

Heuristic Static Analysis of OOXML-Based Word Documents for Malware Identification*

Piotr TARGOWSKI

Military University of Technology, Warsaw, Poland,
ORCID: 0009-0005-0947-7774

Correspondence should be addressed to: Piotr TARGOWSKI, piotr.targowski@wat.edu.pl

*Presented at the 47th IBIMA International Conference, 29-30 June 2026, Madrid, Spain

Abstract

This paper presents a lightweight heuristic framework for static analysis of Microsoft Word Office Open XML (OOXML) documents aimed at early identification of potentially malicious files without executing their content. The proposed method analyses the internal structure of decompressed document containers to extract structural and textual indicators associated with suspicious behaviour. These indicators are aggregated using a weighted heuristic scoring model that estimates the overall risk of a document while accounting for correlations between multiple artifacts. The approach is designed to maintain low computational overhead enabling deployment as a pre execution inspection mechanism close to the end user environment. Experimental demonstrates that the proposed method effectively differentiates between benign and suspicious files while maintaining high detection confidence.

Keywords: malware, DOCX, OOXML, malware document analysis.

Introduction

Modern personal computers store and process unprecedented volumes of sensitive data. These include authentication credentials, financial information, business correspondence, contractual documentation and personal identifiers. Word processing documents constitute a substantial part of this digital ecosystem because they are widely used in professional communication and everyday personal activities. As a result, malicious actors increasingly exploit document-based formats as attack vectors. Attackers often rely on users' trust in seemingly harmless files distributed through email attachments, cloud storage links or collaboration platforms.

Among contemporary document standards Office Open XML (OOXML) formats such as DOCX and DOCM play a central role. Although commonly perceived as ordinary text documents, OOXML files are structured ZIP archives. They contain a hierarchy of XML documents, relationship definitions, media resources, embedded objects and optional executable components such as macros. This containerized architecture improves modularity and interoperability. At the same time, it introduces a complex attack surface. Malicious functionality can be concealed within different structural components of the document in order to evade detection.

Many modern malware detection techniques rely on dynamic analysis. In such approaches the suspicious file is executed in a sandboxed or isolated environment. Although effective in many scenarios these techniques are computationally expensive and time consuming. In addition, document files are commonly used by employees with different levels of cybersecurity awareness. This increases the risk that malicious files will be opened before proper inspection.

OOXML files are internally organized in a structure similar to ZIP archives. They consist of a directory of components that together define the final document. The maliciousness of such containers can be evaluated through static analysis techniques. These methods compare suspicious files with benign versions and search for statistical anomalies. Common indicators include Shannon entropy, byte distribution and chi-square randomness tests (De Gaspari et al., 2020). For complex files that contain partially encrypted malicious payloads the analysis can be performed on smaller segments of the file (Ineza et al., 2025). This segmentation approach allows the detection of concealed code within otherwise benign containers. Regional anomalies within the file structure often indicate hidden malicious content.

Static indicators have historically been important for detecting known malware families. However, the emergence of zero-day attacks has increased the importance of machine learning approaches in malware detection (Vehabovic et al., 2023; Anderson et al., 2018). These methods enable the identification of patterns that may not be visible through traditional rule-based techniques. At the same time, relying on a single indicator or detection method may lead to limited detection performance. Previous studies emphasize that correlating multiple structural and statistical characteristics significantly improves detection accuracy (Targowski and Furtak, 2026). Nevertheless, indicators designed for archive formats such as ZIP or RAR may be less effective for document formats like DOC and DOCX. The internal structure of these documents is standardized which allows malicious actors to conceal harmful content inside legitimate XML components. Execution paths or malicious scripts can therefore be embedded within existing structural elements.

Recent research on document malware detection highlights the importance of multi-perspective feature extraction. Instead of relying on isolated indicators modern approaches combine multiple types of structural and statistical features. For example, the PDFInspect framework integrates structural metadata analysis, statistical text features and temporal inconsistencies into a unified feature space for classification (Sharmila and Tiwari, 2026). This approach demonstrates that combining heterogeneous signals improves robustness against obfuscation and evasion techniques.

A similar direction can be observed in studies focused on malicious document protection and broader cyberattack detection. The work presented in (Surendran, Karthika and Jayalakshmi, 2021) proposes a dynamic scanner for ransomware related document threats using machine learning algorithms and shows that layered inspection can strengthen document protection. Recent studies also emphasize that robust detection requires resistance not only to known attack patterns but also to perturbed or manipulated samples as shown in federated malware detection (Eslamnejad, Taheri and Shojafar, 2025) and uncertainty aware botnet detection (Taheri, 2022). In a broader cybersecurity context ensemble and optimization driven detection models have also been reported as effective for improving classification reliability in complex attack environments (Vaiyapuri et al., 2023). Together these studies support the view that practical malware detection benefits from combining heterogeneous indicators robust decision mechanisms and deployment oriented design.

