# SnifferX: A Digital Forensices System for Detecting Non-Consensual Deepfake Media

Zaid Rakhange[1], Vaibhav Rathod[2]

[1,2] Zaid – Aqdus Department of Information Technology Engineering

Atharva College of Engineering, India

[1] zaidrakhange-inft@atharvacoe.ac.in

[2] vaibhavrathod-inft@atharvacoe.ac.in

*The rapid advancement of artificial intelligence has increased the sophistication and misuse of deepfake technologies, posing significant risks to digital trust and cybersecurity. As cyber threats evolve, traditional security systems struggle to address manipulated visual media used in fraud, misinformation, and identity abuse. This paper introduces* ***Sniffer****; a forensic verification infrastructure designed for deepfake detection and digital evidence integrity. Sniffer integrates cryptographic hashing, structural similarity analysis, pixel-level anomaly localization, severity scoring, and structured chain-of-custody logging to generate evidence-grade forensic reports. The system enables original image registration, comparison with suspected media, and automated documentation suitable for investigative use. By combining AI-assisted analysis with lifecycle-controlled case management, Sniffer aims to bridge the gap between deepfake detection research and practical digital forensic application. The proposed framework emphasizes accountability, traceability, and structured verification to restore confidence in digital media ecosystems.*

***Keywords—****Keywords—Cybersecurity, Deepfake Detection, Digital Forensics, Evidence Integrity, Structural Similarity*

## I. Introduction to Cyber-Security

In the rapidly evolving landscape of computer and information technology, cybersecurity has emerged as a critical area of focus. With the exponential growth of digital connectivity, the volume and sophistication of cyber threats have surged, making cybersecurity an indispensable element of modern technology infrastructure. Cybersecurity involves protecting digital systems, networks, and sensitive data from malicious attacks, unauthorized access, and potential exploitation by cybercriminals.

As technology continues to advance, the digital world has become an integral part of daily life. Individuals, businesses, and governments rely on digital services for communication, financial transactions, healthcare, education, and commerce. However, this dependence on digital platforms has also made cybersecurity threats more prevalent, exposing users to risks such as data breaches, identity theft, and financial fraud. The increased connectivity brought by IoT devices, cloud computing, and artificial intelligence has introduced new vulnerabilities, making it imperative to adopt comprehensive security measures.

### 1.1. Cyber Attack:

A cyberattack is a deliberate attempt to steal, alter, disable, or destroy data by gaining unauthorized

access to digital systems. Threat actors, such as hackers, cybercriminals, and nation-states, use tactics like malware, ransomware, phishing, and password theft to exploit system vulnerabilities. Their motives range from financial gain and espionage to sabotage and political agendas. Cyberattacks can severely disrupt businesses, with the average data breach costing approximately USD 4.35 million, accounting for detection, response, downtime, and reputational damage. These attacks target individuals, businesses, and governments, seeking to access sensitive information, such as intellectual property, customer data, and financial details, causing long-term harm to the victim's operations and brand.

### 1.2. Types of Cyber Attacks:

Cyber attacks come in various forms, targeting individuals, organizations, and even governments. They are often carried out to steal sensitive data, disrupt services, or compromise systems. Below are some of the most common types of cyber-attacks:

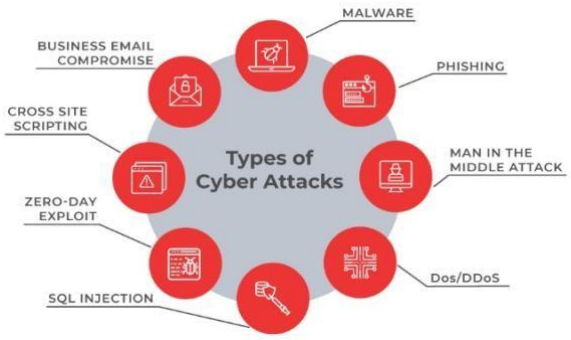



**Fig 1.1 Types of Cyber Attacks**

### 1.3 Common Attacking Techniques:

i. **Brute-forcing:** Attackers systematically try all possible passwords or encryption keys using automated tools to gain unauthorized access. This method is time-consuming and resource-intensive, especially against complex passwords.

