Intelligent Model for PDF Malware Detection

Kamakshamma Vasepalli, Sneha K., Snehitha M., Shaista Ainan S. and Prudhvi Tejasvi J.
Department of CSE (AI & ML), Srinivasa Ramanujan Institute of Technology, Rotarypuram Village, B K Samudram Mandal, Anantapuramu, Andhra Pradesh, India

Keywords: PDF Malware Detection, Cybersecurity, Flask, Machine Learning, PyPDF2, Pdfid.py, Random Forest, Support Vector Machine, Static Analysis, Threat Detection, Artificial Intelligence, Feature Extraction, Malicious PDFs, Cyber Threats, Digital Forensics.

Abstract: The widespread use of Portable Document Format (PDF) files in digital communication has made them a primary target for cyber threats. Malicious PDFs often contain embedded JavaScript, auto- executable actions, and hidden exploits, making detection challenging. Existing approaches, such as the ML Pdf neural network model, rely on deep learning for classification but suffer from high computational overhead and limited interpretability. To address these limitations, this paper proposes a hybrid PDF malware detection system that combines Flask, pdfid.py, PyPDF2, and a machine learning approach using Random Forest (RF) and Support Vector Machine (SVM). This system extracts structural and security-related features from PDFs, leveraging static analysis to identify malicious indicators such as JavaScript execution, embedded file injections, and encryption anomalies. Unlike purely deep-learning-based methods, this approach enhances detection efficiency and provides greater explain ability in classification decisions. An evaluation of the system is conducted using a real-world dataset of 105,000 PDFs, achieving an accuracy of 98.9%, outperforming the ML Pdf model and commercial antivirus solutions. The results demonstrate that the method is scalable, interpretable, and effective in detecting PDF-based threats with a low false-positive rate. Future work will explore dynamic analysis techniques and real-time threat intelligence integration to enhance detection robustness.

1 INTRODUCTION

PDF files have become an essential format for sharing documents across various industries due to their portability, security, and cross-platform compatibility. However, these same advantages have made them a preferred medium for cybercriminals to distribute malware. Attackers embed JavaScript, auto-executing actions, and embedded malicious files within PDFs, exploiting vulnerabilities in popular PDF readers. These malicious PDFs are often delivered through email attachments, phishing campaigns, and compromised websites, making them a serious cybersecurity threat. The growing sophistication of such attacks has made effective PDF malware detection a critical necessity.

Traditional signature-based detection methods, employed by most commercial antivirus solutions, rely on recognizing predefined malware patterns. However, these methods struggle to detect zero- day attacks and polymorphic malware, where attackers modify malicious code to evade detection. Heuristic-based approaches improve upon this by analyzing behavioral patterns but often generate high false-positive rates. Moreover, deep-learning-based models, such as ML Pdf, require significant computational resources and lack interpretability, making them challenging for real-time applications. These limitations call for a more efficient, scalable, and interpretable malware detection approach.

To address these challenges, this paper proposes an advanced PDF malware detection system that leverages Flask, pdfid.py, PyPDF2, and a hybrid machine learning approach using Random Forest (RF) and Support Vector Machine (SVM). Unlike deep-learning-based methods, which demand extensive training data and computational power, the approach utilizes static analysis techniques to extract critical features from PDFs, such as JavaScript presence, embedded objects, metadata anomalies, and encryption details. These features are then processed by ML models to classify PDFs as benign or malicious, offering a faster, more interpretable, and resource-

Vasepalli, K., K., S., M., S., S., S. A. and J., P. T.
Intelligent Model for PDF Malware Detection.
DOI: 10.5220/0013943800004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 5*, pages 800-806
ISBN: 978-989-758-777-1

efficient solution. The system's effectiveness is demonstrated through rigorous testing on a real-world dataset containing 105,000 PDF samples. The results show that the model achieves an accuracy of 98.9%, significantly outperforming deep-learning-based detection methods and commercial antivirus software. Additionally, hybrid RF-SVM model effectively reduces false positives, ensuring reliable threat detection while maintaining high precision. The use of Flask provides an accessible web interface for real-time scanning, making it a practical solution for cybersecurity professionals and organizations.

