WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Defendai Security Suite – An AI-Powered Deepfake Detection and Anti-AI Poisoning Framework

[1]Nishanth S, [2]Sudeepa, [3] Prof Pamela Bera

7th Semester, Dept. of CSE-AIML Engineering.
[1,3] Student, AMC Engineering College, Bangalore – 83, KARNATAKA, INDIA.
[2]Prof., Dept. of CSE-AIML, AMC Engineering College, Bangalore – 83, KARNATAKA, INDIA

ABSTRACT

Deepfakes and unauthorized AI-generated image manipulations are rapidly increasing, creating serious risks related to privacy violation, identity theft, misinformation, and the misuse of personal images in malicious AI pipelines. Traditional detection methods often struggle with accuracy, lack transparency, and fail when exposed to real-world distortions such as compression, cropping, and resizing. DefendAI addresses these challenges by introducing an intelligent, end-to-end framework that combines deepfake detection, forensic explainability, and proactive image protection. The system analyzes images using deep learning and frequency-based forensic cues, generates manipulation heatmaps for transparency, and provides a confidence- based verdict on authenticity. Beyond detection, DefendAI applies imperceptible adversarial perturbations (FGSM/PGD-based) to safeguard user images from being reused in AI training or deepfake generation, ensuring strong resistance even after compression or resizing. This paper presents the system's architecture, implementation methodology, advantages, and limitations while showcasing its contribution to secure digital media authentication. Additionally, the study highlights how integrating cloud deployment, transformers, and scalable APIs strengthens reliability, performance, and robustness against evolving generative threats.

Keywords: Deepfake Detection, Adversarial Defense, AI Security, Image Forensics, CNN, FGSM, PGD, Anti-AI Poisoning, Cyber-Forensics, Explainable AI.

I. INTRODUCTION

In today's digital world, protecting personal images and verifying media authenticity have become critical challenges. Traditional manual inspection, watermarking, and basic detection tools remain unreliable, slow, and easily bypassed by modern generative AI models. These outdated methods lack forensic insight, fail under compression or cropping, and do not provide strong defenses against deepfake misuse. This study introduces DefendAI, an AI-powered security framework designed to overcome these limitations through intelligent deepfake detection and proactive image protection. The system analyzes an uploaded face image, detects manipulation traces, generates forensic heatmaps, and produces an authenticity verdict with confidence scoring. DefendAI improves reliability, transparency, and security by combining CNN-based detection, frequency analysis, explainable Grad-CAM visualization, and adversarial defense mechanisms. Beyond detection, the platform incorporates modules for protective poisoning, digital identity security, usage tracking, dataset integrity checks, and API-based verification for third-party apps. To enhance scalability and availability, this study also explores cloud deployment, containerized microservices, and real-time APIs that support fast processing, distributed storage, and secure digital media authentication.

II. SYSTEM ARCHITECTURE

- **Image Input & Acquisition**

Users upload a face image or any photographic content through the web interface, enabling both individual authenticity checks and batch analysis for larger datasets.

- **Preprocessing & Forensic Feature Extraction**

Using OpenCV, NumPy, and image normalization pipelines, the system enhances image quality, detects regions of interest, and prepares inputs for deepfake detection under varying compression levels, lighting conditions, and resolutions.

- **Deepfake Detection Engine**

CNN-based and forensic-aware deep learning models generate embeddings, detect manipulation artifacts, and classify the image as real or fake. The system also produces confidence scores and forensic heatmaps to highlight suspected tampered regions.

- **Adversarial Poisoning & Protection Module**

FGSM/PGD-driven perturbation methods apply imperceptible protective noise to secure real images against unauthorized deepfake generation, AI training misuse, and identity cloning on third-party generative systems.

- **Database & Cloud Services**

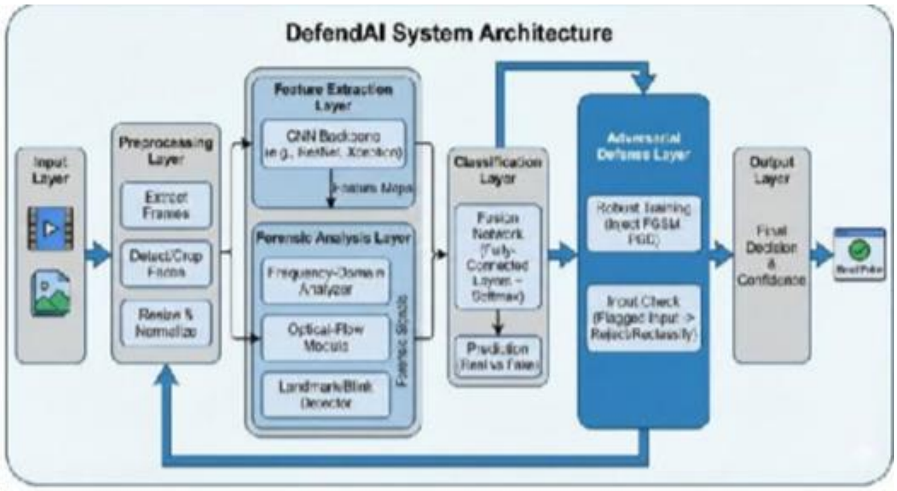
Uploaded images, detection results, poisoning logs, and usage history are securely stored in PostgreSQL or SQLite, with optional cloud integration for scalable storage and remote model inference.

- **User Interface (Frontend Dashboard)**

Security analysts, developers, and users access a React + TailwindCSS dashboard to upload images, view detection results, inspect heatmaps, download poisoned images, and monitor system activity through real-time analytics.

- **Authentication & System Security**

JWT-based authentication, encrypted storage, rate limiting, and strict permission control protect the platform against unauthorized usage, data breaches, and misuse of the detection/poisoning APIs.


DefendAI System Architecture
Input Layer
Preprocessing Layer
Extract Frames
Detect/Crop Faces
Resize & Normalize
Feature Extraction Layer
CNN Backbone (e.g., ResNet, Xception)
Feature Maps
Forensic Analysis Layer
Frequency-Domain Analyzer
Optical-Flow Module
Landmark/Blink Detector
Forensic Signals
Classification Layer
Fusion Network (Fully-Connected Layers - Softmax)
Prediction (Real vs Fake)
Adversarial Defense Layer
Robust Training (Inject FGSM PGD)
Input Check (Flagged Input -> Reject/Reclassify)
Output Layer
Final Decision & Confidence
Real/Fake

III. IMPLEMENTATION AND TECHNOLOGY STACK

Python is used for backend development and AI processing modules. OpenCV, Dlib, NumPy, and Deep Learning models are used for facial detection, recognition, and image

Extend the system to support full video analysis with AI- based frame monitoring, manipulation detection, and automated flagging of suspicious segments.

- **AR/VR Forensic Visualization**

Integrate AR/VR modules to help analysts visualize manipulation regions, heatmaps, and forensic traces through interactive 3D models for enhanced understanding.

- **AI Voice Assistant for Forensic Operations** Introduce a voice-enabled assistant that can explain results, guide users, retrieve logs, answer queries, and support hands-free forensic investigation.

- **Digital Identity and Secure Access Integration**

Use advanced face verification and protected digital IDs to control access, prevent misuse, and secure sensitive detection and poisoning operations.

- **Automated Forensic Report Generation**

Enable automatic generation of authenticity certificates, analysis summaries, poisoning reports, and audit logs based on stored system data

4. CONCLUSION

DefendAI presents a comprehensive and intelligent solution for detecting manipulated media and protecting users from deepfake-related threats. By integrating AI- driven forensic analysis, manipulation detection, adversarial poisoning defense, confidence scoring, and heatmap-based explainability, the system ensures stronger accuracy, transparency, and security compared to traditional manual verification methods. The platform reduces human effort, prevents misuse of personal images, and enhances digital trust for all users.

With its multi-module design—including deepfake detection, protective perturbations, authenticity logs, API verification, and analytics—DefendAI functions as a unified security ecosystem for modern digital environments. The incorporation of cloud deployment, scalable microservices, and secure storage further increases the system's reliability and flexibility. Future improvements such as advanced video detection, blockchain verification, and predictive threat analysis will continue to broaden its capabilities. As generative AI evolves, DefendAI has the potential to transform digital media security entirely, offering a robust, secure, and highly effective alternative to conventional forensic and identity protection systems.

ACKNOWLEDGMENT

We express our heartfelt gratitude to the Department of Computer Science and Engineering at AMC Engineering College for their continuous support, guidance, and motivation throughout the development of this project. We are deeply thankful to our faculty mentors, whose valuable insights, constructive feedback, and technical suggestions greatly helped shape the features and effectiveness of the DefendAI system. We also extend our appreciation to the students, friends, and staff members who participated in testing and shared practical feedback during various stages of development. Their cooperation and involvement played a meaningful role in refining the project and ensuring its real-world applicability.

REFERENCE

1. Zhang, Y. (2020). Applications of Deep Learning in Facial Recognition. AI Research Journal.

2. Patel, R. (2019). Utilizing Computer Vision for Automated Attendance Tracking. International Journal of Computer Applications.

3. Khan, S. (2021). Facial Detection Techniques for Real- Time Classroom Monitoring. IEEE Transactions on Artificial Intelligence.

4. Brown, A. (2022). Cloud-Based Solutions for Managing Attendance Systems. Journal of Emerging Technology Trends.

5. Sharma, R. (2023). Utilization of IoT Technologies to Enhance Educational Infrastructure. Presented at the International Smart Technologies Conference