ii. **Phishing:** Attackers deceive individuals into revealing sensitive information like usernames and passwords by creating fake emails or websites that mimic legitimate ones. It exploits human trust and is a common attack vector.

iii. **Ransomware:** Malware that encrypts a victim's data and demands a ransom for decryption. It disrupts individuals and organizations, and payment doesn't guarantee data recovery. Examples include WannaCry and Ryuk.

iv. **Social Engineering:** Manipulates people into disclosing confidential informationby exploiting psychological tricks, such as impersonating trusted sources.

v. **Deepfake:** Uses AI to create realistic but fake audio, video, or images for malicious purposes like identity theft and disinformation, presenting a growing cybersecurity threat.

## II. Recent Trends in CyberSecurity

**1. The Growing CyberThreat landscape**As technology advances, cybercriminals have developed increasingly sophisticated tactics, leading to a rise in cyberattacks. Ransomware attacks, in particular, have surged, with malicious actors encrypting victims' files and demanding ransom payments for decryption keys. These attacks target organizations of all sizes, including critical infrastructure sectors like healthcare, transportation, and energy. The financial impact of such breaches is significant, with the average cost of a ransomware attack exceeding USD 4 million. This

alarming trend has compelled organizations to enhance their cybersecurity measures and invest in advanced technologies to defend against evolving threats.

**2. Zero Trust Security Model**

In response to the changing threat landscape, one of the most significant trends in cybersecurity is the adoption of the Zero Trust security model. Unlike traditional security approaches that rely on perimeter defenses, Zero Trust operates under the principle that no entity—whether inside or outside the network—is inherently trustworthy. This model emphasizes continuous verification of user identities, device integrity, and access permissions, thereby minimizing the risk of unauthorized access and data breaches. By adopting a Zero Trust approach, organizations can better secure their networks against sophisticated threats, as every request for access is treated as though it originates from an untrusted source, requiring thorough verification before granting permissions.

## III. Understanding Deepfake Technology

**1. What are Deepfakes?:**

Deepfake technology utilizes advanced machine learning techniques, particularly generative adversarial networks (GANs), to create realistic audio and visual content that can manipulate reality. By training algorithms on extensive datasets, deepfakes can swap faces, alter voices, and fabricate entire scenes, resulting in media that appears genuine yet is entirely fabricated. The implications of deepfake technology extend beyond entertainment, as its potential for misuse raises significant ethical and security concerns

**2. Application of Deepfake Technology :**

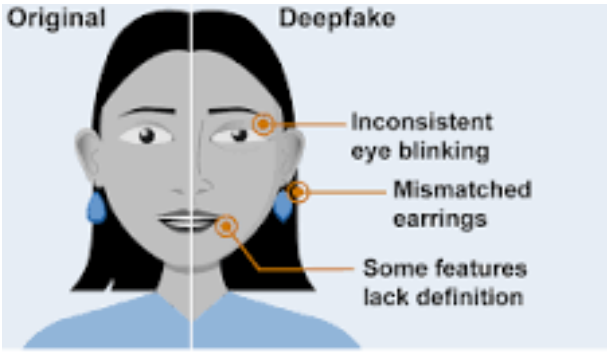
Deepfake technology has found applications across various fields, ranging from entertainment to education and beyond. In film industry, for example, filmmakers use deepfake technology to enhance special effects, allowing for greater creative flexibility in storytelling. Additionally, deepfakes can be employed in virtual reality environments to create immersive training simulations, providing learners with realistic scenarios to practice their skills safely.

**3. How to identify Deepfake Images:**

**3.1 Blurring or Artifacts:** Deepfake videos often show slight blurring or distortion around the face, especially near the edges where the fake face is blended with the original.

**3.2 Unnatural Eye Movements:** The eyes in deepfakes may blink awkwardly or not at all, as early deepfake algorithms struggled with eye movements.

