INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

DIGITAL GUARD

P. Dawood Rehmaan Khan
Department of Cyber Security, School of Engineering and Technology, Dhanalakshmi Srinivasan University, Samayapuram Campus, Tiruchirappalli, Tamil Nadu – 621112, India
Email: khandawood0717@gmail.com

S. Netish
Department of Cyber Security, School of Engineering and Technology, Dhanalakshmi Srinivasan University, Samayapuram Campus, Tiruchirappalli, Tamil Nadu – 621112, India
Email: singupuramnetish123@gmail.com

T. Nikhil Reddy
Department of Cyber Security, School of Engineering and Technology, Dhanalakshmi Srinivasan University, Samayapuram Campus, Tiruchirappalli, Tamil Nadu – 621112, India
Email: tnikhilreddy97@gmail.com

Guided by: MS. Siva Selvi
Assistant Professor, Department of Cyber Security, School of Engineering and Technology, Dhanalakshmi Srinivasan University, Samayapuram Campus, Tiruchirappalli, Tamil Nadu – 621112, India
Email: sivaselvik.set2022@dsuniversity.ac.in

Abstract-Digital piracy has become a major challenge in the entertainment and software industries, leading to significant revenue loss and copyright violations. This project presents a Pirated Content Hunting Tool designed to identify and track illegally distributed digital content across online platforms. The proposed system works by scanning websites and online sources to detect pirated content using keyword matching, metadata analysis, and content similarity techniques. Once suspected pirated content is identified, the tool records relevant details such as source links, file information, and access frequency for further review. By automating the process of piracy detection, the system reduces manual effort and improves the efficiency of copyright enforcement. This tool helps content owners and authorities monitor digital platforms more effectively and supports proactive action against unauthorized content distribution

INTRODUCTION

In the modern digital era, the rapid growth of the internet and online content distribution has significantly increased the risk of digital piracy. Pirated content such as movies, software, e-books, and multimedia files is widely shared across various online platforms without the permission of the original creators. This leads to major financial losses for content creators, companies, and the entertainment industry. Traditional piracy detection methods mainly rely on manual monitoring and keyword-based searching, which are time-consuming, inefficient, and unable to handle the massive amount of data available on the internet. To address this challenge, the proposed project "Digital Guard" aims to develop an automated system for detecting pirated digital content on the web. The system integrates advanced technologies such as web crawling, cryptographic hashing (SHA-256), and machine learning techniques to identify unauthorized copies of digital files across websites. The web crawler automatically scans suspicious websites, extracts content, and generates a unique hash value for each file. These hash values are then compared with an authorized database to detect possible piracy. Additionally, machine learning algorithms are used to verify suspicious files and reduce false positives, improving the overall accuracy of the detection system. The system also generates structured reports and alerts, which can help digital rights owners and authorities take necessary actions against piracy. Overall, this project provides a scalable, automated, and intelligent solution for monitoring digital piracy, helping protect copyrighted content and supporting digital rights enforcement in the cybersecurity domain.

