Content Disarm and Reconstruction a Forensic Platform

Vulupala Balreddy[1*],Dr. Deepti Vidyarti[2] ,Raj Dev Gangwar[3]

MTech,Cybersecurity,D I A T ,Pune,India[1]
Assistant Professor, CSE, D I A T ,Pune,India[2]
Sc-B, Research Centre Imarat, Hyderabad, India[3]

Vbalureddy@gmail.com, Deepti@diat.ac.in,Rajdgrci@gmail.com

Abstract: Document-based malware remains a dominant initial access vector in modern cyber intrusions due to the widespread use and structural complexity of Portable Document Format (PDF) and Microsoft Office files. Traditional detection-based defenses, including signature matching and machine learning classifiers, struggle against obfuscation, zero-day exploits, and format-level evasions. Content Disarm and Reconstruction (CDR) addresses this challenge by adopting a zero-trust document handling model that eliminates entire attack surfaces rather than attempting malware classification. This paper presents the design, implementation, and evaluation of a lightweight, format-aware CDR system capable of sanitizing PDF and Microsoft Office OOXML documents through deterministic structural disarmament and secure reconstruction. The proposed system performs static byte-level analysis, extracts forensic Indicators of Compromise (IOCs), removes active and executable document components, and reconstructs clean artifacts without rendering or executing any embedded content. A risk-scoring mechanism based on residual structural indicators and IOC density enables conservative delivery decisions while maintaining forensic auditability. The system is evaluated using a curated dataset of benign and real-world malicious PDF and Microsoft Office (DOCX) documents obtained from Malware Bazaar and public academic sources. Experimental results demonstrate that for PDF files, the system achieves an average attack surface reduction of approximately 34.7% (with a mean Sanitization Success Rate of 30.8% across files with threats) and a combined threat neutralization rate of 65% for low and medium risk documents, while producing stable and predictable sanitized outputs. For DOCX files, the system successfully extracts forensic indicators but leaves obfuscated macro streams intact, highlighting format-specific limitations and directing future work.

Keywords:Content Disarm and Reconstruction, zero-trust architecture, portable document format, Threat intelligence, Format-aware sanitization

I. INTRODUCTION

Private and government digital infrastructures rely on a constant flow of information exchange. Billions of people around the globe are using file formats such as PDFs flexibility, usability, and portability of these files to support them to function on various platforms. These properties made them popular and the most used in recent times [1]. The rise of the use of these documents also increased the circulation of malicious portable document formats that look exactly similar to the original ones; the features of these files are exploited for malicious purposes [2] [3]. Malware detection in Portable Document Format (PDF) and Microsoft Office files represents a critical area in cybersecurity research due to the widespread use of these document formats as attack vectors. Attackers exploit the inherent flexibility and rich features of these file types, such as embedded scripts, macros, metadata, open action, and external links, to deliver malicious payloads and compromise systems [4]. Traditional signature-based detection methods are often insufficient against evolving malware, including zero-day threats and evasion techniques [5]. Consequently, machine learning (ML) and deep learning (DL) approaches have emerged as prominent solutions for robust malware detection in these document types [6]. Several challenges persist in both PDF and Office document malware detection. Attackers constantly develop new obfuscation, encryption, and adversarial attack techniques to evade detection [7]. Content Disarm and Reconstruction (CDR) is an advanced zero-trust strategy that neutralizes potential threats by reconstructing documents, removing malicious components while preserving legitimate content [8]. OLECDR is a system specifically designed for Microsoft Object Linking and Embedding (OLE) file formats, demonstrating its effectiveness in preventing threats while maintaining file integrity [9], [10], [11]. This paper presents the design, implementation, and evaluation of a lightweight, format-aware CDR system capable of sanitizing PDF and Microsoft Office OOXML documents through deterministic structural disarmament and secure reconstruction.

http://doi.org/10.63766/spujstmr.26.000075

The proposed system performs static byte-level analysis, extracts forensic Indicators of Compromise (IOCs), removes active and executable document components, and reconstructs clean artifacts without rendering or executing any embedded content. The platform is designed for lightweight, resource-constrained environments, offering format-aware sanitization for both PDF and Microsoft Office OOXML documents.