**3.3 Mismatched Lighting:** In many deepfakes, the lighting on the face may not match the rest of the scene, with inconsistent shadows or highlights




Source: GAO, conceived from DARPA image at https://www.darpa.mil/news-events/2019-09-03a. | GAO-20-379SP

**Fig 3.1. Difference between real and deepfake Image**

## IV. Our proposed Technology - Sniffer

**4.1 Overview**

Sniffer is designed as a forensic verification infrastructure rather than a simple detection tool.

The system enables structured registration of original media, comparison of suspected content, and generation of legally structured verification reports. It focuses on evidence integrity, lifecycle management, and tamper localization

**The architecture consists of:**

i. Image Registration Module

ii. Verification & Similarity Engine

iii. Tamper Localization Engine

iv. Severity Scoring Model

v. Case Lifecycle Management

vi. Chain-of-Custody Logging

vii. Evidence Report Generation.

### 4.2 Image Hashing for Digital Integrity

Sniffer uses SHA-256 cryptographic hashing to generate a unique fingerprint for every registered image. Any modification, even at a single pixel level, results in a completely different hash value

**This ensures:**

i. **Detection of direct file alteration**

ii. **Integrity verification**

iii. **Cryptographic evidence traceability**

Hash values are stored securely within the case record to ensure authenticity validation.

### 4.3 Similarity-Based Tamper Detection

In addition to cryptographic hashing, Sniffer performs structural similarity analysis between:

i. **Registered Original Image**
ii. **Suspected Manipulated Image**

**The verification pipeline includes:**

1. Feature Extraction
2. Structural Similarity Index (SSIM) computation
3. Pixel-level difference mapping
4. Region segmentation for anomaly localization

**If divergence exceeds threshold values, the system:**

- Flags manipulated regions
- Generates a tamper localization map
- Assigns a severity score

**This enables detection of:**

- Face swaps
- Region blending
- Synthetic overlays
- Background manipulation

### 4.4 Severity Rating System

Sniffer categorizes manipulation intensity using a structured severity scale:

i. **1-3 (Low Severity):** Minor modifications (brightness adjustments, cropping).

ii. **4-7 (Medium Severity):** Moderate alterations (face blending, background modifications).

iii. **8-10 (High Severity):** Major manipulations (deepfake replacements, synthetic elements).

### 4.5 Case Lifecycle & Forensic Integrity

Each verification is treated as a structured forensic case with defined states:

- **RECEIVED**
- **PROCESSING**
- **COMPLETED**
- **FAILED**

Once analysis is completed, the case is locked to preserve integrity. Post-lock, evidence fields cannot be modified.

### 4.6 Chain of Custody Logging

To ensure accountability, Sniffer maintains an

immutable event log including:

i. IMAGE_UPLOADED
ii. ANALYSIS_STARTED
iii. ANALYSIS_COMPLETED
iv. REPORT_GENERATED
v. CASE_LOCKED

Each event includes timestamp and metadata, forming a traceable audit trail suitable for investigative workflows.

### 4.7 Forensic Report Generation

Sniffer generates structured PDF reports containing:

- **Case ID**
- **Upload timestamp**
- **SHA-256 hash**
- **Image metadata (dimensions, EXIF presence)**
- **Severity score**
- **Classification**
- **Chain of custody log**
- **Legal disclaimer**

This allows the report to be attached to formal complaints or investigative submissions.

## V. How deep faked Content is Undetected?

Detecting deepfakes is increasingly challenging due to the advanced technology used to create them. Deepfake generation tools, like Generative Adversarial Networks (GANs), produce highly realistic content that mimics minute details, making it difficult for detection systems to identify anamoly. Current machine learning models, though widely used, lack the accuracy to reliably spot deepfakes due to limitations like incomplete datasets and the immense time and resources required for training. As deepfake creation tools evolve rapidly, traditional detection models struggle to keep up, often failing to generalize across new and varied deepfakes.

## VI. Prototype Results and System Evaluation

The current implementation of Sniffer demonstrates functional verification capabilities across the following metrics

**i. Integrity Verification**

Cryptographic hash mismatch detection reliably identifies direct file alterations.