LITERATURE REVIEW

Several studies have been conducted to address the problem of digital piracy detection using different technological approaches. Hossain et al. (2021) proposed a deep learning-based piracy detection system that uses content fingerprinting to identify pirated multimedia files. Their approach provides accurate detection of modified media content, but it suffers from high computational costs due to the complexity of deep learning models. Similarly, Karami et al. (2020) introduced a blockchain-based piracy tracking system that utilizes smart contracts to monitor pirated content distribution. This method ensures tamper-proof reporting and improved transparency; however, scalability remains a major challenge when handling large-scale networks. Sharma and Gupta (2022) focused on automated web crawling techniques combined with Natural Language Processing (NLP) to detect pirated web content. Their system enables fast automated scanning of websites, although it may produce false positives in certain situations. In another study, Lee et al. (2023) applied AI-driven network traffic analysis using Deep Packet Inspection (DPI) and machine learning classifiers to detect piracy through network behavior patterns. While effective for peer-to-peer sharing environments, this approach raises potential privacy concerns due to deep analysis of network traffic. Zhang et al. (2022) proposed a perceptual hashing technique to detect duplicate multimedia files by comparing perceptual similarities between media contents. This method allows fast similarity comparison but is sensitive to heavy edits or major modifications in the media. Kumar et al. (2021) developed a text-based copyright violation detection system using cosine similarity and NLP techniques to identify copied textual content efficiently; however, the approach is limited to text content and cannot detect multimedia piracy. Furthermore, Patel et al. (2023) introduced a hybrid machine learning framework combining hashing and Random Forest algorithms to improve piracy detection accuracy and reduce false positives. Despite its effectiveness, the system requires a large training dataset for optimal performance. Additionally, Wang et al. (2022) proposed a distributed web monitoring system that uses distributed crawling techniques to monitor large-scale websites for piracy activities. This architecture offers scalable monitoring capabilities but involves high infrastructure costs. Although these studies contribute significantly

to piracy detection techniques, most existing solutions focus on individual approaches such as crawling, hashing, or machine learning. There is still a lack of a unified system that integrates these techniques for scalable, real-time piracy detection and automated evidence generation.

EXISTING SYSTEM

Keyword-Based Detection:
Many existing systems detect pirated content using keyword matching techniques. However, this method has low accuracy and may fail to identify modified or disguised content.
Blockchain-Based Tracking: Some systems use blockchain technology to track digital content ownership and distribution. While it ensures data integrity and transparency, it suffers from scalability issues when handling large datasets.
Deep Learning Approaches:
Advanced piracy detection systems use deep learning models to analyse multimedia content such as images, videos, and audio. These methods provide high detection accuracy, but they require high computational power and large training datasets.
Network-Based Monitoring:
Network monitoring techniques analyse internet traffic to detect unauthorized sharing of copyrighted content. However, these methods can lead to privacy concerns and may not always identify the original source of piracy.
Manual Monitoring Requirement:
Many existing systems still rely on manual verification and monitoring, which makes the process time-consuming and less efficient for large-scale piracy detection

PROPOSED METHODOLOGY

Automated Piracy Detection:
The proposed system, Digital Guard, automatically detects pirated digital content across multiple websites without requiring continuous manual monitoring.
Web Crawling Mechanism:
The system uses an automated web crawler to scan and monitor websites for suspicious or unauthorized multimedia content.
Content Extraction and Processing:
After crawling, the system extracts relevant content such as images, videos, or files and processes them for further verification.
Secure Content Fingerprinting:
The system generates a SHA-256 cryptographic hash for each piece of content to create a unique digital fingerprint.
Database Comparison:
The generated hash values are compared with the authorized content database to identify whether the content is original or pirated.
Machine Learning Verification:
Machine learning techniques help improve detection accuracy by identifying modified or slightly altered pirated content.
Large-Scale Website Monitoring:
The system is designed to monitor multiple websites simultaneously, enabling large-scale piracy detection.
Alert and Report Generation:
When pirated content is detected, the system automatically generates alerts and structured reports for further action.
Reduced Manual Effort:
By automating crawling, detection, and reporting, the system significantly reduces manual monitoring effort and improves efficiency.

DEPLOYMENT

- The proposed Digital Guard system can be deployed as a scalable cybersecurity monitoring platform for detecting pirated digital content across the internet. The deployment architecture consists of a centralized server integrated with automated web crawlers that continuously scan suspicious websites and online platforms for potential pirated material.
- Initially, the URL input and management module collects suspected website links from administrators or automated threat intelligence sources. The web crawler module then scans these websites and extracts downloadable links and metadata associated with digital content. Extracted files or content features are processed through the content extraction and preprocessing module, which prepares the data for fingerprint generation and similarity comparison.
- The system generates a SHA-256 hash for each detected content file and compares it with the hashes stored in the authorized content database to verify originality. Additionally, text-based and metadata similarities are measured using Cosine Similarity algorithms, while suspicious content is further analyzed using a Random Forest machine learning classifier to reduce false positives and improve detection accuracy.
- Once piracy is confirmed, the alert and reporting module generates structured evidence reports containing the detected URL, hash value, timestamp, and similarity score. These reports can be used by copyright enforcement agencies, digital media companies, or cybersecurity teams to take necessary legal or administrative actions.
- The system can be deployed using cloud infrastructure to support large-scale crawling operations, real-time monitoring, and continuous updates of the authorized content database.

