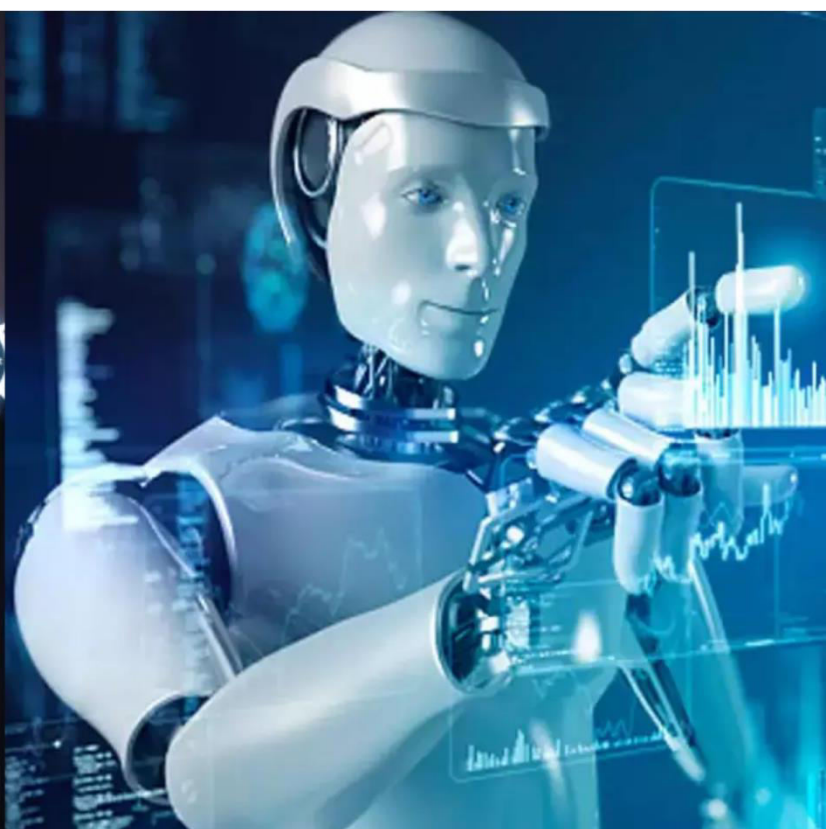




International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

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



Impact Factor: 9.935

Volume 15, Issue 7, July 2026

A Comprehensive Review of Deepfake Detection Techniques: From CNN-Based Models to Explainable Multimodal LLM Frameworks

Er. Shavnam¹, Er. Neha Dhiman²

Department of Computer Science Engineering, Himachal Pradesh Technical University, Hamirpur,
Himachal Pradesh, India¹

Department of Computer Science Engineering, Himachal Pradesh Technical University, Hamirpur,
Himachal Pradesh, India²

ABSTRACT: With the progress of generative adversarial networks (GANs), diffusion models, and large-scale deep learning models, deepfake technology has developed rather quickly and becomes a significant challenge to digital integrity, cybersecurity, journalism, and forensic investigations. While conventional CNN-based deepfake detectors demonstrated high accuracy in detecting visual artifacts; it only fulfills the lack of robustness as well as cross domain generalization and interpretability. The review provides an in-depth survey of fifteen state-of-art methodologies including classical CNN models, temporal-spatial video recognition, transformer-based networks, explainable AI (XAI) models, and models that combine multimodal large language model (LLM) products. The taxonomy classifies the current techniques in terms of representational features, learning models, interpretability procedures and deployment capability. Particular attention is paid to explainable and human-friendly AI systems including Grad-CAM visualizations, captioning systems, and narrative generation modules. Some of the important issues include adversarial robustness, bias in the data sets, computational complexity, cross-dataset generalization, and computational limitations in the real world. Lastly, the release of new research directions in terms of creating reliable, scalable, and understandable deepfake detectors is described.

KEYWORDS: Deepfake Detection, Explainable AI, Vision Transformers, Optical Flow, Multimodal AI, Large Language Models, Grad-CAM, Multimedia Forensics.

I. INTRODUCTION

Deepfakes are artificially produced or altered media that is produced with the help of a deep neural network, specifically GAN- based and transformer-driven generative models. However, as much as they have positive uses in the production of media, virtual communication and entertainment, they have been abused to lead to misinformation, identity frauds, manipulation of politics, and damages to reputations.

