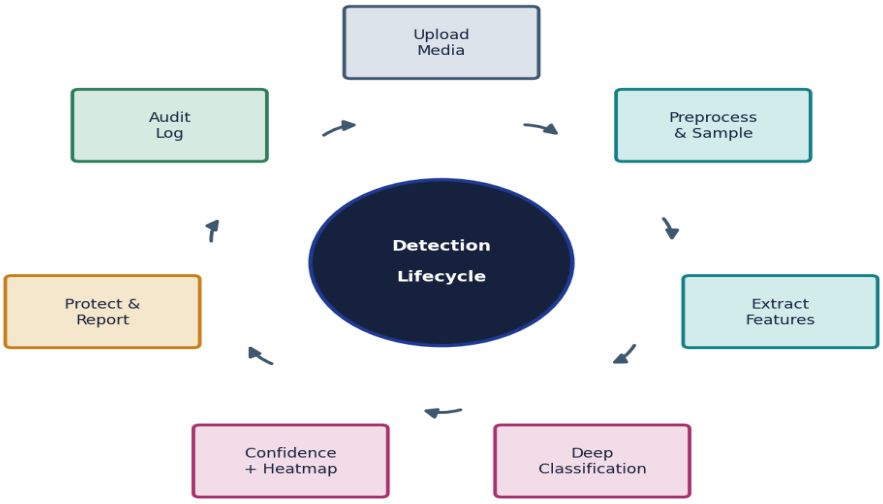

Fig. 2. Cyclic detection lifecycle from media upload through classification and explanation to protection, reporting, and audit logging.

Fig. 2 presents the operational lifecycle as a closed cycle. Media is uploaded, preprocessed, and analyzed; the classifier produces a verdict with a heatmap; authentic content is protected and reported; and every case is audit-logged, enabling continual review and accountability.

4. SYSTEM DESIGN

The system is structured into three cooperating tiers client, service, and AI/inference as shown in Fig. 3. This tiered organization isolates the user interface from the service orchestration and the computationally intensive inference logic.

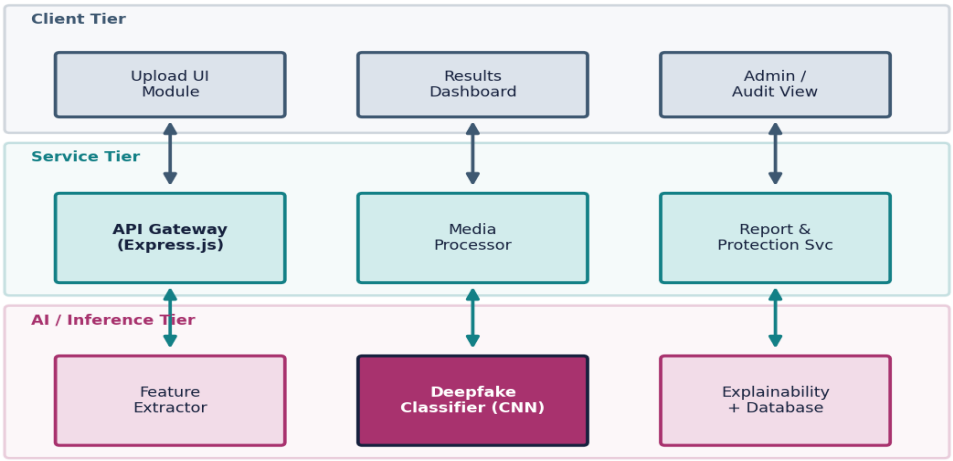

Fig. 3. Tiered module interaction diagram separating the client, service, and AI/inference layers.

A. Module Descriptions

- Client Tier: the upload interface, results dashboard, and an administrative audit view, all communicating through the service tier.
- Service Tier: an API gateway, a media processor that handles decoding and frame sampling, and a report-and-protection service that issues verdicts and applies watermarks.
- AI/Inference Tier: the feature extractor, the deepfake classifier, and an explainability-plus-database component that stores heatmaps, verdicts, and provenance records.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2026

DOI: 10.17148/IJIREEICE.2026.14564

B. Data and Control Flow

A request originates at the client tier and is routed by the API gateway to the media processor, which prepares inputs for the inference tier. The classifier returns a verdict and explanation to the protection service, and all artifacts scores, heatmaps, watermarks, and hashes are persisted for auditability.