In this context the present study proposes a lightweight heuristic static analysis framework for Microsoft Word OOXML documents that supports early identification of potentially malicious files before execution. The method analyses the internal structure of decompressed OOXML containers and inspects XML components metadata relationship definitions embedded binaries and macro related artifacts. It also considers structural inconsistencies entropy related features and packaging anomalies that may signal suspicious behaviour. The novelty of the proposed approach lies in combining structural OOXML inspection textual artifact discovery relationship analysis and interpretable heuristic aggregation within a single pre execution framework. Unlike purely statistical archive screening the method focuses on document specific internal artifacts. Unlike dynamic analysis it does not require execution of the file. Unlike opaque end to end learning models it preserves human readable evidence that can be verified by analysts. The detected indicators are combined in a heuristic scoring model that estimates the overall risk of the analysed document. As a result the framework offers a practical and transparent solution for early stage document malware detection.

Materials and Methods

Malicious software remains one of the most persistent threats to modern digital infrastructures. Contemporary computing environments store large volumes of sensitive data related to both personal and organizational activities. The compromise of such information may result in financial losses, identity theft, operational disruption or long-term reputational damage. According to IBM, the global average cost of a data breach is estimated at approximately USD 4.4–4.9 million per incident excluding ransom payments and indirect losses (IBM, 2025).

A significant portion of malware infections originates from seemingly legitimate files distributed through common communication channels. Email attachments, downloaded documents and compressed archives are frequently used as delivery mechanisms. Attackers often rely on social engineering techniques to encourage users to open such files without verifying their origin. Once opened the file may trigger hidden processes that deploy additional malicious components, establish remote access channels or collect sensitive information. In many cases the initial document serves only as a carrier responsible for delivering a secondary payload.

Office Open XML (OOXML) document formats such as DOCX and DOCM are particularly attractive targets for such attacks. These files are implemented as compressed archive packages that contain multiple XML documents, media resources and relationship definitions describing interactions between components. While this modular architecture improves flexibility and interoperability it also introduces potential security risks. Malicious code, macros, embedded objects or external references can be concealed within the document structure while the file itself appears legitimate to the user. Because office documents are widely exchanged in professional and personal communication completely blocking them is not a practical security measure.

Static analysis provides an important mechanism for identifying potentially malicious documents without executing them. By examining the internal structure of OOXML containers and their embedded artifacts it is possible to identify anomalies associated with malicious behaviour. Combining multiple indicators through heuristic evaluation enables an overall risk assessment of the analysed document.

In this context, the purpose of this study is to investigate whether lightweight static inspection of OOXML Word documents can support early detection of potentially malicious files before execution. The research focuses on analysing the internal structure of decompressed document packages and identifying indicators that may signal hidden malicious functionality.

Results

The objective of the proposed approach is to design and implement a lightweight static analysis mechanism capable of identifying potentially malicious Microsoft Word OOXML documents prior to execution. The methodology focuses on structural inspection of document containers and the identification of suspicious artifacts embedded within internal XML resources. The proposed solution emphasizes early detection while maintaining low computational complexity and operational safety. The developed tool performs analysis without executing the inspected file which eliminates the risk associated with opening potentially harmful documents during the inspection process.

The developed method operates by decomposing the OOXML document structure and analysing its internal components. Since OOXML documents are implemented as containers the analysis process begins with decompression of the archive and enumeration of its internal resources. The program extracts all files contained within the document package and processes them sequentially in order to identify structural artifacts that may indicate malicious behaviour. Particular attention is given to XML documents and relationship files that define document behaviour, content dependencies and references to internal or external resources. These files provide valuable contextual information about potential execution paths, embedded scripting constructs, and network interactions that may be triggered when the document is opened.

After decompression, the program performs parsing of the extracted XML resources and relationship definitions. During this stage the tool scans the document content for textual patterns and structural elements associated with known attack techniques used in malicious documents. These artifacts may include command execution instructions, encoded scripting constructs, references to system utilities, macro containers, external hyperlinks and remote resource references embedded within document relationships. The analysis process therefore focuses on identifying features that may enable hidden execution paths, remote payload retrieval or automated actions triggered during document interaction.