Deepfakes include the manipulation of images, video, audio or text-based content with the help of artificial intelligence. Whereas in images, deepfaking can produce artificial images or fake videos, typical manipulations include face swapping, lip-sync manipulation, face synthesis, reenactment, or even modifying face characteristics. Deepfakes in audio can be based on text-to-speech or voice-swapping algorithms, which generate artificial voices and make a person sound like saying what they have never stated. Text-based deepfakes involve artificial scripts, dialogues or messages written in the style of one person or another which is created by AI. In general, deepfakes can be used to produce very realistic and AI-assisted fake content in the video and image as well as audio and text spheres that is a potent technology in terms of both creative and malevolent use. Deepfake techniques can be broadly categorized into image and video-based, audio-based, and text-based manipulations, as illustrated in Figure 1.

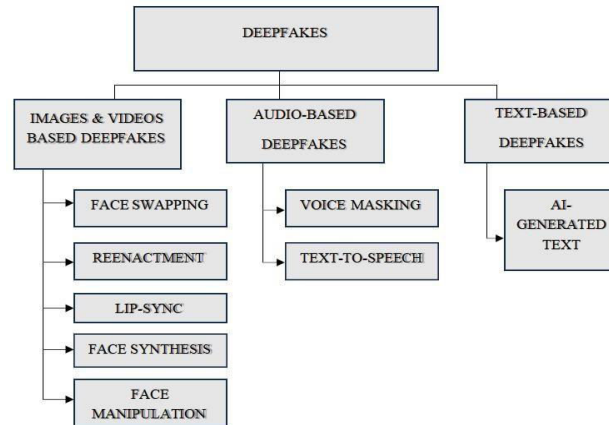


Figure.1. Classification of Deepfake

Early deepfake detection methods were mostly based on spatial artifact detection policies with convolutional neural networks (CNNs) [6,8,12]. Through extensive reviews of early detection methods, the prevailing trends revealed CNN-based models as the most successful ones in detecting texture anomalies and facial distortions [11]. Nevertheless, with the growing sophistication of deepfake generation techniques, purely spatial models were problematic in cross-dataset generalization as well as robustness [2,9].

To overcome these constraints, the efforts of researchers included spatio-temporal modeling techniques that examine motion irregularities, and frame-level consistency throughout the videos [1,3]. These strategies enhanced robustness by extracting temporal features over sense features. Later, transformer-based structures proposed global self-attention networks to improve contextual modeling and learning long-range dependencies [4,10].

Although the accuracy of detection has been increased, several advanced systems are run as black-box models, which restrict the trust a user has, as well as, forensic transparency. Explainable AI (XAI) frameworks were developed to promote interpretability and visualization of manipulated areas [5,13]. In addition, systematic quantitative assessment frameworks were suggested to test the explanation fidelity and robustness [15]. Multimodal models that allow the use of Visual AI to deliver explanations along with semantic reasoning and narrative generation have been added to the transparency of recent multimodal frameworks that incorporate Large Language Models (LLMs) [14].

This review systematically analyzes existing studies, and categorizing them into:

- Classical CNN-based detection
- Spatio-temporal video modelling
- Transformer-based architectures
- Explainable AI frameworks
- Multimodal LLM-based explainable systems

II. LITERATURE SURVEY

With the improvement of artificial intelligence and multimedia processing, a deepfake detection system has progressed and continues to do so. The current architectures are determined to detect insidious spatial, temporal, and semantic inconsistencies that arise when synthetic content is generated. The majority of detection frameworks have a systematic pipeline with elements of preprocessing, feature extraction, model learning, and classification. More recent methods incorporate explainable AI and multimodal learning solutions to increase robustness as well as explainability of the detection outputs.

Deepfake Detection Using CNN in a classical way

The methods of early deepfake detection were based mainly on Convolutional Neural Networks (CNNs) in detecting spatial artifacts in manipulated images and video frames. Hu et al. [6] suggested a detection and localization system that could deal with randomly placed tampered spots with CNN-based feature detection. On the same note, Kumar et al. [8] proposed a new CNN structure to enhance the accuracy of deepfake classification using the best spatial features learning.

In terms of application, Sahu et al. [12] created a deployable deepfake detector system with focus on optimization of performance and feasibility in real-time. Besides, a detailed overview of the fake image and video detection methods was also given by Ruby et al. [11] with CNN-based pipelines leading in early detection systems. Even though, CNN based models proved to be good in terms of spatial feature extraction, their results usually deteriorated when they were tested on unknown datasets or high-quality distortions.

