International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

AI-Enhancing Security with AI: Media Fabrication Detection Through Advanced Neural Architectures

G. Lingaswamy, G. Sudheer, V. Akhil, Asst. Prof. D. Manasa
Department of Information Technology, ACE Engineering College, Hyderabad, Telangana, India

ABSTRACT

Purpose: With the widespread adoption of AI-based content generation tools, the boundaries between authentic and synthetic media have blurred considerably. Fabricated video and image content — commonly referred to as deepfakes — now represent a tangible threat to individual privacy, electoral integrity, and institutional trust. This study addresses that challenge by proposing a detection framework capable of identifying forged visual media with high precision. **Design/Methodology:** A hybrid architecture is developed that unites the spatial discriminative power of Convolutional Neural Networks (CNNs) with the sequential reasoning capability of Long Short-Term Memory (LSTM) units. The CNN backbone — instantiated as a fine-tuned ResNeXt model — extracts rich frame-level feature representations, while the LSTM layer models temporal patterns across video frames. **Findings:** On the FaceForensics++ benchmark, the integrated architecture attains an overall detection accuracy of approximately 94%, surpassing both standalone CNN and standalone LSTM baselines by a clear margin. Explainability mechanisms are additionally incorporated to make predictions interpretable for end users and security analysts. **Practical Implications:** The system is designed for deployment in cybersecurity platforms, digital forensics workflows, and social media content moderation pipelines. Its modular design supports both cloud hosting and edge inference, making it suitable for diverse real-world environments.

Keywords: *Synthetic Media Detection, Deepfake Identification, Hybrid CNN-LSTM, Transfer Learning with ResNeXt, Temporal Forgery Analysis, Digital Forensics, Explainable Artificial Intelligence, Cybersecurity.*

I. INTRODUCTION

Synthetic media generation has evolved at a remarkable pace. Techniques grounded in deep generative modelling — particularly Generative Adversarial Networks (GANs) — can now produce photorealistic facial videos in which one person's likeness is seamlessly superimposed onto another's body. While such capabilities have legitimate uses in film production, virtual communication, and assistive technology, their potential for harm is equally considerable. Fabricated footage has already been weaponised for political disinformation, non-consensual intimate imagery, corporate fraud, and targeted harassment, and the problem is accelerating as generative tools become more accessible and easier to use.

Despite this growing threat, automated countermeasures have lagged. Early detection attempts relied on hand-crafted visual cues — visible blurring at facial boundaries, abnormal eye-blink frequencies, or colour-space inconsistencies — that newer generation methods have largely rendered obsolete. More recent machine learning approaches have shown greater promise, yet most suffer from a fundamental architectural limitation: they either model spatial artefacts within individual frames or temporal patterns across frame sequences, but seldom both in a unified and jointly optimised manner. This architectural gap is significant because high-quality deepfakes are specifically designed to avoid obvious per-frame anomalies; it is the subtle temporal incoherence across frames that most reliably betrays them.

This paper introduces a detection system built around a two-stage deep learning pipeline that closes this gap. In the first stage, a ResNeXt-based CNN processes individual frames and extracts spatial feature maps sensitive to texture irregularities, blending artefacts, and boundary inconsistencies. In the second stage, an LSTM network receives the frame-wise feature sequences and learns to identify temporal anomalies — unnatural motion trajectories, flickering identity signals, and inconsistent lighting transitions — that emerge when a sequence of generated frames is viewed holistically. The result is a detection model that exploits both what is wrong in each frame and how the wrongness evolves.

The principal contributions of this work are:

- A jointly trained CNN-LSTM pipeline that models spatial and temporal forgery cues within a single end-to-end framework.
- Fine-tuned ResNeXt integration that transfers rich feature knowledge from large-scale image recognition tasks to the deepfake detection domain.
- Deployment-oriented system design with support for both real-time inference and batch processing of image and video inputs.
- Explainability overlays that highlight the spatial regions most influential in each detection decision, supporting analyst review.
- Rigorous evaluation against FaceForensics++ and Celeb-DF datasets using accuracy, precision, recall, and F1 score metrics.

II. LITERATURE REVIEW

Understanding the limitations of prior work is essential to contextualising the contributions presented here. Research in deepfake creation and detection has followed an adversarial trajectory over the past decade, with each generation of detection methods being quickly challenged by more sophisticated generation techniques.

A. Foundational Generative Methods

The seminal contribution of Goodfellow and colleagues in 2014 established the GAN framework, demonstrating that a generator network trained in competition with a discriminator could produce highly realistic images from random noise [1]. This work initiated a wave of generative research that progressively improved the visual fidelity of synthetic content. Karras and co-authors subsequently introduced progressive training strategies for GANs, enabling the synthesis of high-resolution facial images with fine-grained detail [2]. Critically, this improvement in generation quality had a direct and adverse effect on detection difficulty, as the coarse artefacts that earlier detectors relied upon were largely eliminated.

