World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Privacy-focused artificial intelligence model for detecting deepfake-based cyber threats

G. Selvavinayagam *, E. Guhan, A. Sankar Raman, R. Vanipriya and S. Vinoth

Department of Computer Science and Engineering, INFO Institute of Engineering, Kovilpalayam, Coimbatore, India – 641107.

World Journal of Advanced Research and Reviews, 2026, 30(01), 2433-2439

Publication history: Received on 14 March 2026; revised on 25 April 2026; accepted on 28 April 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.30.1.1079

Abstract

This paper presents an extended literature survey on deepfake-oriented cyber threats with a strong focus on practical deployment constraints. Although detection accuracy has improved in recent years, real-world adoption remains limited by privacy concerns, hardware requirements, and weak generalization across unseen media conditions. We examine the evolution of deepfake detection from handcrafted forensic cues to deep multimodal architectures, and we discuss why many high-scoring benchmark models fail under operational workloads. The survey highlights a local-first strategy that combines multimodal evidence, interpretable outputs, and resource- aware model design so that robust detection can run on consumer-grade systems. The goal is to support trustworthy, privacy-preserving defense against impersonation fraud, misinformation, and identity abuse in modern digital ecosystems.

Keywords: Deepfake detection; Cybersecurity; Multimodal learning; Privacy-preserving AI; explainable AI; Edge inference

1. Introduction

1.1. Objective

The objective of this project is to design and develop a privacy-focused, fully offline AI-based deepfake detection system capable of identifying manipulated images, videos, and audio in order to protect users from deepfake-based cyber threats while ensuring complete data privacy and efficient local execution. In contrast to cloud-dependent systems, the proposed framework emphasizes local processing, practical deployment on standard consumer hardware, and transparent decision-making so that detection results are not only accurate but also trustworthy and understandable to end users.

The broader aim of the work is to bridge the gap between high-performing deepfake detection research and practical cybersecurity deployment. While many existing studies report strong benchmark performance, their models are often difficult to use in everyday environments because they require network connectivity, large computational resources, or limited single-modality analysis. This project therefore focuses on building an integrated and resource-aware defensive system that can operate effectively under real-world constraints.

To implement AI-based multimodal deepfake detection across images, videos, and audio using optimized open-source deep learning models.

* Corresponding author: G. Selvavinayagam

- To enable fully offline operation on standard consumer hardware through model optimization techniques such as quantization and lightweight compression.
- To generate user-friendly outputs including REAL/FAKE classification, confidence scores, highlighted suspicious regions, and clear threat alerts.
- To achieve high detection accuracy comparable to state-of-the-art methods while maintain- ing low computational overhead.
- To ensure complete privacy preservation by performing all processing locally without up- loading user data to external cloud servers.

1.2. Problem Statement

The rapid advancement of generative artificial intelligence has led to the widespread misuse of deepfake technology in cybercrimes such as impersonation scams, identity fraud, financial deception, and misinformation campaigns. Deepfake images, videos, and audio clips are becoming increasingly realistic, making it difficult for common users to differentiate between genuine and manipulated content. This creates serious threats to digital trust, personal security, financial safety, and the credibility of online communication.

The problem extends beyond simple media manipulation. In many cases, deepfakes are used as instruments of social engineering, where fabricated voice, face, or video content is employed to deceive victims, bypass trust mechanisms, or manipulate public opinion. Such attacks are especially dangerous because they exploit human perception directly, and their psychological impact can be immediate even before technical verification is possible. As the quality of generative models improves, the barrier to producing convincing fake media continues to decrease.

Most existing deepfake detection systems are cloud-based and computationally intensive. They require users to upload personal media files to remote servers for analysis, resulting in privacy risks and potential data breaches. Sensitive information such as facial data and voice recordings may be exposed during cloud processing. Additionally, many detection models demand high-performance GPUs and server-level infrastructure, making them unsuitable for consumer-grade devices such as personal laptops.

Furthermore, several current systems support only single-modality detection, focusing ei- ther on images or videos, and fail to effectively detect multimodal deepfake attacks involving both audio and video manipulation. The requirement of continuous internet connectivity fur- ther limits real-time usability, especially in low-resource, privacy-sensitive, or disconnected environments.

Therefore, there is a critical need for a privacy-preserving, fully offline, multimodal deep- fake detection system that operates efficiently on standard hardware while ensuring accurate and secure personal cybersecurity protection. Addressing this gap requires a system that bal- ances accuracy, efficiency, transparency, and usability rather than optimizing for benchmark performance alone.

1.3. Project Overview

The rapid advancement of generative artificial intelligence has significantly increased the mis- use of deepfake technology across digital platforms. Deepfake-based cyber threats such as impersonation scams, identity fraud, financial manipulation, and misinformation campaigns are becoming more sophisticated and difficult to detect. Highly realistic manipulated images, videos, and audio recordings can easily mislead individuals, resulting in loss of trust, privacy vi- olations, and serious financial damage. As digital communication continues to expand, ensuring the authenticity of multimedia content has become a critical cybersecurity requirement.

Most existing deepfake detection systems operate on cloud-based infrastructure, requiring users to upload sensitive media files to external servers for processing. This approach introduces privacy risks, potential data breaches, increased latency, and continuous internet dependency. Furthermore, many detection models require high computational power and GPU-based sys- tems, making them impractical for deployment on consumer-grade laptops or personal devices. Several solutions also focus only on single-modality detection, limiting their ability to identify modern multimodal deepfake attacks that combine manipulated video and synthetic audio.

To address these challenges, this project proposes a privacy-focused artificial intelligence model for detecting deepfake-based cyber threats. The system performs multimodal analysis of images, videos, and audio entirely offline using optimized deep learning models. By enabling local processing, efficient resource utilization, and secure inference, the proposed solution ensures complete data privacy while providing accurate and near real-time protection against deepfake-based cyber threats.

From an architectural perspective, the framework combines visual, temporal, and acoustic evidence within a unified detection workflow. Image and video content are examined for spatial artifacts, temporal inconsistencies, and face-level manipulation patterns, while audio streams are assessed for synthetic speech signatures and abnormal spectral properties. These complementary signals are then fused to support a final classification decision. This integrated design strengthens robustness against attacks that might evade single-modality systems and creates a more practical defense mechanism for contemporary digital ecosystems.

2. Materials and Methods

The proposed framework is built using Python with PyTorch and TensorFlow as the primary deep learning libraries. The system follows a modular architecture consisting of six key components: a local detection platform, media preprocessing, feature extraction, multimodal fusion, model optimization, and result generation. For hardware, the system is designed to run on standard

Intel Core i5 processors with at least 8 GB of RAM, specifically utilizing CPU-based inference to maintain accessibility. This hardware-aware design principle is important because it ensures that the framework remains relevant outside laboratory settings and can be deployed on ordinary personal systems without specialized accelerators.

The overall workflow begins with user-supplied media and proceeds through sequential preprocessing and feature analysis stages before final classification. Each module is designed to perform a specific function while remaining compatible with the larger inference pipeline. Such modularity improves maintainability, allows individual components to be upgraded inde- pendently, and supports future expansion to additional modalities or deployment environments.