**ii. Structural Similarity Detection**

Similarity-based comparison successfully highlights modified regions between registered and suspected images.

**iii. Processing Performance**

The verification pipeline processes images within 2–5 seconds under prototype testing conditions.

**iv. Evidence Report Generation**

Forensic reports are generated instantly upon case completion, providing structured documentation suitable for investigative use.

## VII. Challenges and Limitations

### I. Computational Complexity:

Blockchain Similarity analysis and pixel-level comparison require processing resources, particularly for high-resolution images. Optimization techniques are required for large-scale deployment. expenses.

### II. Adversarial Manipulation:

Advanced attackers may attempt minimal perturbations or noise injection techniques to evade similarity thresholds. Continuous model refinement is necessary.

### III. Dataset Generalization

Deepfake detection models may struggle when exposed to unseen manipulation techniques. Regular retraining and validation against updated datasets is essential..

## VIII. Future Enhancements

### i. Image Video Deepfake Detection

Expanding Sniffer AI's capabilities to detect video manipulations using frame hashing and AI-driven analysis.

### ii. AI-Enhanced Hash Verification

Integrating deep learning for pattern recognition. Detecting manipulated areas within an image.

### iii. Integration with Law Enforcement

Assisting forensic investigators in verifying digital evidence.

## IX. Ethical and Legal Considerations

- **Data Privacy Concerns**

  Images are not stored—only hashes are kept for security.

  Ensures user privacy compliance (GDPR, CCPA).

- **Preventing Misuse**

  Blockchain preventsunauthorized modifications, but ethical use must be monitored..

## X. Conclusion

The rapid evolution of deepfake technologies demands a shift from purely detection-based systems to structured forensic verification infrastructures. Sniffer presents a practical framework for deepfake integrity validation through cryptographic hashing, structural similarity analysis, tamper localization, lifecycle control, and chain-of-custody logging.

By combining AI-assisted detection with evidence-grade documentation and integrity preservation mechanisms, Sniffer bridges the gap between digital media analysis and investigative application. Future work includes large-scale benchmarking, video deepfake detection expansion, and institutional deployment pilots.

## XI. References


1. P. Agrawal S. Johnson, L. Smith, and M. Patel, "A Survey of Machine Learning Techniques in Intrusion Detection Systems," IEEE Xplore. [Online]. Available: https://ieeexplore.ieee.org/document/8589012. Accessed: Oct. 10, 2023.
2. A. Gupta, B. Verma, and T. Sharma, "Blockchain Applications in Cybersecurity: A Comprehensive Review," IEEE Xplore. [Online]. Available: https://ieeexplore.ieee.org/document/9145472. Accessed: Jun. 21, 2024.
3. Y. Kim, J. Park, and D. Choi, "Deep Learning Approaches for Anomaly Detection in Network Traffic," IEEE Xplore. [Online]. Available: https://ieeexplore.ieee.org/document/9481115. Accessed: Aug. 11, 2023
4. C. Williams and H. Martinez, "Cyber Threat Intelligence: Enhancing Detection and Response," ACM Digital Library. [Online]. Available: https://dl.acm.org/doi/10.1145/1234567. Accessed: Jan. 5, 2024.
5. M. Anderson, R. Thompson, and E. Brown, "The Impact of AI on Cybersecurity Defenses," Journal of Cybersecurity Research, vol. 12, no. 3, pp. 45-67, 2023.
6. T. Nakamura and S. Lee, "Zero Trust Security Model: Principles and Implementation Challenges," International Journal of Cybersecurity, vol. 9, no. 2, pp. 112-130, 2023.

7. J. Roberts, K. Zhao, and L. Fernandez, "Deepfake Detection Using Neural Networks: A Comparative Analysis," IEEE Transactions on Information Security, vol. 15, no. 4, pp. 87-102, 2024.
8. D. Patel, "Ethical and Legal Considerations in Cybersecurity Policies," Cyber Law Review, vol. 18, no. 1, pp. 22-38, 2024.