B. Detection Approaches

Sabir and collaborators were among the first to incorporate recurrent modelling into deepfake detection, pairing a CNN with a recurrent neural network to capture inter-frame dependencies in video sequences [3]. This represented a meaningful architectural advance, though the resulting model exhibited limited generalisation when applied to forgery methods not represented in its training data. Afchar and colleagues took a different approach with MesoNet, designing a compact architecture targeting mesoscopic image statistics — the intermediate-scale texture patterns that differ between authentic and synthetically altered faces [4]. MesoNet demonstrated good performance on first-generation deepfakes but proved brittle when evaluated against more recent generation techniques that better preserve mesoscopic consistency. Li and colleagues explored a behavioural signal — the frequency and naturalness of eye blinking — as a passive indicator of fabrication [5]. The rationale was sound: early deepfake pipelines did not model blinking accurately, leaving a detectable biological signature. However, this approach became unreliable as generative models matured to replicate natural eye-movement patterns.

C. Identified Research Gaps

Reviewing the literature reveals three recurring deficiencies. First, most published systems treat spatial and temporal feature extraction as separate problems, even when working with video data. This separation forfeits the complementary information that arises when both are modelled jointly. Second, generalisation across unseen generation methods and datasets remains poor; models trained on one forgery type typically fail on another. Third, explainability is rarely addressed, leaving detection decisions as black-box outputs that are difficult to validate or challenge in forensic contexts. The approach proposed in this paper is designed specifically to address each of these gaps.

III. PROBLEM STATEMENT

The core technical problem is the reliable automated identification of AI-generated facial media in conditions that closely resemble real-world deployment scenarios. Several properties of the detection problem make this non-trivial.

Contemporary deepfake videos often exhibit no frame-level artifacts that are visible to human observers, and increasingly little that automated spatial detectors can reliably find. The distinguishing information is instead distributed across time — encoded in the subtle inconsistency of how identity, lighting, and motion co-vary from one frame to the next. A detection system that examines frames in isolation will therefore fail on a growing proportion of modern forgeries, regardless of how sophisticated its spatial feature extractor may be.

At the same time, practical deployment imposes constraints that purely research-oriented work often ignores. A useful detection system must process both images and video with low latency; it must maintain acceptable false positive rates to avoid flagging authentic content; and its outputs must be interpretable enough to support downstream review by human analysts or automated moderation systems. The system presented here is designed to satisfy all of these requirements simultaneously.

Formal Objective: To construct an automated deepfake detection system that jointly models spatial and temporal forgery signals, achieves high accuracy on established benchmarks, maintains low false positive and false negative rates, and produces human-interpretable detection evidence.

IV. PROPOSED SYSTEM ARCHITECTURE

The proposed detection framework is organised as a sequential two-stage pipeline in which a CNN-based spatial encoder feeds into an LSTM-based temporal reasoner, culminating in a binary classification output with an associated confidence estimate.

A. End-to-End Processing Flow

- Input acquisition: the system accepts a raw image or video file through a web-based upload interface.
- Preprocessing module: for video inputs, frames are extracted at a configurable sampling rate; faces are localised, cropped, aligned, and resized to a standard resolution; pixel values are normalised to the unit range.
- Spatial feature extraction: the preprocessed frame sequence is passed through a fine-tuned ResNeXt CNN, which produces a high-dimensional feature vector for each frame.
- Temporal sequence modelling: the ordered series of per-frame feature vectors is presented to a stacked LSTM, which learns to detect motion-level and identity-level inconsistencies across the sequence.
- Classification head: a fully connected layer with softmax activation maps the LSTM output to a probability distribution over the binary output classes (authentic / fabricated).
- Output delivery: the predicted class label and the associated confidence score are returned to the user interface, along with an explainability heatmap indicating the most informative spatial regions.

B. System Advantages

The hybrid architecture confers several properties that neither component alone can provide. The CNN's ability to detect per-frame spatial artifacts complements the LSTM's sensitivity to temporal anomalies; together, they produce a detection signal that is considerably more robust than either in isolation. The use of a pre-trained ResNeXt backbone accelerates convergence and reduces the volume of labelled training data required. The modular design of the pipeline means that individual components — the face detector, the CNN backbone, or the LSTM depth — can be swapped or upgraded without redesigning the system from scratch.

V. METHODOLOGY

A. Datasets

Two standard benchmarks are used for training and evaluation. FaceForensics++ is a large-scale dataset containing thousands of manipulated videos produced by four distinct forgery methods (DeepFakes, Face2Face, FaceSwap, and NeuralTextures) at multiple compression levels, making it a demanding test of generalisation across manipulation types. Celeb-DF supplements this with higher visual quality deepfakes of celebrity subjects, providing additional challenge from a more recent generation pipeline. The combined dataset spans a wide range of forgery strategies, resolutions, and compression artefacts.