The analytical process is based on a rule-based heuristic framework implemented within the program. A collection of detection rules was defined to identify patterns commonly associated with malicious document activity. Each rule corresponds to a specific indicator that can be identified within the internal document structure. These rules search for textual and structural indicators within the decompressed document resources. Indicators include explicit command execution constructs, encoded scripting commands, references to system utilities frequently abused in attack scenarios, as well as document features enabling dynamic content inclusion or external resource loading. The program records all detected indicators together with their location within the document structure, which allows further inspection and interpretation by security analysts.

Each detected indicator contributes to an overall risk score that reflects the likelihood of malicious behaviour. The scoring mechanism is designed as a weighted aggregation of indicator contributions. Formally the risk score R assigned to a document can be expressed as:

$$R = \min (100, \sum_{i=1}^n w_i I_i + \sum_{j=1}^m c_j C_j)$$

Where:

- R - risk score
- I_i - single indicator occurrence rate
- w_i - weight of indicator
- C_j - indicators' combination occurrence rate
- c_j - weight of combination

Indicator combinations play an important role in improving detection reliability. Certain malicious techniques rely on the interaction of multiple document features. For example the simultaneous presence of external resource references and encoded scripting constructs may significantly increase the likelihood of malicious intent. By incorporating combination rules the method allows correlated indicators to contribute additional weight to the final score.

While the calculated risk score provides a quantitative measure of potentially suspicious characteristics, the interpretation of this value is equally important. The presence of individual indicators does not necessarily imply malicious intent as some structural elements commonly appear in legitimate documents as well. During the analysis of reference files it was observed that even benign documents produced non-zero risk scores due to the presence of such elements. Therefore, the risk score of combinations is higher and allows to negate false positives. The final score must be interpreted within a defined classification framework rather than treated as a direct indicator of maliciousness. The following formula describes the aggregation of indicators used to calculate the overall document risk score:

$$D = \begin{cases} \text{Malicious}, & R \geq T_h \\ \text{Suspicious}, & T_l \leq R < T_h \\ \text{Benign}, & R < T_l \end{cases}$$

Where:

- T_h - high risk threshold = 80
- T_l - low risk threshold = 25

The threshold values were established empirically on the basis of observations made during the experimental analysis. Benign reference documents usually produced risk scores not exceeding 25. This was mainly due to the presence of structural elements and relationship definitions that are inherent to legitimate Microsoft Word OOXML files and may still activate low weight indicators. For this reason the lower threshold was set to separate normal document structure from genuinely suspicious behaviour. Higher scores observed in malicious or clearly suspicious samples supported the adoption of the upper threshold for malicious classification.

The program produces a structured report describing the results of the analysis including risk score histogram Figure 1, information of most occurring single indicators Table 1 and combinations Table 2. The report includes the calculated risk score, the detected indicators and their locations within the document structure. This information enables further inspection of suspicious artifacts and provides transparency in the decision-making process. The generated output can therefore support security analysts in understanding which structural elements contributed to the final risk assessment.

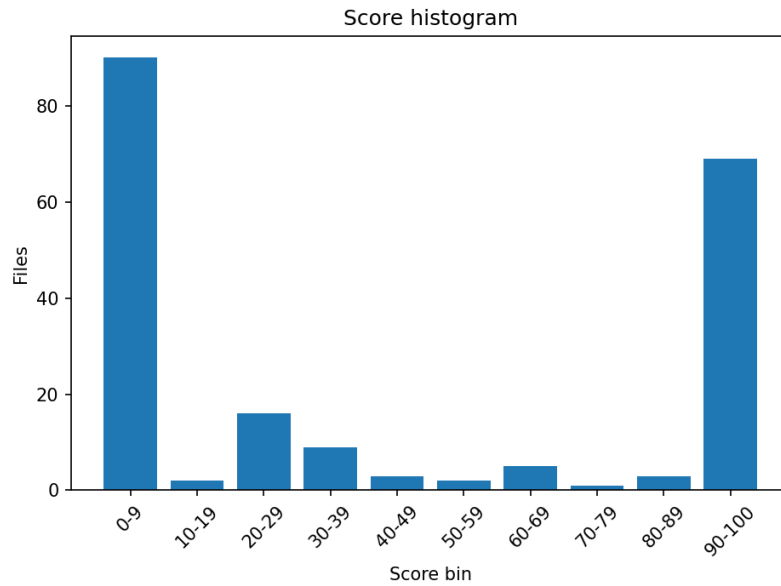


Figure 1. Risk score histogram of samples
Table 1. Top 10 single indicators by occurrence.