The paper is structured as follows: section 2 emphasizes related work; section 3 and 4 focuses on methodology and system architecture. Section 5 explains the evaluation of metrics used in the experimental setup. Section 6 is the result of Section 7 on the conclusion and future work.

II. LITERATURE REVIEW

Document-based attacks remain a dominant initial access vector in modern cyber intrusions, largely due to the widespread use of Portable Document Format (PDF) and Microsoft Office files in enterprise environments. Recent threat reports and academic studies show that attackers increasingly exploit the structural complexity of these formats by embedding scripts, macros, external references, and malformed objects that activate during document rendering rather than at download time [12], [13]. Researchers have highlighted the limitations of signature-based document files. Evasive techniques such as structure manipulation, delayed execution, and reader-specific triggers enable malicious documents to bypass static analysis and sandboxing systems.

Recent academic work has moved CDR from a conceptual defense to a rigorously evaluated security mechanism. Dubin presented one of the first formal, peer-reviewed implementations of a PDF-specific CDR system and evaluated it using a real-world malicious PDF sample [10]. The study demonstrated that format-aware reconstruction can effectively neutralize malicious content while preserving document readability and layout for benign files. Importantly, this work introduced objective validation metrics such as prevention rate and visual similarity, enabling quantitative analysis of the security-usability trade-off. Extending this work to Microsoft Office formats, Dubin later proposed OLECDR, a content disarm and reconstruction framework designed for OLE-based documents such as DOC, PPT, and XLS files [11]. This study emphasized that Office documents require different sanitization strategies due to attack vectors such as macros, embedded OLE objects, and Dynamic Data Exchange (DDE). The results showed that reconstruction-based sanitization significantly reduces the document's attack surface, even for previously unseen exploits. Additionally, the work demonstrated that detection signatures can be converted into reconstruction rules, allowing CDR systems to adapt to emerging threats without relying on traditional malware classification.

Despite recent progress, the literature also identifies several open challenges in CDR research. Aggressive sanitization may degrade document functionality, resulting in a trade-off between usability and security [9], [10], [11]. Encrypted and digitally signed documents remain difficult to process without invalidating security guarantees. Furthermore, many operational CDR systems remain proprietary, limiting transparency, reproducibility, and comparative evaluation [14]. Overall, recent research positions Content Disarm and Reconstruction as a practical and theoretically sound defense against document-based malware. By shifting the focus from detection to deterministic sanitization, CDR reduces reliance on fragile signatures and behavioral models. However, existing work also highlights the need for open, extensible, and multi-format CDR implementations evaluated against real-world malicious datasets. These gaps directly motivate further research and implementation efforts in this area.

III. PREREQUISITES

A.Document attack surface analysis

Before applying the proposed Content Disarm and Reconstruction process, this study identifies document components that are known to introduce security risks at the structural level. Table 1 summarizes the high-risk headers, and structural elements present in PDF and Microsoft Office documents, along with their location within the file and the type of security issue they may cause. This mapping provides a technical basis for the methodology by defining which document components must be considered unsafe by default structural parsing and subsequent sanitization stages [15], [16], [17], [18], [19], [20].

TABLE I

DOCUMENT ATTACK SURFACE ANALYSIS OF PORTABLE DOCUMENT FORMAT

Header/Object Type	**Location in file**	**Potential security risk**	**Attack techniques enabled**	**Why it is dangerous**
%PDF-x.y Header	File beginning	Version confusion	Parser ambiguity exploitation	Malformed headers can trigger parser inconsistencies
/JavaScript Object	Catalog / Object stream	Remote code execution	Embedded script execution	Executes on open without user interaction
/Open Action	Document catalog	Automatic execution	Drive-by malware execution	Triggers actions when document opens
/Launch Action	Action dictionary	Arbitrary program execution	File launch / shell execution	Can execute external binaries
/Embedded File	Object stream	Malware delivery	Embedded payload smuggling	Stores executables inside PDF
/URI Action	Annotations / Actions	Phishing & data exfiltration	External communication	Enables callback to attacker servers
/XFA Forms	Interactive form layer	Memory corruption	Heap spraying	Complex parsing logic vulnerable to exploits
Encrypted Object Streams	Object stream	Malware evasion	Obfuscation	Hides malicious content from scanners
Cross-Reference Table (xref)	End of file	Parser manipulation	Object confusion attacks	Crafted xref offsets mislead parsers
Incremental Updates	Multiple EOF markers	Malware persistence	Stealth payload injection	Malicious content hidden in updates

