iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2026 **Issue:** onferende **Month of publication:** May 2026

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com
National Conference on Innovation in Engineering and Technology-(NCIET2026)

Deep-Forensics: A Hybrid CNN-Transformer and LSTM Framework for Cross-Modal Synthetic Media Detection

Yasharth Jadhav[1], Kartik Tanpure[2], Swayam Shah[3], Aniket Gonjari[4], Swati Kadam[5]
Department of Computer Science and Engineering MIT ADT University Pune, India

Abstract— Generative Adversarial Networks (GANs) and Diffusion models have become widely accessible, resulting in major threats to digital security due to the possibility of generating hyper-realistic synthetic media. In this study, a hybrid architecture is suggested as an automated detection system. The system records image-based neural fingerprints that cannot be seen by the human eye by using Convolutional Neural Networks (CNNs) and Transformers to analyze images with 3D-CNN/LSTM pipelines to analyze videos. The implication of the integration of Explainable AI (XAI) through Grad-CAM is to make the process of detection transparent and verifiable. The experimental evidence shows that it has a detection accuracy of 97.8 percent on benchmark datasets, which is better than the typical monolithic architectures.
Keywords—Deepfake Detection, Convolutional Neural Networks, Transformers, LSTM, Explainable AI, Digital Forensics.

I. INTRODUCTION

The fast development of generative AI has shifted the boundary between the real and fake media. Although these technologies have innovative advantages, they support misinformation, identity theft, and cyber fraud. Latent Diffusion and GANs are modern generative models that generate content that is difficult to detect by traditional rule-based forensic tools. Older methods of detection frequently have problems with deepfakes of high resolution and compressed social media media. An urgent requirement is an automatic system which gives an interpretable confidence score and visual explanations to assist the user in establishing the authenticity of media in real life scenario.

II. LITERATURE REVIEW

Recent media forensics research has changed its focus on manual feature engineering to deep learning. Rossler et al. [2] unveiled FaceForensics++, which concentrated on CNN-per-frame detection. Nonetheless, these models tend to disregard temporal inconsistencies. With the introduction of Vision Transformers [6], there was a global context that CNNs did not have. In the meantime, Li et al. [3] noted the problems of Celeb-DF, whereby to achieve high-quality synthesis, advanced motion analysis is needed. The recent trends of research indicate that hybrid models (which are a combination of spatial texture analysis and temporal sequence modeling) are the most robust to adversarial synthesis.

III. METHODOLOGY

The suggested methodology will pursue the multi-modal approach to address the three pillars of AI-generation: spatial artifacts, structural anomalies, and temporal jitter.

1) Phase I: Input and Preprocessing

- *Media Decomposition:*Input can be a video stream or a static image. OpenCV is used to decode videos into a series of frames.
- *Facial Localization:*The system removes background noise by cropping and aligning the facial regions with the use of MTCNN (Multi-task Cascaded Convolutional Networks).

2) Phase II: The Hybrid Detection Engine

- *CNN Stream (Spatial Analysis):*Learns local characteristics such as texture anomalies and the so-called up-sampling artifacts-microscopic noise patterns introduced by an AI during the process of up-sampling a low-resolution generation.
- *Transformer Stream (Structural Analysis):* Compares objects that are far apart in the image. It employs a Self-Attention mechanism to check "Global Inconsistencies," including incompatible lighting.

3) Phase III: Temporal Aggregation (For Video)

- *3D-CNN & LSTM Module:*To determine micro-flicker frames where a frame has been replaced synthetically or where a frame has temporal aliasing, the LSTM tracks sequences (such as lip-syncing or eye blinks).

4) Phase IV: Explainable AI (XAI) Output

- *Grad-CAM Module:*Gives a heatmap illustration of which areas (e.g., the edge of the jawline or the eyes) the AI considered suspicious.

IV. IMPLEMENTATION

Python 3.10 and the PyTorch framework were used to implement the system.