Rule ID	Description	Total Matches
SOFT_HTTP	http:// present	22088
MED_LOLBINS	LOLBins strings	496
MED_WEBCLIENT	System.Net.WebClient / DownloadString	197
MED_POWERSHELL	powershell(.exe) string	99
HARD_CMD	cmd.exe or cmd /c marker	50
HARD_DDEAUTO	DDEAUTO field marker	47
HARD_MSMT	ms-msdt protocol marker	40
HARD_IEX	Invoke-Expression / IEX	38
HARD_PS_ENC	PowerShell encoded (-enc/-encodedcommand)	38
HARD_ALTCHUNK	w:altChunk usage	35

Table 2. Top 10 combinations by accumulated occurrence.

Combination ID	Description	Triggers
COMBO_REMOTE_TEMPLA TE	External relationship + http://	73
COMBO_EXTERNAL_WITH POWERSHELL	External relationship + powershell	40
COMBO_PS_ENC	PowerShell + encoded	38
COMBO_PS_DOWNLOAD_I EX	PowerShell + WebClient/DownloadString + IEX	38
COMBO_DDE_MSMT_EXTE RNAL	DDEAUTO/ms-msdt + external relationship	32

An important design objective of the proposed approach is maintaining minimal computational overhead. The analysis tool is intended to operate as close to the end user as possible potentially as part of endpoint security mechanisms or local document inspection workflows. In many real world environments office workstations used by non-technical personnel are equipped with limited computational resources. Consequently the analysis methodology was designed to avoid resource intensive techniques such as full behavioural emulation or dynamic sandbox execution. Instead the approach relies on lightweight operations limited to heuristic scoring of malicious content occurrence. These operations can be performed efficiently even on modest hardware allowing the analysis to be executed locally without introducing noticeable performance degradation for the user.

Another important aspect of the proposed methodology is the balance between detection sensitivity and operational usability. Excessively restrictive filtering mechanisms may lead to a large number of false positives which can disrupt normal organizational workflows and reduce trust in automated detection systems. At the same time insufficient sensitivity may allow malicious documents to bypass detection mechanisms. In practical security environments it is generally preferable to temporarily block or flag a benign file for further inspection than to allow a malicious document to reach the end user undetected. The proposed scoring approach attempts to maintain this balance by differentiating between weak indicators and strong exploitation markers while allowing the aggregation of multiple signals to influence the final risk assessment.

Another important design consideration concerns the role of human oversight in the decision making process. The proposed tool is intended to support security analysts and system administrators by providing structured information about suspicious artifacts identified during static analysis. The generated risk score and indicator reports should therefore be interpreted as support mechanisms rather than fully autonomous security controls. Human verification remains an essential component of the analysis process particularly in environments where documents may legitimately contain advanced features or external references.

Discussion

The results of this study demonstrate that static inspection of structural characteristics of OOXML document containers can effectively support the early identification of potentially malicious Microsoft Word files. The experimental evaluation indicates that the heuristic aggregation of multiple structural indicators enables reliable differentiation between benign documents and files containing suspicious artifacts. The proposed framework achieved an estimated confidence level of approximately 97% in identifying the presence of a potential threat. These findings support the working hypothesis that structural anomalies and statistical irregularities within OOXML containers can provide meaningful indicators of malicious activity even without executing the analysed file.

The primary objective of this work was to develop and validate a practical approach to office document malware detection based solely on static analysis of internal OOXML components. The results suggest that correlating multiple structural indicators within a heuristic scoring model provides a more reliable assessment than relying on individual artifacts in isolation. During the analysis it was also observed that legitimate office documents may contain elements that resemble indicators commonly associated with malicious files. This finding highlights an important challenge in document analysis and confirms that strict binary classification may lead to false positives. Instead, a threshold-based interpretation of aggregated indicators allows more accurate differentiation between benign and suspicious documents.

These observations are consistent with previous research emphasizing the importance of combining heterogeneous indicators when analysing complex file formats. Earlier studies have shown that statistical characteristics such as entropy distribution and structural irregularities can reveal concealed malicious content within archive like containers. Similarly, research on document malware detection suggests that integrating multiple perspectives of analysis improves detection reliability. The results of the present study further support this view by demonstrating that the correlation of structural features within OOXML documents can provide meaningful signals of potential malicious behaviour.