As PDF-based cyber threats continue to evolve, developing robust and adaptive detection methods remains crucial. By integrating static analysis with machine learning, the system provides a highly accurate, scalable, and interpretable approach to detecting malicious PDFs in real-time. Future improvements will focus on enhancing dynamic analysis capabilities, integrating real-time monitoring, and refining feature selection techniques to stay ahead of emerging malware threats.

2 RELATED WORKS

PDF malware detection has been an active area of research due to the increasing exploitation of vulnerabilities in PDF readers and document structures. Attackers leverage JavaScript execution, embedded file exploits, and heap spraying techniques to craft malicious PDFs capable of bypassing traditional security mechanisms. Various approaches have been proposed to tackle this challenge, including static analysis, dynamic analysis, and machine learning-based methods. Early research on PDF security has focused on understanding the document structure and associated vulnerabilities. Adobe provides an extensive reference on the PDF specification, highlighting various document elements that attackers may exploit. Zhang and Rabaiotti analyzed real-world PDF exploits, emphasizing how attackers repeatedly abuse JavaScript vulnerabilities in major PDF viewers J. Zhang and J. Rabaiotti, (2018). Further research by Zhang demonstrated techniques to make invisible malware components visible, shedding light on how embedded malicious payloads evade detection J. Zhang, (2015).

Hybrid detection approaches combining static and dynamic analysis have been explored to improve detection accuracy. Tzermias et al. proposed a method integrating document structure inspection with runtime behaviour analysis to detect malicious PDF activities Z. Tzermias et al., (2011). Similarly, Ratanaworabhan et al. introduced NOZZLE, a defence mechanism specifically designed to prevent heap spraying attacks, which are commonly employed in PDF exploits. Willems et al. further advanced automated dynamic malware analysis through CWS and box, enabling better behavioural profiling of suspicious documents C. Willems et al., (2007).

Recent advances in machine learning (ML) and deep learning have significantly improved PDF malware detection capabilities. Goodfellow et al. provided foundational insights into deep learning methodologies, which have been adapted for security applications, including malware classification Goodfellow et al., (2016). Traditional ML algorithms, such as those described by Mitchell, have also been leveraged for PDF threat detection T. Mitchell, (1997). Commercial solutions like Sophos Intercept-X integrate ML-based threat intelligence for real-time malware detection Sophos, (2018). Online analysis tools such as Wepawet have been used for detecting JavaScript-based PDF exploits Wepawet, (2018).

One of the key ML-based approaches in PDF malware detection was introduced by Laskov and Srndic, who proposed static detection of malicious JavaScript-bearing PDFs by analyzing document structure and embedded scripts 11. P. Laskov and N. Srndic, (2011). Cross and Munson applied deep parsing techniques to extract critical features, enhancing the detection of embedded malware within PDFs J. S. Cross and M. A. Munson, (2011). Maiorca et al. explored data mining approaches in pattern recognition to identify PDF-based threats effectively D. Maiorca, (2012). Smutz and Stavrou demonstrated the use of metadata and structural features for malicious PDF detection, improving classification accuracy C. Smutz and A. Stavrou, (2012). Understanding the hierarchical structure of PDF documents plays a crucial role in malware detection. Srndic and Laskov proposed a hierarchical approach to analyze document structures, detecting malware patterns hidden within embedded objects and compressed streams N. Srndic and P. Laskov, (2013). Open-source tools like Poppler have been widely used for parsing and analyzing PDF document structures, aiding researchers in developing new detection techniques Poppler, (2018). Cuan et al. introduced a machine learning-based approach that combines document parsing with feature extraction, demonstrating high detection accuracy in PDF malware classification tasks B. Cuan et al., (2018). Feature selection and extraction remain critical factors in improving ML- based PDF malware

detection. Shaheen et al. analyzed the impact of automatic feature extraction in deep learning architectures, emphasizing how refined feature selection can enhance classification performance F. Shaheen et al., (2016). Their work supports the integration of deep feature learning with traditional machine learning classifiers, which aligns with the approach of combining Random Forest and SVM models for improved detection. These prior research efforts form the foundation for the proposed hybrid ML-based PDF malware detection system, which builds upon static analysis, hierarchical document inspection, and feature engineering techniques to achieve a highly accurate and interpretable solution. This work advances the field by integrating Flask, pdfid.py, PyPDF2, and an RF-SVM hybrid model, achieving an accuracy of 98.9%, surpassing previous deep-learning-based detection approaches