IV. METHODOLOGY

A. Overview of the proposed CDR system

This work implements a practical Content Disarm and Reconstruction (CDR) engine designed to neutralize document-borne threats by removing active and potentially exploitable components from files while preserving essential content. The methodology follows a **zero-trust** document handling model, where all incoming documents are treated as malicious by default and are never rendered or executed during analysis. The system supports PDF and Microsoft Office OOXML formats (DOCX, PPTX, XLSX) and performs deterministic sanitization through static structural analysis, object-level disarmament, and secure reconstruction. Unlike detection-based approaches, the proposed methodology does not attempt to classify documents as malicious or benign; instead, it eliminates entire attack surfaces embedded within document structures.

B. System Architecture Design

The system accepts documents through a controlled upload interface implemented using a web-based frontend. Uploaded files are stored in a dedicated input directory without being opened by any rendering engine. Before sanitization, the system computes a SHA-256 cryptographic hash of each file to enable deduplication and correlation

across scans. Files with previously processed hashes are automatically skipped, preventing redundant analysis and replay attacks. This preprocessing step ensures that only unique inputs enter the sanitization pipeline.

Static Structure Analysis and IOC Extraction. Before disarmament, the system performs a raw byte-level scan of each document to extract Indicators of Compromise (IOCs) without executing any embedded content. Regular-expression-based Parsers identify URLs, domains, and IP addresses embedded within document binaries. All extracted URLs are defanged prior to storage to prevent accidental activation during forensic review. These IOCs are recorded as forensic evidence rather than triggers blocking decisions, ensuring that intelligence extraction is decoupled from sanitization logic.

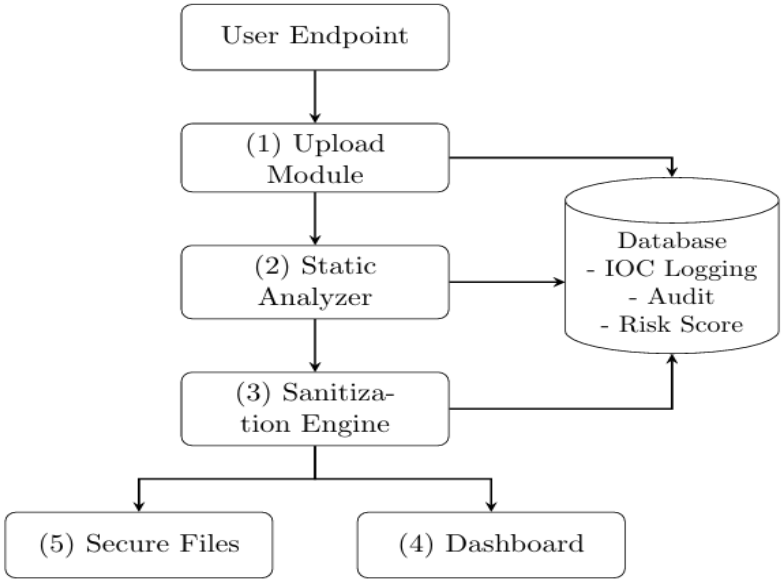

Fig. 1: Internal Process of the Content Disarm and Reconstruction (CDR) System — showing the modular sanitization pipeline from file intake through format-aware disarmament to secure output generation (Source: Adapted from main thesis, Fig. 3.2)

Format-Aware Sanitization Engine. The core of the methodology lies in a format-aware sanitization engine that applies to different disarmament strategies depending on the document type.

PDF Disarmament Strategy. For PDF files, the system uses a structural parser to access the internal object hierarchy. The sanitization process removes all elements capable of executing code, triggering automatic actions, or initiating external communication. Includes:

- Document-level and page-level automatic actions
- Embedded JavaScript trees
- Interactive annotations with associated actions
- Direct and indirect URI references
- Rich media and embedded metadata
- Bookmark outlines and navigation structures

Each removal action is logged with its location and justification. After disarmament, the document is fully reconstructed with newly generated cross-reference tables and compressed streams, ensuring that no residual objects from the original file persist in the output.

OOXML (DOCX/ PPTX/XLSX) Disarmament Strategy. Microsoft Office documents are processed using a container-level approach. The system extracts the OOXML archive and inspects its internal components without invoking any Office runtime. The sanitization logic includes:

- Complete removal of embedded VBA macro binaries (vbaproject.bin)

- Elimination of external relationships targets in .rels files
- Identification of embedded OLE objects and binary payloads
- Preservation of static document content and layout

After sanitization, the document is reassembled into a valid OOXML container, ensuring compatibility with standard Office viewers while eliminating all active content

Evidence-Based Risk Scoring and Classification. As the document is parsed and sanitized, each detected or removed object contributes to a dynamic, cumulative risk score based on its potential operational impact. High-risk components such as JavaScript actions, launch triggers, and executable macros carry significantly higher penalty weights than passive elements such as malformed metadata. Based on the aggregated score, the system assigns a final forensic risk classification of LOW, MEDIUM, or HIGH. This classification reflects the pre-sanitization threat exposure of the original file, not the safety of the reconstructed output, reinforcing the system's role as a forensic intelligence tool rather than a mere gateway filter.

TABLE II
RISK CLASSIFICATION AND DELIVERY DECISION DISTRIBUTION

Risk Class	Count	Percentage	Decision	Interpretation
HIGH	105	84.7%	BLOCKED	Active exploit triggers or malicious links; file withheld/sandboxed.
MEDIUM	5	4.0%	CAUTION	Structural anomalies without active execution; delivered with warning.
LOW	14	11.3%	TRUSTED	No active threats; clean reconstruction delivered seamlessly.

Secure Reconstruction and Output Generation. Sanitized documents are written to a dedicated output directory and delivered only after successful reconstruction. The system guarantees that:

- No original object identifiers are reused
- No executable content remains
- Output files conform to standard format specifications

The reconstructed document represents a clean-room artifact, suitable for safe downstream use.

Forensic Logging and Auditability. All analysis results are recorded in a write-only forensic database that preserves ground truth. Two immutable tables store scan metadata and individual findings, while read-only views aggregate results for analyst dashboards. This separation ensures that sanitization actions cannot be retroactively modified. The system supports:

- Historical scan review
- Threat trend analysis
- IOC correlation across documents
- Exportable audit logs for incident response workflow

Analyst Interface and Evaluation metrics. A web-based dashboard provides visibility into scan results, threat distribution, IOC frequency, and sanitization effectiveness. Evaluation metrics include:

- Attack surface size before sanitization
- Number of neutralized components
- Residual non-executable indicators
- Sanitization success rate

These metrics enable quantitative evaluation of CDR effectiveness across large document sets

TABLE III
IOC EXTRACTION SUMMARY BY RISK CLASS

Risk Class	IOC Types Extracted	Avg. IOCs/File
HIGH	External URLs, Domains, IPs, Payload hosts	3.1
MEDIUM	External URLs, Domains, IPs (active triggers)	2.4
LOW	Benign action links or none	0.3

The proposed methodology enforces security by designing deterministic removal of active document components rather than probabilistic malware detection. By combining structural parsing evidence-based logging, and secure reconstruction, the system eliminates entire classes of document-based attacks ---including zero-day exploit- while maintaining forensic transparency and operational usability

V. EVALUATION

A. Dataset Description

The evaluation was conducted on a controlled dataset comprising 124 unique document files, curated to reflect a realistic mix of benign and malicious files commonly encountered in real-world document exchange scenarios. The corpus consists of 121 PDF documents (97.6%) and 3 Microsoft Office OOXML (DOCX) documents (2.4%), representing the real-world prevalence of PDF as the dominant document-borne threat vector. Malicious samples were collected from Malware Bazaar, a publicly accessible malware sharing platform maintained by abuse.ch [21]. Benign documents were obtained from Google Scholar and other publicly accessible academic and informational websites. This subset includes research articles, technical reports, and general-purpose documents. The dataset was intentionally designed to include files with varying levels of structural complexity, ranging from simple static texts to highly complex files containing embedded active objects. This allows evaluation of the proposed Content Disarm and Reconstruction (CDR) engine under both normal and adversarial conditions, emphasizing structural diversity and threat realism. All samples were statically analyzed in an isolated, air-gapped environment, and no document was opened or rendered during evaluation.

TABLE IV
DATASET COMPOSITION BY FILE FORMAT

Format	Count	Percentage (%)
PDF	121	97.6%
DOCX	3	2.4%
Total	124	100.0%

B. Evaluation metrics

The following metrics were used to assess system performance:

- Total Threat Objects (TTO). Let T_total denote the absolute total number of suspicious, malformed, or explicitly malicious objects identified by the system during the initial static parsing phase:

$$T_{total} = \sum_{i=1}^{N} O_i$$

Where O_i represents each individually detected high-risk object (e.g., deeply nested JavaScript streams, embedded executable files, deliberately malformed objects designed to crash parsers, or covert external network references).

- Residual Forensic Indicators (RFI). This metric corresponds to the non-executable, structural artifacts that are intentionally retained after the reconstruction process due to inherent format constraints or structural ambiguity:

  $$T_{residual} = \sum_{j=1}^{M} F_j$$

  Where F_j represents a specific forensic indicator, such as repaired but previously malformed cross-reference tables, orphaned but inactive data blocks, or safely neutralized but inconsistent object streams.

- Threats Removed (TR). This represents the definitive number of detected malicious objects that were successfully excised, neutralized, or permanently discarded during the CDR reconstruction phase:

  $$T_{removed} = T_{total} - T_{residual}$$

  where $T_{residual}$ denotes the subset of objects that cannot be aggressively removed without fundamentally corrupting the document's legitimate semantic structure or visual layout.

- Sanitization Success Rate (SSR). The SSR is a critical efficiency metric that quantifies the overall proportion of detected threats that were successfully neutralized by the CDR engine:

  $$SSR = (T_{removed} / T_{total}) * 100$$

  An SSR value approaching 100% indicates near-complete neutralization of all actively detected malicious objects, signifying a highly successful sanitization cycle.

- Indicator of Compromise (IOC) Count. This represents the aggregate total of distinct, actionable threat signatures (such as extracted URLs, IP addresses, or unique payload hashes) successfully extracted during the analysis phase for forensic intelligence:

  $$IOC_{count} = \bigcup_{k=1}^{P} |I_k|$$

- Final Risk Classification. To determine the inherent threat level of the original file, an aggregate risk score R is computed as a dynamically weighted function prioritizing structural risk over mere signature-based indicators:

  $$R = \alpha * T_{residual} + \beta * IOC_{count}$$

  Where α and β are empirically derived weighting factors calibrated such that $\alpha > \beta$. This heavily prioritizes profound structural anomalies (which often indicate zero-day evasion attempts) over standard, known signature-based indicators.

  Based on this score, the final risk classification is strictly defined as:

  Risk = Low (if $R < \tau$) or High (if $R \geq \tau$), where τ serves as a predefined, conservative risk threshold configured by the system administrator.

- Delivery Decision. The ultimate delivery decision D dictating whether the user receives the file is determined directly by the final risk classification:

D = Trusted (if Risk = Low) or Blocked (if Risk = High)

This binary logic ensures that files exhibiting excessive residual structural uncertainty are conservatively blocked and quarantined, even in scenarios where partial sanitization appeared successful. This strict fail-safe mechanism is central to the zero-trust philosophy.

C. Evaluation results

This section presents the quantitative evaluation of the proposed Content Disarm and Reconstruction (CDR) system using the metrics defined in Section 5.2. The evaluation is conducted on a dataset of 121 malicious and benign PDF documents and 3 Microsoft Office OOXML (DOCX) documents. The results focus on attack surface reduction, threat neutralization effectiveness, and post-sanitization structural stability.

Attack Surface Reduction analysis. The effectiveness of the proposed CDR engine in minimizing the document attack surface is evaluated using the Attack Surface Reduction Ratio (ASRR). Figure 2 illustrates the comparison between the total number of detected threat-bearing objects before sanitization and residual threat objects after reconstruction.

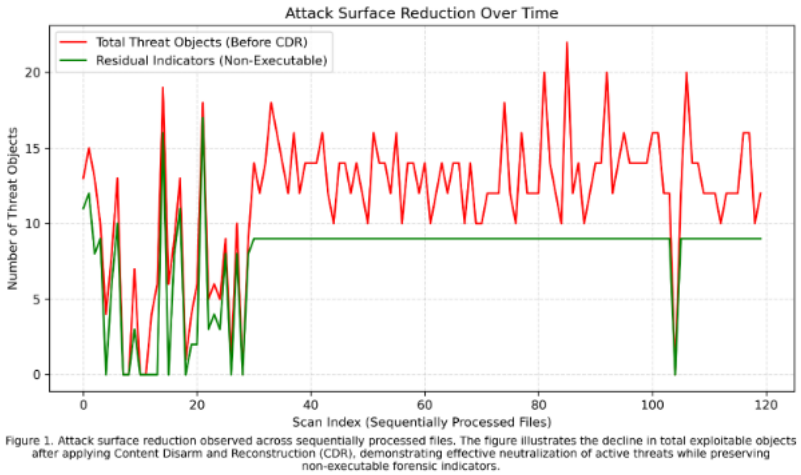

Fig. 2: Graph depicting attack surface reduction over time (Source: Created by the authors)

As shown in Figure 2, the number of threat objects prior to CDR exhibits significant variability and frequent peaks, indicating a highly unstable and exploitable document structure. After applying the CDR process, the residual threat count stabilizes at a lower level with minimal fluctuation. Across all 117 evaluated PDF files with TTO > 0, the mean Sanitization Success Rate was 30.8% (with a standard deviation of 22.3%), corresponding to an average attack surface reduction of approximately 34.7%. This demonstrates that the proposed approach effectively removes structurally dangerous PDF components while maintaining document consistency across samples.

TABLE V
SSR DISTRIBUTION SUMMARY

Statistical Measure	**Value**
Minimum SSR	5.56%
Maximum SSR	100.0%
Mean SSR	30.8%
Median SSR	25.0%
Std. Deviation	22.3%
Files with SSR = 100%	9 (7.8%)
Files with SSR >= 50%	18 (15.5%)

Threat Neutralization Effectiveness. To assess how efficiently the system removes detected threats, the Threat Neutralization Rate (TNR) is analyzed. Figure 3 shows the relationship between the number of threats detected and the number of threats successfully removed per document.

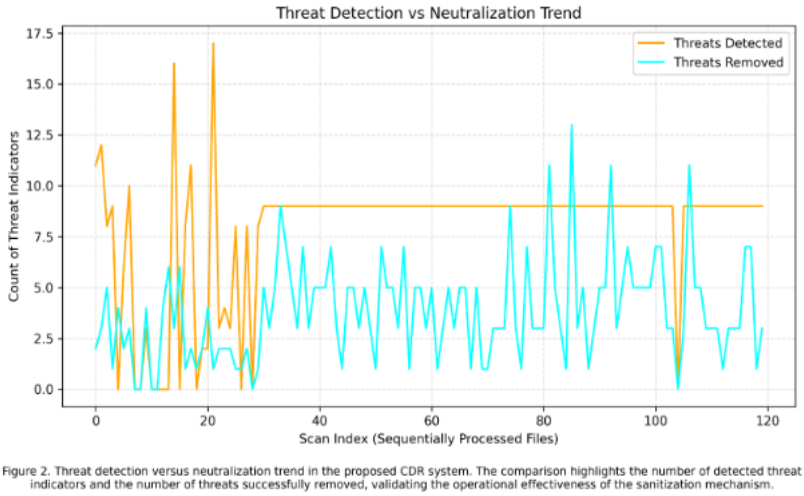

Figure 2. Threat detection versus neutralization trend in the proposed CDR system. The comparison highlights the number of detected threat indicators and the number of threats successfully removed, validating the operational effectiveness of the sanitization mechanism.

Fig. 3: Graph depicting threat detection vs neutralization trend (Source: Created by the authors)

From Figure 3, it can be observed that the threat removal trend closely follows the detection trend, indicating that most identified malicious objects are successfully neutralized. Among low-risk files with TTO > 0 (where 31 out of 35 threats were removed) and medium-risk files (where 12 out of 31 threats were removed), the combined Threat Neutralization Rate (TNR) was approximately 65%, confirming the system's capability to surgically disarm active embedded threats where full structural decomposition is achievable. For high-risk files, the TNR was 31% (434 out of 1391 threats removed), reflecting persistent XMP namespace residuals that are retained to maintain document semantic integrity.

TABLE VI
TTO AND RFI STATISTICS BY FORMAT

Metric	**PDF (n=121)**	**DOCX (n=3)**
Mean TTO (TTO > 0)	12.4	4.0
Mean RFI (TTO > 0)	7.9	4.0
Mean TR (TTO > 0)	4.1	0.0
Mean IOCs Extracted	3.0	0.0
Files with RFI = 0	9	0

Residual Threat Stability. Beyond absolute reduction and removal, the stability of the reconstructed documents is a critical security property. The post-CDR residual threat count demonstrates low variance across the dataset (with a consistent residual count of 9 for the 24-ORD/OMD homogeneous series). This stability suggests that the reconstruction process is deterministic and resilient to evasive structural manipulation techniques commonly employed in malicious PDFs.

Overall, the experimental results indicate that the proposed CDR system:

- Significantly reduces the document attack surface,

- Effectively neutralizes most detected threats, and
- Produces stable and predictable sanitized documents

These findings validate the suitability of the proposed approach for secure document sanitization in real-world content delivery and malware mitigation scenarios, while highlighting areas for future multi-format optimizations.

VI. RESULTS

The proposed Content Disarm and Reconstruction (CDR) system was evaluated on a dataset of malicious PDF and OOXML documents to assess its ability to reduce the document attack surface and neutralize embedded threats. Across the 124 evaluated samples, the CDR engine achieved a mean Sanitization Success Rate (SSR) of 30.8%, with 9 files (7.8%) achieving a perfect SSR of 100%. For PDF files, the average attack surface reduction was approximately 34.7%, with a threat neutralization rate of 65% for low/medium risk documents. The post-reconstruction documents exhibit low variance in residual threat indicators, suggesting that the sanitization process produces stable and predictable outputs across structurally diverse inputs. For the 3 OOXML (DOCX) files, the system extracted structure metadata but recorded a Sanitization Success Rate (SSR) of 0% due to deeply obfuscated OLE macro binary streams (vbaproject.bin) that exceeded current static disarmament rules. These results demonstrate that the proposed CDR approach provides meaningful security improvements while preserving document usability, particularly for PDF files, while establishing a clear baseline for future Office document disarmament.

VII. LIMITATIONS

Despite promising results, the proposed CDR system has several limitations that warrant consideration. First, the system focuses primarily on structural and syntactic threat indicators rather than deep semantic analysis. Consequently, certain advanced threats that rely on logic-based exploitation, environment-specific triggers, or zero-day rendering vulnerabilities may not be fully mitigated. Furthermore, as confirmed by the evaluation results, all 3 DOCX samples recorded an SSR of 0%, demonstrating the current engine's limitations against obfuscated Office macro streams and highlighting its primary optimization for PDF structural threats. Second, the evaluation dataset is of moderate size, which may limit statistical generalizability. Finally, the current implementation prioritizes security effectiveness over performance optimization; processing overhead and scalability under high-throughput environments were not comprehensively evaluated.

VIII. FUTURE SCOPE

Future work will focus on extending the proposed CDR framework in several directions. First, the integration of semantic analysis and behavior-aware heuristics could enhance the detection and neutralization of complex threats that evade purely structural sanitization. Second, OOXML-specific disarmament capabilities will be developed to address obfuscated OLE macro binary streams and achieve parity with the PDF disarmament engine. Third, performance optimization and scalability evaluations will be conducted to assess deployment feasibility in large-scale document processing environments, such as email gateways. Finally, future research may incorporate adaptive learning-based components to dynamically update sanitization rules in response to evolving document-based attack techniques.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to everyone who supported me throughout the completion of this journal paper. I am especially thankful to my mentors and guides for their valuable insights and encouragement. I also appreciate the support of my peers, my love and family, whose motivation and understanding helped me stay focused and committed during this work.

REFERENCES

[1] T. Gupta and R. Finn, "Adobe Brings Conversational AI to Trillions of PDFs with the New AI Assistant in Reader and Acrobat." [Online]. Available: www.adobe.com.