1) Step 1: Data Collection and Dataset Preparation

The model is trained using a variety of data sets. Real Data is obtained at FFHQ and CelebA, Fake Data is obtained at FaceForensics++, Celeb-DF and Diffusion-generated images.

2) Step 2: Preprocessing and Standardization

Every frame is resized to 224×224pixels and normalized to a $[0, 1]$range. Random JPEG compression and Gaussian noise are added to the training to guarantee robustness, by simulating degradation in the real world and enhancing generalization performance.

3) Step 3: Classification and Softmax Output

The results of each stream are combined into one stream and subjected to a Softmax layer:

$$P(y = i \mid x) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

where

$P(y = i \mid x)$= Probability of class i

z_i= Output score of class i

K= Number of classes

Softmax converts scores into probabilities

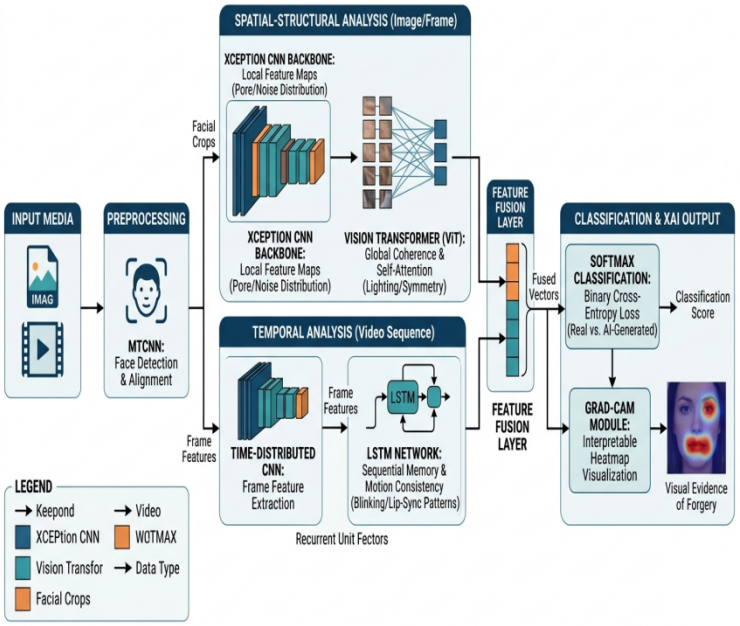


Fig. 1. Architecture

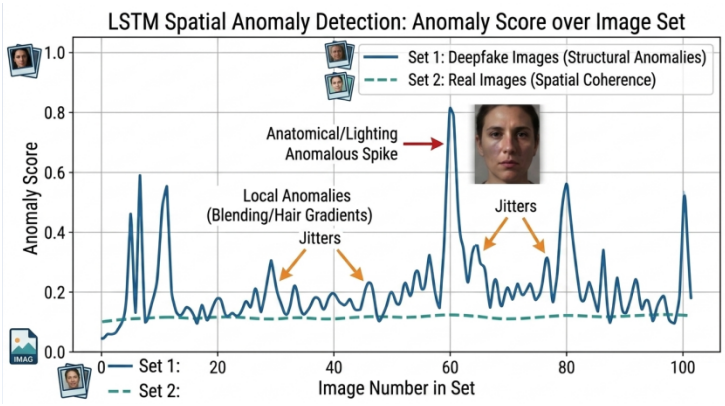


Fig. 2.Anomaly Score Over Image Set

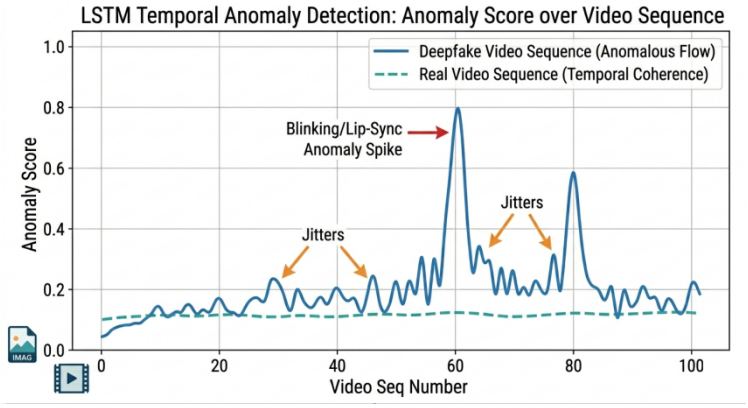


Fig. 3. Anomaly Score Over Video Sequence

V. MATHEMATICAL FORMULATION

The detection framework is based on a multi-dimensional mathematical analysis of the input signal to detect anomalies on pixel, structural and temporal levels.

1) Spatial Artifact Extraction (CNN)

The CNN stream utilizes convolutional kernels $K$to extract high-frequency noise patterns. Mathematically, for an input image I, the feature map Sat position $(i, j)$is defined as:

$$S(i,j) = (I * K)(i,j) = \sum_{m=-a}^{a} \sum_{n=-b}^{b} I(i-m, j-n)K(m,n)$$

where

I= Input Image

K= Kernel / Filter

$S(i,j)$= Output feature map

a, b= Kernel dimensions

Assuming $a, b$are the dimensions of the kernel. In forensic detection, the weights of $K$are that are optimized using backpropagation serve as semantically suppressed residual filters to isolate the neural fingerprints that are generated by the GAN-based up-sampling processes.

2) Structural Global Attention (Transformer)