Video Detection using Temporal-Spatial and Optical Flow

Researchers have included spatio-temporal modeling and artifact-based learning to increase strength. Chen et al. [1] came up with a spatio-temporal consistency and attention-based model to detect frame-level discrepancies in video deepfakes. On the same note, Dagar et al. [3] presented a dual branch network that takes advantage of noise and edge artifacts in shallowfake and deepfake localization to enhance detection accuracy.

Coccomini et al. [2] have studied the significance of using cross-dataset generalization and have found out that a lot of detection models do not retain their performance on different datasets. Further, Tariq et al. [9] provided the case study-based analysis of deepfake generation and detection discussing the practical difficulties in the real life.

These papers have highlighted that the fact of incorporating time is complemented with the ability of artifacts, which enhances detection effectiveness as opposed to the absolute spatial approach.

Transformer-Based Deepfake Detection

In the advent of Vision Transformers (ViTs), deepfake detector systems became more oriented towards global context modeling. Mahendrain et al. [10] enhanced Video Vision Transformer models by adding facial landmark driven depth wise separable convolution combined with augmented attention functions to ensure a better representation of the spatial-temporal features. Subsequent discoveries on explainability using transformers include Drews et al. [4], who presented a visual transformer model with interpretable AI methods to improve detection and explainability. Transformer-based systems can offer better modeling of long-range dependency at the cost of greater computing resources than traditional CNNs.

Deepfake Detection Explainable AI

With more advanced detection systems, interpretability was becoming a key necessity. Tsunkari and Srinagesh [13] also improved the transparency of deepfake detection with visual explanations in the form of Grad-CAM that are included into CNN models. Feng et al. have published a structured dataset to analyze the explainable deepfake detectors [5], which makes it easy to analyze the model explanations in a systematic manner. Moreover, Tsigos et al. [15] offered a quantitative framework to assess the XAI techniques based on perturbation-based metrics used to assess the fidelity and robustness of explanations. These papers represent the shift toward interpreting and reliable AI systems as the black-box models.

Explainable Deepfake Detection Multimodal and LLM-Based

Recently, the most recent development to deepfake detection is multimodal and human-based frameworks. To enhance the interpretability of the visualized saliency, Tariq et al. [14] suggested a prediction-to-explanation pipeline that incorporated saliency visualization, saliency-focused caption generation, semantic reasoning modules.

Mixed visual and textual reasoning systems can offer both multi modal explanations - one simple and one more complex - to increase belief and utilisation in forensic or media verification scenarios.

Table.1. Summary of major benchmark datasets used in deepfake detection research

Dataset Name	Media Type	Manipulation Type	Research Usage Focus
FaceForensics++	Video	Face swapping, reenactment, compression variants	Benchmark evaluation & spatial artifact detection
DFDC	Video	Real-world deepfake videos	Large-scale performance evaluation
Celeb-DF	Video	High-quality celebrity deepfakes	Robustness testing against realistic forgeries
Kaggle Deepfake Dataset	Image/Video	General deepfake samples	CNN-based classification experiments
UADFV	Video	Early face swap videos	Baseline and comparative analysis

Comparison of Deepfake Detection Techniques

The table below provides a comparative analysis of some of the deepfake detection methods in recent papers. It underlines the progress of approaches, data sets used and publicly shared performance indicators. The analogy points out the development of architecture, as well as increased focus on explainability and path-resistance in contemporary systems of detecting objects.

Table.2. Comparative summary of deepfake detection techniques and performance trends

Ref.	Year	Model	Dataset	Accuracy/AUC	Key Focus
[1]	2024	Deep CNN (Spatiotemporal Feature Fusion)	Benchmark Video Dataset	~95%	Performance enhancement using feature fusion
[2]	2022	CNN + Attention	FaceForensics++	90.1%	Spatio-temporal consistency modelling
[3]	2023	Deep Learning Models	Cross-Dataset Evaluation	AUC 71.2%	Generalization Study
[4]	2024	Dual-Branch CNN	Video Dataset	AUC 99.79%	Noise and edge-based localization
[5]	2024	CNN	Face Manipulation Dataset	AUC 99.1%	Detection and Localization
[6]	2025	Deep CNN	Video Dataset	~95%	Improved video classification
[7]	2025	CNN + Optical Flow	Video Dataset	95.45%	Motion-based detection (EVDO)
[8]	2025	CNN Variants	Case Study Dataset	N/A (Survey Paper)	Generation and detection challenge
[9]	2024	Video Vision	FaceForensics++	87.18%	Transformer-based

