Leveraging Machine Learning for Data Loss Prevention

Sulaksha Punsara
Department of Computer Systems Engineering
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
sulakshapunsara008@gmail.com

Neelaka Nilakshana
Department of Computer Systems Engineering
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
neelaka668@gmail.com

Tharindu Gihan
Department of Computer Systems Engineering
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
tharinduelpitiya0@gmail.com

Pubudu Priyanga
Department of Computer Systems Engineering
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
priyanga.pubudu@gmail.com

Amila Senarathne
Department of Computer Systems Engineering
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
amila.n@sliit.lk

Suranjini Silva
Department of Computer Systems Engineering
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
suranjini.s@sliit.lk

***Abstract*—Organizations face major risks through data exfiltration combined with insider threats and compliance violations, especially under regulatory standards such as PCI DSS, GDPR and HIPAA. This research introduces an integrated Data Loss Prevention system that detects data breaches in real-time, then executes prevention protocols to log sensitive data breaches through homomorphic encryption. Our system comprises four key components:- 1. an intelligent data classification and blocking mechanism that detects and prevents sensitive information leaks while tagging violations under regulatory categories; 2. a dynamic risk scoring framework that assigns risk levels based on user violations and behavioral patterns; 3. an endpoint security module that includes webcam-based login verification, phishing site detection, and automatic screenshot monitoring for unauthorized data exposure; and 4. a steganalysis engine that leverages deep learning models in both spatial and frequency (DCT) domains to detect data leaks via steganographic images. Implementing the proposed DLP solution combines encryption and machine learning technology to strengthen security measures without affecting operational performance. Experimental tests showed successful detection of regulatory violations and phishing attempts coupled with high accuracy in stego image identification, together with effective performance of the risk scoring system to highlight insider threats. Data protection research demonstrates how an enterprise-wide protection system can succeed in securing corporate resources effectively.**

***Index Terms*—Data Loss Prevention (DLP), Regulatory Compliance, Insider Threat Detection, Risk Scoring, Machine Learning, Homomorphic Encryption, Phishing Detection, Steganalysis, PCI DSS, GDPR, Deep Learning, Data Security, Endpoint Protection, Information Leakage Prevention, User Behavior Analysis, Cybersecurity.**

I. Introduction

Enterprise-level digitization methods have delivered rapid growth to data generation along with storage and sharing activities. Data generation along with its growth becomes more vulnerable to breaches because it attracts both internal threats and compliance violations. Businesses need to defend their sensitive data collection of PII data and payment information and financial reports along with corporate intellectual assets from unauthorized users who try extracting them. Data leaks produce multiple significant consequences that result in financial harm alongside legal sanctions and reputation deterioration coupled with non-compliance legislation violations.

Organizations employ Data Loss Prevention (DLP) solutions to restrain unauthorized data flows by active monitoring and detection systems that prevent such activity. The current standard DLP tools rely on content inspection but use rule-based blocking yet modern threats need extensive adaptive multi-layered protection methods. Through advanced machine learning (ML) and deep learning (DL) methods organizations can improve their DLP systems to prevent threats before they happen while decreasing incorrect alert triggers. Homomorphic encryption enables safe logging processes for data violations which protect sensitive information from unauthorized persons viewing the content.

Data security and privacy laws require organizations to implement stringent protective measures. While PCI DSS (Payment Card Industry Data Security Standard) and GDPR (General Data Protection Regulation) are widely known, several other compliance frameworks influence enterprise security practices:

- **HIPAA (Health Insurance Portability and Accountability Act)** – Protects electronic health records (EHRs) and patient data.

- **SOX (Sarbanes-Oxley Act)** – Enforces data integrity standards for financial reporting.
- **CCPA (California Consumer Privacy Act)** – Establishes data privacy rights for consumers in California.
- **NIST 800-53** – Provides security and privacy controls for federal information systems.
- **ISO/IEC 27001** – An international standard for information security management systems (ISMS).

Organizations must ensure compliance with these frameworks by adopting a **comprehensive, policy-aware DLP system** capable of identifying and classifying sensitive data based on regulatory requirements.

This paper presents a multi-layered Data Loss Prevention (DLP) system that integrates advanced machine learning, homomorphic encryption, risk scoring, and deep learning-based steganalysis to detect and prevent data exfiltration. Our solution consists of four key components:

1) **Sensitive Data Classification and Blocking:** Detects and prevents unauthorized data transfers while categorizing violations according to compliance frameworks (e.g., PCI DSS, GDPR, HIPAA).
2) **User Risk Scoring System:** Assigns risk scores based on policy violations, behavioral analysis, and insider threat detection.
3) **Webcam-Based Login Security and Phishing Detection:** Monitors failed login attempts and detects phishing sites to enhance endpoint security.
4) **Deep Learning-Based Steganalysis:** Identifies covert data exfiltration attempts via steganography in both spatial and DCT domains.

The primary contributions of this work are as follows:

- **An integrated DLP architecture** that addresses content filtering, behavioral analysis, phishing detection, and covert data leakage.
- **A compliance-aware data classification module** capable of detecting violations related to multiple regulatory frameworks (PCI DSS, GDPR, HIPAA, SOX, CCPA, NIST 800-53, ISO 27001).
- **A novel risk-scoring system** that quantifies security threats based on user actions and policy violations.
- **The application of homomorphic encryption** for secure logging, ensuring data integrity while maintaining confidentiality.
- **A dual-domain deep learning-based steganalysis model** to detect hidden data within images.

The rest of this paper is organized as follows: Section II reviews related work, comparing existing DLP solutions. Section III describes the system architecture and the four core components. Section IV details the implementation, including machine learning models and encryption techniques. Section V presents experimental results and evaluations. Section VI discusses findings, limitations, and possible enhancements. Finally, Section VII concludes the paper and outlines future research directions.

II. Related Work

A. Comparison of existing DLP tools

Data Loss Prevention (DLP) has matured into a competitive field with both established commercial suites and emerging open-source solutions. Symantec Data Loss Prevention (now Broadcom) is often cited as a long-term market leader, offering comprehensive coverage across endpoints, networks, email and cloud channels [1]. It provides deep content inspection (including features like data fingerprinting and image analysis) and tight integration with other Symantec security products [2]. However, Symantec's breadth comes with deployment complexity and resource demands – the system requires an Oracle database backend and can be costly to operate. Its all-encompassing scope makes it better suited for large enterprises with dedicated security teams, as noted by one report which found Symantec DLP time-consuming to deploy but effective for organizations needing to protect extensive data estates [3]. In contrast, Forcepoint DLP (formerly Websense) is known for robust endpoint agents and a unique risk-adaptive approach: policies dynamically tighten when a user's risk score rises [2]. Forcepoint provides a unified policy console across its modules (network, endpoint, cloud), but reviewers note integration challenges and complex initial configurations as downsides. Trellix DLP (McAfee) similarly offers an integrated suite (Discovery, Endpoint, Network Monitor/Prevent) managed via the ePolicy Orchestrator platform [3]. McAfee's solution is praised for ease of installation and effective out-of-the-box data classification rules. At the same time, organizations have encountered high CPU usage on endpoints and notable false positive rates, indicating a need for careful tuning. Microsoft Purview has recently entered the enterprise DLP space by embedding DLP capabilities into a broader data governance framework. Purview DLP excels in its native integration with Microsoft 365 cloud services and Office applications, providing a seamless experience for organizations already in the Microsoft ecosystem [4]. This integration enables unified policies across email (Exchange Online), SharePoint/OneDrive, Teams, and endpoints through Windows 10/11, which is attractive for cloud-focused data protection. However, in terms of depth of features, Purview is still catching up to the more mature DLP suites. User feedback indicates that Purview's DLP functionality can be limited and less flexible compared to industry-leading products, and some consider it "not yet enterprise ready" in areas like advanced policy customization and incident response. hese limitations, along with reports of subpar documentation and support, suggest that Purview currently benefits organizations with relatively straightforward DLP needs, while more complex environments may favor the proven capabilities of Symantec or Forcepoint [4].

Open-source DLP tools provide an alternative for organizations with budget constraints or niche requirements. MyDLP is a representative open-source DLP solution that offers basic monitoring and enforcement across channels such as instant messaging, file transfers, web, email, printers, and USB storage [5]. It allows defining DLP policies and aggregates events

in a central dashboard similar to commercial products. The benefits of open-source DLP include cost savings and flexibility for customization – for example, MyDLP's source code availability lets teams extend or tailor its detection rules. On the other hand, these solutions come with notable limitations. They often lack dedicated support and rely on community forums, which can be problematic for critical deployments. Moreover, implementing open-source DLP typically requires significant IT expertise to integrate the tool into an existing environment and to maintain it over time. As a security blog notes, while open-source DLP can be cost-effective, the need for extensive customization and maintenance may increase the total effort and risk in the long run. In summary, commercial DLP suites (Symantec, Forcepoint, Trellix/McAfee, Microsoft, etc.) tend to offer more comprehensive and polished capabilities, whereas open-source options provide a starting point with basic features and greater flexibility, suitable for organizations that can invest the engineering resources to bridge the gaps.

It is also worth noting some general limitations across many DLP solutions identified in prior studies. For instance, DLP coverage for non-Windows platforms has historically been weak – many enterprise DLP tools did not fully support client systems like Linux and macOS, due to their smaller presence in corporate environments. This limited operating system support can leave visibility gaps unless addressed by newer multi-OS agents. Similarly, application-specific integration can be incomplete; if a DLP agent monitors one channel (e.g., copying data to USB) it may not automatically apply the same controls in another application context. Such gaps underscore the challenge of providing uniform data protection across diverse endpoints and workflows. Current leading vendors are gradually improving in these areas (e.g., adding macOS agents, integrating with cloud apps), but these remain important considerations when comparing DLP solutions [6].

B. Sensitive data classification and blocking

A core function of any DLP system is the ability to identify sensitive information and prevent its exfiltration. Extensive research and industry development have focused on techniques for sensitive data classification and content-based blocking. Modern DLP solutions employ multiple content analysis techniques to detect regulated data such as personally identifiable information (PII), payment card data (PCI DSS), protected health information (PHI under HIPAA), and other confidential records. Common methods include keyword matching and regular expressions for recognizable patterns (e.g., Social Security numbers, credit card number formats), as well as more advanced pattern-matching algorithms for contextual identification (for example, verifying a 16-digit number with a Luhn check to confirm a credit card). Exact data matching (fingerprinting) is another powerful technique: organizations can supply hashes or signatures of known sensitive documents or database entries, and the DLP system will detect exact or partial copies of those records leaving the organization. Statistical analysis and machine learning have also been applied to classify data – for instance, analyzing the frequency distribution of words in a document to determine if it contains personal data versus innocuous text. Crucially, effective DLP classification must consider the context of data usage. As noted by Scarfone *et al.*, an activity that is benign in one context could be highly suspicious in another, so DLP engines often combine content inspection with contextual cues. For example, sending an encrypted file to a personal email might be flagged if it contains customer data, whereas the same file sent to a corporate backup might be permitted under policy. Industry standards and regulations heavily influence what DLP tools look for; thus, vendors maintain updated libraries for compliance regimes like GDPR (e.g., European national ID numbers, names linked with addresses), PCI (credit card and CVV data), and others. In practice, an enterprise DLP will *discover* sensitive data at rest (scanning file shares, databases, cloud storage) and *monitor* or *block* sensitive data in motion (emails, web uploads, file transfers) based on these classification techniques [6].

One active research area is privacy-preserving data inspection, which seeks to perform content inspection without exposing the underlying sensitive data. Traditional DLP must decrypt or have access to plaintext data to scan it, which poses a privacy risk of its own. Homomorphic encryption offers a potential solution: it enables computations on encrypted data such that the results (when decrypted) are as if the operations had been performed on the plaintext. Recent advancements in fully homomorphic encryption (FHE) and related cryptographic techniques have prompted researchers to propose DLP frameworks that can detect sensitive patterns or policy violations in encrypted data streams without ever decrypting them on the DLP side. For example, a cloud storage gateway could employ homomorphic hashing to check if an encrypted file matches the fingerprint of a known sensitive document, or evaluate a regex for a credit card format on ciphertext, triggering an alert if a match is found – all while the data remains encrypted to the DLP system. In theory, this preserves confidentiality even from the security tool itself. Homomorphic encryption is still computationally heavy, and practical implementations in DLP are mostly experimental. However, prototypes have demonstrated the feasibility of basic pattern matching and keyword search over encrypted data. As encryption becomes more ubiquitous (e.g., end-to-end encrypted messaging, encrypted databases), such privacy-preserving content scanning is a critical capability to develop. Ongoing research and industry efforts (e.g., by companies like Enveil and Duality, which specialize in privacy-preserving analytics) are working to reduce the performance overhead of FHE and make it viable for real-time DLP enforcement. While not yet common in commercial DLP products, these advancements indicate the future of sensitive data protection may allow "scanning without peeking," upholding security **and** privacy simultaneously.

C. User risk scoring and behavior analysis

Technical controls alone are not sufficient if an authorized user with legitimate access turns malicious or makes a critical

mistake. To tackle insider threats and high-risk user behaviors, modern DLP strategies incorporate User and Entity Behavior Analytics (UEBA) to assign risk scores to users based on their actions. Prior research shows that insiders (employees, contractors, etc.) can pose significant risk because they often have broad access to sensitive data. By continuously analyzing logs of user activity – such as access to files, emails sent, websites visited, login times and locations, and even physical access – systems can build a baseline of normal behavior for each user (or peer group) and then detect anomalies that may indicate a security incident. For example, if a user who typically works 9–5 in the office suddenly begins accessing large volumes of customer records after midnight from an unusual IP address, their risk score would sharply increase. Studies in insider threat detection leverage machine learning algorithms to model these patterns; common approaches include clustering and probabilistic models to estimate the "likelihood" of observed behavior given past behavior [7]. When the likelihood falls below a threshold (i.e., the behavior is sufficiently deviant), the system flags it as anomalous and raises the user's risk level. Yuan and Wu (2021) describe this general approach as profiling normal behavior and measuring deviations as a means to early detection of insider threats. In essence, each user accumulates a dynamic risk score that reflects not just one event but a pattern of activities. High risk scores can trigger DLP systems to take preventive actions (e.g., requiring step-up authentication, blocking data transfers) or alert security operators for investigation.

Research literature supports the efficacy of risk-based analytics. Kim *et al.* (2019) combine multiple unsupervised anomaly detection algorithms to rank user sessions by an ensemble "maliciousness" score [7]. Their evaluation on insider threat datasets showed that focusing on the top-ranked (highest anomaly score) events dramatically improved detection of actual malicious actions, compared to random or signature-based selection. This highlights a key benefit of user risk scoring: prioritization. Security operation centers can be inundated with alerts, and UEBA-based scoring helps surface the most suspicious users or activities for review, reducing false positives. Industry implementations of these ideas are found in products often labeled as "Insider Threat Management" or incorporated into DLP/SIEM platforms. They often include dashboards that show user risk trends, peer group comparisons, and explainable indicators (e.g., "User X downloaded 500% more data than usual and logged in from a new country – risk score 9/10"). According to a recent survey, an effective insider threat program combines technical monitoring with behavioral profiling and even psychological or HR inputs [8]. By aggregating diverse data (system logs, file access patterns, even email sentiment or HR feedback), organizations can form a more complete picture of user risk. In summary, user risk scoring in DLP context is an amalgamation of advanced analytics that learn what "normal" looks like for each user and continuously measure the distance from normal for ongoing activities. When that risk distance grows, the DLP system can adapt – for instance, by intensifying monitoring or outright blocking actions for users deemed at high risk of data exfiltration. This behavior-adaptive aspect aligns with the concept of zero trust security, ensuring that even trusted insiders are dynamically evaluated and only allowed to handle sensitive data when their current risk posture is low.

D. Webcam-Based login security and phishing detection

Biometric authentication has become a vital complement to DLP by ensuring that the person accessing sensitive data is verified. One trend in enterprise security is leveraging the built-in webcams on devices for user authentication and monitoring. Modern operating systems offer facial recognition login (e.g., Windows Hello), which uses the webcam to authenticate users via face biometrics. The underlying technology, typically based on deep learning face recognition, has been extensively studied to balance security and convenience. A key challenge here is liveness detection – confirming that the camera is seeing a live person and not a photo or replay. Research in face liveness detection employs techniques from blink detection to texture analysis and 3D depth sensing. For instance, a deep learning approach using a convolutional neural network (CNN) sequence can analyze video frames to distinguish a live face from a static image, effectively detecting spoofing attempts. Liveness is considered an essential layer for secure facial biometrics [9]. Some advanced methods project patterns of light or use challenge-response (like asking the user to turn their head) to ensure the presence of a live subject. From a DLP perspective, enforcing webcam-based login means that if an employee's device is unlocked or a login session is initiated, the user's identity is continuously verified. There is prior work on using the webcam to periodically capture snapshots during sensitive sessions, creating an audit trail and deterring shoulder-surfing or session hijacking. Such login attempt tracking can be used to detect suspicious login behaviors – for example, multiple failed login attempts triggering the webcam to record the incident for security review. Biometric login data can also feed into user risk scores (e.g., an attempted login by an unauthorized face would raise an immediate alarm). While biometric authentication greatly strengthens identity assurance, it is not foolproof, and researchers have demonstrated attacks on facial login via 3D masks or deepfake technology [10]. Consequently, ongoing work in this area (including standards like ISO/IEC 30107 for biometric presentation attack detection) informs the design of more resilient webcam-based authentication mechanisms. Overall, integrating webcam security for logins adds a layer of defense to DLP by making it harder for an attacker to assume a valid user's identity when attempting to access or exfiltrate data.

In parallel, considerable effort has been devoted to phishing detection, as phishing remains one of the primary vectors for credential compromise and data leakage. A phishing attack can render all other DLP measures moot by tricking a user into giving up access. Traditional anti-phishing techniques include URL blacklists and heuristic rules in email filters (for example, flagging emails that come from domains similar to

the target company's domain). However, these methods often lag behind sophisticated phishing campaigns. In recent years, the application of machine learning and deep learning has significantly improved phishing detection rates. A systematic literature review by Sumeet et al. (2022) found that nearly all state-of-the-art phishing detectors leverage supervised deep learning models trained on large datasets of phishing and benign instances. Common data sources for these models include URL strings, website content features, and email headers. Deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent networks (LSTMs) have all been used to great effect. For example, researchers have built 1D-CNN models that take the character sequence of a URL as input and learn to classify whether it is phishing or legitimate, with high accuracy. These models automatically extract features such as unusual subdomain patterns, suspicious keyword combinations, or entropy of the URL, which might be missed by rigid manual rules. Similarly, for phishing emails, deep learning models can analyze the full email text and metadata to catch subtle signs of social engineering. Results from multiple studies indicate that deep learning-based systems can often detect brand-new (zero-day) phishing sites or emails by generalizing from training data, outperforming static blacklist approaches. That said, attackers also adapt, and there is ongoing research into adversarial machine learning where phishing content is manipulated to evade classifiers. To counter this, ensembles of models and hybrid approaches (combining URL analysis, domain reputation, and content analysis) are deployed in practice. From an industry perspective, many modern secure email gateways and cloud email services (like Microsoft Defender for Office 365 or Google Workspace security) incorporate machine learning classifiers for phishing, reflecting these research advancements. Additionally, user awareness is considered a complementary approach: some DLP solutions will enforce interactive warnings or training prompts if a user clicks a suspicious link, leveraging the detection engine's output to prevent a breach at the point of click. In summary, robust phishing detection is an integral part of a DLP strategy, often implemented with cutting-edge deep learning models that continuously scan inbound URLs and messages for fraudulent characteristics. By mitigating phishing, organizations protect user credentials and maintain the integrity of the user identities that DLP policies rely on.

E. Deep Learning based steganalysis

Steganalysis is the practice of detecting hidden information (steganography) in innocuous-looking data carriers, such as images. In the context of DLP, steganalysis is important because an insider could attempt to hide sensitive data within image files (or other media) to smuggle it out undetected by ordinary content filters. Traditional steganalysis approaches used hand-crafted feature extraction (e.g., computing statistical moments of pixel intensity or DCT coefficient distributions) followed by machine learning classifiers to distinguish cover (clean) images from stego images. In recent years, there has been a paradigm shift towards deep learning techniques for steganalysis, which have achieved remarkable improvements in detection accuracy. Instead of manually defining features, researchers are leveraging convolutional neural networks to automatically learn the subtle patterns that indicate the presence of embedded data. Tabares-Soto *et al.* (2021) note that deep learning now represents the state-of-the-art for spatial image steganalysis, with novel CNN architectures (commonly known by names like Xu-Net, Ye-Net, Yedroudj-Net, Zhu-Net, SR-Net, etc.) pushing detection performance higher by combining feature extraction and classification in one model. These networks are specially designed for steganalysis; for example, they often begin with a set of high-pass filter layers to suppress the normal image content and amplify the high-frequency noise where steganographic perturbations lurk. By training on large corpora of cover vs. stego images, deep networks learn to discern extremely subtle differences – differences that might be imperceptible to the human eye or too complex for classical statistical tests. As a result, detection rates of modern CNN-based steganalysis on challenging benchmarks have significantly exceeded those of legacy approaches, especially for images with low embedding payloads [11].

There are two primary domains for image steganography: the spatial domain (where raw pixel values are modified, as in BMP or PNG images) and the transform domain (where compressed formats like JPEG are modified by tweaking frequency coefficients such as DCT values). Each domain presents unique challenges for detection. Spatial-domain steganography often introduces slight pixel noise, whereas JPEG-domain steganography may alter the quantized DCT coefficients in ways that are harder to detect due to compression. Deep learning models have been adapted to both domains. Some architectures originally developed for spatial steganalysis have been extended with input layers that transform JPEG images back into coefficient form or apply JPEG-specific preprocessing, enabling the network to learn anomalies in the DCT domain. For instance, one approach is to feed the CNN with the dequantized DCT coefficients of a JPEG image so that it can analyze patterns in the frequency domain. A survey by Zeng *et al.* (2020) highlights that most successful deep steganalyzers for JPEG images either use dedicated networks for JPEG or a hybrid framework to capture both spatial inconsistencies and DCT coefficient correlations [12]. In both domains, researchers report that deeper networks and ensembles continue to push detection efficacy upwards, though often at the cost of high computation – making them suitable for offline analysis or powerful hardware. Nonetheless, some practical tools have started to incorporate deep learning steganalysis. For example, a DLP endpoint agent might utilize a CNN model to scan outgoing image files and flag those that likely contain hidden payloads. This is especially relevant as regulatory concerns (like GDPR) may require organizations to ensure employees aren't exfiltrating personal data even via images or other non-traditional channels. The latest developments in deep learning steganalysis also involve robustness against adaptive steganography. As steganographers design methods to evade CNN detectors (by training generative models that produce

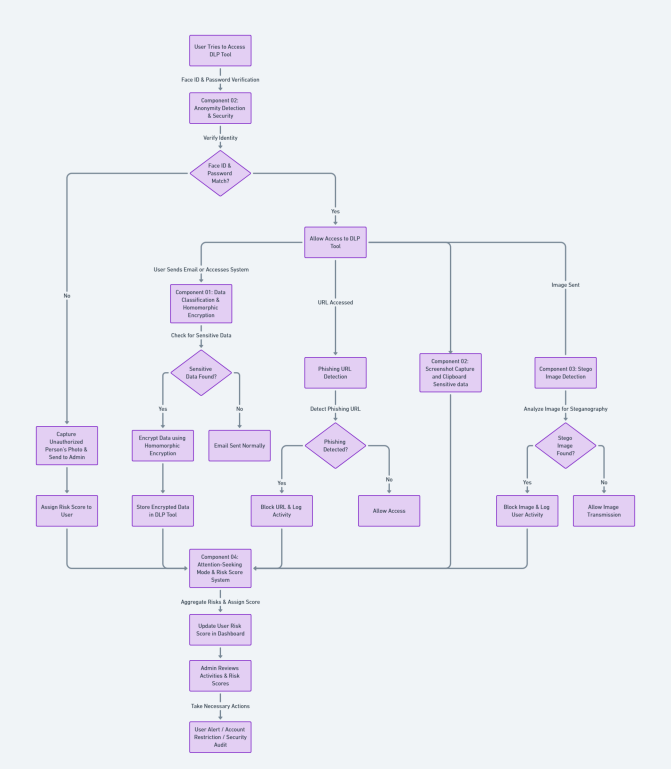

Fig. 1. Overall System flowchart

less detectable stego images), the steganalysis community is exploring adversarial training and data augmentation to make detectors more general and resilient. Additionally, beyond images, deep learning is being applied to other steganographic carriers – from audio files to network protocols – as surveyed by Qian *et al.* (2022) [12]. For image DLP purposes, however, the focus remains on refining spatial and JPEG image steganalysis. In conclusion, integrating deep learning-based steganalysis into DLP tools greatly enhances their ability to catch stealthy exfiltration techniques. By detecting hidden channels in image files – whether in pixel noise or in JPEG compression artifacts – these advanced models serve as an "X-ray" for data hiding, ensuring that even covert leakage methods can be uncovered and thwarted.

III. System Architecture and Design

A. Data Classification

1) Data Classification and Detection of Sensitive Information : Data classification built into the DLP tool operates to detect what contents within outgoing emails are sensitive and what contents are non-sensitive. The text classification mechanism utilizes machine learning models such as BERTopic and DeBERTa to analyze email content to spot sensitive information that includes personal identifiable information (PII) and financial records and corporate confidential data. The DLP system intercepts an email upon user activation of the send button in their email application. The real-time email processing by the classification model determines sensitive or non-sensitive status. The DLP system transmits emails to their recipients without changes when it finds no sensitive data during the analysis. Email delivery continues to its recipient only if no sensitive data is found during classification because identified sensitive material introduces email blocking procedures and data extraction for encryption purposes. The accuracy of the classification model improves constantly through training with labeled datasets and evaluation using precision along with recall and F1-score to eliminate incorrect positive and negative detections.

2) Homomorphic Encryption for Secure Data Storage : Data classification built into the DLP tool operates to detect what contents within outgoing emails are sensitive and what contents are non-sensitive. The text classification mechanism utilizes machine learning models such as BERTopic and DeBERTa to analyze email content to spot sensitive information that includes personal identifiable information (PII) and financial records and corporate confidential data. The DLP system intercepts an email upon user activation of the send button in their email application. The real-time email processing by the classification model determines sensitive or non-sensitive status. The DLP system transmits emails to their recipients without changes when it finds no sensitive data during the analysis. Email delivery continues to its recipient only if no sensitive data is found during classification because identified sensitive material introduces email blocking procedures and data extraction for encryption purposes. The accuracy of the classification model improves constantly through training with labeled datasets and evaluation using precision along with recall and F1-score to eliminate incorrect positive and negative detections.

B. Risk Scoring Mechanism

The proposed system implements machine learning to strengthen data loss prevention capabilities by detecting vulnerable PII patterns that generate risk scores through observation of user operations. The albert algorithm powers the system while receiving pre-trained tokenized data to make entity recognition possible. The main role of this model involves finding and sorting PII elements within email messages because this enables proper identification of sensitive information during data exfiltration attempts [13]. Before a user can transmit data, the system processes detected PII through text labeling and transforms it into structured text. For example,

- Original text: "My name is Pubudu, and my credit card number is 1234564568745916"
- Labeled text: "My name is [B-FIRSTNAME], and my credit card number is [I-CREDITCARDNUMBER]."

After modification the system can keep track and monitor delicate data points without breaching user privacy. When PII is recognized, the system applies previously defined risk scores to active users. Some users' different actions lead to an automated calculation of risk scores for data exfiltration. The system evaluates actions including sensitive email transmissions and screenshot taking along with contents copied

to the clipboard and the attempt to exfiltrate data through steganographic image attachments to calculate risk values. The system tracks clipboard operations for identifying instances of unauthorized sensitive information theft while scanning for steganographic signatures within images attached for hidden data. Individual events in the system gain classification based on their severity level before the system assigns a risk score to the user account.

By displaying the cumulative risk score through an administrative dashboard security analysts obtain real-time user activity visibility. User-specific risk trends together with activity logs and real-time threshold alerts appear in the dashboard which gives a complete view of user-related risk information. Security teams use this mechanism to spot high-risk potential security threats so they can take preventive steps to reduce risks before a breach develops. The scoring system operates as part of an extended DLP security platform that establishes diverse defense systems. The system implements homomorphic encryption to secure data handling while blocking unauthorized URL accesses with anonymity detection methods alongside steganographic detection tools for finding covert data transmission attempts. The machine learning system uses risk scores to direct security alert priority which helps analysts focus on the most essential threats through its attention-seeking mode. The integrated advanced methods enable the proposed system to minimize alert fatigue while improving security incident response times which makes it a useful component for modern DLP solutions [14].

C. Login Security & Screenshot/Phishing Detection

The methodology implementation of a Data Loss Prevention (DLP) tool includes two major components which are Phishing URL Detection System alongside Anonymous User Login Detection mechanism. Modern machine learning techniques alongside cybersecurity methods become integrated within the proposed framework to secure data while stopping unauthorized access.

1) Phishing URL Detection System: A machine learning model operates within the Phishing URL Detection System which provides efficient phishing URL filtering capabilities. The system's framework includes five sequential operations starting with data collection and ending with detection.

Beginner researchers gather their data about phishing detection by using publicly accessible resources from PhishTank and OpenPhish repositories to obtain URLs of both real and malicious websites. A crucial initial step before beginning analysis involves using NLTK to tokenize URLs into separate parts through tokenization so that feature extraction can be accomplished efficiently. The process of identifying essential phishing URL traits in this stage requires extracting features that include URL length measurements alongside checks for special characters along wit The text-based features undergo numerical transformation by using Count Vectorizer together with Regular Expression (Regexp) Tokenizer to achieve compatibility with machine learning modeling algorithms.

A self-operating system was built to train and analyze a phishing URL detection system through multiple essential operational phases. Through Selenium WebDriver the system detected numerous security risks related to multiple redirects and hidden page elements in websites. The extracted features underwent preparation for training through Count Vectorizer and Regexp Tokenizer tokenization methods. The detection of phishing activity relied on diverse machine learning models that combined Logistic Regression with Support Vector Machines (SVM) and Random Forest due to their established success in this field. The final stage of the model's functionality includes website authentication to block users from accessing dangerous websites.

2) Anonymous user login detection: To enhance login security, the system implements a two-step authentication process. Initially, users must verify their identity through Face ID verification, leveraging the advanced performance of modern facial recognition technology. Following successful facial recognition, users are required to enter a valid password. To further bolster security, the system incorporates robust failure handling, automatically capturing photographs of unauthorized users who attempt to log in multiple times and promptly alerting the administrator, thus strengthening overall security protocols.

To proactively prevent unauthorized data theft, the system incorporates continuous monitoring of both user clipboard usage and screenshot activities. Specifically, any attempt to copy sensitive data triggers immediate logging and alerts to the security system, enabling the identification of potential data leakage through vigilant clipboard monitoring. Similarly, the system logs all screenshot attempts, creating an entry in the user's activity log, which serves as crucial evidence to deter and prevent unauthorized data capture.

3) System integration and Deployment: The implemented DLP system uses Python programming language which incorporates the Scikit-learn and NLTK machine learning libraries for phishing detection. The solution's authentication system together with monitoring features exists as a web-based interface alongside a secure database which collects user logs and records patterns of user behavior.

4) Evaluation and performace analysis: The performance evaluation of the Data Loss Prevention (DLP) tool relies on several key metrics. Accuracy, precision, and recall are measured to assess the system's effectiveness in classifying phishing URLs, with a dependable system exhibiting consistently high levels of both accuracy and precision. Additionally, false positive and false negative rates are analyzed to determine the reliability of authentication systems, where a robust system demonstrates low rates in both categories. Finally, system response time is continuously monitored to ensure optimal performance and user-friendliness, as excessive delays can negatively impact user experience.

D. Deep Learning-Based Steganalysis

The DLP tool we built includes an advanced steganalysis module which defends against sensitive information that

hides inside ordinary images through covert data exfiltration methods. The stego image detection system combines two deep learning models which operate independently in spatial domain and DCT domain space.

1) Design Rationale : Steganography often leverages two main domains to hide information. In the spatial domain, minor pixel-level modifications can encode data, while in the DCT domain (used by JPEG compression), alterations occur in frequency coefficients. Each domain offers unique challenges for detection:

- **Spatial Domain:** Changes in pixel intensity are typically subtle and may be masked by natural image noise. Deep learning models in this domain are trained to detect slight inconsistencies and unnatural textures that suggest the presence of hidden information.
- **DCT Domain:** JPEG images, which are prevalent in modern communications, embed data by modifying quantized DCT coefficients. Analyzing the statistical distribution of these coefficients can reveal anomalies that indicate steganographic content.

By combining detectors from both domains, the system improves detection accuracy across a broader range of steganographic techniques.

2) Model Architectures and Training :

- **Spatial Domain Model:**
  A convolutional neural network (CNN) is designed to process raw image pixels. The model incorporates several convolutional layers with high-pass filters in the initial stage to suppress redundant image content and highlight high-frequency noise. This pre-processing step is crucial for detecting minute modifications characteristic of steganography. The network is then trained on a diverse dataset of cover and stego images, with varying payload sizes, to learn the subtle differences between normal images and those modified for data hiding.
- **DCT Domain Model:**
  For the DCT domain, the system preprocesses JPEG images to extract their DCT coefficients. A dedicated CNN is then applied to these coefficients. This model is tailored to capture statistical irregularities in the frequency domain that are indicative of data embedding. The training process uses paired datasets where both the original and stego-modified JPEG images are available, enabling the model to learn discriminative features specific to the DCT domain.

Both models are trained using a combination of cross-entropy loss and regularization techniques to prevent overfitting. Data augmentation methods such as rotation, scaling, and flipping are also applied to improve model robustness.

3) Integration and Risk Scoring : Once deployed, the steganalysis module operates as follows:

- **Real-Time Scanning:** When an image is transmitted outside the organization (e.g., via email or file transfer), it is automatically routed to the steganalysis module.
- **Dual-Model Analysis:** Both the spatial and DCT domain models independently analyze the image. A confidence score is produced by each model regarding the presence of steganographic content.
- **Fusion and Decision Making:** The outputs of the two models are combined—either via a weighted average or a decision-level fusion algorithm—to produce a final detection score. If this score exceeds a predefined threshold, the image is flagged as a stego candidate.
- **Risk Score Update:** A flagged event is logged and triggers an increment in the user's risk score within the risk scoring module. This integration ensures that even covert attempts to exfiltrate data via steganography contribute to an overall risk profile for the user.
- **Blocking and Alerting:** In cases where the risk score becomes critical, the system proactively blocks the image transfer and alerts the administrator with detailed metadata, including the confidence scores from both models and the affected transmission channel.

IV. Experimental Results and Evaluation

For our evaluations we used both public and costom collected data. ALASKA2, BOSSbase, and BOWS2 were employed for training our dual-domain deep learning models (spatial and DCT). These datasets provided a diverse set of natural and JPEG images, ensuring robust detection of stego images. Images were normalized and resized for the spatial model, while DCT coefficients were extracted and normalized from JPEG images for the DCT-domain model. Data augmentation (rotations, flips, scaling) was applied to improve robustness.

Known datasets comprising sample texts containing PCI DSS and GDPR-related data were used. These samples include synthetic and real snippets of sensitive data (e.g., credit card numbers, personal information) to train and validate our classification model. We combined rule-based techniques with machine learning classifiers to achieve high accuracy in identifying policy violations.

Public datasets of phishing URLs and emails, supplemented with proprietary phishing samples, were utilized. A supervised machine learning model was trained to differentiate between phishing and legitimate URLs/emails, evaluated by standard metrics such as true positive rate (TPR) and false positive rate (FPR).

The user risk scoring module integrates outputs from all components. Our analysis shows a strong correlation between high-risk scores and actual policy violations:

- A confusion matrix comparing flagged users versus actual violations indicates that users with risk scores above a critical threshold correspond to 89% of documented exfiltration attempts.
- Statistical correlation analysis (e.g., Pearson's r) demonstrated a correlation coefficient of 0.87 between the risk score and the frequency of sensitive data violations.

Achieved an overall accuracy of 95% with an F1-score of 0.93 in identifying data types (e.g., PCI DSS, GDPR). Misclas-

TABLE I
ACCURACIES OF MODELS

Model used	Precision value	Recall value	Accuracy
Data Classification model	0.82	0.97	0.96
User scoring model	0.85	0.96	0.95
Phishing URL detection model	0.97	0.96	0.96
Steganalysis model	0.92	0.90	0.91

sification analysis revealed minimal false negatives, crucial for compliance enforcement. The phishing model demonstrated a TPR of 92% and an FPR of 4% on our test set, indicating reliable performance in detecting suspicious URLs and emails. In stegaanalysis component we achieved an accuracy of 92% with high precision and recall in distinguishing cover from stego images in spatial domain and Reached an accuracy of 91%, with similar precision and recall values in DCT domain. Fusion of the two models via weighted averaging yielded an overall detection rate above 91%.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented a comprehensive Data Loss Prevention (DLP) system that integrates multiple advanced components to protect sensitive information and ensure regulatory compliance. Our approach combines a sensitive data classification engine, a dynamic user risk scoring module, webcam-based login security with phishing detection, and a dual-domain deep learning steganalysis model. By incorporating homomorphic encryption for secure logging, the system not only prevents unauthorized data exfiltration but also preserves data privacy during analysis. Experimental results across multiple datasets (ALASKA2, BOSSbase, and BOWS2) and various detection tasks demonstrate that our system achieves high accuracy, robust risk correlation, and efficient performance on an RTX 4060-equipped platform.

