WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Enhancing Document Security: A Dual Approach with NLP And Cryptographic Techniques

Mr. R. Makendran [a], S. Agilan [b], M. Jagathkishore [c], N. Madhanbabu [d] , M. Murugan [e]

[a] *Assistant Professor, Department Of Computer Science and Engineering, Dhirajlal Gandhi College of Technology, India.*
[b, c, d, e] *Student, Department Of Computer Science and Engineering, Dhirajlal Gandhi College of Technology, India.*

ABSTRACT

In today's digital era, the risk of document theft and unauthorized access to sensitive information has become increasingly prevalent, posing serious threats to individuals, organizations, and governments. Traditional security measures such as access control and basic encryption often fall short in addressing these challenges, especially with the growing complexity of cyber threats. This project proposes a dual-layered defense system that integrates Natural Language Processing (NLP) and Cryptography to enhance document security. NLP techniques are utilized to analyze, classify, and monitor document content in realtime, enabling the system to identify sensitive data, generate unique document fingerprints, and detect unauthorized changes or content extraction that traditional systems might miss. To strengthen protection further, the system incorporates Elliptical curve cryptography encryption to ensure that documents remain secure during storage and transmission, making them unreadable to unauthorized users. Additionally, the project includes a deception strategy by generating realistic fake documents and storing them in alternative repositories to mislead potential attackers. This hybrid approach of combining content-based monitoring, strong encryption, and decoy deployment creates a comprehensive and proactive solution for securing sensitive information. It offers a significant advancement in combating document theft and unauthorized modifications, making it highly applicable across industries that prioritize data confidentiality and integrity.

Keywords: Document Security, Deceptive Repository, Natural Language Processing,, Fake Document Storage, Cryptography, Unauthorized Access Prevention, Elliptic Curve Cryptography.

TF-IDF	— Term Frequency–Inverse Document Frequency
ECC	— Elliptic Curve Cryptography
IP	— Intellectual Property
NLP	— Natural Language Processing

1. INTRODUCTION

In the modern digital landscape, protecting sensitive documents from theft, unauthorized access, and tampering is a major concern across various sectors. Traditional document security measures such as passwords and basic encryption often fall short when faced with sophisticated cyber threats. As the volume and importance of digital documents grow, there is an increasing need for a comprehensive security system that goes beyond access control and static protection mechanisms. This project, DARD – Defense Against IP Theft, addresses these challenges by introducing a multi-layered strategy to safeguard critical information. The proposed solution integrates Natural Language Processing (NLP) and cryptographic techniques to ensure both proactive and reactive defense mechanisms. NLP plays a key role in analyzing document contents using TF-IDF (Term Frequency–Inverse Document Frequency), helping identify sensitive keywords and understand contextual relevance. This enables intelligent monitoring, classification, and alerting of potential risks related to unauthorized modifications or content extraction. Simultaneously, the system creates realistic decoy documents to deceive attackers and divert them from accessing actual sensitive files, adding an extra layer of protection. To further strengthen the system, Elliptical curve cryptography encryption is used to secure the storage and transmission of real documents, making them accessible only to authorized users. This combination of content analysis, encryption, and deception creates a holistic defence mechanism that is adaptive, secure, and efficient. By implementing this model, organizations can better manage and protect their digital documents, reduce the chances of data breaches, and ensure the integrity and confidentiality of valuable information.

2. LITERATURE SURVEY

2.1. THE ROLE OF CYBERSECURITY IN PROTECTING INTELLECTUAL PROPERTY

This paper explores the critical role of cybersecurity in safeguarding intellectual property (IP) in the digital era. As businesses increasingly rely on digital infrastructures, the protection of proprietary assets like patents, trade secrets, and source codes becomes paramount. The study highlights various cybersecurity threats that endanger IP, including cyber espionage, insider threats, and ransomware attacks. The authors review current defense mechanisms and emphasize the need for layered security frameworks, incorporating firewalls, intrusion detection systems, encryption, and access controls. The paper presents real-world case studies where cybersecurity lapses led to IP theft, reinforcing the importance of proactive defense strategies. It also evaluates the impact of government policies and legal frameworks on IP protection. The researchers recommend integrating cybersecurity awareness training into corporate culture. Moreover, the study outlines how emerging technologies like AI and blockchain can bolster IP protection by enabling real-time threat detection and secure auditing. A key contribution is the proposed cybersecurity-IP integration model, which aligns IT security protocols with IP asset classification. The paper underscores the need for ongoing vulnerability assessments and robust incident response plans.