		Transformer			detection
[10]	2022	Survey (Various Models)	Multiple Datasets	N/A (Survey Paper)	Comprehensive Review
[11]	2025	CNN-Based Detection System	Kaggle Dataset	Not Reported	Real-time detection System
[12]	2024	CNN + Grad-CAM (DeepExplain)	Video Dataset	~97%	Explainable AI-based detection
[13]	2024	CNN + LLM (EDVD-LLaMA)	Multimodal video Dataset	77.15%	Explainable video detection
[14]	2025	CNN + LLM + Captioning	Multimodal Dataset	AUC 91.3%	Human-centred explainability
[15]	2024	EfficientNet + XAI	FaceForensics++	98%	Quantitative XAI evaluation

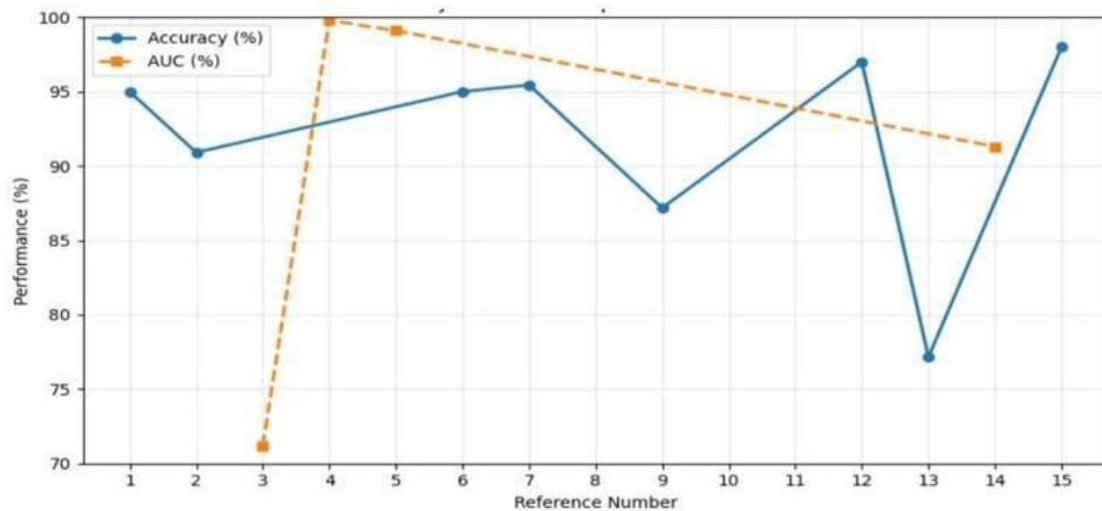


Figure.2. Performance Comparison of Deepfake Detection Models using Accuracy or AUC Metrics

As shown in Figure 2, the comparative analysis of the deepfake detection studies chosen basing on the reported Accuracy and AUC values. The current models show high performance around 90 per cent with some reaching up to 95 per cent meaning, they will be able to detect.

III. ARCHITECTURE OF DEEPPAKE DETECTION SYSTEM

The previous techniques were predominantly based on CNN-based spatial analysis [6], [8], [12], and the new ones combines temporal modeling, transformer models, and explainable AI infrastructure to enhance the strength and transparency [1,4,15].

