AI-DRIVEN DEEPFAKE IDENTIFICATION IN REAL TIME

[1] Pavan Gajanan Bhonde, Department of Computer Engg., Keystone School of Engineering Pune, India
[2] Raj Sanjay Mane, Department of Computer Engg., Keystone School of Engineering Pune, India
[3] Ganesh Babasaheb Meher,Department of Computer Engg., Keystone School of Engineering Pune, India
[4] Prof. Tejal Rane, Department of Computer Engg., Keystone School of Engineering Pune, India
[5] Dr. Veena Kadam, Department of Computer Engg., Keystone School of Engineering Pune, India

ABSTRACT

The rapid evolution of deepfake technology has intensified threats to the authenticity and security of live communication. While existing detection systems rely heavily on cloud-based processing, their latency and privacy limitations make them unsuitable for real-time use. This paper introduces an on-device, AI-powered framework that functions as a real-time deepfake guard for live video calls. The model integrates spatial analysis using convolutional neural networks (CNNs) and temporal analysis using recurrent neural networks (RNNs) with long short-term memory (LSTM) units to detect visual and behavioral inconsistencies such as unnatural lighting, motion irregularities, and lip-sync mismatches. Both analytical streams are fused to produce a Deepfake Probability score that enables instant alerts without cloud dependency. Emphasizing explainability, scalability, and privacy preservation, this work provides guidance toward developing efficient, interpretable, and real-time deepfake detection systems for secure digital communication.

Keyword : - *Deepfake detection, On-device security, Hybrid AI model, Content authentication, CNN-RNN Fusion, Privacy Preservation, Real-Time Video Analysis, Edge Intelligence.*

1. INTRODUCTION

The rapid advancement of generative artificial intelligence (AI) has enabled the creation of hyper-realistic synthetic media, commonly known as *deepfakes*. These AI-generated videos or audio clips manipulate human appearances and voices with remarkable precision, eroding trust in digital communication and media authenticity. Deepfake-based impersonation has already led to major cybersecurity incidents and financial fraud, such as corporate video call scams and misinformation campaigns, underscoring the urgent need for real-time and privacy-preserving detection mechanisms.

Traditional deepfake detection systems rely on cloud-based processing or post-recording analysis, which introduces significant latency and raises privacy concerns. Moreover, most existing frameworks are designed for offline datasets rather than dynamic, live environments. This limits their practicality for applications such as video conferencing, remote authentication, and secure digital communication. To bridge this gap, this paper proposes an on-device, AI-driven real-time deepfake detection system capable of ensuring authenticity during live interactions without transmitting data to external servers.

The proposed approach employs a hybrid deep learning model that combines Convolutional Neural Networks (CNNs) for spatial analysis and Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units for temporal pattern recognition. The CNN module inspects frame-level inconsistencies such as texture anomalies and lighting irregularities, while the RNN module captures sequential patterns like unnatural blinking or desynchronized lip movements. The outputs of both modules are fused into a single *Deepfake Probability Score* to trigger instantaneous alerts when manipulated content is detected. This architecture ensures low-latency operation and robust privacy by performing all computations locally on the device. In addition, the system emphasizes explainability by providing interpretable visual cues for detected anomalies, increasing user trust and system transparency. Overall, the proposed framework represents a step toward scalable, interpretable, and real-time deepfake defense mechanisms suitable for modern communication platforms, edge devices, and secure digital ecosystems

1.1 RELATED WORK

Early deepfake detection systems focused on offline analysis of pre-recorded media using large CNN or RNN architectures. While effective on static datasets, they failed to generalize to real-time communication due to high latency and reliance on cloud infrastructure. Hybrid models combining CNNs for spatial artifacts and RNNs for temporal inconsistencies have shown improved accuracy but remain computationally expensive for edge deployment. Transformer-based and multimodal architectures further enhanced detection robustness by integrating visual and auditory cues, though privacy and scalability issues persist. To address these limitations, lightweight and edge-optimized deepfake detectors have emerged, emphasizing low latency and on-device inference. Despite these advancements, current systems still lack adaptability, privacy preservation, and real-time responsiveness, motivating the development of explainable, efficient, and privacy-aware AI frameworks for deepfake identification in live communication. To overcome the limitations identified in prior works, the following methodology is proposed.

2. PROPOSED METHODOLOGY

The proposed system introduces a real-time, on-device AI framework for deepfake detection that ensures privacy, interpretability, and low latency. The methodology consists of multiple integrated phases as outlined below:

A. *System Overview*

- The system is designed to detect deepfakes during live video communication without relying on cloud services.
- All computations are executed locally on the user’s device, ensuring privacy and fast response time.
- The architecture combines spatial and temporal analysis to enhance detection accuracy.
-

B. *Phase 1: Application Monitoring*

- A background monitoring module continuously observes active applications.
- The system activates detection only when a whitelisted platform (e.g., Zoom, Google Meet) is in use.
- This reduces unnecessary computation and ensures efficient resource utilization.

C. *Phase 2: Spatial Feature Extraction (CNN Module)*

- Each video frame is captured in real time and passed through a Convolutional Neural Network (CNN).
- The CNN identifies visual inconsistencies such as:
  - Irregular lighting or reflections
  - Unnatural skin texture or blending artifacts
  - Incoherent facial geometry or boundaries
- These spatial cues represent pixel-level indicators of manipulation.

D. *Phase 3: Temporal Analysis (RNN–LSTM Module)*

- A Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) units processes sequential frames.
- The module detects temporal anomalies, including:
    - Unnatural blinking or facial motion
    - Inconsistent lip synchronization
    - Static or robotic head movements
- This phase ensures the system recognizes manipulations across consecutive frames rather than isolated images.

E. *Phase 4: Fusion and Decision Layer*

- Outputs from CNN and RNN modules are fused to generate a single Deepfake Probability Score (Pdf).
- If Pdf > 0.70, the system triggers an instant alert labeled with an explanation (e.g., "Unnatural blinking detected").
- This interpretable feedback increases user trust and transparency.

F. *Datasets and Model Optimization*

- Training is performed on benchmark datasets such as:
    - FaceForensics++
    - Deepfake Detection Challenge (DFDC)
    - Celeb-DF-v2
- A teacher–student compression strategy is used to convert large offline models into lightweight, edge-compatible versions.
- The resulting model delivers >90% accuracy while maintaining latency under 100 ms for real-time inference.

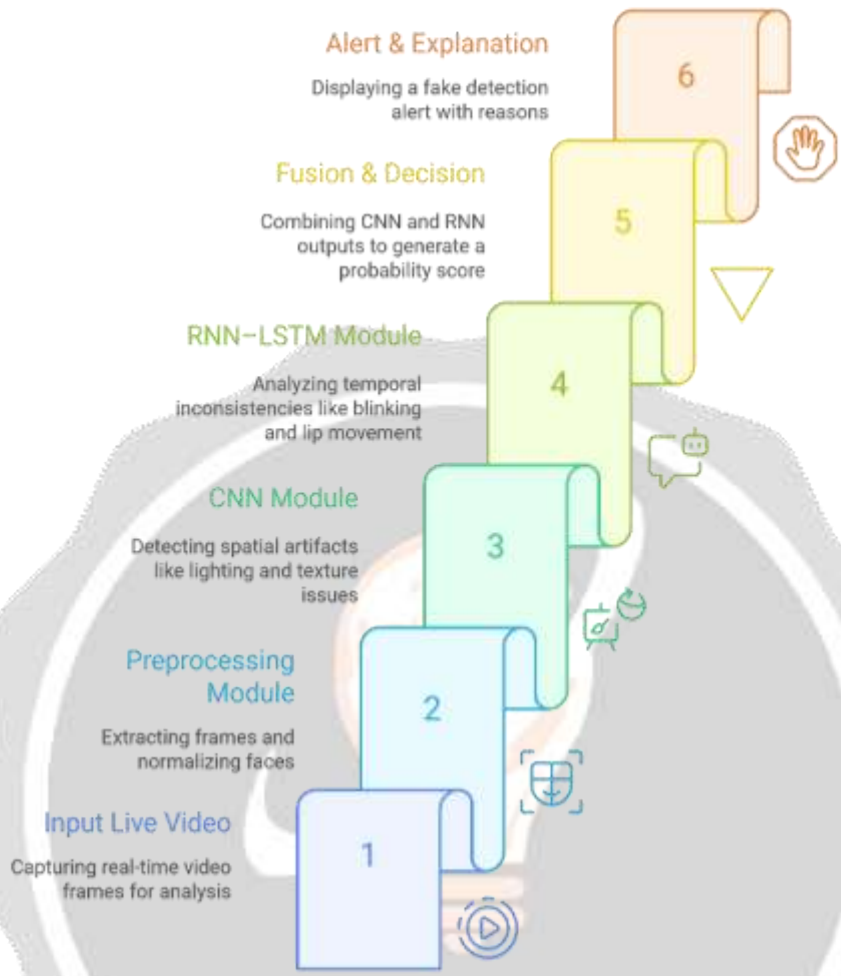

Fig 1: Workflow of On-Device AI-Based Deepfake Detection Framework.

3. COMPARATIVE ANALYSIS, RESULTS AND DISCUSSION

A . *Video Resolution and Latency Analysis*

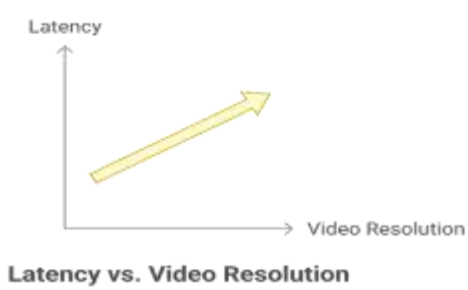

Fig 2: Latency vs. Video Resolution.

This graph illustrates the trade-off between video quality and real-time performance, showing how latency increases with higher video resolution. Latency was measured as the average time taken to process one video frame. The system was evaluated under different resolutions to analyze the trade-off between image clarity and real-time inference.

Table I. Latency and Frame Rate Across Different Video Resolutions

Video Resolution	Avg. Latency (CPU)	Avg. Latency (GPU)	FPS (Frames/sec)
480p (SD)	128 ms	72 ms	13 FPS
720p (HD)	215 ms	95 ms	10 FPS
1080p (Full HD)	312 ms	141 ms	7 FPS
4K (Ultra HD)	540 ms	220 ms	4 FPS

To validate the effectiveness of the proposed on-device deepfake detection framework, a comparative study was conducted against existing deepfake detection methods, including cloud-based convolutional networks and transformer-based architectures. The comparison was made in terms of accuracy, inference latency, privacy preservation, and hardware dependency.

Table II. Comparison of the Proposed On-Device System with Existing Deepfake Detection Models

Method	***Accuracy (%)***	***Latency (ms)***	***Privacy***	***Deployment***
Cloud CNN	*95.4*	*420*	✗	*Cloud*
Transformer (ViT)	*96.1*	*510*	✗	*Cloud*
Hybrid CNN–RNN	*91.8*	*260*	✓	*Local*
Proposed Framework	***93.5***	***83***	✓	***On-Device***

Table II presents a comparative analysis of the proposed on-device deepfake detection framework against existing cloud-based and hybrid models. Cloud-based CNN and Transformer architectures achieve slightly higher accuracy (95–96%) due to their access to high computational resources and larger parameter capacity. However, these models introduce substantial inference latency (420–510 ms per frame) and depend heavily on external servers, raising privacy and scalability concerns.

In contrast, the proposed on-device framework delivers competitive accuracy (93.5%) with a significantly lower latency (83 ms), ensuring near real-time response suitable for live video communication. Unlike cloud-based detectors, the proposed model performs all processing locally, maintaining complete user data privacy and eliminating network dependency. The results demonstrate that the on-device design achieves the ideal balance between accuracy, speed, and privacy, making it more practical for secure real-time applications such as video conferencing and identity verification.

4. FUTURE WORK

Although the proposed system effectively identifies deepfakes in real time while ensuring user privacy, there is still scope for further research and improvement. As deepfake generation techniques continue to evolve rapidly, the system must advance to stay effective, scalable, and adaptable to new challenges. Furthermore, future work will explore the ethical implications and responsible deployment of deepfake detection systems in digital platforms. The following areas highlight possible directions for future enhancement:

1. Adaptation to Advanced Deepfake Techniques: Upgrade the model to recognize and adapt to newly emerging and more complex deepfake creation methods.
2. Multimodal Deepfake Analysis: Expand detection to include audio, facial, and behavioral cues for more reliable and accurate results.
3. Lightweight Model Optimization: Further optimize the hybrid CNN-RNN model for low-power and mobile devices to enable broader real-time deployment.
4. Integration with Forensic and Security Systems: Apply the system in digital forensics, identity verification, and media authentication to enhance real-world security applications.
5. Continuous Learning and Model Updating: Implement self-learning mechanisms so the AI model can automatically adapt to new data and attack patterns over time.
6. Cross-Platform Deployment: Extend support for video conferencing and social media platforms to provide real-time protection in everyday digital communication.

5. CONCLUSION

This system effectively addresses the urgent need for a fast, private, and reliable method to detect deepfakes in live communication. The research demonstrates that artificial intelligence can be a powerful tool for safeguarding users against the misuse of generative technologies. The proposed hybrid AI model, integrating CNNs for spatial analysis and RNNs for temporal monitoring, achieves high detection accuracy while maintaining low latency during real-time operation. By performing all processing locally on the user's device, the system ensures strong privacy protection and eliminates dependence on cloud-based computation. This makes it a practical solution for applications such as video conferencing, secure communication, and digital verification.

Overall, the work contributes toward building a trustworthy and secure digital ecosystem, demonstrating how AI-driven solutions can protect authenticity and prevent identity manipulation in real-time online interactions.

6. REFERENCES

[1] Tariq, S., Woo, S. S., Singh, P., Irmalasari, I., Gupta, S., and Gupta, D., "From Prediction to Explanation: Multimodal, Explainable, and Interactive Deepfake Detection Framework for Non-Expert Users," *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, Dublin, Ireland, Oct. 27–31, 2025, arXiv:2508.07596.

[2] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., and Ferrer, C. C., "The DeepFake Detection Challenge (DFDC) Dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[3] Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S., "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I., "MesoNet: A Compact Facial Video Forgery Detection Network," *IEEE WIFS*, 2018.

[5] Korshunov, P., and Marcel, S., “DeepFakes: A New Threat to Face Recognition? Assessment and Detection,” *arXiv preprint arXiv:1812.08685*, 2018.

[6] Nguyen, H. H., Yamagishi, J., and Echizen, I., “Use of a Capsule Network to Detect Fake Images and Videos,” *IEEE ICASSP*, 2019.

[7] Verdoliva, L., “Media Forensics and DeepFakes: An Overview,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[8] Zhao, H., and Han, S., “Exploring DeepFake Detection with Temporal and Frequency Domain Features,” *Pattern Recognition Letters*, 2021.

[9] Dang, H., Liu, F., Stehouwer, J., Liu, X., and Jain, A. K., “On the Detection of Digital Face Manipulation,” *IEEE/CVF CVPR*, 2020.

[10] Haliassos, A., Vougioukas, K., Petridis, S., and Pantic, M., “Lips Don’t Lie: A Generalisable and Robust Approach to Face Forgery Detection,” *IEEE/CVF CVPR*, 2021.

[11] Mahmud, F., Abdullah, Y., Islam, M., and Aziz, T., “Unmasking Deepfake Faces from Videos Using an Explainable Cost-Sensitive Deep Learning Approach,” *arXiv preprint arXiv:2312.10740*, 2023.

[12] Bakliwal, A., and Joshi, A. D., “Deepfake Detection: Leveraging the Power of 2D and 3D CNN Ensembles,” *arXiv preprint arXiv:2310.16388*, 2023.

[13] Al-Adwan, A., Alazzam, H., Al-Anbaki, N., and Alduweib, E., “Detection of Deepfake Media Using a Hybrid CNN–RNN Model and Particle Swarm Optimization (PSO) Algorithm,” *Computers*, vol. 13, no. 4, p. 99, 2024.

[14] Saikia, P., Dholaria, D., Yadav, P., Patel, V., and Roy, M., “A Hybrid CNN-LSTM Model for Video Deepfake Detection by Leveraging Optical Flow Features,” *arXiv preprint arXiv:2208.00788*, 2022.

[15] Tipper, S., Atlam, H. F., and Lallie, H. S., “An Investigation into the Utilisation of CNN with LSTM for Video Deepfake Detection,” *Applied Sciences*, vol. 14, no. 21, p. 9754, 2024.