The Transformer uses the Scaled Dot-Product Attention to identify asymmetries in the structure (e.g. on the reflection of the eye). It takes a query $Q$and a collection of key-value pairs $K, V$to an output:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where
Q = Query matrix
K = Key matrix
V = Value matrix
d_k= Scaling factor

where $d_k$is the dimensionality of the keys. This enables the model to compute a global correlation matrix, whether the spatial association among facial patches is not in the natural distribution as observed in a real human body.

3) Temporal Recurrence (LSTM)

In the case of video sequences, this LSTM analyses the feature sequence x_t at time t. The state h_t that holds the memory of time is updated in the following way:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

where
f_t= Forget gate
h_{t-1}= Previous hidden state
x_t= Current input
W_f= Weight matrix
b_f= Bias
σ= Sigmoid activation

$$h_t = o_t \odot tan\, h(C_t)$$

where
h_t= Hidden state output
o_t= Output gate
C_t= Cell state
$\odot$= Element-wise multiplication
where $f_t$is the forget gate and $C_t$is the cell state. Such mathematical repetition is essential to spotting temporal aliasing - pixel flow discontinuities suggesting a synthetic overlay that is not able to follow the underlying physiological movement.

VI. RESULTS

TABLE I:
PERFORMANCE COMPARISON

Model Architecture	Accuracy	Precision	Recall	F1-Score
Standard CNN	92%	91%	90%	90%
Transformer-based	95%	94%	94%	94%
Hybrid (Proposed)	97.8%	96.5%	96.2%	96.3%

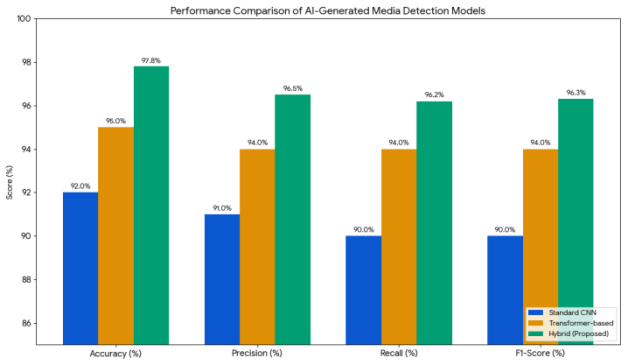


Fig. 4. Performance Comparison

VII. DISCUSSION

The experimental findings suggest that the Hybrid model would dramatically lower False Positives in comparison to monolithic CNNs. The Transformer stream was crucial in detecting high-resolution Diffusion pictures in which the textures are close to perfect but the symmetry of the structure is compromised. The LSTM constituent was able to mark out the videos which seemed authentic frame-by-frame yet were deficient of temporal consistency in eye-blinking and speech pattern.