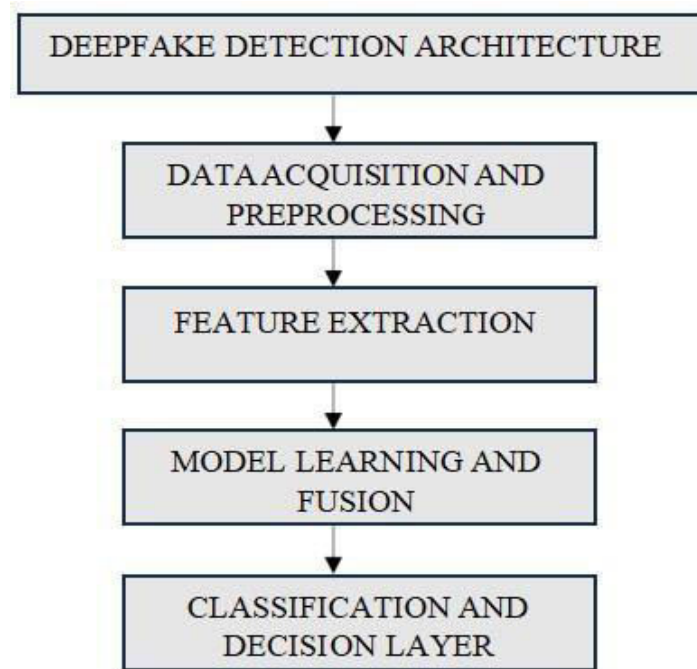


Figure.3. Architecture of deepfake detection system

Step 1: Data Acquisition and Preprocessing

Dataset Collection: FaceForensics++ and other curated video datasets are generally utilized as benchmark datasets to train and test [5,15].

Frame Extraction: Videos are sliced into frames so that they can be analysed in terms of time and motion [1,3].

Face Detection and Alignment: Face parts are localized and normalized to promote similar feature learning [6,12].

Normalization: This resizes images, boosts augmentation, and preprocessing to generalize models across datasets [2,9].

Step 2: Feature Extraction

Spatial Features: CNN models analyze different texture irregularities, noise artifacts, and edge-level differences in manipulated information [6,8].

Spatio-Temporal Features: Attention mechanisms and modeling of motion differences and changes in frames [1,3].

Transformer-Based Features: Vision Transformer networks learn global contextual and long-range dependencies with a self-attention mechanism [4,10].

Explainability Maps: Grad-CAM and saliency-based approaches indicate regions of manipulation in a face that are easier to understand [13,14].

Step 3: Model Learning and Fusion

CNN-Based Models: Binary classifiers are trained using extracted spatial features to differentiate real and fake content [8,12].

Attention Mechanisms: Focused learning strategies emphasize manipulation-sensitive regions to improve robustness [1,10].

Multimodal Fusion: Integration of spatial, temporal, and semantic embeddings enhances detection performance and interpretability [14,15].

Step 4: Classification and Decision Layer

Binary Classification: Most systems classify media as real or fake using fully connected layers followed by Softmax or Sigmoid activation [6,8].

Explainable Output: Advanced frameworks generate visual heatmaps and textual explanations to improve transparency and user trust [13,14].

IV. EVALUATION METRICS

The performance assessment is a critical activity of making comparisons between deepfake detection models. The selected articles employ the popular classification metrics to consider predictability and strength.

Accuracy: The overall proportion of instances correctly classified divided by all the instances tested is the accuracy. Accuracy is a general performance measure of the model.

Precision: Precision is the number of fake instances that are correctly identified divided by the number of instances which are predicted to be fake. The exactness comes into play when it is necessary to reduce false positives.

Recall: Recall is the ratio of correctly identified manipulated instances to all actual fake instances. High recall ensures fewer deepfakes are missed.

F1-score: F1-score is a balanced measure of precision and recall. F1-score is useful when working with imbalanced datasets where fake and real instances are not equally represented.

Specificity: Specificity is a measure of the model's ability to correctly identify real instances. Specificity is a measure of the system's ability to avoid false positives, that is, to avoid labeling real instances as fake.

Area Under the ROC Curve: AUC is a measure of the model's discriminative power to distinguish between real and fake instances at various thresholds. A higher AUC value indicates better discriminative power.

Loss Value: Loss is a measure of the error between predicted outputs and actual outputs during training. Lower loss values indicate better model convergence and learning performance.

V. SUMMARY AND OPINION

This review has covered the recent trends in deepfake detection and noticed a shift from CNN-based spatial models [3,7,12] to spatio-temporal and attention-based models [6,11], and then to transformer and optical flow-based models [5,8]. Although initial research focused on increasing the accuracy of classification, recent research has given importance to robustness and generalization. One of the major improvements in recent research is the incorporation of Explainable AI models [1,13], which are more transparent than the traditional black-box models [4]. In addition, multimodal and LLM-based models [2,14,15] are the recent trends, which combine visual detection with semantic reasoning. Deepfake detection research is currently heading towards the development of unified, explainable, and multimodal models. Future research should aim at robustness, scalability, and feasibility of real-world deployment, in addition to accuracy.

VI. CONCLUSION

The present review will offer an overview of the existing developments in the field of deepfake detection. It denotes the movement towards spatio-temporal and attention models [6,11] of CNN- based spatial models [3, 7, 12]. This is succeeded by transformer and optical flow-based models [5,8]. First detection methods of deepfakes relied mainly on classification performance. Nevertheless, the recent studies are concentrated on the further strengthening, generalization, and interpretability. The use of Explainable Artificial Intelligence strategies [1,13] and multimodal LLM-based systems [2,14,15] is one of the better steps towards the creation of easier to understand and interpret deepfake detection systems. Despite the huge improvements in the performance of deepfake detectors, certain improvements are still required. These include adversarial robustness, cross-dataset generalization, and real-time detection, which are yet to be explored. The future deepfake detection system should strive to develop single architectures that promote accuracy, interpretability, and scalability to assure reliable application in the real world.

REFERENCES

1. E. Almetekawy, M. Elhoseny, G. M. El-Banby, and A. E. Hassanien, "Deepfake detection: Enhancing performance with spatiotemporal texture and deep learning feature fusion," *Egyptian Informatics Journal*, vol. 28, pp. 1–15, 2024.
2. Y. Chen, N. Akhtar, A. Mian, and N. A. H. Haldar, "Deepfake detection with spatio-temporal consistency and attention," in *Proc. DICTA, IEEE*, pp. 1–8, 2022.
3. D. A. Coccomini, R. Caldelli, F. Falchi, and C. Gennaro, "On the generalization of deep learning models in video deepfake detection," *Journal of Imaging*, vol. 9, no. 5, p. 89, 2023.
4. D. Dagar and D. K. Vishwakarma, "A noise and edge extraction-based dual-branch method for shallowfake and deepfake localization," 2024.
5. J. Hu, X. Liao, D. Gao, S. Tsutsui, Q. Wang, Z. Qin, and M. Z. Shou, "Delocate: Detection and localization for deepfake videos with randomly-located tampered traces," in *Proc. IJCAI-24*, pp. 5862–5870, 2024.
6. M. Kumar, P. Rai, and P. Kumar, "A novel approach for deepfake video detection," in *Proc. ICDT 2025, IEEE*, pp. 218–230, 2025.
7. M. Melouk, M. Shao, A. Basit, and M. Shafique, "EVDO: An enhanced framework for deepfake detection in videos through optical flow and temporal-spatial analysis," *Journal of Visual Communication and Image Representation*, vol. 95, p. 103902, 2025.
8. Y. Patel, S. Tanwar, R. Gupta, P. Bhattacharya, I. E. Davidson, R. Nyameko, S. Aluvala, and V. Vimal, "Deepfake generation and detection: Case study and challenges," *IEEE Access*, vol. 11, pp. 143296–143324, 2023.
9. K. N. Ramadhani, R. Munir, and N. P. Utama, "Improving video vision transformer for deepfake video detection using facial landmark, depthwise separable convolution and self-attention," *IEEE Access*, vol. 12, pp. 8932–8945, 2024.
10. C. Ruby, R. Popli, and I. Kansal, "A comprehensive review on fake images/videos detection techniques," in *Proc. ICRITO*, pp. 1–7, 2022.
11. C. Sahu, "Deepfake detection system," *International Journal of Engineering Research & Technology (IJERT)*, vol. 14, no. 5, 2025.
12. V. Sunkari and A. Srinagesh, "DeepExplain: Enhancing deepfake detection through transparent and explainable AI model," *Informatica*, vol. 48, no. 8, pp. 103–110, 2024.
13. H. Sun, C. Cai, H. Zhuang, K. A. Lee, L.-P. Chau, and Y. Wang, "EDVD-LLaMA: Explainable deepfake video detection via multimodal large language model reasoning," *arXiv preprint arXiv:2510.16442*, 2024.
14. S. Tariq, S. S. Woo, P. Singh, I. Irmalasari, S. Gupta, and D. Gupta, "From prediction to explanation: Multimodal, explainable, and interactive deepfake detection framework for non-expert users," in *Proc. ACM Multimedia (MM '25)*, pp. 1–11, 2025.
15. K. Tsigos, E. Apostolidis, S. Baxevanakis, S. Papadopoulos, and V. Mezaris, "Towards quantitative evaluation of explainable AI methods for deepfake detection," in *Proc. ACM MAD '24, Phuket, Thailand*, pp. 1–9, 2024.
16. N. Dhiman, H. Singh, and A. Thakur, "Exploring diverse techniques in image and video forgery," in *Machine Learning Algorithms (Proc. ICMLA 2024)*, CCIS, vol. 2238, Springer, Cham, pp. 318–328, 2025.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details