2.2. A PSYCHOLINGUISTICS-INSPIRED METHOD TO COUNTER IP THEFT USING FAKE DOCUMENTS

This innovative study presents a novel psycholinguistics-based approach to intellectual property theft mitigation by strategically deploying fake documents. The method is grounded in cognitive deception and linguistic manipulation to mislead adversaries attempting to steal proprietary data. These decoy documents are crafted using psycholinguistic principles to appear authentic, increasing the likelihood of interception while leaving real IP untouched. The authors detail an algorithm that dynamically generates fake documents embedded with behavioral triggers to alert organizations when accessed. The study presents a threat model of adversaries engaging in document exfiltration and how psychological misdirection through linguistic mimicry can delay or divert them. Realistic evaluation scenarios demonstrate the effectiveness of the method in both protecting real assets and enabling intrusion detection. A highlight is the system's adaptability—fake documents evolve with the context of the organization's operations. The authors also address ethical implications, noting how the technique aligns with deterrence strategies in cybersecurity. This work contributes a fresh interdisciplinary angle, merging language science with cybersecurity practices. It calls for more research into deception technologies and their role in modern data protection. Experimental results confirm increased timeto-theft and confusion among attackers. This approach also reduces reliance on perimeter security.

3. SYSTEM STUDY

3.1. EXISTING SYSTEM

Traditional document security systems primarily rely on reactive defense mechanisms such as firewalls, intrusion detection systems (IDS), and basic encryption protocols. These methods are often focused on protecting network perimeters or restricting unauthorized access through passwords and user permissions. However, with the evolving nature of cyber threats, these approaches are increasingly proving insufficient. Attackers today are more sophisticated, utilizing advanced tactics to bypass access controls and infiltrate sensitive document repositories. The lack of active monitoring and contextual content analysis in these systems means that breaches often go undetected until the damage is already done.

An emerging solution in the cybersecurity landscape is the concept of deceptive repositories, which aim to proactively mislead and trap attackers by creating decoy assets. These repositories are filled with fake but realistic-looking documents, databases, and credentials designed to mimic genuine data. When attackers interact with these decoys, alerts are triggered, enabling cybersecurity teams to monitor and analyze the behavior of potential threats in a controlled environment. One such tool, WE-Forge, employs word embeddings to generate convincing fake documents by preserving semantic and syntactic relationships between words. This ensures that the decoy content is contextually relevant to the targeted domain, effectively confusing and slowing down attackers. Despite their innovative approach, deceptive repository systems have notable limitations. For instance, in highly specialized fields, generating fake documents that convincingly mimic real ones can be challenging without in-depth domain knowledge. Additionally, if the training corpus for tools like WE-Forge is outdated or not representative of the actual domain, the realism of the decoy documents suffers, reducing their effectiveness. There's also the risk that legitimate users may mistakenly interact with decoy files, leading to confusion or operational disruption. Thus, while deception-based defenses offer a new layer of security, they must be carefully maintained and supplemented with more robust content analysis and encryption mechanisms to ensure complete and reliable document protection.

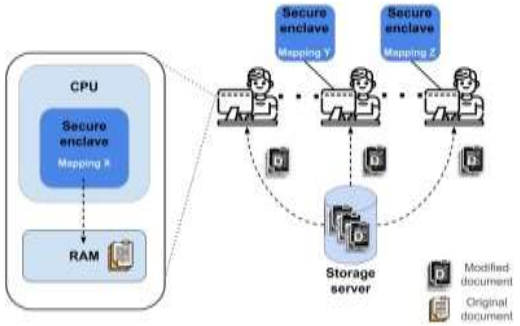

Figure 3.1.1: Secure Document Processing Using Enclave-Based Isolation and Mapping

3.2. PROPOSED SYSTEM

The proposed system introduces a dual-layered defense mechanism by integrating Natural Language Processing (NLP) and cryptographic techniques to ensure robust document security. NLP is used to analyze document content using the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm, enabling the system to identify and classify sensitive data based on keyword relevance and contextual significance. This intelligent analysis allows real-time monitoring of user interactions with documents, helping to detect any unauthorized attempts to extract, modify, or misuse critical information. By understanding the content and its importance, the system ensures that high-priority documents are given greater protection.

In addition to content analysis, the system incorporates deception techniques by generating fake documents and storing them in alternative repositories. These decoy documents mimic the structure and context of genuine documents but contain no real sensitive data. The purpose of this deception layer is to mislead unauthorized users or potential attackers, causing them to engage with false information instead of actual intellectual property. This not only protects the genuine assets but also allows cybersecurity teams to track intrusion attempts and gather insights into attacker behavior. To further fortify the security, ECC encryption is applied to all sensitive documents. ECC is a symmetric-key block cipher known for its high speed and strong security, ensuring that only users with the correct decryption key can access the protected data. This encryption layer guarantees the confidentiality and integrity of documents during both storage and transmission. By combining NLP for intelligent content monitoring, deception tactics for misdirection, and ECC encryption for secure access control, the proposed system delivers a comprehensive and proactive solution for defending against document theft and unauthorized access.

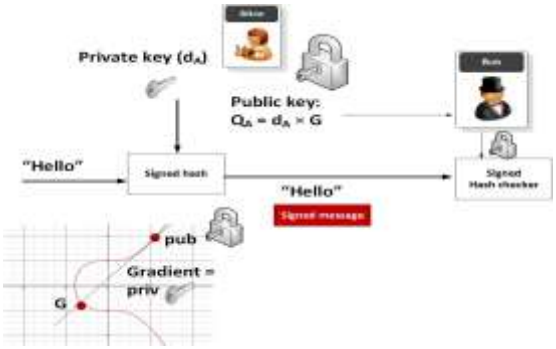

Figure 3.2.1: ECC ENCRYPTION PROCESS

Feature	Existing System	Proposed System
Security Type	**Basic Encryption**	**ECC + NLP + Decoy Layer**
Content Analysis	**No**	**Yes (TF-IDF)**
Deceptive Documents	**No**	**Yes**
Real-time Monitoring	**No**	**Yes**
Authentication Level	**Password Only**	**Multi-Factor**

4. METHODOLOGY

4.1. NATURAL LANGUAGE PROCESSING (NLP)

Natural Language Processing (NLP) plays a central role in enhancing document security by enabling intelligent, content-aware analysis and monitoring. In the proposed system, NLP techniques such as TF-IDF (Term Frequency–Inverse Document Frequency) are employed to scan and understand the content of documents in real time. This allows the system to identify sensitive keywords, phrases, and contextual patterns that could indicate critical or confidential information. By quantifying the importance of specific terms relative to a corpus of documents, TF-IDF helps in flagging content that may need additional protection. This enables proactive classification and tagging of documents based on their sensitivity level, rather than relying solely on static metadata or user-defined labels. Moreover, the NLP engine can detect unusual textual patterns or modifications that may signal an attempt at tampering or unauthorized content extraction. Beyond content classification, NLP also supports the generation of unique document fingerprints essentially semantic signatures that allow the system to track changes, verify integrity, and detect plagiarism or data leakage. By continuously monitoring document interactions and comparing live content against previously stored fingerprints, the system can issue alerts when discrepancies are found. This real-time contextual analysis surpasses traditional security systems that only control access or encrypt content without understanding its meaning. Additionally, NLP is employed in crafting decoy documents that are contextually realistic. These fake documents are generated using linguistic patterns similar to genuine sensitive files, misleading attackers and increasing the likelihood of engagement with non-critical data. Thus, NLP not only strengthens security through intelligent content analysis but also enhances deception strategies, making the overall defense system robust, adaptive, and highly efficient.

4.2. ELLIPTIC CURVE CRYPTOGRAPHY (ECC)

Elliptic Curve Cryptography (ECC) is a form of public-key cryptography that offers high security with relatively smaller key sizes compared to traditional cryptographic methods such as RSA. The core principle of ECC lies in the mathematical properties of elliptic curves over finite fields, which provide a more efficient and robust method for securing digital communications. ECC enables secure encryption, digital signatures, and key exchange mechanisms with shorter key lengths, making it computationally faster and more energy-efficient, especially in environments with limited resources like mobile devices or IoT systems. The small key sizes in ECC do not compromise security, as they offer the same level of protection as longer RSA keys, making ECC a preferred choice for modern cryptographic applications. In the context of document security, ECC is utilized to enhance the encryption of sensitive data during storage and transmission. By using elliptic curve-based algorithms, documents are encrypted in a way that makes them unreadable to unauthorized users. The ECC keys allow for secure data exchange without the need for heavy computational resources, which is crucial in scenarios where large volumes of sensitive information need to be transmitted securely. ECC also facilitates the generation of digital signatures, which ensure both the authenticity and integrity of documents. By employing ECC for document protection, organizations can achieve a high level of security while maintaining efficiency and reducing the processing overhead associated with other encryption methods. This makes ECC an ideal solution for safeguarding intellectual property, ensuring that only authorized users can access or modify critical documents.

$$y2=x3+ax+b$$

5. MODULES IMPLEMENTATION

5.1 LIST OF MODULES

- User interface module
- Document sensitivity classification module
- Document encryption and authentication module
- Document access control and audit module
- Document tracking and integrity monitoring module
- Alert and notification module

5.2 MODULES DESCRIPTION

5.2.1 USER INTERFACE MODULE

This module provides a user-friendly interface for both administrators and authorized users to interact with the system efficiently. It allows users to upload their Intellectual Property (IP) documents directly through the platform. Once uploaded, users can encrypt documents and manage their metadata securely. Administrators can monitor document activity and real-time security events. The interface supports the customization of access permissions based on user roles. Users can also set their preferences, including notification types, language, and interface themes. The design emphasizes simplicity while ensuring full functionality. Navigation is intuitive, making document handling seamless. This module acts as the control hub for all user-level operations.

5.2.2 DOCUMENT SENSITIVITY CLASSIFICATION MODULE

The Document Content Analysis module uses advanced Natural Language Processing (NLP) techniques to examine and classify content based on its sensitivity. It identifies proprietary, financial, or personal data within the uploaded document. Named Entity Recognition (NER) is used to highlight sensitive fields like names, account numbers, and locations. The module generates a unique fingerprint for every document by analyzing its structure, keywords, and patterns. This fingerprint is securely stored for future verification and integrity checks. It continuously scans for unauthorized modifications to the document content. Any suspicious change triggers alerts in real-time. The goal is to ensure that data within the document remains intact and confidential. This module enhances document intelligence and traceability.

5.2.3 DOCUMENT ENCRYPTION AND AUTHENTICATION MODULE

This module ensures data confidentiality by applying ECC encryption to all uploaded documents before storage or transmission. ECC provides strong security using smaller keys, making it faster and efficient compared to other encryption methods. The encrypted files are difficult to decrypt without the proper key, ensuring unauthorized users are blocked. Digital signatures are added to verify the authenticity of the sender and detect tampering. This module also manages the decryption process using the recipient's private key, enabling only authorized users to view the content. It protects the document during both transit and rest. Security policies are enforced consistently. Encryption and decryption are seamless to users. This layer is crucial for IP protection.