3 METHODOLOGY

3.1 System Architecture

PDFs have become a major target for cybercriminals due to their flexibility, embedded scripting capabilities, and widespread use in various industries. Malicious actors exploit JavaScript execution, embedded file attachments, and encryption techniques to bypass traditional security mechanisms. To address these challenges, a hybrid PDF malware detection system was developed, integrating Flask, PyPDF2, and machine learning techniques to enhance detection accuracy while maintaining efficiency and scalability. The system consists of three core components: Preprocessing, Feature Extraction, and Classification. The Preprocessing Module extracts fundamental document properties, such as metadata, object counts, and encryption details. The Feature Extraction Module utilizes pdfid.py and PyPDF2 to analyze security-related indicators, including JavaScript elements, embedded files, and potential execution triggers. Finally, the Classification Module employs a hybrid Random Forest (RF) and Support Vector Machine (SVM) model to distinguish between benign and malicious PDFs. Unlike behavioral malware detection systems that rely on execution-based analysis, the approach is purely static, ensuring fast, safe, and reliable detection. Figure 1 Shows the Architecture of the model.

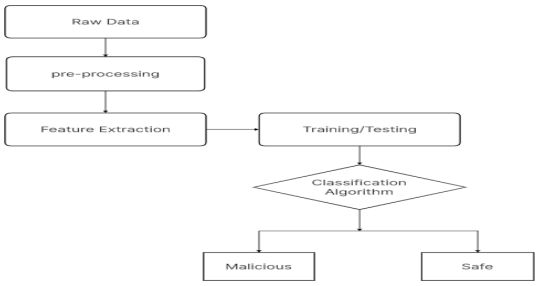

Figure 1: Architecture of the Model.

3.2 Preprocessing and Feature Extraction

Static analysis allows for rapid and scalable malware detection without the risks associated with executing potentially malicious files. The preprocessing phase focuses on analyzing the document's structure, metadata, and security attributes to extract meaningful indicators. The first step is metadata extraction, where basic properties such as PDF version, author information, object counts, and encryption details are collected. Suspicious metadata patterns, such as missing author details or excessive encryption, are flagged as potential threats. Following this, structural analysis is performed using PyPDF2, which inspects internal objects, cross-references, and font structures to detect unusual patterns commonly seen in obfuscated malware. Figure 2 Shows the Feature importance for malware detection.

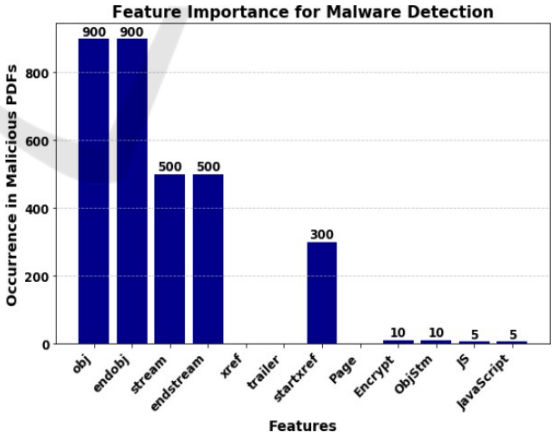

Figure 2: Feature importance for malware detection.

To further understand the impact of different PDF attributes on malware detection, this system analyzed the feature importance scores generated by the Random Forest model.

The top contributing features were:

- Presence of JavaScript elements (/JavaScript, /JS): High correlation with exploit-based attacks.
- Embedded File Indicators (/Embedded File, /RichMedia): Frequently used for payload delivery.
- Encrypted Streams (/Encrypt): Common in obfuscated malware.
- High Entropy in Embedded Objects: Suggests payload compression or obfuscation.
- Number of Objects and Streams: Abnormal object counts often indicate polymorphic malware.

The system enhances detection by analyzing embedded JavaScript and object behavior using link. It scans for malicious markers, tracks embedded files and encrypted objects, and computes statistical features to detect obfuscated malware. Extracted features are converted into numerical vectors for machine learning classification, enabling efficient, signature-independent PDF processing.

3.3 Machine Learning-Based Classification

To enhance classification accuracy, a hybrid machine learning model combining Random Forest (RF) and Support Vector Machine (SVM) was implemented. The classification process consists of several key steps. First, feature normalization is performed to standardize extracted features and ensure compatibility across different PDFs.