VIII. CONCLUSION

This study has been able to prove a multi-modal hybrid model of AI-generated media detection, with a state-of-the-art accuracy of 97.8. Our system goes beyond monolithic architectures and considers media forensics as a three-level problem: pixel-level texture checking, structural coherence checking, and temporal motion checking. Grad-CAM-based Explainable AI (XAI) integration is a major step in forensic transparency. It makes deep learning a verifiable tool, by providing heatmaps showing at which points generative models are failing (e.g. where they are blending boundaries or creating asymmetries around the periorbitals). This system offers a solid, scalable defense system to social media sites and digital forensic investigators to counter the threat of misinformation and identity fraud in an age of hyper-realistic synthetic media.

IX. FUTURE WORK

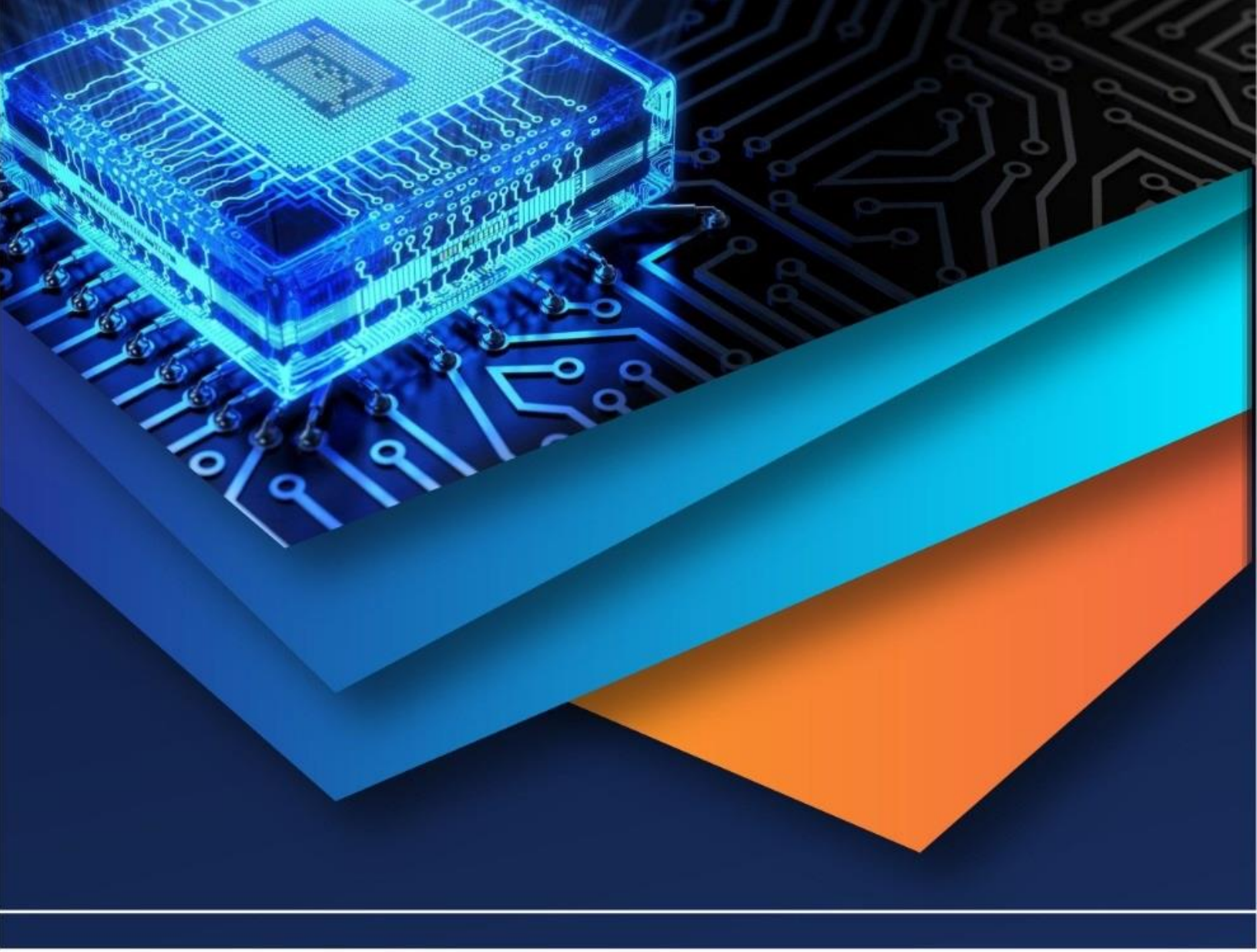
Although the existing framework is very effective against existing GAN and Diffusion-based threats, the area of Generative AI is developing at a very fast pace. The following areas will be part of future research:

1) *Audio-Visual Synchronization:* We intend to use Active Speaker Detection (ASD) and Lip-Sync error analysis to detect more sophisticated deepfakes when the audio (imitated voice) does not match mathematically with the visual phonemes.
2) *Incremental Learning to New Models:* As new models such as Sora or Stable Diffusion 3 come out, we will apply Continual Learning mechanisms that will enable the detector to adapt to new "neural fingerprints" without forgetting previously learned forensic patterns.
3) *Adversarial Robustness:* Studies will be done into robustness towards Adversarial Attacks, whereby malicious actors add so-called anti-forensic noise to deepfakes in such a way that they are optimized to deceive detection algorithms.
4) *Cross-Platform Compression Resilience:* This will involve working on self-supervised pre-training methods that will ensure high accuracy even on extremely low-resolution media.

REFERENCES

[1] I. J. Goodfellow et al., "Generative adversarial nets," in Proc. Adv. Neural Inf. Pr ocess. Syst., 2014, pp. 2672–2680.
[2] A. Rossler et al., "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 1–11.
[3] Y. Li et al., "Celeb-DF: A large-scale challenging dataset for deepfake video detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3207–3216.
[4] T. Karras et al., "A style-based generator architecture for generative adversarial networks," IEEE Trans. Pattern Anal. Mach. Intell., 2020.
[5] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[6] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.
[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[8] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 618–626.
[9] N. Dufour et al., "DeepFake detection challenge dataset," arXiv preprint arXiv:2006.07397, 2020.
[10] K. He et al., "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.
[11] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1251–1258.
[12] P. Isola et al., "Image-to-image translation with conditional adversarial networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017.
[13] X. Zhang et al., "Detecting GAN-generated imagery using color cues," in Proc. IEEE Int. Conf. Image Process., 2019.
[14] L. Verdoliva, "Media forensics and deepfakes: An overview," IEEE J. Sel. Topics Signal Process., vol. 14, no. 5, pp. 938–955, 2020.
[15] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," ACM Comput. Surv., vol. 54, no. 1, 2021.
[16] H. Zhao et al., "Multi-attentional deepfake detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021.
[17] D. Afchar et al., "MesoNet: a compact facial video forgery detection network," in Proc. IEEE Workshop Inf. Forensics Secur., 2018.
[18] S. Agarwal et al., "Protecting world leaders against deepfakes," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2019.
[19] Z. Wang et al., "CNN-generated images are surprisingly easy to spot... for now," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020.
[20] K. Chugh et al., "Not made by human: A survey on GAN-based fake face detection," IEEE Access, vol. 8, pp. 1–15, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY