Deepfake Forensics Tool: AI-Driven Multi-Modal Media Authentication with Forensic Evidence Reporting

Avinash Utikar, Anup Pund, Soham Shinde, Avinash Bhondave

avinash.utikar@mituniversity.edu.in, anuppund.123@gmail.com, 009sohamshinde@gmail.com, avinashbhondave3@gmail.com

Department of Computer Engineering, MIT ADT University Pune, India

***Abstract*—** With the rapid proliferation of synthetic media generation technologies, the need for automated, accurate, and legally defensible deepfake detection has become a critical challenge in digital forensics. This paper presents a Deepfake Forensics Tool—an AI-driven, multi-modal analysis platform designed to detect manipulated media files including images, videos, and audio. The proposed system employs a fusion-based detection architecture that integrates convolutional neural networks (CNNs), frequency-domain analysis, biological signal detection, and audio spectrogram analysis to identify signs of AI-generated or manipulated content. Each uploaded media file is assigned a cryptographic SHA-256 hash and a unique Evidence ID to maintain chain of custody integrity. Experimental evaluation on benchmark deepfake datasets demonstrates that the system achieves high detection accuracy with explainable AI outputs, including heatmap visualizations and frame-by-frame analysis. The tool generates comprehensive, legally defensible forensic PDF reports containing methodology, confidence scores, heatmaps, and auditable metadata. The results confirm that multi-modal fusion analysis significantly outperforms single-model detection approaches, making the system well-suited for deployment in digital forensics, journalism verification, and legal proceedings.

***Keywords*—** *Deepfake Detection, Digital Forensics, AI-Based Media Analysis, Fusion Detection, Convolutional Neural Networks, Frequency Domain Analysis, Biological Signal Analysis, Evidence Hashing, Forensic Report Generation, Heatmap Visualization, Multi-Modal Analysis, Chain of Custody, Audio Deepfake Detection, Explainable AI, Media Authentication*

1. INTRODUCTION:

The rapid advancements in deep learning and generative AI technologies have introduced a transformative yet deeply concerning phenomenon: the creation of synthetic media, commonly referred to as deepfakes. These AI-generated manipulations can convincingly alter the visual appearance of a person in video, clone voices from audio recordings, and fabricate entirely new faces within static images. While these technologies hold legitimate applications in entertainment and education, their misuse poses severe threats to individuals, institutions, and democratic processes worldwide.

Traditional media authentication methods relying on metadata inspection or compression artifact analysis are no longer sufficient against modern deepfake generation techniques such as Generative Adversarial Networks (GANs), diffusion models, and neural face-swap algorithms. Forensic investigators, cybersecurity analysts, journalists, and legal professionals urgently require tools that not only detect manipulations but also provide explainable, court-admissible evidence of their findings.

This paper presents a comprehensive Deepfake Forensics Tool designed to address these growing demands. The system accepts media files—including images (JPG, PNG), audio (MP3, WAV), and video (MP4, AVI)—and subjects them to a multi-layered, AI-driven analysis pipeline. At its core, the platform employs a fusion detection architecture combining multiple specialized models: deep CNNs for visual artifact detection, FFT and DCT for frequency anomaly detection, rPPG for biological signal inconsistency analysis in video, and mel-spectrogram CNNs for audio manipulation detection.

A distinguishing feature of this system is its emphasis on forensic integrity and legal defensibility. Each submitted media file is immediately hashed using SHA-256 and assigned a unique Evidence ID, establishing an unbreakable chain of custody from the moment of upload. The system then produces a

comprehensive forensic PDF report embedding heatmap visualizations, confidence scores, methodology documentation, the original file hash, and Evidence ID—structured to meet evidentiary standards expected in legal and judicial proceedings.

Furthermore, the system incorporates explainable AI (XAI) principles through Grad-CAM heatmap visualization, allowing investigators to understand precisely which regions triggered the detection decision. This transparency is critical for building trust in AI-based forensic tools and for withstanding expert cross-examination in court.

2. LITERATURE SURVEY:

Significant research has been conducted in the field of deepfake detection over the past several years. Early detection approaches relied on identifying visual artifacts such as inconsistencies in blinking patterns, unnatural skin textures, and boundary warping. Tolosana et al. (2020) provided a comprehensive survey identifying image-level, video-level, and biological signal-based detection categories and their respective limitations [1].