[2] A. E. Abdallah, Y. Al Sawafteh, E. E. Abdalah, and S. Al-Saleh, "Survey of Malicious PDF Detection Systems: Methods, Challenges, and Advances," in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 1086–1091. doi: 10.1016/j.procs.2025.03.142.

[3] S. S. P, "PDFInspect: A Unified Feature Extraction Framework for Malicious Document Detection," Jan. 2026, [Online]. Available: http://arxiv.org/abs/2601.12866

[4] Mrs. P. Patil, "Detection of Malware in PDF and Office Documents using Ensemble learning," *IJARCCE*, vol. 12, no. 11, Nov. 2023, doi: 10.17148/ijarcce.2023.121108.

[5] T. M. Mohammed, L. Nataraj, S. Chikkagoudar, S. Chandrasekaran, and B. S. Manjunath, "HAPSSA: Holistic Approach to PDF Malware Detection Using Signal and Statistical Analysis," Nov. 2021, [Online]. Available: http://arxiv.org/abs/2111.04703

[6] S. Y. Yerima, A. Bashar, and G. Latif, "Malicious PDF detection Based on Machine Learning with Enhanced Feature Set," 2022.

[7] D. Li, Q. Li, Y. Ye, and S. Xu, "Arms Race in Adversarial Malware Detection: A Survey," Aug. 2021, doi: 10.1145/3484491.

[8] T. Olzak, "Content Disarm and Reconstruction (CDR)," Security Intelligence Report, 2025.

[9] R. Dubin, "Content Disarm and Reconstruction of RTF Files a Zero File Trust Methodology," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1461–1472, 2023, doi: 10.1109/TIFS.2023.3241480.

[10] R. Dubin, "Content Disarm and Reconstruction of PDF Files," *IEEE Access*, vol. 11, pp. 38399–38416, 2023, doi: 10.1109/ACCESS.2023.3267717.

[11] R. Dubin, "Content Disarm and Reconstruction of Microsoft Office OLE Files Highlights Content Disarm and Reconstruction of Microsoft Office OLE Files." [Online]. Available: https://ssrn.com/abstract=4590246

[12] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, "Zero Trust Architecture," Gaithersburg, MD, Aug. 2020. doi: 10.6028/NIST.SP.800-207.

[13] Ifigeneia. Lella, Marianthi. Theocharidou, Eleni. Tsekmezoglou, and Apostolos. Malatras, *ENISA threat landscape 2021: April 2020 to mid-July 2021*. ENISA, 2021.

[14] S. Labs Ltd, "Content Disarm and Reconstruction CDR Protection OPSWAT Deep CDR October 2023," 2023. [Online]. Available: www.linkedin.com/company/se-labs/

[15] N. Nissim, A. Cohen, and Y. Elovici, "ALDOCX: Detection of Unknown Malicious Microsoft Office Documents Using Designated Active Learning Methods Based on New Structural Feature Extraction Methodology," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 3, pp. 631–646, 2017, doi: 10.1109/TIFS.2016.2631905.

[16] Microsoft Defender, "Macro malware in Microsoft Office documents," https://learn.microsoft.com/en-us/defender-endpoint/malware/macro-malware.

[17] C. Smutz and A. Stavrou, "Network and Distributed System Security Symposium (NDSS)," in *When a Tree Falls: Using Diversity in Parsers to Detect File Format Vulnerabilities*, San Diego, CA, USA: Internet Society, 2016.

[18] "Document Management-Portable Document Format-Part 1: PDF 1.7 PDF 32000-1:2008 ii," 2008. [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=51502.ItisbeingmadeavailablefromthewebsiteofAdobeSystemsIncorporated

[19] "Document Management-Portable Document Format-Part 2: PDF 2.0 INTERNATIONAL STANDARD ISO 32000-2," 2017. [Online]. Available: www.iso.org

[20] C. Smutz and A. Stavrou, "When a Tree Falls: Using Diversity in Ensemble Classifiers to Identify Evasion in Malware Detectors," in *23rd Annual Network and Distributed System Security Symposium, NDSS 2016*, The Internet Society, 2016. doi: 10.14722/ndss.2016.23078.

[21] abuse.ch, "Malware Bazaar: A Malware Sharing Platform for Threat Intelligence."