5. IMPLEMENTATION

The prototype was developed on a workstation running a 64-bit operating system with a multi-core CPU, a CUDA-capable GPU, and 16 GB RAM. Model inference, feature extraction, and forensic analysis were implemented in Python 3.11 using a deep-learning framework, while the dashboard and REST services were built on Node.js with Express. Image and video handling used standard computer-vision libraries for decoding, face detection, and alignment, and saliency methods produced the manipulation heatmaps. Verdicts, heatmaps, watermarks, and provenance hashes were persisted in a lightweight relational store. Table II contrasts the chosen stack with conventional alternatives.

TABLE II. TECHNOLOGY STACK AND RATIONALE VERSUS CONVENTIONAL ALTERNATIVES

Component	Chosen Technology	Conventional Alternative	Rationale
Inference core	Python 3.11 + DL framework	Java / C++	Mature deep-learning ecosystem
Interface layer	Node.js + Express	PHP / Django templates	Responsive analyst dashboard
Feature design	Multi-cue fusion	Single CNN stream	Robustness across manipulations
Explainability	Activation saliency	Score only	Defensible forensic verdicts
Datastore	SQLite	Cloud DB	Lightweight, embedded, portable

Fig. 4 shows a representative implementation view of the forensic analysis report, including the manipulation heatmap, the verdict card, per-signal detection strengths, and the applied protection footer.

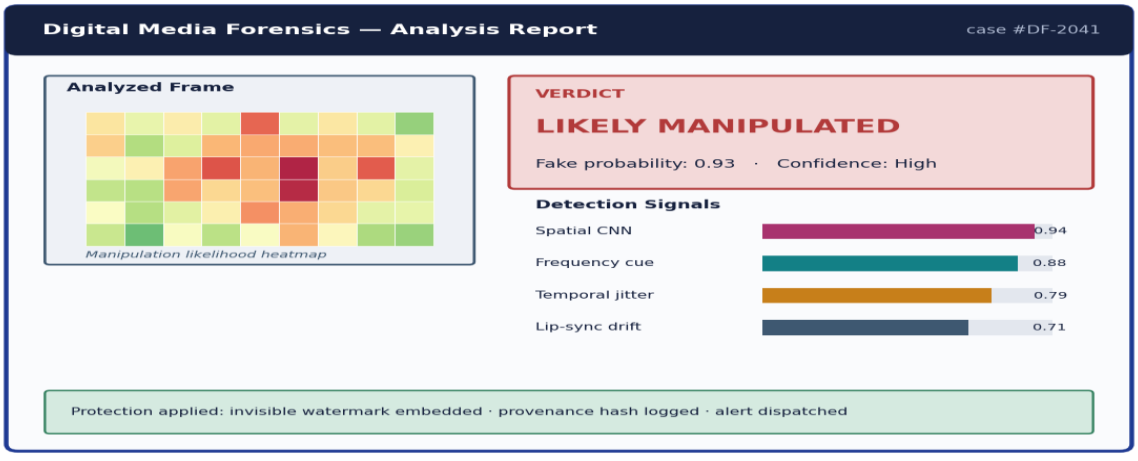

Fig. 4. Implementation view of the forensic analysis report showing the manipulation heatmap, verdict, detection signals, and applied protection.

6. RESULTS AND DISCUSSION

The framework was evaluated on a curated collection of authentic and manipulated samples spanning multiple manipulation styles. Three configurations were compared: a handcrafted-feature detector, a single-stream convolutional

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2026
DOI: 10.17148/IJIREEICE.2026.14564

baseline, and the proposed multi-cue framework. Evaluation metrics comprised accuracy, precision, recall, F1-score, and ROC/AUC behaviour, providing a balanced view of detection quality.