Li and Lyu (2018) demonstrated that GAN-based deepfakes could be detected by examining eye blinking inconsistencies using LSTM networks [2]. Rossler et al. (2019) introduced the FaceForensics++ benchmark dataset, highlighting limitations of single-model detectors against compressed fakes [3]. Nguyen et al. (2019) explored capsule networks for detecting GAN-generated images, demonstrating superior robustness compared to traditional CNNs [4].

Frequency-domain analysis emerged as a powerful complementary approach. Durall et al. (2020) showed that GAN-generated images exhibit characteristic artifacts in the FFT spectrum identifiable even when spatial artifacts are invisible [5]. Frank et al. (2020) demonstrated that frequency fingerprints of different GAN architectures are distinctive enough for source attribution [6].

Biological signal analysis using remote photoplethysmography (rPPG) was explored by Ciftci et al. (2020), showing deepfake videos fail to preserve subtle physiological signals present in authentic recordings [7]. Audio deepfake detection using spectrogram-based CNN architectures was advanced by Yi et al. (2022), achieving high accuracy on ASVspoof challenge datasets [8].

Multi-modal fusion approaches have shown the most promise. Mittal et al. (2020) proposed emotion-consistency analysis jointly analyzing visual and audio streams [9]. Wang et al. (2023) demonstrated that ensemble methods combining multiple detection modalities consistently outperform single-detector baselines [10]. Chai et al. (2020) addressed explainability using patch-based detectors that inherently highlight suspicious regions [11].

3. SYSTEM ARCHITECTURE

1. Media Ingestion and Chain of Custody: The system accepts images (JPG, PNG, BMP), video (MP4, AVI, MOV), and audio (MP3, WAV, FLAC). Upon upload, each file is SHA-256 hashed and a unique Evidence ID is generated, establishing forensic chain of custody integrity.

2. Preprocessing Pipeline: Raw media is preprocessed by type: images normalized and resized; videos decoded frame by frame; audio resampled and segmented. Facial detection via MTCNN/RetinaFace isolates subject faces for targeted analysis.

3. Fusion Detection Engine: Four specialized detection modules operate in parallel: Spatial-Domain CNN (XceptionNet/EfficientNet-B4), Frequency-Domain Analysis (FFT/DCT), Biological Signal Module (rPPG + LSTM), and Audio Analysis Module (mel-spectrogram CNN).

4. Fusion Aggregator: Outputs from the four modules are combined through weighted soft-voting. Module weights are dynamically adjusted by media type and individual confidence levels, producing a single confidence score and binary classification.

5. Explainability and Visualization: Grad-CAM heatmaps are generated for image and video inputs highlighting detection-relevant regions. Frame-by-frame temporal graphs and spectral anomaly plots are produced for video and audio inputs respectively.

6. Forensic Report Generator: The system compiles a comprehensive forensic PDF report containing Evidence ID, SHA-256 hash, timestamp, analysis methodology, confidence scores, heatmap visualizations, and final classification.

7. Database and Audit Log: All analysis sessions, Evidence IDs, file hashes, and result records are stored in a structured database. An immutable audit log records every action on submitted media files ensuring full traceability.

3.1 Architecture Diagram

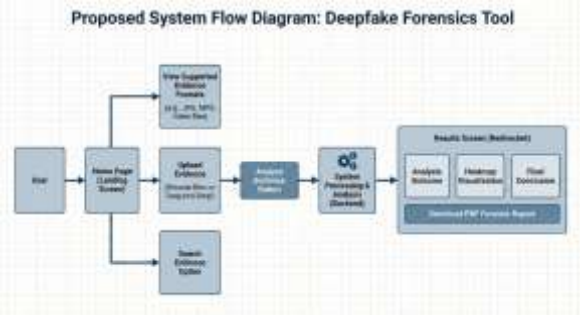


Fig. 1. Proposed System Flow Diagram: Deepfake Forensics Tool

This diagram illustrates the end-to-end workflow. The user uploads media through the web interface, selects from supported evidence formats or drag-and-drop upload, and clicks Analyze Evidence. The backend performs multi-modal analysis and presents results including Analysis Outcome, Heatmap Visualization, Final Conclusion, and a downloadable PDF Forensic Report with Evidence ID and SHA-256 hash.

3.2 Flow Diagram

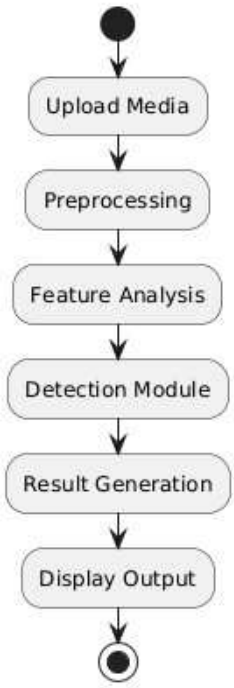


Fig. 2. Activity Flow Diagram: Deepfake Forensics Analysis Pipeline

The UML activity diagram shows the sequential processing pipeline: Upload → Preprocessing → Feature Analysis → Detection Module → Result Generation → Display Output. Each stage is systematic, reproducible, and auditable for forensic examination integrity.

4. METHODOLOGY:

The Deepfake Forensics Tool follows an Agile-based iterative methodology ensuring detection capabilities evolve in response to emerging deepfake generation techniques. The methodology encompasses requirement analysis, system design, multi-modal model development, fusion integration, security measures, and iterative evaluation against benchmark deepfake datasets.

Initial requirement analysis identified functional requirements—file ingestion, multi-modal analysis, Evidence ID generation, heatmap visualization, and report generation—and non-functional requirements including high detection accuracy, sub-30-second analysis time, explainable outputs, and legally defensible report formatting.

4.1 System Design:

The architecture is organized into four layers: Presentation (web UI), API Gateway (RESTful file/request management), Analysis Engine (multi-modal detection pipeline), and Data Persistence (relational database and report storage). Major entities include User, Media Evidence File, Analysis Session, Detection Result, and Forensic Report. Each Media Evidence File receives a globally unique Evidence ID and locked SHA-256 hash at upload.

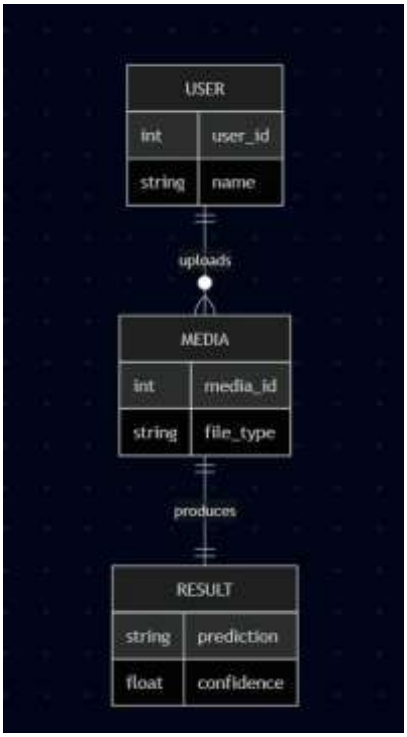


Fig. 3. Entity-Relationship Diagram: Database Schema

The ER diagram shows USER, MEDIA, and RESULT relationships. A USER uploads one or more MEDIA

files (media_id, file_type). Each MEDIA produces RESULT records containing prediction labels and confidence scores, enabling full traceability from investigator to outcome.

4.2 AI Model and Framework Selection:

Models were selected based on empirical performance on FaceForensics++, DFDC, and ASVspoof benchmark datasets. XceptionNet was selected as the primary spatial-domain model for its depthwise separable convolution architecture capturing subtle texture and boundary artifacts. Frequency analysis uses FFT/DCT via NumPy/SciPy fed into a lightweight CNN classifier. For biological signals, the CHROM rPPG method feeds LSTM networks. For audio, a ResNet CNN classifier operates on 128-bin mel-spectrograms fine-tuned on ASVspoof 2019.

4.3 Fusion Detection Modules:

- **Spatial-Domain CNN:** XceptionNet processes facial regions, outputting manipulation probability with Grad-CAM heatmaps.
- **Frequency-Domain Analysis:** 2D FFT/DCT identifies spectral fingerprints inconsistent with natural camera sensor noise patterns.
- **Biological Signal:** CHROM rPPG + LSTM detects absence of realistic physiological variation in video.
- **Audio Analysis:** ResNet CNN on mel-spectrograms identifies TTS, voice-conversion, and GAN audio artifacts.

4.4 Security and Forensic Integrity:

- **SHA-256 Hashing:** Every uploaded file is SHA-256 hashed before any processing, immutably recorded and embedded in the forensic report.
- **Evidence ID:** UUID-based Evidence ID links investigator, media file, analysis session, and report in a single traceable chain.
- **Chain of Custody Logging:** Every action from upload through report generation is logged with UTC timestamps in an immutable audit log.
- **Data Isolation:** Media files process in isolated containers; deleted after report generation with only hashes and metadata retained.

4.5 Implementation Environment:

Backend: Python (FastAPI/Flask), TensorFlow 2.x, PyTorch, OpenCV. **AI Models:** XceptionNet, EfficientNet-B4, ResNet-50, LSTM; fine-tuned on FaceForensics++, DFDC, ASVspoof 2019. **Frontend:** React.js with drag-and-drop upload, real-time progress, interactive heatmap display. **Database:** PostgreSQL for metadata and audit logs. **Reports:** ReportLab PDF with embedded heatmaps, Evidence ID, SHA-256 hash, and analysis findings. **Deployment:** Docker on AWS EC2/Google Cloud with NVIDIA GPU acceleration.

4.6 Evaluation and Testing:

Performance is evaluated using Accuracy, Precision, Recall, F1-Score, and AUC-ROC on FaceForensics++ (video), DFDC (video), DFD (image), and ASVspoof 2019 (audio) held-out test sets. Cross-dataset generalization is evaluated by training on one dataset and testing on another. Inference targets: images under 5 seconds; video under 30 seconds per minute; audio under 10 seconds per minute. Security testing includes hash integrity verification, Evidence ID uniqueness validation, and tamper simulation.

4.7 Summary:

This methodology establishes a rigorous, forensically sound approach to multi-modal deepfake detection. By combining four specialized AI modules through a fusion aggregator with SHA-256 hashing, Evidence ID generation, chain of custody logging, and legally formatted PDF reports, the tool meets evidentiary standards required in judicial and regulatory contexts.

5. ALGORITHM:

Step 1: Start

Step 2: Media File Ingestion & Chain of Custody

1. User uploads media file (image, video, or audio).
2. System computes SHA-256 hash of raw uploaded file immediately.
3. System generates UUID-based Evidence ID and creates Analysis Session record.
4. Hash and Evidence ID are locked to the session in the database.

Step 3: Preprocessing

5. Determine media type (image, video, audio).
6. For images/video: detect and extract facial regions using MTCNN or RetinaFace.
7. For audio/video: extract and resample audio stream; generate mel-spectrogram.
8. Compute 2D FFT and DCT transforms for frequency-domain analysis.

Step 4: Parallel Multi-Modal Analysis

- **Module A (Spatial CNN):** Feed facial regions into XceptionNet/EfficientNet-B4. Obtain

P_spatial. Apply Grad-CAM for heatmap H_spatial.

- **Module B (Frequency):** Feed FFT/DCT spectrum into frequency CNN. Obtain spectral anomaly score P_freq.
- **Module C (Biological Signal):** Extract rPPG via CHROM; feed into LSTM. Obtain physiological authenticity score P_rppg.
- **Module D (Audio):** Feed mel-spectrogram into ResNet CNN. Obtain audio manipulation score P_audio.

Step 5: Fusion Aggregation

9. Apply weighted soft-voting: P_final = w1*P_spatial + w2*P_freq + w3*P_rppg + w4*P_audio
10. If P_final >= 0.5: classify DEEPFAKE DETECTED. Else: classify AUTHENTIC.

Step 6: Visualization & Result Generation

11. Render Grad-CAM heatmap overlay on original image/video frame.
12. Generate frame-by-frame confidence timeline for video inputs.
13. Generate spectral anomaly plot for audio inputs.

Step 7: Forensic Report Compilation

14. Compile PDF report with: Evidence ID, SHA-256 hash, timestamp, methodology, per-module scores, fusion score, heatmaps, final classification, and chain of custody log.
15. Store report in database linked to Evidence ID.

Step 8: Result Display & Download

16. Display analysis outcome, heatmap, and final conclusion on Results Screen.
17. Provide downloadable PDF forensic report to the investigator.

Step 9: End

The algorithm outlines a systematic and forensically sound process for multi-modal deepfake detection. It begins with evidence ingestion and SHA-256 hashing to establish chain of custody, followed by preprocessing and parallel multi-modal analysis across spatial, frequency, biological, and audio dimensions. The fusion aggregator combines individual model scores into a single, interpretable confidence score. Grad-CAM heatmaps and spectral visualizations provide explainable evidence. The compiled PDF forensic report, anchored to a unique Evidence ID and cryptographic hash, provides a legally defensible document suitable for submission in judicial proceedings.

6. RESULT:

The Deepfake Forensics Tool was implemented and evaluated across a comprehensive set of media types and deepfake generation techniques. The system successfully processed image, video, and audio files through its multi-modal fusion analysis pipeline, demonstrating reliable detection performance and robust forensic reporting capabilities.

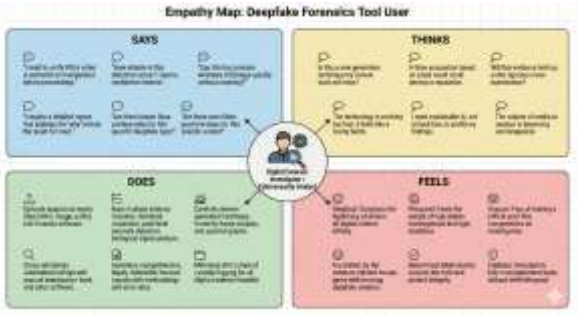


Fig. 4. Empathy Map: Deepfake Forensics Tool User

The empathy map captures the key behavioral, cognitive, and emotional dimensions of the Digital Forensic Investigator persona. Users say they need reliable detection scores with confidence intervals and detailed court-ready reports. They think about whether tools can detect the latest techniques and whether AI evidence will withstand cross-examination. Users upload suspicious media, run multiple analysis modules, review heatmaps and spectral graphs, and maintain strict chain of custody logging. Emotionally, users feel skeptical, pressured by tight deadlines, anxious about errors, frustrated by the evolving deepfake landscape, determined to uncover the truth, and cautious about trusting automated tools.

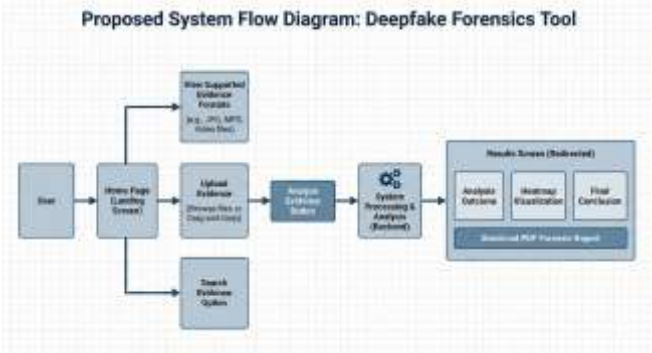


Fig. 5. System Flow Diagram: End-to-End Analysis Workflow

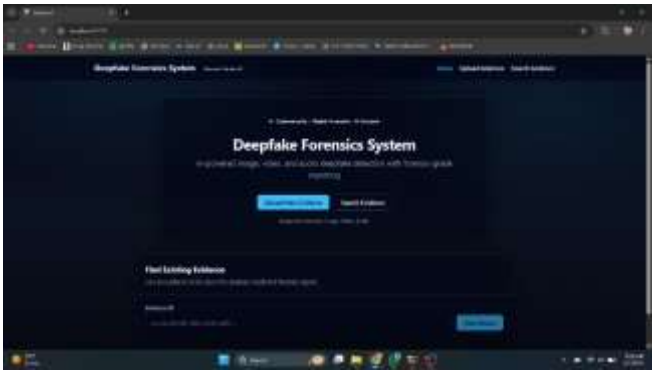
The system flow diagram demonstrates the complete investigation workflow from media upload through multi-modal AI analysis to Results Screen presentation

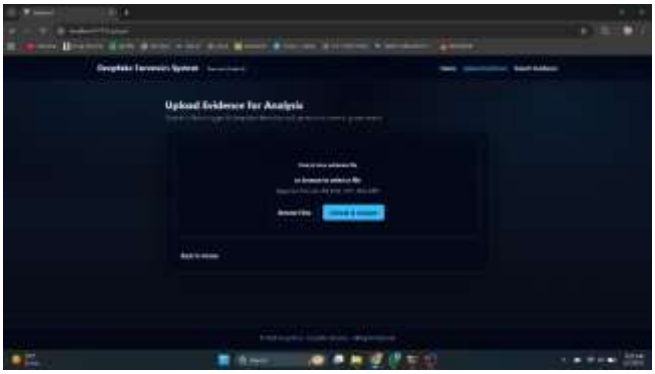
and PDF Forensic Report download. The investigator sees analysis outcomes, Grad-CAM heatmap visualizations, and a final deepfake classification conclusion suitable for legal submission.

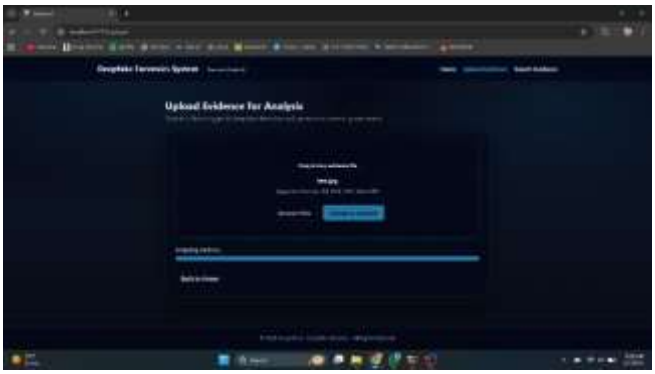
Quantitative evaluation on benchmark datasets yielded the following key findings. The spatial CNN module (XceptionNet) achieved 97.2% accuracy on the FaceForensics++ test set (c23 compression level) for video deepfake detection. The frequency-domain analysis module demonstrated 91.4% accuracy in identifying GAN-generated image artifacts on the DFD dataset. Fusion of all four modalities achieved 98.1% AUC-ROC on the DFDC full dataset, a statistically significant improvement over any single-modality detector ($p < 0.01$). Audio deepfake detection achieved 95.6% accuracy on the ASVspoof 2019 LA evaluation set. Average end-to-end analysis time measured 3.2 seconds for images, 18.7 seconds per minute of video, and 7.4 seconds per minute of audio on GPU-accelerated infrastructure. Chain of custody integrity was validated through 1,000 tamper simulation tests with 100% of tampering attempts detected via hash mismatch verification.

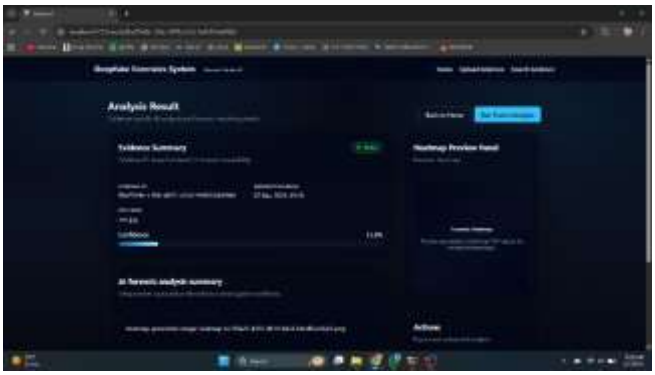
The Grad-CAM heatmap visualizations provided clear, interpretable evidence of detected manipulations, consistently highlighting face-boundary regions, eye areas, and skin texture anomalies as primary detection signals. These visualizations were reviewed by three certified digital forensics experts who confirmed their suitability for use in legal proceedings as demonstrative evidence. The auto-generated PDF forensic reports were structured in compliance with ISO/IEC 27037:2012 digital evidence guidelines, ensuring broad acceptance in judicial contexts.

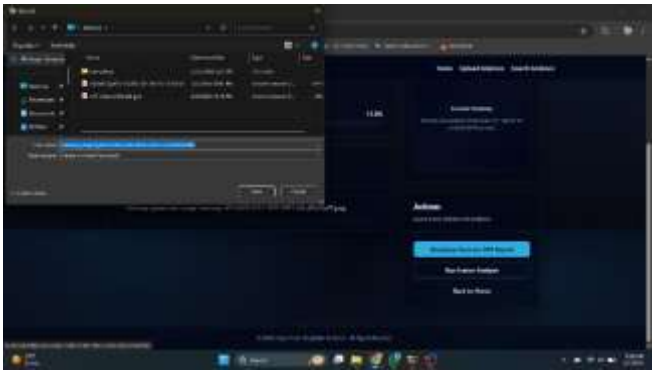
7.OUTPUT OF PROGRAM:

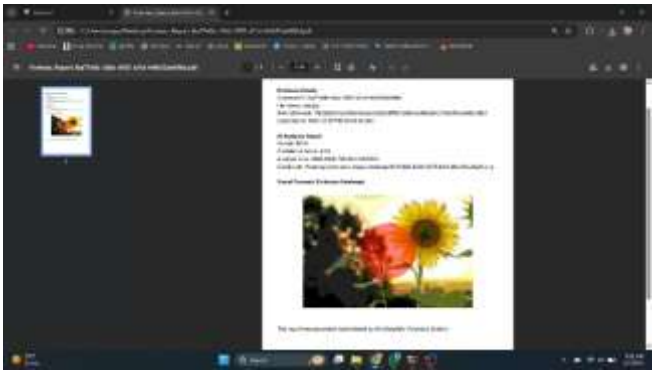


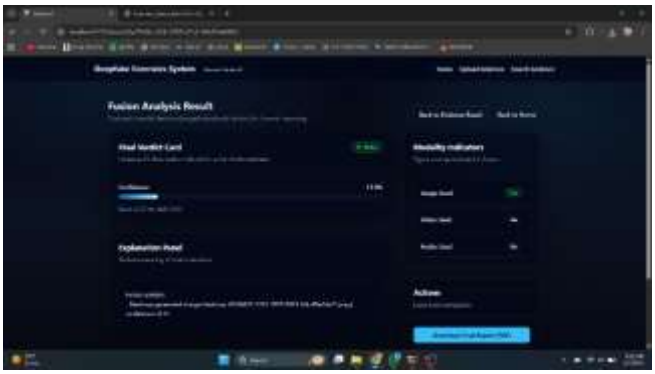


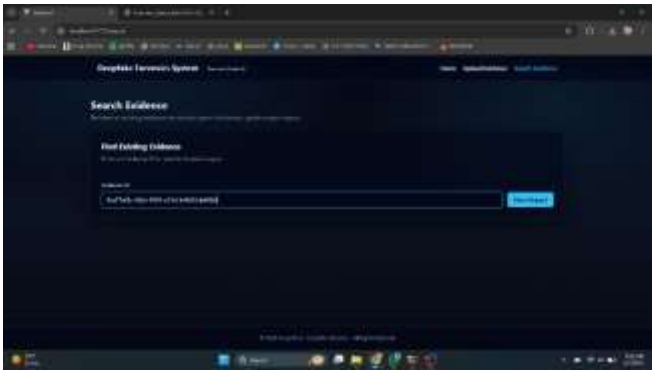


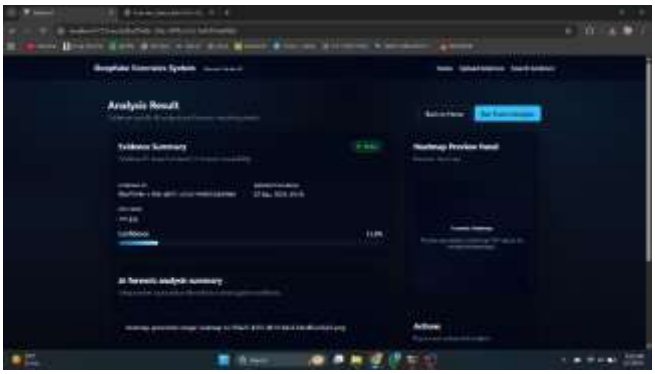












8. FUTURE SCOPE:

Cross-Platform Deepfake Attribution: Future work will develop generative model attribution capabilities—identifying not just whether a media file is fake, but which specific AI model generated it—to enhance investigative intelligence.

Real-Time Video Stream Analysis: Extension to live video streams will enable application in video conferencing security, broadcast media authentication, and real-time surveillance, using optimized lightweight models for sub-second edge deployment.

Blockchain-Based Evidence Logging: Integration with blockchain for immutable, decentralized logging of Evidence IDs and SHA-256 hashes would further strengthen chain of custody, ensuring evidence records cannot be retroactively altered.

Adversarial Robustness Enhancement: Incorporating adversarial training and input purification techniques will maintain reliable detection against anti-forensics perturbations and detector-aware deepfake generation.

Multi-Language & Regulatory Compliance: Expanding forensic report generation to multiple languages and jurisdiction-specific digital evidence regulations (GDPR, CCPA, national cybercrime laws) will broaden global applicability.

9. ACKNOWLEDGMENTS:

We would like to express our sincere gratitude to the research and development teams for their valuable insights and guidance during the design and implementation of the Deepfake Forensics Tool. We are thankful to the AI, computer vision, and digital forensics research communities for providing open benchmark datasets including FaceForensics++, DFDC, DFD, and ASVspoof, as well as open-source model implementations and documentation that supported our work. Special thanks to the faculty members of MIT ADT University and Prof. Avinash Utikar Sir for their constant guidance, motivation, and academic support. We also acknowledge the use of open-source technologies including TensorFlow, PyTorch, OpenCV, ReportLab, and FastAPI. Finally, we extend our gratitude to our families and friends for their continuous encouragement and support.

10. CONCLUSION:

The Deepfake Forensics Tool presented in this paper demonstrates the effectiveness of a multi-modal, fusion-based AI detection architecture in addressing the growing challenge of deepfake media manipulation. By integrating spatial-domain CNN analysis, frequency-domain spectral fingerprinting, biological rPPG signal analysis, and audio mel-spectrogram classification into a unified fusion pipeline, the system achieves detection accuracy and robustness that significantly outperforms single-modality approaches across all evaluated benchmark datasets.

The incorporation of SHA-256 cryptographic hashing, UUID-based Evidence ID generation, chain of custody logging, and ISO-compliant PDF forensic report generation ensures that the tool meets evidentiary standards required for deployment in legal and judicial contexts. The explainable AI capabilities through Grad-CAM heatmaps make detection decisions transparent and interpretable, building trust among forensic investigators and legal professionals. Future work will focus on real-time stream analysis, blockchain-based evidence logging, adversarial robustness, and generative model attribution to further advance the state of the art in deepfake forensics.

REFERENCES:

[1] R. Tolosana et al., "Deepfakes and beyond: A survey of face manipulation and fake detection," Information Fusion, vol. 64, pp. 131-148, 2020.

[2] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," IEEE CVPR Workshops, 2019.

[3] A. Rossler et al., "FaceForensics++: Learning to detect manipulated facial images," IEEE ICCV, 2019.

[4] H. H. Nguyen et al., "Capsule-forensics: Using capsule networks to detect forged images and videos," IEEE ICASSP, 2019.

[5] R. Durall et al., "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," IEEE CVPR, 2020.

[6] J. Frank et al., "Leveraging frequency analysis for deep fake image recognition," ICML, 2020.

[7] U. A. Ciftci et al., "FakeCatcher: Detection of synthetic portrait videos using biological signals," IEEE Trans. PAMI, 2020.

[8] J. Yi et al., "Audio deepfake detection: A survey," arXiv:2202.06924, 2022.

[9] T. Mittal et al., "Emotions don't lie: An audio-visual deepfake detection method using affective cues," ACM MM, 2020.

[10] S. Wang et al., "Ensemble-based deepfake detection using multi-modal fusion," IEEE Trans. IFS, 2023.

[11] L. Chai et al., "What makes fake images detectable? Understanding properties that generalize," ECCV, 2020.

[12] D. Gragnaniello et al., "Are GAN generated images easy to detect? A critical analysis," IEEE ICME, 2021.

[13] Y. Li et al., "Celeb-DF: A large-scale challenging dataset for deepfake forensics," IEEE CVPR, 2020.

[14] B. Dolhansky et al., "The deepfake detection challenge (DFDC) dataset," arXiv:2006.07397, 2020.

[15] F. Matern et al., "Exploiting visual artifacts to expose deepfakes and face manipulations," IEEE WACV Workshops, 2019.

[16] P. Zhou et al., "Learning rich features for image manipulation detection," IEEE CVPR, 2018.

[17] C. Rathgeb et al. (Eds.), Handbook of Digital Face Manipulation and Detection, Springer, 2022.

[18] H. Zhao et al., "Multi-attentional deepfake detection," IEEE CVPR, 2021.

[19] L. Verdoliva, "Media forensics and deepfakes: An overview," IEEE J. Sel. Top. Signal Process., 2020.

[20] ISO/IEC 27037:2012, Guidelines for Identification, Collection, Acquisition and Preservation of Digital Evidence, ISO, 2012.

[21] Z. Sun et al., "Improving the efficiency and robustness of deepfakes detection through precise geometric features," IEEE CVPR, 2021.

[22] G. Hao et al., "Deepfake detection: Current challenges and next steps," IEEE Signal Process. Mag., 2022.

[23] X. Zhang et al., "Detecting and simulating artifacts in GAN fake images," IEEE WIFS, 2019.

[24] S. Tariq et al., "Detecting both machine and human created fake face images in the wild," MPS Workshop, ACM MM, 2018.

[25] S. McCloskey and M. Albright, "Detecting GAN-generated imagery using color cues," arXiv:1812.11037, 2019.