B. Preprocessing Pipeline

Video inputs are decoded and frames are sampled at a fixed rate to produce sequences of consistent temporal density. The MTCNN face detector is applied to each frame to localise and crop the facial region, which is then aligned using five facial landmarks to a canonical pose. Cropped faces are resized to 224 x 224 pixels to match the ResNeXt input specification. Pixel intensities are normalised using ImageNet channel statistics to facilitate transfer learning. During training, online augmentation is applied, including random horizontal flipping, rotation within ±10 degrees, colour jitter, and Gaussian noise injection to improve robustness to compression and post-processing artefacts.

C. Feature Extraction via ResNeXt

ResNeXt-50 is used as the CNN backbone, initialised with weights pre-trained on ImageNet. The model is fine-tuned end-to-end on the target dataset. ResNeXt extends the standard residual network by introducing grouped convolutions parameterised by a cardinality dimension, which increases representational capacity without a proportional increase in computational cost. The final classification head of the original model is removed and replaced with a feature projection layer that outputs a 512-dimensional vector per frame. These vectors serve as the input to the LSTM stage.

D. Temporal Modelling via LSTM

Sequences of 16 consecutive frame feature vectors are presented to a two-layer LSTM with 256 hidden units per layer. The LSTM processes the sequence autoregressively, updating its cell and hidden states at each timestep. Dropout with probability 0.3 is applied between LSTM layers to mitigate overfitting. The hidden state at the final timestep is taken as the sequence-level representation and passed to the classification head. This design allows the network to accumulate evidence across the entire clip before committing to a prediction.

E. Training Configuration

The model is trained end-to-end using binary cross-entropy loss and the Adam optimiser with an initial learning rate of 1e-4, decayed by a factor of 0.5 every five epochs without improvement on the validation set. Training proceeds for a maximum of 40 epochs with early stopping based on validation accuracy. Class weights are applied during loss computation to account for any imbalance between authentic and fabricated samples in the training split.

VI. ALGORITHMS AND DESIGN RATIONALE

A. ResNeXt Convolutional Neural Network

ResNeXt employs aggregated residual transformations — effectively a set of grouped convolutions applied in parallel within each residual block — parameterised by a cardinality value that controls the number of parallel paths. Empirically, increasing cardinality is a more effective way to boost accuracy than increasing depth or width alone, while remaining computationally tractable. In the context of deepfake detection, ResNeXt's rich multi-scale feature representations are particularly well-suited for capturing the subtle texture and boundary anomalies that characterise synthetically altered faces.

B. Long Short-Term Memory Networks

Standard recurrent networks suffer from the vanishing gradient problem, which limits their ability to learn dependencies over long time horizons. LSTMs address this through a gating mechanism comprising an input gate, a forget gate, and an output gate, together with a dedicated cell state that can preserve information over many timesteps with minimal gradient degradation. For video deepfake detection, this translates to an ability to detect inconsistencies in how facial attributes — identity, pose, lighting, and expression — evolve, even when no individual frame is obviously manipulated in isolation.

C. Soft max Classification Layer

The final classification layer applies the softmax function to convert the raw output logits into a proper probability distribution over the two output classes. The class with the higher probability is taken as the prediction, while its probability value provides a calibrated confidence estimate that can be thresholded to control the trade-off between sensitivity and specificity for different deployment contexts.

D. Explainability via Gradient-Weighted Activation Mapping

To support interpretable detection, Gradient-weighted Class Activation Mapping (Grad-CAM) is applied to the final convolutional layer of the ResNeXt backbone. This technique computes the gradient of the classification score with respect to the feature maps and uses the resulting gradient magnitudes as weights to produce a spatially localised heatmap highlighting the image regions most influential in the model's decision. In practice, these heatmaps typically highlight boundary artefacts, texture discontinuities, or facial regions where the blending operation was imperfect.

VII. RESULTS AND ANALYSIS

The proposed system is evaluated on the FaceForensics++ test partition using a 70:15:15 train/validation/test split stratified by video identity to prevent data leakage. Table I reports the four principal performance metrics for three experimental configurations: a standalone CNN baseline, a standalone LSTM baseline operating on raw pixel sequences, and the proposed CNN-LSTM hybrid.

TABLE I: Performance Comparison of Detection Configurations

Model / Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
CNN Only (ResNeXt)	88.1	85.7	86.9	86.3
LSTM Only (Pixel Sequences)	89.0	87.2	88.1	87.6
Proposed CNN-LSTM Hybrid	94.3	92.1	93.4	92.7

A. Interpretation of Results

The hybrid model delivers a 6.2 percentage point accuracy gain over the CNN-only baseline and a 5.3 percentage point gain over the LSTM-only baseline. More informatively, the precision and recall gap between the hybrid and either single-stage model is consistent across both metrics, which rules out the possibility that the improvement is an artifact of the operating threshold and confirms that the gains reflect a genuine improvement in discrimination ability.

The confusion matrix analysis reveals that the primary source of improvement in the hybrid model is a reduction in false negatives — fabricated samples that the spatial-only model failed to flag. This is consistent with the design rationale: temporal modelling catches forgeries where individual frames appear clean but the across-frame identity or motion signal is inconsistent.

Performance on lower compression quality variants of FaceForensics++ (where compression artefacts mask spatial forgery cues) shows the largest relative improvement from temporal integration, confirming that LSTM-based temporal analysis provides a meaningful complementary signal precisely in the conditions where spatial-only detection is most challenged.

VIII. IMPLEMENTATION DETAILS

A. Technology Stack

The model is implemented in Python 3.10 using Py Torch as the primary deep learning framework. The ResNeXt backbone is sourced from the torch vision model zoo with ImageNet-pretrained weights. MTCNN face detection is implemented via the facenet-pytorch library. OpenCV handles video decoding and frame-level preprocessing operations. The web application layer is built with Django, exposing a REST API that accepts image or video uploads and returns structured JSON detection responses. The frontend is implemented as a lightweight HTML/CSS/JavaScript interface that renders detection results and explainability heatmaps in the browser.

B. Infrastructure and Deployment

All training experiments were conducted on an NVIDIA RTX 3090 GPU with 24 GB VRAM. The trained model is exported to Torch Script format for deployment, enabling inference without a full Python environment. The inference server is containerised using Docker, allowing deployment on cloud platforms (AWS, GCP, Azure) or on-premises hardware. Average inference latency for a 10-second video clip on the target GPU is under 1.2 seconds, satisfying the real-time processing requirement. Model artefacts, detection logs, and uploaded media are stored in a relational database (PostgreSQL) with audit-trail support for forensic review workflows.

IX. CONCLUSION

This paper has presented a deepfake detection system that addresses the central limitation of prior work: the failure to model spatial and temporal forgery cues jointly within a unified, end-to-end trainable architecture. By coupling a fine-tuned ResNeXt CNN with a stacked LSTM, the proposed system achieves 94.3% detection accuracy on FaceForensics++, outperforming both spatial-only and temporal-only baselines by clear margins. The integration of Grad-CAM explainability overlays enhances the system's operational value by making detection evidence visible to analysts, not just to the model.

Beyond the quantitative results, the system is designed with deployment practicality in mind. Its containerised architecture, REST-based API, and sub-second inference latency position it as a viable foundation for real-world applications in digital forensics, social media content moderation, and enterprise cybersecurity. The modular pipeline design ensures that the system can be updated as generation techniques evolve, without requiring a full architectural redesign. In an information landscape increasingly shaped by synthetic media, robust and interpretable detection infrastructure of this kind is not merely a research objective — it is a societal necessity.

X. FUTURE WORK

Several directions offer natural extensions to the work presented here. In the near term, adapting the detection pipeline for live video streams — rather than pre-recorded clips — would extend its applicability to broadcast monitoring and real-time social media analysis. Simultaneously, developing a compressed model variant suitable for mobile deployment would make detection accessible in edge contexts where cloud connectivity is unavailable or undesirable.

Over a longer horizon, extending the detection scope to synthesised audio — voice cloning and speech synthesis forgeries, and developing cross-modal detectors that reason over video, audio, and metadata jointly would significantly broaden the system's coverage. Federated learning approaches present an interesting avenue for improving model accuracy across diverse data distributions while respecting the privacy constraints that often prevent direct data sharing between organisations. Finally, incorporating adversarial training into the detection pipeline — explicitly exposing the model to adaptively generated forgeries during training — may improve robustness against detection-aware deepfake generation methods that are already beginning to emerge.

REFERENCES

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *"Generative Adversarial Nets,"* in *Advances in Neural Information Processing Systems (NeurIPS),* vol. 27, Montreal, Canada, 2014, pp. 2672–2680.

[2] T. Karras, T. Aila, S. Laine, and J. Lehtinen, *"Progressive Growing of GANs for Improved Quality, Stability, and Variation,"* in *Proceedings of the International Conference on Learning Representations (ICLR),* Vancouver, Canada, 2018.

[3] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, *"Recurrent Convolutional Strategies for Face Manipulation Detection in Videos,"* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* Long Beach, CA, USA, 2019.

[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, *"MesoNet: A Compact Facial Video Forgery Detection Network,"* in *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS),* Hong Kong, China, 2018.

[5] Y. Li, M.-C. Chang, and S. Lyu, *"In Ictu Oculi: Exposing AI-Generated Fake Face Videos by Detecting Eye Blinking,"* in *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS),* Hong Kong, China, 2018.