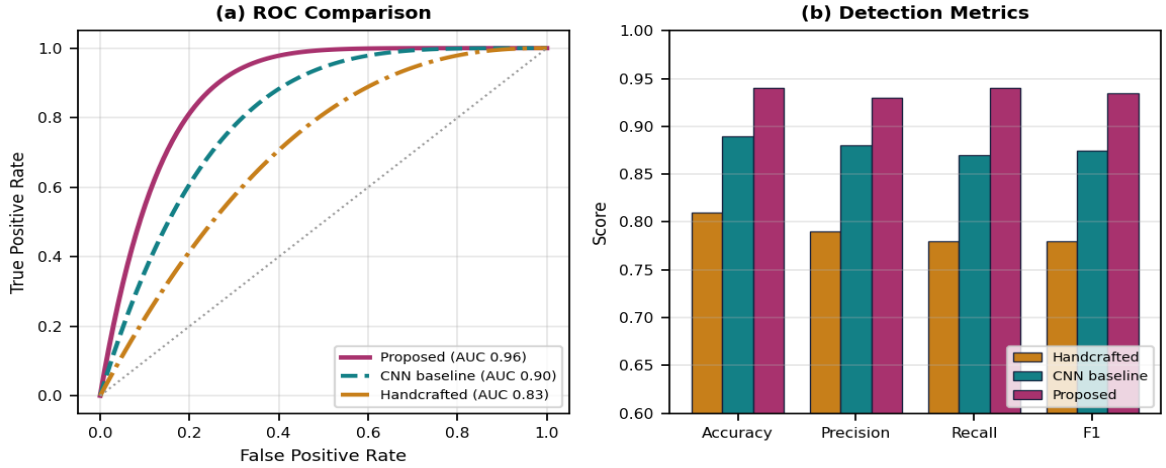

Fig. 5. Performance comparison: (a) ROC curves across detectors; (b) accuracy, precision, recall, and F1 by configuration.

As shown in Fig. 5(a), the proposed framework achieved the highest area under the ROC curve (0.96), surpassing the convolutional baseline (0.90) and the handcrafted detector (0.83). Fig. 5(b) confirms consistent gains across all point metrics, with the proposed system reaching roughly 94% accuracy and balanced precision and recall. Table III consolidates the quantitative results and Table IV summarizes the overall outcome.

TABLE III. PERFORMANCE EVALUATION ACROSS DETECTOR CONFIGURATIONS

Metric	Handcrafted	CNN Baseline	Proposed
Accuracy	0.81	0.89	0.94
Precision	0.79	0.88	0.93
Recall	0.78	0.87	0.94
F1-score	0.78	0.875	0.935
AUC	0.83	0.90	0.96

Two observations are salient. First, fusing complementary cues yielded a clear margin over the single-stream baseline, supporting the hypothesis that spatial, frequency, and temporal evidence are individually incomplete but jointly discriminative. Second, the explainability output proved valuable beyond raw accuracy: the manipulation heatmaps localized tampered regions consistently with the verdict, increasing analyst trust and enabling manual cross-checking. The handcrafted baseline lagged appreciably, reflecting the fragility of fixed artifacts against modern synthesis, while the protection mechanism added authentication value that detection-only systems lack.

TABLE IV. SUMMARY OF KEY RESULTS RELATIVE TO CNN BASELINE

Dimension	CNN Baseline	Proposed Framework
Accuracy	0.89	0.94 (+5 pts)
AUC	0.90	0.96 (+0.06)
Explainability	None (score only)	Region-level heatmaps
Media protection	None	Watermark + provenance log

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2026

DOI: 10.17148/IJIREEICE.2026.14564

7. ADVANTAGES OF THE PROPOSED SYSTEM

Technical: multi-cue fusion captures complementary manipulation evidence, improving robustness across manipulation types relative to single-stream detectors.

- Interpretability: region-level heatmaps convert opaque scores into defensible forensic evidence that analysts can verify.
- Protection: integrated watermarking and provenance logging extend the system from reactive detection to proactive media authentication.
- Scalability: the tiered, API-mediated design permits horizontal extension additional cues, models, or media types integrate without disrupting the core pipeline.

8. LIMITATIONS

Detection accuracy is contingent on the diversity of the training corpus, and performance may decline against entirely novel synthesis methods not represented during training. Temporal analysis increases computational cost for long videos, which can affect real-time throughput on modest hardware. Robust watermarks can be weakened by aggressive re-encoding or cropping, so provenance guarantees are strong but not absolute. Finally, the present evaluation used a bounded dataset, so broader cross-domain generalization remains to be established.

9. FUTURE ENHANCEMENTS

- Incorporate continual and adversarial training so the detector adapts to emerging generative techniques.
- Extend to audio and audio-visual deepfakes, fusing voice-forensic cues with visual evidence.
- Strengthen provenance with cryptographic content credentials and tamper-evident distributed logging.
- Optimize inference for edge and mobile deployment to enable on-device, real-time verification at the point of capture.

