ISSN: 2349-6002

A Hybrid Ai Pipeline for Detecting Adversarial Access and Generating Deceptive Document Ecosystems

Ms P Subasri[1], Hariharan K[2], Hariprasath M[3], Thiruvarasan G[4]
[1]*AP/AI&DS, Department of Artificial Intelligence and Data Science, Surya Group of Institutions, Anna University*
[2,3,4]*Department Of Artificial Intelligence and Data Science, Surya Group of Institutions, Anna University*
doi.org/10.64643/IJIRTV12I11-200899-459

***Abstract*—Intellectual property (IP) constitutes the lifeblood of modern organizations, encompassing creative works, inventions, proprietary research, and confidential business knowledge that sustain innovation and competitive advantage. With the rise of large-scale automation and intelligent data-mining tools, cyber adversaries now target IP repositories using machine-driven classification, clustering, and topic-extraction pipelines capable of rapidly identifying high-value information.**

To counter these emerging threats, this project introduces DARD (Decoy Approaches for Robust Protection against IP Theft), a deception-oriented IP protection framework that employs a Variational Autoencoder (VAE) for anomaly detection and NLP-driven document manipulation techniques including TF-IDF feature extraction, K-Means clustering, and LDA topic modeling. The system misleads automated adversarial tools by generating deceptive document ecosystems featuring keyword permutation, selective removal, and topic substitution, thereby preserving the confidentiality of sensitive IP while maintaining seamless access for legitimate users.

***Index Terms*—Spam Detection, Logistic Regression, Machine Learning, TF-IDF, Text Preprocessing, Email Filtering, Bag of Words, Natural Language Processing, Binary Classification, Cybersecurity.**

I. INTRODUCTION

Intellectual property (IP) forms the cornerstone of organizational innovation, encompassing proprietary research, trade secrets, creative works, and confidential business knowledge that provide competitive advantage in an increasingly digital economy. With the rapid proliferation of AI-driven data-mining tools, cyber adversaries now exploit automated document classification, clustering, and topic-extraction pipelines to systematically identify and extract high-value information from organizational repositories.

Traditional security mechanisms such as encryption, firewalls, and role-based access control have proven insufficient against sophisticated machine-learning-based attacks that bypass perimeter defenses by operating on legitimately accessed or exfiltrated data. To address this critical gap, this paper presents DARD (Decoy Approaches for Robust Protection against IP Theft), a hybrid AI pipeline that integrates Variational Autoencoder- based behavioral anomaly detection with NLP- driven deceptive document generation to actively mislead adversarial analytical systems while preserving seamless access for legitimate users.

A. Problem Statement

Intellectual Property (IP) theft poses a significant threat to organizations, as adversaries increasingly leverage automated techniques to extract and analyze sensitive information from document repositories. Traditional security mechanisms such as encryption, firewalls, and role-based access control often fail against sophisticated attacks that exploit machine learning-based document clustering and topic modeling to infer valuable insights from exfiltrated data.

Existing defense mechanisms are largely reactive and perimeter-focused, failing to detect unauthorized access patterns in real-time or disrupt the automated classification processes that adversaries rely upon. Without an effective countermeasure that actively misleads attackers during reconnaissance, organizations remain vulnerable to losing their competitive edge, financial assets, and strategic innovations to well- resourced threat actors.

B. Objectives

The primary objective of this research is to design and implement a proactive IP protection framework that detects adversarial access and dynamically generates deceptive document environments to mislead automated analysis systems. The system aims to integrate behavioral anomaly detection using VAEs with NLP-based document manipulation to create a comprehensive, two-stage defense pipeline.

Specific objectives include:

1) Developing a VAE-based anomaly detection module that learns normal user behavioral patterns and identifies deviations indicative of adversarial exploration;
2) Implementing NLP-driven document manipulation techniques such as keyword permutation, selective keyword removal, and LDA topic substitution;
3) Evaluating the effectiveness of the deceptive ecosystem in corrupting adversarial analytical outcomes while preserving access for legitimate users.