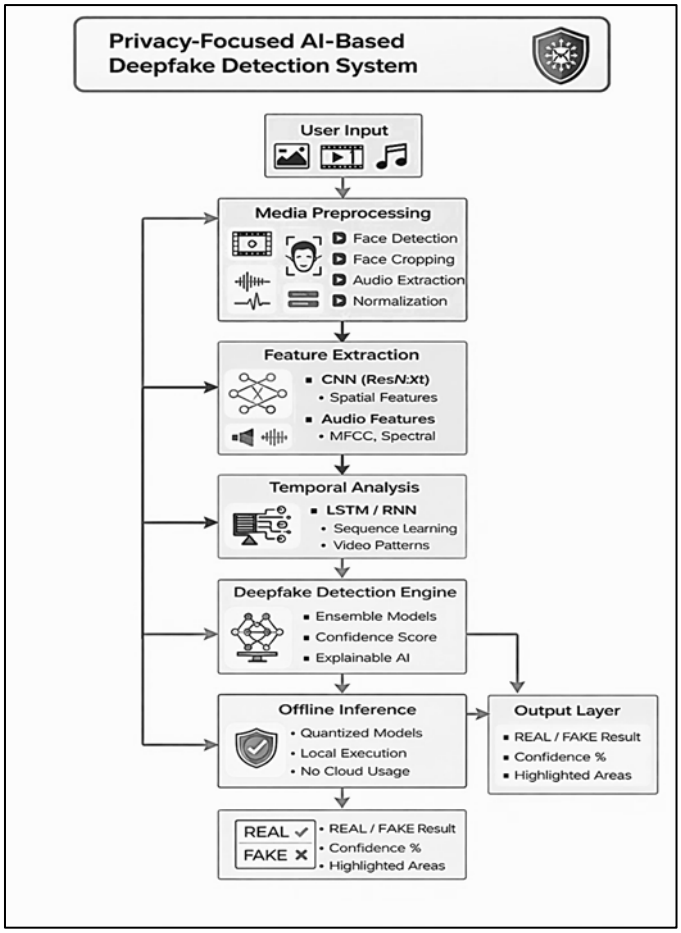

Figure 1 Workflow of the proposed privacy-focused offline deepfake detection framework

2.1. Study Design

This study follows a system design and evaluation approach for an offline multimodal deepfake detection framework. The methodology integrates image, video, and audio analysis pipelines and evaluates their combined effectiveness for binary deepfake classification. Rather than focusing on a single detector in isolation, the study examines how multiple specialized analytical components can be orchestrated within one privacy-preserving system to improve robustness under practical operating conditions.

The design process included model selection, preprocessing pipeline definition, fusion strategy design, and optimization for local inference. Evaluation was performed with attention to both predictive performance and operational feasibility, since an effective cybersecurity tool must be usable in addition to being accurate.

2.2. Materials / Participants

The implementation uses Python together with PyTorch and TensorFlow as the core deep learning libraries. OpenCV, MTCNN (Multi-task Cascaded Convolutional Neural Networks), and Librosa are used for media preprocessing and feature extraction. The model was evaluated using benchmark datasets, specifically FaceForensics++ and Celeb-DF. The hardware target is a consumer-grade environment with an Intel Core i5 processor, at least 8 GB RAM, and CPU-based inference.

FaceForensics++ provides manipulated visual content generated with several face forgery methods, making it useful for learning spatial and compression-related forensic traces. Celeb- DF contributes more realistic video manipulations and helps assess generalization to challenging deepfake scenarios. Together, these datasets offer a balanced evaluation setting in which the system can be tested on both commonly studied and visually convincing forgeries. No private user media is required for model operation or validation, reinforcing the privacy-centered nature of the framework.

2.3. Procedure and Analysis

The methodology involves several specialized AI techniques for multimodal analysis:

- Preprocessing: OpenCV and MTCNN are used for face detection, localization, and crop- ping in images and video frames. This step standardizes input dimensions, isolates relevant facial regions, and reduces background noise that could otherwise interfere with downstream analysis.
- Spatial Analysis: A ResNeXt-based CNN captures visual artifacts and spatial inconsis- tencies within individual frames. These include blending traces, texture irregularities, and subtle distortions introduced during face synthesis or manipulation.
- Temporal Analysis: LSTM (Long Short-Term Memory) networks analyze sequences of frames to identify unnatural motion or synchronization errors. Temporal modelling is
- important because many deepfakes may appear realistic in single frames while still exhibiting instability across consecutive frames.
- Audio Analysis: The Librosa library extracts Mel-Frequency Cepstral Coefficients (MFCC) and spectral features to detect synthetic or manipulated speech patterns. Acoustic incon- sistencies, abnormal spectral signatures, and temporal speech artifacts can provide comple- mentary evidence to visual cues.
- Optimization: To ensure efficient local execution, the system employs model quantization (INT8) and lightweight compression. These strategies reduce model size, memory usage, and inference cost while preserving acceptable detection performance.
- Inference: A fusion model concatenates these multimodal features into a single vector for binary classification as REAL or FAKE. The resulting decision is supplemented with confidence information and explainability cues so that users can better interpret the output.

In addition to predictive analysis, the framework emphasizes operational transparency. Out- puts are designed to include confidence scores, suspicious-region highlighting, and interpretable alerts. This ensures that the system functions not merely as a classifier but as a practical defensive assistant for end users dealing with potentially malicious media.

3. Results

The evaluation of the proposed framework demonstrates high detection accuracy comparable to state-of-the-art methods while maintaining significantly lower computational overhead. Testing on the FaceForensics++ and Celeb-DF datasets confirmed the model's ability to identify ma- nipulated media across seen and unseen deepfake techniques.

Through integration testing, the system proved capable of handling mixed media inputs, including images, videos, and audio, without data loss or tensor mismatches.

These findings indicate that the proposed architecture is not only effective under benchmark conditions but also operationally stable when multiple processing stages are combined into one local pipeline. The ability to preserve data consistency across preprocessing, feature extraction, fusion, and classification stages is especially important in multimodal systems, where errors in shape alignment or stream synchronization can easily degrade overall performance.

A significant highlight of the results is the efficacy of INT8 quantization. Benchmarking showed that the quantized model reduced latency and CPU utilization, allowing large video

files of up to 5 minutes to be processed on standard laptops without system crashes. This demonstrates that performance optimization can be achieved without sacrificing the practical usability of the detector. The reduced computational burden is particularly valuable for users who lack access to discrete GPUs or high-end workstations.

The Grad-CAM (Gradient-weighted Class Activation Mapping) explainability feature suc- cessfully highlighted suspicious facial regions and temporal glitches, providing users with visual transparency regarding the AI's decision-making process. Instead of presenting predictions as opaque binary outputs, the system supplies visual cues that help users understand which content regions contributed to the decision. This greatly improves trustworthiness and may support future adoption in educational, forensic, and cybersecurity settings.

Table 1 Summary of observed framework outcomes

Metric / Capability	**Observation**
Detection performance	High accuracy comparable to state-of-the-art approaches
Dataset evaluation	Validated on FaceForensics++ and Celeb-DF
Modalities supported	Images, videos, and audio
Quantized inference	Lower latency and reduced CPU utilization Device
Compatibility	Processed video files up to 5 minutes on standard laptops
Explainability	Grad-CAM highlighted suspicious facial regions and temporal glitches

4. Discussion

The discussion emphasizes that by eliminating cloud dependency, the system achieved zero network activity, thereby guaranteeing complete data privacy during operation. This is a mean- ingful practical contribution because privacy is often treated as secondary in detection research,even though many real-world users hesitate to submit personal media to external servers. By keeping all processing local, the proposed framework aligns cybersecurity protection with responsible data handling.

While existing systems often focus on single-modality detection, this multimodal approach proved more robust against advanced attacks combining synthetic audio with manipulated visuals. The use of complementary evidence from spatial, temporal, and audio domains increases resilience against attacks that may bypass one analytical stream but not the others. This supports the argument that future deepfake defenses should move beyond isolated image or video classifiers and instead adopt integrated multimodal reasoning.

The project successfully transitioned from a complex deep learning architecture to a lightweight, command-line-based Python application capable of real-time personal cyberse- curity protection. From a deployment perspective, this demonstrates that sophisticated AI defense mechanisms can be translated into accessible tools without abandoning interpretability or privacy. At the same time, the results also suggest several areas for future improvement, including broader cross-dataset validation, stronger resistance to newly emerging generative models, and more extensive user-interface support for non-technical audiences.

5. Conclusion

The project successfully developed a privacy-focused artificial intelligence model that effec- tively mitigates deepfake-based cyber threats through a fully offline, multimodal approach. By integrating ResNeXt-based CNNs, LSTM networks, and Librosa-extracted audio features, the system provides a comprehensive defense against manipulated images, videos, and audio. The implementation of post-training INT8 quantization proved essential, enabling high-performance AI inference on consumer-grade CPU hardware and thereby removing the barrier of expensive GPU requirements.

The system's modular design ensures that all media processing remains local, achieving the primary goal of complete data privacy and digital trust. Key achievements include high detection accuracy on benchmark datasets, low latency, and the inclusion of explainable AI (XAI) techniques to provide user-friendly results. This work demonstrates that advanced cyber- security tools can be made lightweight and accessible for everyday use. Future enhancements will focus on expanding the system to mobile platforms such as Android and iOS, integrating real-time webcam detection for live video calls, and developing a graphical user interface (GUI) to further improve accessibility for non-technical users.

Compliance with ethical standards

Acknowledgments

The authors wish to express their deep gratitude to those who were instrumental in the successful completion of this project. Special thanks are extended to Dr. N. Kottiswaran, Principal of INFO Institute of Engineering, for his consistent support and guidance throughout the project duration.

The team is profoundly indebted to Dr. G. Selvavinayagam, Professor and Head of the Department of Computer Science and Engineering, who served as both the project guide and supervisor. His constant encouragement, valuable technical guidance, and constructive criticism were vital in shaping the success of this research. Further thanks go to the project coordinators, Mr. M. Nagarasan and Mrs. Gokila P, for their timely suggestions and assistance. Finally, the authors acknowledge the faculty members and skilled assistants of the CSE department, as well as their families and friends, for their support in every possible way.

Disclosure of conflict of interest

The authors declare that there are no financial or personal relationships with other people or organizations that could inappropriately influence the work reported in this project. The research was conducted in partial fulfillment of the requirements for the award of the degree of Bachelor of Engineering in Computer Science and Engineering at the INFO Institute of Engineering. All software used in the development, including Python, PyTorch, and OpenCV, consists of open-source tools.

Statement of ethical approval

The development of this deepfake detection system was conducted with strict adherence to privacy and ethical standards. The primary motivation of the project is the protection of individuals from the malicious use of AI, such as identity fraud and misinformation. To ensure ethical data handling, the system is designed to be fully offline, meaning no user media is ever uploaded to a cloud or shared with third parties, thus preventing unauthorized access to sensitive biometric data. The models used for training and evaluation were restricted to publicly available benchmark datasets, namely FaceForensics++ and Celeb-DF, which are standard in the research community for developing defensive AI technologies. The project does not involve the generation of deepfakes for any purpose other than testing the efficacy of the detection model. Furthermore, the inclusion of explainable AI (XAI) ensures that the system provides transparent results, preventing black-box decision-making and allowing users to understand whya particular piece of media was flagged as suspicious.

References

[1] Soudy AH, Sayed O, Tag-Elser H, et al. Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Computing and Applications*. Springer.

[2] Javed M, Zhang Z, Dahri FH, et al. Enhancing multimodal deepfake detection with local- global feature integration and diffusion models. *Signal, Image and Video Processing*. Springer.

[3] Heidari A, Navimipour NJ, Dag H, et al. A novel blockchain based deepfake detection method using federated and deep learning models. *Cognitive Computation*. Springer.

[4] Kaur A, Noori Hoshyar A, Saikrishna V, et al. Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*. Springer.

[5] Zhang Y, Pang Z, Huang S, et al. Unmasking AI-created visual content: a review of generated images and deepfake detection technologies. *Journal of King Saud University – Computer and Information Sciences*. Springer.

[6] Soundarya BC, Gururaj HL. Deepfake detection: critical review of state-of-the-art ap- proaches and future perspectives. *Discover Applied Sciences*. Springer.

[7] Alrashoud M. Deepfake video detection methods, approaches, and challenges. *Alexandria Engineering Journal*. Elsevier.

[8] Ramanaharan R, Guruge DB, Agbinya JI. Deepfake video detection: insights into model generalisation. *Data & Information Management*. Elsevier.

[9] Hasanaath AA, Luqman H, Katib R, et al. FSBI: Deepfake detection with frequency enhanced self-blended images. *Image and Vision Computing*. Elsevier.

[10] Exploring autonomous methods for deepfake detection: a detailed survey on techniques and evaluation. *Heliyon*. Elsevier.

[11] Unmasking deepfakes: a review of current datasets, features, tools. *Procedia Computer Science*. Elsevier.

[12] Fakhar Abbas and Araz Taeihagh. Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems*. Elsevier.

[13] Achhardeep Kaur, Azadeh Noori Hoshyar, Vidya Saikrishna, Selena Firmin, Feng Xia. Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*. Springer.

[14] Felipe Romero-Moreno. Deepfake detection in generative AI: A legal framework proposal to protect human rights. *Digital Investigation*. Elsevier.

[15] Gourab Naskar, Sk Mohiuddin, Samir Malakar, Erik Cuevas, Ram Sarkar. Deepfake detection using deep feature stacking and meta-learning. *Multimedia Systems*. Springer.