CHALLENGES

- Although the proposed system improves piracy detection automation, several technical and operational challenges remain.
- One of the primary challenges is the dynamic nature of piracy websites. Many illegal websites frequently change their domain names or hosting servers to avoid detection, which makes continuous monitoring difficult. Additionally, some piracy platforms use encrypted file sharing or obfuscated links, preventing easy access to downloadable content.
- Another challenge is large-scale data processing. Web crawling across multiple websites generates massive amounts of data, requiring efficient storage, processing power, and bandwidth to analyze content fingerprints and similarity scores in real time.
- The system may also face false positive or false negative detection issues when content is slightly modified, compressed, or re-encoded. Although machine learning classification reduces these errors, the model still requires well-labeled training datasets for accurate predictions.
- Furthermore, legal and ethical considerations must be addressed when scanning external websites, as automated crawling must comply with website policies and data protection regulations.

CONCLUSION AND FUTURE WORK

- Digital piracy continues to be a major concern for content creators, media companies, and digital distribution platforms. Traditional manual detection techniques are inefficient and incapable of monitoring the rapidly growing number of online piracy sources.
- This research proposed Digital Guard, an automated pirated content hunting system that integrates web crawling, cryptographic hashing, similarity analysis, and machine

learning techniques to detect unauthorized digital content across online platforms. The system improves scalability, reduces manual monitoring efforts, and enhances piracy detection accuracy by combining multiple verification methods.

- By generating structured evidence reports and supporting large-scale monitoring, the proposed approach provides an effective tool for digital rights enforcement and cybersecurity monitoring. Overall, the system demonstrates a practical and automated solution for addressing the challenges of online digital piracy detection.
- Although the current system provides an effective solution for piracy detection, several improvements can be implemented in future research.
- Future versions of the system can incorporate deep learning techniques for more advanced multimedia piracy detection, especially for modified video, audio, and image files. Techniques such as convolutional neural networks (CNNs) can help detect heavily altered or compressed media content.
- The system can also be extended with distributed web crawling frameworks to improve scalability and enable monitoring of a larger number of websites simultaneously. Integration with blockchain-based evidence storage may further enhance the reliability and tamper-proof nature of piracy detection reports.
- Another potential improvement is the development of a real-time piracy monitoring dashboard that provides visual analytics, alert notifications, and automated reporting features for cybersecurity analysts.
- Finally, integrating the system with law enforcement and copyright management platforms could streamline the process of reporting piracy violations and enforcing digital rights more efficiently.

REFERENCES

[1] Hossain, M., Rahman, S., and Islam, M., “Deep Learning-Based Multimedia Piracy Detection,” Journal of Artificial Intelligence and Applications, vol. 12, no. 4, pp. 221–235, 2021.
[2] Karami, A., and Shafiee, M., “Blockchain-Based Digital Content Piracy Tracking System,” IEEE Transactions on Information Forensics and Security, vol. 15, pp. 3895–3906, 2020.
[3] Sharma, R., and Gupta, P., “Automated Web Crawling for Detecting Pirated Digital Content,” Neural Computing and Applications, Springer, 2022.
[4] Lee, J., Kim, H., and Park, S., “AI-Based Network Traffic Analysis for Piracy Detection,” Applied Sciences, vol. 13, no. 4, 2023.
[5] Zhang, Y., Liu, X., and Chen, T., “Perceptual Hashing Techniques for Multimedia Content Protection,” IEEE Access, vol. 10, pp. 34562–34574, 2022.