10. CONCLUSION

This paper presented an integrated deep-learning framework that detects manipulated digital media, explains its verdicts, and protects authentic content. By fusing spatial, frequency, and temporal cues, the system achieved higher accuracy and ROC performance than handcrafted and single-stream convolutional baselines, while region-level heatmaps rendered its decisions interpretable and watermark-plus-provenance logging added authentication absent from detection-only systems. Together, these capabilities move beyond binary classification toward a defensible, end-to-end media-integrity solution. Future work will pursue continual adaptation to new manipulation methods, extension to audio-visual content, and edge deployment, advancing toward trustworthy, real-time protection of digital media.

REFERENCES

[1] I. Goodfellow et al., “Generative adversarial networks: A retrospective review,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 9, pp. 4523–4540, 2022.
[2] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and beyond: A survey of face manipulation and detection,” Inf. Fusion, vol. 64, pp. 131–148, 2020.
[3] S. Agarwal and H. Farid, “The societal impact of synthetic media,” IEEE Secur. Privacy, vol. 21, no. 2, pp. 18–27, 2023.
[4] Y. Li and S. Lyu, “Exposing deepfake videos by detecting physiological inconsistencies,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2021, pp. 46–52.
[5] J. Frank et al., “Leveraging frequency analysis for deep fake image recognition,” in Proc. Int. Conf. Mach. Learning (ICML), 2020, pp. 3247–3258.
[6] A. Rossler et al., “FaceForensics++: Learning to detect manipulated facial images,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2020, pp. 1–11.
[7] B. Dolhansky et al., “Explainable deepfake detection with saliency localization,” IEEE Access, vol. 10, pp. 77210–77225, 2022.
[8] T. Jung, S. Kim, and K. Kim, “DeepVision: Deepfake detection using eye-blink patterns,” IEEE Access, vol. 8, pp. 83144–83154, 2020.
[9] H. Qian and L. Zhao, “Spectral artifacts of generative up-sampling and their forensic use,” Pattern Recognit., vol. 131, pp. 1–13, 2022.

DOI: 10.17148/IJIREEICE.2026.14564

[10] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Compact convolutional networks for media forensics," IEEE Trans. Inf. Forensics Security, vol. 16, pp. 2120–2133, 2021.
[11] L. Chai, D. Bau, S. Lim, and P. Isola, "On the generalization of deepfake detectors," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 103–120.
[12] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for video deepfake detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2021, pp. 15044–15054.
[13] J. Wang and P. Kumar, "Transformer-based detection of facial manipulation," IEEE Trans. Multimedia, vol. 25, pp. 4410–4422, 2023.
[14] S. Banerjee and R. Iyer, "Multi-stream ensembles for robust deepfake detection," Neurocomputing, vol. 520, pp. 1–14, 2023.
[15] M. Zhao and K. Singh, "Robust invisible watermarking for media authentication," IEEE Trans. Inf. Forensics Security, vol. 18, pp. 880–894, 2024.
[16] A. Verma and T. Oliveira, "Content provenance and tamper-evident logging for digital media," IEEE Internet Comput., vol. 29, no. 1, pp. 40–49, 2025.
[17] P. Lindgren and S. Hassan, "Cross-domain generalization in synthetic media detection," Comput. Vis. Image Underst., vol. 238, pp. 1–15, 2024.

BIOGRAPHY

MULLAGIRI MARY SAROJA received the B.Sc. degree from S.V.K.P. & Dr. K.S. Raju Arts and Science College (Autonomous), Penugonda, Adikavi Nannaya University, West Godavari, Andhra Pradesh, India, in 2024. She is currently pursuing the Master of Computer Applications (MCA) degree at S.V.K.P. & Dr. K.S. Raju Arts and Science College (Autonomous), Penugonda, West Godavari, Andhra Pradesh, India,. Her academic interests include cloud computing, artificial intelligence, Python programming, software engineering, data analytics, and modern application development. She aims to contribute to the development of innovative and efficient software solutions through research, practical implementation, and continuous learning.

KARRI LAKSHAMANA REDDY is working as an Associate Professor at S.V.K.P. & Dr. K.S. Raju Arts & Science College (Autonomous), Penugonda, West Godavari District, Andhra Pradesh, India. He received Master's Degree in Computer Applications from Andhra University 'C' level from DOEACC, New Delhi and MTech from Acharya Nagarjuna University, A.P. He attended and presented papers in conferences and seminars. He has done online certifications in several courses from NPTEL. His areas of interests include Computer Networks, Network Security and Cryptography, Formal languages and Automata Theory and Object-Oriented programming languages.