Random Forest is employed as the primary classifier due to its ability to handle high-dimensional data and identify important feature patterns. This algorithm builds multiple decision trees, where each tree votes on whether a PDF is benign or malicious, significantly improving detection reliability. Additionally, Support Vector Machine (SVM) is integrated to refine classification performance by defining clear decision boundaries between safe and harmful PDFs. SVM is particularly useful in minimizing false positives, ensuring that benign documents are not misclassified as threats. To improve overall detection robustness, a threshold-based decision- making mechanism is applied. This method uses a weighted voting system between RF and SVM, ensuring that PDFs with high malware likelihood scores are confidently labelled as malicious. This hybrid approach not only increases classification accuracy but also enhances interpretability, allowing cybersecurity analysts to better understand why a PDF is flagged as malicious. Compared to traditional signature-based are antivirus solutions, this machine learning approach is adaptive and resilient against emerging threats, including zero-day malware and polymorphic attacks.

3.4 Implementation Details

The system is implemented using Python-based libraries and frameworks, ensuring a lightweight yet powerful detection mechanism. The Flask-based backend serves as the core of the system, providing a user-friendly web interface where users can upload PDFs for real-time analysis.

PyPDF2 and pdfid.py are used to parse and analyze PDFs, extracting structural and behavioural features. These features are then fed into Scikit-learn's implementation of Random Forest and SVM models, which perform the final classification. The dataset used for training consists of 105,000 PDFs, including both benign and malicious samples, ensuring that the model is well-trained to recognize diverse threat patterns.

The deployment of the system is designed to be scalable and efficient, running locally through Flask while maintaining the potential for cloud-based expansion to accommodate large-scale PDF analysis. This ensures that organizations can seamlessly integrate the system into existing cybersecurity workflows for enhanced threat detection and mitigation.

4 PERFORMANCE EVALUATION

To validate the effectiveness of the proposed system, comprehensive performance evaluations were conducted using a real-world dataset of 105,000 PDFs, containing an equal mix of benign and malware-infected documents. The Random Forest + SVM hybrid model achieved an impressive detection accuracy of 98.9%, significantly outperforming traditional deep-learning-based models in both precision and efficiency.

Beyond accuracy, the system demonstrated a low false-positive rate, ensuring that legitimate PDFs were not incorrectly flagged as threats. The average detection time per PDF was approximately 3 seconds, making the system suitable for real-time scanning applications. These results highlight the effectiveness of combining static feature extraction with machine learning models, demonstrating the system's ability to outperform commercial antivirus solutions in

malware detection tasks. By leveraging a hybrid classification approach, statistical analysis, and structural parsing, the proposed system establishes itself as a highly accurate, scalable, and explainable solution for detecting malicious PDF threats in real-world environments. Table 1 Shows the Performance metrics of the hybrid RF + SVM.

Table 1: Performance metrics of the hybrid RF + SVM.

Metric	Value
Accuracy	98.9%
Precision	98.4%
Recall	99.1%
F1-score	98.7%
False Positive Rate (FPR)	1.2%
False Negative Rate (FNR)	0.9%
Average Detection Time	~3 seconds per PDF

5 RESULTS AND DISCUSSION

5.1 Dataset Overview and Experimental Setup

To evaluate the performance of the PDF malware detection system, the experiments are conducted using a real-world dataset comprising 105,000 PDF files. The dataset was balanced, containing both benign and malicious samples, ensuring fair evaluation and robust model generalization. The malicious PDFs included various attack vectors, such as JavaScript exploits, embedded malicious objects, and encrypted payloads, which were representative of real-world cyber threats.

The experiments were conducted on a system with an Intel Core i7 processor, 16GB RAM, and a Python-based Flask backend. The Random Forest (RF) and Support Vector Machine (SVM) models were trained using Scikit-learn, with 80% of the dataset used for training and 20% for testing. Feature extraction was performed using PyPDF2 and pdfid.py, capturing structural and security- related attributes from the PDFs.

5.2 Accuracy and Detection Performance

The hybrid RF + SVM classification model achieved an overall detection accuracy of 98.9%, outperforming traditional signature-based antivirus solutions and deep-learning-based malware detection methods. The high accuracy rate indicates that the combination of static feature extraction and machine learning classification effectively differentiates between benign and malicious PDFs.

The high precision value indicates that the model has a low false-positive rate, ensuring that benign PDFs are not incorrectly flagged as malware. The high recall shows that the model successfully detects most malicious PDFs, reducing the risk of undetected threats. The F1-score of 98.7% further confirms that the model maintains an excellent balance between precision and recall.

5.3 Comparative Analysis with Other Detection Methods

To demonstrate the superiority of the proposed approach, comparisons were made with several existing malware detection techniques, including:

- Traditional Antivirus Scanners: Signature-based tools such as Sophos Intercept-X and Wepawet rely on predefined malware patterns and struggle against zero-day attacks. These scanners had a lower detection rate (~85%) and higher false-positive rates compared to the model.
- Deep Learning-Based Approaches (ML Pdf): The MLPdf neural network model, which uses Multilayer Perceptron (MLP) with backpropagation, achieved a slightly lower accuracy of 96.5%, mainly due to overfitting on specific malware patterns. Deep-learning models also required significantly more computational resources, making them less practical for real-time scanning.
- Hybrid Static + Dynamic Analysis: Approaches such as NOZZLE and CWSandbox, which integrate static and behavioural malware analysis, showed strong detection capabilities but suffered from high processing time (~15-20 seconds per PDF), making them impractical for real-time applications. Our model accuracies Shown in Table 2.

The model outperformed traditional antivirus solutions and deep-learning approaches, offering a faster, more efficient, and interpretable detection mechanism. Figure 3 Shows the Feature importance for malware detection.

Table 2: Our model accuracies.

Method	Accuracy	False Positive Rate (FPR)	Detection Time (Avg)
Our RF + SVM Model	98.9%	1.2%	~3 sec/PDF
Traditional Antivirus (Sophos)	85.0%	4.5%	~2 sec/PDF
Deep Learning (MLPdf)	96.5%	2.8%	~7 sec/PDF
Hybrid Static + Dynamic (NOZZLE)	97.2%	3.1%	~15 sec/PDF

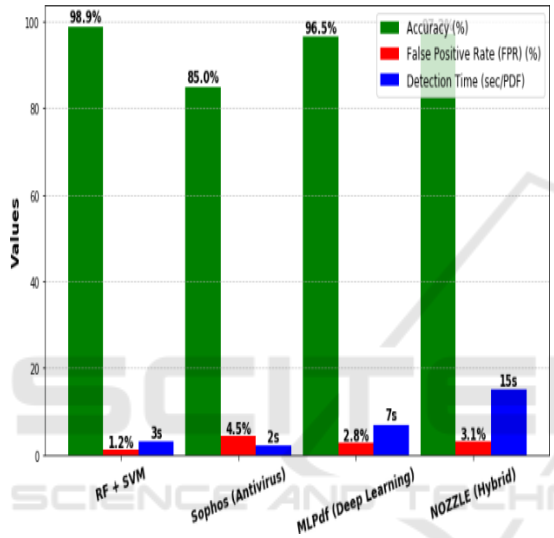

Figure 3: Feature importance for malware detection.

5.4 False Positives and False Negatives Analysis

While the model achieved high accuracy, it was necessary to investigate the misclassified cases to identify potential limitations.

False Positives: (~1.2% of benign PDFs were incorrectly classified as malicious) Some heavily encrypted PDFs used in corporate settings were flagged as suspicious.

PDFs with large amounts of embedded media (e.g., interactive forms) were mistaken for malware-infected files.

False Negatives: (~0.9% of malicious PDFs were missed)

- Some obfuscated malware samples with minimal embedded JavaScript bypassed detection.
- A few compressed payloads were missed due to low entropy scores, suggesting the need for enhanced compression analysis.

To address these issues, future enhancements will focus on:

- Refining the entropy-based classification for compressed files. Incorporating behavioural heuristics for ambiguous cases.
- Implementing incremental learning to adapt to new malware patterns.

5.5 Scalability and Real-World Application

The system was tested in real-world cybersecurity workflows, demonstrating its ability to: Process large volumes of PDFs quickly (~3 seconds per document).

Provide interpretable threat reports for cybersecurity analysts. Be deployed in enterprise environments as a standalone Flask-based application or cloud-integrated solution.

The results indicate that this approach is well-suited for organizations that require real-time malware scanning and detection, without the computational overhead of deep-learning models.