5.2.4 DOCUMENT ACCESS CONTROL AND AUDIT MODULE

The Access Control and User Authentication module safeguards the system from unauthorized access by implementing multi-factor authentication (MFA). Biometric, password, and private key-based methods can be used based on system settings. It defines roles and permissions clearly, so different users have different levels of access. For example, only an admin may edit while others can view. It tracks every access, login, and document modification with accurate timestamps. Logs are stored for auditing and forensic analysis. This module ensures accountability for all actions within the system. It restricts sensitive operations to specific roles. Tampering or unusual access patterns are flagged instantly. The module helps prevent insider threats and breaches.

5.2.5 DOCUMENT TRACKING AND INTEGRITY MONITORING MODULE

This module tracks the lifecycle of each document, including when and where it is accessed, edited, or shared. It captures metadata such as IP address, timestamp, and user credentials for every activity. Integrity verification is done by comparing the current state of the document with its previously stored fingerprint. If any deviation is found, it implies possible tampering, and an alert is generated. This ensures that even subtle unauthorized changes are not missed. Document flow across systems is monitored in real-time. It supports audit trails for legal and compliance requirements. It helps in quickly identifying the point of breach. Tracking is continuous and automated. This module maintains transparency and trust.

5.2.6 ALERT AND NOTIFICATION MODULE

The Alert and Notification System provides real-time updates regarding document activity and potential threats. It sends notifications when unauthorized users attempt to access or alter any document. Alerts are also triggered by suspicious content extraction or structural changes in a document. The system can be configured to notify through emails, dashboard pop-ups, or SMS. Periodic reports are sent to administrators and users summarizing document security events. It ensures immediate awareness of ongoing breaches. Notification settings can be customized based on user roles and preferences. The system logs all alerts for future reference. It strengthens the response time during attempted IP thefts. This module enhances proactive protection and oversight.

Figure 5.1.1: Home page

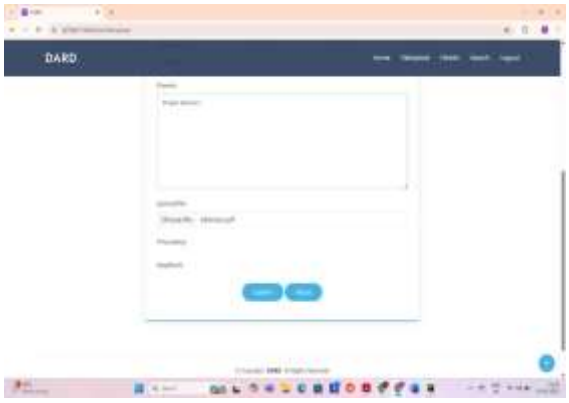

Figure 5.1.2: Document upload page

Figure 5.1.3: Document download page

Figure 5.1.4: Original document

Figure 5.1.5: Fake document

5.3 SYSTEM ARCHITECTURE

The architecture of the proposed system consists of several key components working together to ensure document security. At the core, Natural Language Processing (NLP) analyzes document content to identify sensitive information, using techniques like TF-IDF for keyword extraction and content classification. Documents are then secured using ECC encryption, ensuring that only authorized users with the correct decryption keys can access them. Additionally, a deception layer is integrated, where fake documents are generated and stored in alternative repositories to mislead unauthorized users. Real-time monitoring tools continuously track interactions with documents, triggering alerts for any unauthorized access or tampering. This layered architecture provides a comprehensive and proactive defense mechanism for safeguarding sensitive information.