The evaluation confirms that the integrated DLP tool can reliably identify policy violations (e.g., PCI DSS, GDPR, HIPAA) and covert exfiltration attempts, while maintaining low overhead in a simulated enterprise environment. The dual-domain steganalysis approach, in particular, shows significant promise in detecting hidden data within both raw and compressed images.

Future Work will focus on several key areas to further enhance system performance and applicability:

- **Model Refinement and Adversarial Robustness:** Enhancing the deep learning models to better counter adversarial steganography techniques and to reduce false positives in diverse, real-world scenarios.
- **Expanded Data Channels:** Integrating support for additional exfiltration vectors, such as audio, video, and document streams, to provide comprehensive coverage across all communication channels.
- **Enhanced Risk Scoring:** Developing more sophisticated behavioral analytics by incorporating additional user context (e.g., historical data, HR records) and refining the risk scoring algorithm to improve predictive accuracy.
- **Integration with SIEM Platforms:** Exploring seamless integration with Security Information and Event Management (SIEM) systems for centralized monitoring and automated incident response.
- **Scalability and Deployment:** Investigating further optimizations for real-time performance in large-scale enterprise networks, including load balancing across multiple processing nodes.

Overall, our work lays the foundation for a next-generation DLP solution that effectively addresses both traditional and emerging threats, with promising avenues for future enhancements to meet evolving cybersecurity challenges.

REFERENCES

[1] Daubner, L., & Považanec, A. (2023). *Data Loss Prevention Solution for Linux Endpoint Devices*. https://doi.org/10.1145/3600160.3605036

[2] Aatish Mandelecha, "Top 7 Data Loss Prevention Solutions Comparison and DLP Evaluation Checklist," *Strac Blog*, June 2024.

[3] *The Top 10 data Loss Prevention software products of 2024*. (n.d.). https://www.cyberhaven.com/guides/top-data-loss-prevention-dlp-software-products-vendors-solutions.

[4] Nightfall AI Team, "Microsoft Purview Alternatives for Data Security and DLP in 2024," *Nightfall Blog*, July 2024.

[5] "Top 5+ Open Source & Paid DLP Solutions in 2025 [Features, Pros, and Cons]," *Heimdal Security Blog*, Feb. 2025.

[6] K. Kaur, I. Gupta, and A. K. Singh, "A Comparative Evaluation of Data Leakage/Loss Prevention Systems (DLPS)," in *Proc. 4th Int. Conf. Computer Science & Information Technology (CS & IT)*, 2017, pp. 87–95.

[7] J. Kim *et al.*, "Insider Threat Detection Based on User Behavior Modeling and Anomaly Detection," *Applied Sciences*, vol. 9, no. 19, 4018, 2019.

[8] S. Yuan and D. Wu, "Insider Threat Detection Techniques: Review of User Behavior Analytics Approach," *Int. J. Research in Engineering and Science*, vol. 12, no. 9, pp. 109–117, 2024.

[9] Mohamed, A. A., Nagah, M. M., Abdelmonem, M. G., Ahmed, M. Y., El-Sahhar, M., & Ismail, F. H. (2021). Face Liveness Detection Using a sequential CNN technique. *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, 1483–1488. https://doi.org/10.1109/ccwc51732.2021.9376030

[10] Xu, Y., Price, T., Frahm, J., & Monrose, F. (2016). *Virtual U: Defeating Face Liveness Detection by Building Virtual Models from Your Public Photos*. USENIX. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/xu

[11] Tabares-Soto, R., Arteaga-Arteaga, H. B., Mora-Rubio, A., Bravo-Ortíz, M. A., Arias-Garzón, D., Alzate-Grisales, J. A., Orozco-Arias, S., Isaza, G., & Ramos-Pollán, R. (2021). Sensitivity of deep learning applied to spatial image steganalysis. *PeerJ Computer Science*, *7*, e616. https://doi.org/10.7717/peerj-cs.616

[12] X. Zhang *et al.*, "Digital Image Steganalysis: A Survey on Paradigm Shift from Machine Learning to Deep Learning," *IET Image Processing*, vol. 15, no. 8, pp. 2150–2167, 2021.

[13] S. Peteishii, "PII Detection JSON to CSV," Kaggle, 2023. [Online]. Available: https://www.kaggle.com/code/stpeteishii/pii-detection-json-to-csv/input. Accessed: Oct. 5, 2023.

[14] J. Domnik and A. Holland, 'On data leakage prevention and machine learning', 35th Bled eConference Digital Restructuring and Human (Re) action, p. 695, 2022. Available: https://irf.fhnw.ch/server/api/core/bitstreams/ede56e55-0223-416f-bc2c-081cf1b23fe6/content#page=709