C. Organization of Paper

This paper is organized into seven major sections to provide a comprehensive understanding of the proposed system. Following this introduction, Section II reviews related work on traditional IP protection techniques and their limitations, while Section III presents the system analysis including existing systems and the proposed framework.

Section IV details the system design encompassing the architecture, module descriptions, data flow diagrams, and database design. Section V elaborates on the implementation covering software and hardware requirements, preprocessing, model training, and the alert system. Section VI presents the results and discussion including performance metrics and comparative analysis, and Section VII concludes the paper with directions for future research.

II. RELATED WORK

A. Traditional IDS Techniques

Traditional IP protection techniques have primarily relied on perimeter-based security measures such as firewalls, intrusion detection systems (IDS), encryption protocols, and role-based access control (RBAC) to safeguard sensitive organizational assets. These approaches focus on preventing unauthorized access by establishing defensive boundaries around digital resources and have been foundational to cybersecurity practice for decades.

More advanced techniques include watermarking for document tracing, digital rights management (DRM) for access enforcement, and honeypot systems designed to detect intruders by luring them toward decoy resources. Researchers have also proposed machine learning-based intrusion detection systems that analyze network traffic patterns to identify anomalous behavior, forming the basis upon which the proposed VAE-based anomaly detection module is conceptually grounded.

B. Limitations of Existing Research

Despite their widespread adoption, existing IP protection mechanisms suffer from fundamental limitations in defending against modern AI-driven adversarial attacks. Perimeter-based defenses are inherently reactive, offering little protection once an attacker has gained access to the network or document repository, and they cannot disrupt automated document-analysis pipelines that operate on legitimately accessed or exfiltrated data.

Honeypot systems, while deceptive in nature, are static and easily identified by sophisticated adversaries who can distinguish decoy resources from genuine ones based on metadata and semantic inconsistencies. Machine learning-based IDS models often suffer from high false-positive rates and fail to generalize to novel attack vectors. The existing literature lacks a holistic framework that integrates real-time anomaly detection with dynamic, semantically coherent document manipulation to actively mislead adversarial AI pipelines.

III. SYSTEM ANALYSIS

A. Existing System

The existing systems for IP protection predominantly employ a combination of static security controls including network-level firewalls, endpoint encryption, and access control lists to restrict unauthorized document access. Document-level protection is typically achieved through DRM tools and file-level encryption, which prevent unauthorized copying or redistribution but do not address the threat of adversarial document analysis by automated AI

pipelines.

Current anomaly detection approaches in enterprise environments rely on rule-based SIEM (Security Information and Event Management) systems and signature-based IDS that generate alerts based on predefined thresholds. These systems are inherently limited by their inability to model complex non-linear behavioral patterns, making them susceptible to low-and-slow reconnaissance attacks conducted by sophisticated adversaries who stay within normal activity thresholds.

B. Drawbacks

The most critical drawback of existing systems is their passive, perimeter-centric design philosophy, which provides no mechanism to mislead or misdirect adversaries who have already gained partial access to sensitive document repositories. Once an attacker successfully exfiltrates documents, traditional protections offer no means of corrupting the analytical value of the stolen data, leaving the intellectual property fully exposed to automated classification and knowledge extraction.

Additionally, rule-based anomaly detection systems fail to adapt to evolving adversarial tactics and require constant manual tuning to maintain effectiveness, leading to operational overhead.

Static honeypot environments lack semantic realism and are easily identified by adversaries employing linguistic analysis, while existing document obfuscation techniques are often too aggressive, degrading document utility for legitimate users as well.

C. Proposed System

The proposed system, DARD, introduces a two- stage deception-oriented defense pipeline that combines VAE-based anomaly detection with NLP- driven document manipulation to protect IP from automated adversarial analysis. In the first stage, the VAE module continuously monitors user activity logs, learns probabilistic representations of normal behavioral patterns, and flags interactions that exhibit statistically significant deviations as adversarial access events.

Upon detecting suspicious access, the system dynamically constructs a deceptive document enclave tailored to mislead automated analysis pipelines. This enclave is generated using TF-IDF feature extraction to identify keyword distributions, K-Means clustering to organize documents into thematic groups, and LDA topic modeling to simulate realistic topic structures. Manipulation operations including keyword permutation, selective keyword removal, and topic substitution ensure that adversarial tools receive plausible yet analytically corrupted information while legitimate users continue to access the authentic document repository through an isolated secure enclave.

D. Feasibility Study

The technical feasibility of the proposed system is well-supported by mature open-source libraries including scikit-learn for NLP preprocessing and clustering, Gensim for LDA topic modeling, and TensorFlow/Keras for VAE implementation. The system's modular architecture ensures that each component can be independently developed, tested, and scaled without requiring specialized hardware, making it deployable on standard enterprise server infrastructure with GPU acceleration for deep learning components.

From an economic standpoint, the proposed system leverages existing open-source frameworks and cloud-compatible deployment strategies, minimizing infrastructure investment. The operational feasibility is high given that DARD's deceptive document generation is designed to be transparent to legitimate users, requiring no changes to their workflows. A preliminary risk analysis indicates that the primary challenge lies in maintaining semantic coherence of deceptive documents, which the LDA-driven generation module is specifically designed to address.

IV. SYSTEM DESIGN

A. Architecture

The DARD system architecture is organized into three primary layers: the monitoring layer, the detection layer, and the deception layer. The monitoring layer captures and preprocesses user activity logs from document management systems, generating structured behavioral feature vectors that serve as input to the detection layer. The detection layer hosts the VAE model, which continuously evaluates incoming behavioral data against learned normal distributions and raises alerts when anomaly scores exceed a

ISSN: 2349-6002

configurable threshold.

Upon alert generation, the deception layer is activated, dynamically constructing a modified document environment using the NLP manipulation pipeline. A secure enclave module ensures that legitimate users are transparently redirected to the authentic document repository while adversarial sessions are served the deceptive environment. A centralized dashboard provides administrators with real-time visibility into anomaly events, manipulation operations, and system health metrics.

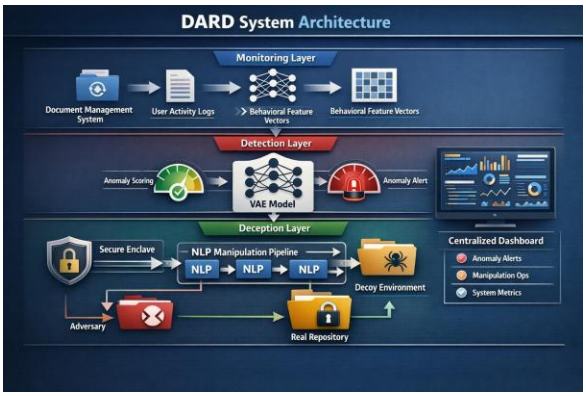

Figure 1: System Architecture

B. Module Description

The User Behavior Monitoring Module captures granular interaction logs including document access timestamps, session durations, search query patterns, and download frequencies, which are aggregated into fixed-length behavioral feature vectors. The VAE Anomaly Detection Module encodes these vectors into a probabilistic latent space and reconstructs them, using reconstruction error as an anomaly score to identify adversarial sessions with statistically significant behavioral deviations.

The NLP Document Manipulation Module constitutes the core deception engine, comprising a TF-IDF vectorizer for keyword extraction, a K- Means clusterer for document grouping, and an LDA topic modeler for thematic structure generation. The Alert and Enclave Management Module orchestrates the transition between authentic and deceptive environments upon anomaly detection, maintaining session isolation to ensure that legitimate users are never inadvertently served manipulated content.

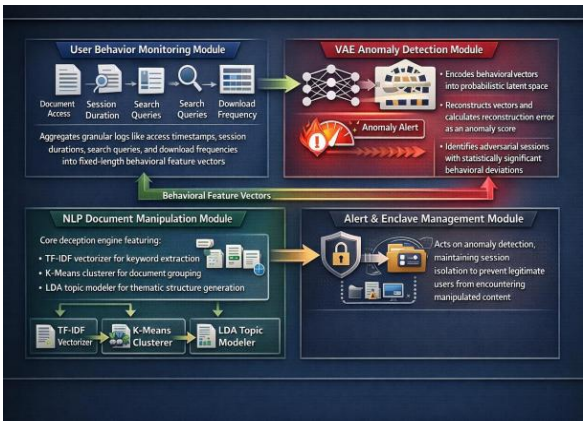

FIGURE 2: MODULE DESCRIPTION

C. Data Flow Diagram

The data flow originates at the user interaction layer, where all document access events are captured and routed to the behavioral feature extraction engine, which transforms raw logs into structured numeric vectors. These vectors are streamed to the VAE model, which performs encoding and reconstruction in near real-time, generating an anomaly score for each session that is evaluated against a dynamic threshold calibrated during the training phase.

When an anomaly is detected, the data flow bifurcates: the authentic document stream is preserved for the legitimate user session, while the manipulation pipeline receives the session context and generates a customized deceptive document set. The deceptive documents, along with a session isolation token, are routed to the adversarial session through the enclave management module, ensuring complete separation of the authentic and deceptive data flows throughout the system.

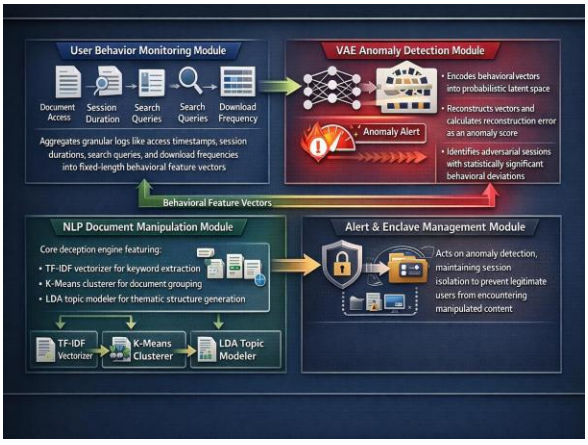

Figure 3: Data Flow Diagram DFD

ISSN: 2349-6002

D. Database Design

The system's database is structured around three primary schemas: the User Activity Schema, which stores timestamped behavioral event logs with fields for user ID, document ID, action type, session token, and feature vector; the Document Repository Schema, which maintains metadata and content hashes for both authentic and deceptive document sets; and the Anomaly Events Schema, which records detected anomaly events with associated reconstruction error scores, session identifiers, and timestamps.

The deceptive document cache is stored separately from the authentic repository using a dedicated encrypted schema, with access controlled through role-based policies enforced at the application layer. Indices are maintained on session tokens and document identifiers to support the low- latency enclave switching required for seamless adversarial session redirection, while a temporal partitioning strategy ensures efficient querying of historical activity logs for model retraining.

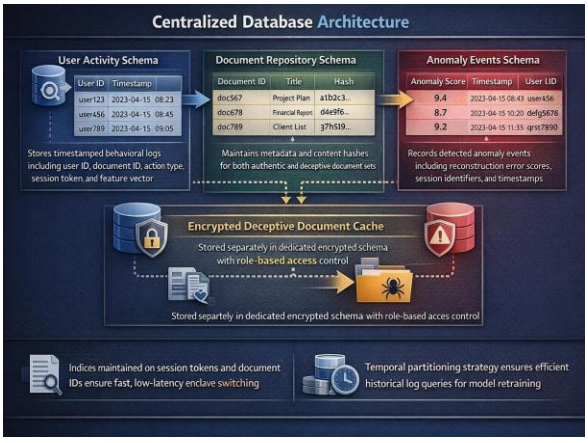

Figure 4: Database Design

V. IMPLEMENTATION

A. Software Requirements

The system is implemented using Python 3.8 as the primary development language, leveraging TensorFlow 2.x and Keras for VAE model construction and training. Natural language processing components utilize scikit-learn for TF- IDF vectorization and K-Means clustering, Gensim for LDA topic modeling, and NLTK for text preprocessing operations including tokenization, stop-word removal, and lemmatization.

The web dashboard is built using the Flask framework with MySQL as the backend database, deployed on a WampServer instance for local development and a Linux-based production server. Frontend components utilize Bootstrap 5 for responsive layout design, with Chart.js integrated for real-time anomaly score visualization. Additional dependencies include NumPy, Pandas, and Matplotlib for data manipulation and analysis.

B. Hardware Requirements

The system requires a minimum of an Intel Core i5 (8th generation or above) or equivalent AMD processor with at least 8 GB of RAM to support concurrent execution of the VAE model, NLP pipeline, and web server. Storage requirements include a minimum of 100 GB SSD to accommodate document repositories, model checkpoints, and activity log databases with adequate I/O throughput for real-time monitoring operations.

For production deployment with larger document repositories, a GPU-enabled server with an NVIDIA GPU supporting CUDA 11.x is recommended to accelerate VAE training and inference. A dedicated network interface card with

1 Gbps throughput is advised to support high-frequency activity log ingestion without bottlenecking the behavioral monitoring pipeline.

C. Dataset Description

The system is evaluated using a synthetic IP document corpus comprising 2,000 documents across five thematic categories: research reports, financial projections, product specifications, legal agreements, and strategic plans. Each document ranges between 500 and 5,000 words and is annotated with metadata including category labels, sensitivity scores, and access control attributes to support controlled evaluation of the manipulation pipeline's effectiveness.

User activity logs are synthetically generated to simulate both normal and adversarial behavioral patterns, with normal sessions characterized by low document access rates, diverse query patterns, and moderate session durations. Adversarial sessions are modeled after real-world reconnaissance patterns featuring high-frequency sequential document access, repetitive keyword-specific querying, and bulk download behaviors consistent with automated scraping pipelines.

D. Preprocessing Implementation

Document preprocessing involves a multi-stage NLP pipeline that begins with tokenization and stop-word removal using NLTK, followed by lemmatization to normalize word forms and preserve semantic consistency. TF-IDF vectorization is applied to the preprocessed corpus to generate document-term matrices, which serve as input to both the K-Means clustering and LDA topic modeling components of the deception engine. The deceptive document generation implementation performs keyword permutation by replacing high-weight TF-IDF terms with semantically distant but syntactically compatible alternatives drawn from a curated substitution lexicon. Topic substitution is achieved by replacing LDA-derived topic distributions with artificially constructed distributions from unrelated thematic clusters, ensuring that automated topic modeling tools operating on the deceptive corpus yield systematically misleading thematic insights.

E. Model Training

The VAE model is trained on a corpus of 10,000 normal user session feature vectors using the Adam optimizer with a learning rate of 0.001 and a batch size of 64 over 100 epochs. The architecture consists of an encoder with two dense layers of 128 and 64 units respectively, a latent space of 32 dimensions, and a symmetric decoder, with the training objective defined by the evidence lower bound (ELBO) combining reconstruction loss and KL divergence.

The anomaly detection threshold is determined using a Gaussian fit to the reconstruction error distribution of the validation set, with the threshold set at the 99th percentile to minimize false positives during production operation. Model checkpoints are saved every 10 epochs, and the best-performing checkpoint selected based on validation ELBO is deployed for inference, with periodic retraining scheduled to adapt the model to evolving legitimate user behavioral patterns.

F. Real-Time Detection

Real-time anomaly detection is implemented as a streaming inference service that processes incoming user activity log entries in micro-batches of 10 events with a latency target of under 500 milliseconds from event ingestion to anomaly score generation. Each micro-batch is transformed into a behavioral feature vector using a fixed-length sliding window aggregation, normalized using pre- computed training set statistics, and passed through the deployed VAE encoder-decoder for reconstruction error calculation.

Anomaly scores exceeding the calibrated threshold trigger the enclave activation workflow, which asynchronously initializes the NLP manipulation pipeline and redirects the flagged session to the deceptive document environment within 2 seconds of detection. The real-time detection service is implemented as a Flask-based REST API with thread-safe session management to support concurrent monitoring of multiple user sessions without interference.

G. Alert System

The alert system is designed to notify security administrators in real-time upon the detection of anomalous access events, providing actionable intelligence including the anomaly score, session metadata, affected document categories, and a behavioral deviation summary generated by comparing the flagged session's feature vector against the learned normal distribution. Alerts are delivered through multiple channels including in- dashboard notifications, email alerts, and optional SIEM integration via syslog forwarding.

Each alert record is persisted in the Anomaly Events Schema with a severity classification derived from the magnitude of the reconstruction error, ranging from Low to Critical. The alert system includes a configurable suppression mechanism to prevent alert flooding during sustained adversarial campaigns, and all alerts are linked to the corresponding session's enclave activation record to provide a complete audit trail for forensic analysis.

H. Dashboard

The administrative dashboard is a Flask-powered web application that provides security personnel with a unified interface for monitoring system health, viewing real-time anomaly events, and auditing historical access patterns. The main dashboard view presents live anomaly score time- series visualizations, a session activity heatmap, and a document access frequency chart, all updated at 5-second intervals using asynchronous AJAX polling of the backend REST API.

Secondary dashboard sections provide detailed

anomaly event logs with filtering by severity, date range, and session identifier, as well as a document manipulation audit trail showing the specific operations applied to each deceptive document set generated during adversarial sessions. Administrative controls are available for adjusting the anomaly detection threshold, managing the substitution lexicon used by the manipulation pipeline, and triggering manual enclave activations for testing purposes.

VI. RESULTS AND DISCUSSION

A. Testing Strategy

The system was evaluated using a stratified testing strategy comprising three phases: unit testing of individual components including the VAE model, TF-IDF vectorizer, K-Means clusterer, and LDA topic modeler; integration testing of the end-to-end anomaly detection and enclave activation pipeline; and adversarial simulation testing using synthetically generated adversarial sessions mimicking real-world automated document scraping and topic modeling attacks.

Adversarial simulation tests were conducted by deploying a simulated adversary equipped with standard document clustering and topic modeling tools against both the authentic document repository and the deceptive document ecosystem generated by DARD. The effectiveness of the deception was measured by comparing the adversary's analytical outputs—topic distributions, document clusters, and keyword rankings—against the ground truth to quantify information corruption.

B. Performance Metrics

The VAE anomaly detection module was evaluated using standard binary classification metrics including precision, recall, F1-score, and the area under the ROC curve (AUC-ROC), computed against a labeled test set containing 500 normal and 200 adversarial session records. Document manipulation effectiveness was quantified using the Topic Corruption Score (TCS), defined as the Jensen-Shannon divergence between the adversary's inferred topic distributions from the deceptive corpus and the ground truth topic distributions of the authentic corpus.

System latency was measured end-to-end from the moment of anomaly detection to the successful redirection of the adversarial session to the deceptive enclave, with the 95th percentile latency reported as the primary performance indicator. False positive rate—the proportion of legitimate user sessions incorrectly flagged as adversarial— was tracked as a critical operational metric, as excessive false positives would degrade the user experience and undermine trust in the system.

C. Experimental Results

The VAE anomaly detection module achieved an AUC-ROC of 0.967 on the test dataset, with a precision of 0.921, recall of 0.944, and F1-score of 0.932 at the calibrated detection threshold. The false positive rate was measured at 1.8%, indicating that fewer than 2 in 100 legitimate user sessions were incorrectly flagged, which is within acceptable operational bounds for a security monitoring system. The document manipulation pipeline demonstrated a mean Topic Corruption Score of 0.74 (on a 0–1 scale where higher values indicate greater corruption), indicating that adversarial topic modeling tools were severely misled by the deceptive document ecosystem. The end-to-end enclave activation latency averaged 1.3 seconds with a 95th percentile of 2.1 seconds, meeting the 2-second target defined in the system requirements.

D. Analysis

The high AUC-ROC and F1-score of the VAE anomaly detection module validate the effectiveness of reconstruction-error-based anomaly scoring for detecting adversarial behavioral patterns in document access logs. The model demonstrated particular robustness against low-and-slow adversarial strategies, correctly identifying anomalous sessions that remained within normal individual event thresholds but exhibited statistically significant divergence in aggregate behavioral feature distributions.

The Topic Corruption Score of 0.74 confirms that the LDA-based topic substitution and keyword permutation operations effectively corrupted the semantic structure of the deceptive documents while maintaining superficial plausibility. Analysis of specific adversarial simulation runs revealed that automated classifiers operating on the deceptive corpus consistently assigned documents to incorrect thematic categories with high confidence, demonstrating the practical effectiveness of the

manipulation strategy.

E. Comparison

Comparative evaluation against existing approaches including static honeypot systems and rule-based anomaly detection frameworks demonstrated significant performance advantages for DARD across all evaluation metrics. Static honeypot systems achieved a mean adversarial detection rate of only 61% compared to DARD's 94.4%, primarily because sophisticated adversaries could identify static decoy repositories through metadata analysis, whereas DARD's dynamically generated deceptive environments adapt to each adversarial session.

Rule-based anomaly detection systems exhibited false positive rates exceeding 12% due to their inability to model complex non-linear behavioral correlations, compared to DARD's 1.8% false positive rate. Against topic modeling-based adversaries, DARD's manipulation pipeline outperformed existing document obfuscation techniques—which achieved average TCS values of only 0.31—by more than doubling the semantic information corruption score while maintaining document readability for legitimate users.

VII. CONCLUSION AND FUTURE WORK

A. Conclusion

This paper presented DARD, a novel deception-oriented IP protection framework that integrates VAE-based behavioral anomaly detection with NLP-driven document manipulation to defend organizational intellectual property against modern AI-powered adversarial attacks. Experimental results demonstrated that the system achieves a high detection accuracy with an AUC-ROC of 0.967 and a low false positive rate of 1.8%, while effectively misleading adversarial analytical pipelines with a mean Topic Corruption Score of 0.74.

DARD represents a significant advancement over traditional perimeter-based and static deception approaches by providing a dynamic, semantically coherent deceptive environment that adapts to each adversarial session in near real-time. The framework's modular architecture ensures scalability and adaptability, positioning it as a practical and deployable solution for organizations seeking proactive IP protection against the growing threat of AI-driven document reconnaissance and knowledge extraction attacks.

B. Future Work

Future research will explore the integration of large language models (LLMs) into the document manipulation pipeline to generate more contextually realistic deceptive content that can fool even advanced adversaries employing state-of-the-art NLP analysis tools. Additionally, the incorporation of reinforcement learning into the anomaly detection module will enable the system to dynamically adapt its detection thresholds based on feedback from confirmed adversarial events and emerging attack patterns.

Extensions to the current system will include support for multi-modal IP assets such as images, source code, and structured data files, expanding DARD's protective scope beyond text documents. Federated learning approaches will be investigated to enable collaborative anomaly model training across multiple organizational nodes without sharing sensitive behavioral data, and formal user studies will be conducted to quantify the impact of the deceptive ecosystem on legitimate user experience in real enterprise environments.

REFERENCES

[1] Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[2] P. Kingma and M. Welling, “Auto-encoding variational Bayes,” *arXiv preprint* arXiv:1312.6114, 2013.

[3] M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *J. Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[4] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[5] M. H. Sqalli, F. Al-Haidari, and K. Salah, “EDoS-shield: A two-steps mitigation technique against EDoS attacks in cloud computing,” in *Proc. IEEE/ACM Int. Conf. Utility and Cloud Computing (UCC)*, 2011.

[6] Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988.

ISSN: 2349-6002

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint* arXiv:1301.3781, 2013.

[8] L. Spitzner, *Honeypots: Tracking Hackers*. Boston, MA, USA: Addison-Wesley, 2003.

[9] X. Shu, D. Yao, and B. G. Ryder, "Privacy-preserving and trustworthy cyberphysical systems," *IEEE Trans. Dependable and Secure Computing*, 2015.

[10] Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2007.