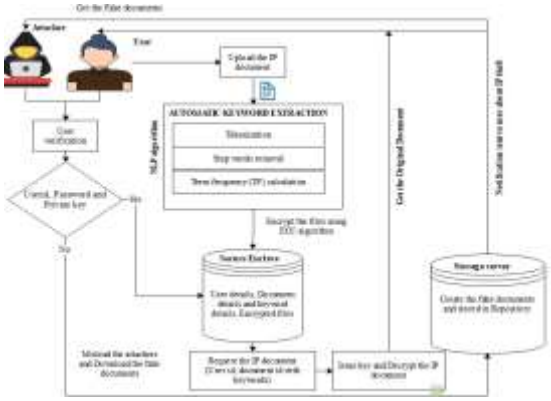

Figure 6.1: System architecture

6. CONCLUSION AND FUTURE ENHANCEMENTS

CONCLUSION

In conclusion, this project offers a robust and intelligent system for protecting Intellectual Property (IP) documents from unauthorized access, tampering, and theft. By integrating advanced technologies such as Natural Language Processing (NLP), ECC encryption, and secure user authentication methods, the system ensures that only verified users can access or modify sensitive content. The automatic keyword extraction and document fingerprinting enhance content monitoring, allowing the system to detect suspicious changes or data leakage effectively. Furthermore, the deployment of decoy documents misleads potential attackers and prevents them from obtaining the original sensitive data, adding another layer of proactive security. Additionally, the system's modular structure enables seamless integration of user management, document tracking, and real-time alert mechanisms. The implementation of multi-factor authentication and detailed logging of user activity enhances accountability and allows for accurate auditing. The notification module keeps users informed of any suspicious activities, reinforcing user awareness and security. Overall, this project represents a comprehensive solution to modern digital threats faced by IP document holders, combining automation, cryptographic protection, and intelligent analysis to ensure privacy, integrity, and secure document sharing in an increasingly connected world.

FUTURE ENHANCEMENTS

- **Integration of AI-Based Threat Detection**: Implement advanced AI and machine learning models to predict and prevent potential security breaches by analyzing user behavior patterns and access logs.
- **Blockchain-Based Document Logging**: Use blockchain technology to create immutable logs of document access and modifications, enhancing transparency, traceability, and tamper-proof auditing.
- **Cloud Storage and Remote Access**: Enable secure cloud-based document storage with end-to-end encryption to support remote access and collaboration without compromising security.
- **Support for Multiple File Formats**: Extend support for various document types including spreadsheets, presentations, multimedia files, and code repositories, providing broader coverage.
- **Mobile Application Integration**: Develop a companion mobile app for real-time notifications, quick document access, and biometric authentication on-the-go, improving user convenience and accessibility.

REFERENCES

1. Mavani, Chirag, et al. "The Role of Cybersecurity in Protecting Intellectual Property." International Journal on Recent and Innovation Trends in Computing and Communication 12.2 (2024): 529-538.

2. Denisenko, Natalia, et al. "A Psycholinguistics-Inspired Method to Counter IP Theft using Fake Documents." ACM Transactions on Management Information Systems (2024).

3. Zhu, Hongyu, et al. "Reliable Model Watermarking: Defending Against Theft without Compromising on Evasion." arXiv preprint arXiv:2404.13518 (2024).

4. Zhang, Ruisi, and Farinaz Koushanfar. "EmMark: Robust Watermarks for IP Protection of Embedded Quantized Large Language Models." arXiv preprint arXiv:2402.17938 (2024).

5. Wang, Zhenyi, Yihan Wu, and Heng Huang. "Defense against Model Extraction Attack by Bayesian Active Watermarking." Forty-first International Conference on Machine Learning, (2024)

6. Pagnotta, Giulio, et al. "Dolos: A novel architecture for moving target defense." IEEE Transactions on Information Forensics and Security (2023).

7. Sayeed, Sarwar, et al. "TRUSTEE: Towards the creation of secure, trustworthy and privacy-preserving framework." Proceedings of the 18th International Conference on Availability, Reliability and Security. 2023.

8. Ajmal, Abdul Basit, et al. "Toward effective evaluation of cyber defense: Threat based adversary emulation approach." IEEE Access 11 (2023): 7044370458.

9. Martínez, Antonio López, Manuel Gil Pérez, and Antonio Ruiz-Martínez. "A Comprehensive Model for Securing Sensitive Patient Data in a Clinical Scenario." IEEE Access 11 (2023): 137083-137098.

10. Gambarelli, Gaia, Aldo Gangemi, and Rocco Tripodi. "Is your model sensitive? SPEDAC: A New resource for the automatic classification of sensitive personal data." IEEE Access 11 (2023): 10864-10880.