Despite the promising results several limitations of the proposed approach should be acknowledged. Because the method relies exclusively on static structural indicators it may fail to detect threats that depend entirely on runtime behaviour or dynamically retrieved payloads. Documents that activate malicious functionality only after external interaction or delayed execution may therefore bypass static inspection mechanisms. In addition, the heuristic rule set is derived from observable patterns present in known malicious documents. This dependency may reduce the ability of the system to identify highly novel or previously unseen attack techniques.

Future research should therefore focus on extending the proposed framework with adaptive analytical methods. In particular, the integration of machine learning techniques may significantly improve the robustness of the detection process. Unlike purely heuristic approaches that rely on predefined indicators, machine learning models can identify patterns in how malicious artifacts are introduced into document structures. Pattern-based analysis may allow the detection of subtle structural correlations that are difficult to capture through manually defined rules. Combining heuristic structural inspection with data-driven models could therefore reduce false positives and improve resilience against zero-day attacks.

In a broader security context, the findings of this study highlight the value of lightweight pre execution inspection mechanisms. Static analysis of document containers requires minimal computational resources and can be deployed close to the end-user environment. Such mechanisms can therefore serve as an additional defensive layer that identifies suspicious files before they are opened or executed. Although static inspection cannot replace more advanced behavioural analysis techniques, it can significantly reduce the number of potentially harmful files reaching later stages of the security pipeline. Consequently, the proposed approach contributes to the ongoing development of practical and scalable methods for detecting malware embedded in widely used office document formats.

References

- Anderson, H.S., Kharkar, A., Filar, B., Evans, D. and Roth, P. (2018) Learning to evade static PE machine learning malware models via reinforcement learning. arXiv preprint arXiv:1801.08917.
- Baldini, G. (2025) On the application of entropy measures with sliding window for intrusion detection in automotive in-vehicle networks. *Entropy*, 22(9), 1044.
- De Gaspari, F., Hitaj, D., Pagnotta, G., De Carli, L. and Mancini, L.V. (2020) Encod: Distinguishing compressed and encrypted file fragments. In: *International Conference on Network and System Security*, pp. 42-62. Cham: Springer International Publishing.
- Eslamnejad, M., Taheri, R. and Shojafar, M. (2025) Federated learning-based robust android malware detection: label-flipping attacks and defenses. *Neural Computing and Applications*, 37, 27057-27082.
- IBM (2025) Cost of a Data Breach Report 2025. Available at: <https://www.ibm.com/reports/data-breach> (Accessed: 4 March 2026).
- Ineza, Y., Jackson, G., Niyonkuru, P., Kevil, J. and Serwadda, A. (2025) Intermittent file encryption in ransomware: Measurement, modeling, and detection. arXiv preprint arXiv:2510.15133.
- Malware Bazaar (2026) Malware Bazaar. Available at: <https://bazaar.abuse.ch/browse/> (Accessed: 5 March 2026).
- Sgandurra, D., Muñoz-González, L., Mohsen, R. and Lupu, E.C. (2026) Automated dynamic analysis of ransomware: Benefits, limitations and use for detection. arXiv preprint arXiv:1609.03020.
- Sharmila, S.P. and Tiwari, A. (2026) PDFInspect: A unified feature extraction framework for malicious document detection. arXiv preprint arXiv:2601.12866.
- Surendran, R., Karthika, R. and Jayalakshmi, B. (2021) Implementation of dynamic scanner to protect the documents from ransomware using machine learning algorithms. In: *2021 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, pp. 65-70.
- Taheri, R. (2022) UNBUS: uncertainty-aware deep botnet detection system in presence of perturbed samples. arXiv preprint arXiv:2204.09502.
- Vaiyapuri, T., Shankar, K., Rajendran, S., Kumar, S., Gaur, V., Gupta, D. and Alharbi, M. (2023) Automated cyberattack detection using optimal ensemble deep learning model. *Transactions on Emerging Telecommunications Technologies*, 35. doi: 10.1002/ett.4899.
- Vehabovic, A., Zanddizari, H., Ghani, N., Shaikh, F., Bou-Harb, E., Pour, M.S. and Crichigno, J. (2023) Data-centric machine learning approach for early ransomware detection and attribution. In: *NOMS 2023 - 2023 IEEE/IFIP Network Operations and Management Symposium*, pp. 1-6.
- VirusShare (2026) VirusShare. Available at: <https://virusshare.com/> (Accessed: 5 March 2026).
- Targowski, P., Furtak, J., (2026) No Extraction Archive Ransomware Detection Using Machine Learning. *The 25th International Conference on Artificial Intelligence and Soft Computing* (accepted).