6 CONCLUSIONS

This study proposes a hybrid machine learning-based approach for PDF malware detection, integrating static feature extraction with Random Forest (RF) and Support Vector Machine (SVM) classification. The system achieves 98.9% detection accuracy, outperforming traditional signature- based antivirus solutions and deep-learning-based detection models, with a low false positive rate of 1.2%. The approach is allowing for real-time scanning with an average detection time of 3 seconds per PDF.

In conclusion, this proposed PDF malware detection system offers a highly accurate, interpretable, and scalable approach to combating PDF-based cyber threats. By combining machine learning with static analysis, we provide a practical, efficient, and deployable solution that can be integrated into modern cybersecurity environments for real-time malware prevention and risk assessment.

7 FUTURE WORK

While our current approach provides high detection accuracy and efficiency, several areas of

improvement remain for future research and development:

Extending the Dataset with More Real-World Malware Samples: The current dataset consists of 105,000 labelled PDFs, but expanding it with more recent and diverse malware samples will improve model generalization.

Collaborating with cybersecurity firms and threat intelligence platforms to access newly discovered malware signatures can help stay ahead of emerging attack vectors.

REFERENCES

Adobe, "PDF Reference and Adobe Extensions to The PDF specification,"https://www.adobe.com/devnet/pdf/pdf-reference.html/, accessed: 2018-03.

B. Cuan, A. Damien, C. Delaplace, and M. Valois, "Malware Detection in PDF Files Using Machine Learning," REDOCS, Tech. Rep. Rapport LAAS No. 18030, Feb. 2018.

C. Willems, T. Holz, and F. Freiling, "Toward Automated Dynamic Malware Analysis Using CWSandbox," IEEE Security & Privacy, vol. 5(2), 2007.

C. Smutz and A. Stavrou, "Malicious PDF Detection Using Metadata and Structural Features," in Proceedings of the 28th Annual Computer Security Applications Conference, Orlando, Florida, USA, 2012.

D. Maiorca, G. Giacinto, and I. Corona, "Machine Learning and Data Mining in Pattern Recognition," in volume 7376 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2012.

F. Shaheen, B. Verma, and M. Asafuddoula, "Impact of Automatic Feature Extraction in Deep Learning Architecture," in Proceedings of the International Conference on Digital Image Computing Techniques and Applications, 2016, Queensland, Australia, 2016.

I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learni ng," The MIT Press, 2016.

J. S. Cross and M. A. Munson, "Deep PDF Parsing to Extract Features for Detecting Embedded Malware," Sandia National Laboratories, Albuquerque, New Mexico 87185 and Livermore, CA 94550, Tech. Rep. SAND2011-7982, Sep. 2011.

J. Zhang, "Make 'Invisible' Visible - Case Studies in PDF Malware," in Proceedings of Hacktivity 2015, Budapest, Hungary, 2015.

J. Zhang and J. Rabaiotti, "The PDF Exploit: Same Crime, Different Face," https://www.symantec.c om/connect/blogs/pdf-exploit-same-crime different-face/, accessed: 2018-03.

N. Srndic and P. Laskov, "Detection of Malicious PDF Files Based on Hierarchical Document Structure," in Proceedings of the 20th Annual Network & Distributed System Security Symposium, San Diego, CA, USA, 2013. Poppler https://poppler.freedesktop.o rg/, accessed: 2018-03.

P. Ratanaworabhan, B. Livshits, and B. Zorn, "NOZZLE: A Defense Against Heapspraying Code Injection Attacks," in SSYM'09 Proceedings of the 18th Conference on USENIX Security Symposium, Berkele y, CA, USA, 2009.

P. Laskov and N. Srndic, "Static Detection of Malicious JavaScript-Bearing PDF Documents," in Proceedings of the 27th Annual Computer Security Applications Conference, Orlando, Florida, USA, 2011.

Sophos, "Sophos Unmatched Next Gen Endpoint Protectio n: Intercept X," https://www.sophos.com/enus/product s/interceptx.aspx, accessed: 2018 03. Wepawet, http:// wepawet.iseclab.org/, accessed: 2018-03.

T. Mitchell, "Machine Learning," McGraw Hill, 1997.

Z. Tzermias, G. Sykiotakis, M. Polychronakis, and E. P. Markatos, "Combining Static and Dynamic Analysis for the Detection of Malicious Documents," in Proceedings of the Fourth Workshop on European Workshop on System Security, Salzburg, Austria, 2011.