# Malicious Document Detection and Adversarial Analysis based on Machine Learning

**Abstract -** Nowadays, with the highly rapid development of information technology, it is becoming more and more important to perform detection on malicious documents (such as PDFs). But due to the diversity of the document structure, attackers can gradually have larger attack vector. This research project aims to construct a robust AI document classifier both for industry and academia. Around 200,000 samples have been collected and the AI model have been trained and optimized. The experimental results show that the Accuracy of the model is as high as 99.82% while the False Positive Rate is only as low as 0.01%. More, through the study of adversarial ML, the model has certain capability to resist attacks and enjoys good robustness. At last, we demonstrate our model can be widely deployed in typical scenarios such as security products or mail servers.

**Key Words：**AI Security; Machine Learning; Maldoc Detection; Adversarial ML

## Introduction

Machine learning models are popular in security tasks such as malware detection，PDF specifications and styles have become more diversity. The new version adds scripting and JavaScript to make documents and executables work in almost the same way, such as connecting to the Internet, running processes, and effecting with other files/programs. This increase in complexity provides the attacker with more weapons to launch attacks, and is more flexible in hiding malicious payloads and escape detection. Because enterprises and individuals generally respond to such security vulnerabilities slowly and their security awareness is insufficient,a large number of user systems are not updated with the latest version, and these attacks are ultimately successful.

Of the 24 0-days discovered in 2014, 16 were for Adobe Reader and Flash Player. It is also clearly observed that the vulnerability of Adobe Reader's discovery has been growing rapidly since 2015 in CVE, which opens the door to PDF-based document attacks.

For a variety of document-based attacks encountered in recent years, the traditional detection methods for PDF malicious files include shellcode-based detection [15], signature-based detection [13], and so on. These methods all have common problems such as low recognition rate and inability to update malicious code in time. Machine learning-based PDF malware detection provides a new direction for this purpose. The first method of using machine learning was 2011 Nedim Srndic et al., mainly extracting and analyzing javascript. Afterwards, many researchers extracted static files based on content and structure. Feature [8], or extracting file features based on metadata and structure [24] uses SVM and decision tree to classify files. After AI algorithm tuning, it can achieve good results. Our method combines the features they mention to extract, including structure, content, javascript, metadata information, and more. Through research, we found that our algorithm using random forests is more accurate than using SVM. At the same time, we also considered the security of the AI ​​model. Nedim Srndic et al. focused on the accuracy of the model not only in the later research, but also in the anti-escape and robustness of the AI ​​model. They were addressed at the IEEE conference [4]. Several assumptions were made in the escape of the AI ​​model. Finally, the classifier was successfully escaped. In our experiment, four of them were also used to verify whether our model has this ability to resist escaping. We passed some of the model. Feature selection and algorithm optimization, found that a part of the sample that was previously escaped can be detected, indicating that our model has a certain resistance to escape.

Paper Contributions. In summary, this paper makes the following contributions:

* 一个PDF数据集，总样本数201368 个，其中恶意样本173036个，正常样本28332个；
* 精心选取了一套静态特征集（133个）以用于刻画PDF恶意文档形象，以用于区分恶意与良性样本；
* 模型准确率高达99.82%，误报率0.01%，单个文件检测时间维持在毫秒水平；
* 成功使用自己生成的变种病毒对分类器发动逃逸攻击，分类器根据攻击进行自我修复，重新训练得出一个鲁棒性强，抗逃逸能力强的模型。

 A PDF dataset with a total number of 201368 samples, of which 173036 malicious samples and 28332 normal samples;

 A set of static feature sets (133) was carefully selected to describe the image of a PDF malicious document for distinguishing between malicious and benign samples;

 The accuracy of the model is as high as 99.82%, the false alarm rate is 0.01%, and the detection time of a single file is maintained at the millisecond level.

 Successfully uses its own generated variant virus to launch escape attacks on the classifier. The classifier self-repairs according to the attacks and retrains them to obtain a model with strong robustness and strong anti-escape capability.

## Background

## Related work

## Design and Implementation of ML Model

## Adversarial Analysis

## Applications

## Conclusion

**References**

1. Nedim ˇ Srndic and Pavel Laskov. Detection of Malicious Pdf Files Based on Hierarchical Document Structure. In 20th Network and Distributed System Security Symposium (NDSS), 2013
2. Nedim ˇ Srndic and Pavel Laskov. Mimicus: A Library for Adversarial Classifier Evasion. [https://github.com/srndic/ mimicus](https://github.com/srndic/mimicus).
3. Nedim Šrndic and Pavel Laskov . Hidost: a static machine-learning-based detector of malicious files, Šrndi′c and Laskov EURASIP Journal on Information Security (2016) 2016
4. Nedim Srndic and Pavel Laskov. Practical Evasion of a Learning- Based Classifier: A Case Study. In Proceedings of the 35th IEEE Symposium on Security and Privacy (Oakland), San Jose, CA, May 2014
5. Pavel Laskov and Nedim Srndic. Static Detection of Malicious JavaScript-Bearing PDF Documents. In Proceedings of the Annual Computer Security Applications Conference (ACSAC), 2011
6. Davide Balzarotti, Marco Cova, Christoph Karlberger, Christopher Kruegel, Engin Kirda, and Giovann Vigna. Efficient Detection of Split Personalities in Malware. In Proceedings of the 17th Annual Network and Distributed System Security Symposium (NDSS), San Diego, CA, February–March 2010
7. Igino Corona, Davide Maiorca, Davide Ariu, and Giorgio Giacinto. Lux0R: Detection of Malicious PDF-embedded JavaScript Code through Discriminant Analysis of API References. In Proceedings of the Artificial Intelligent and Security Workshop (AISec), 2014.PDFrate
8. Davide Maiorca, Davide Ariu, Igino Corona, and Giorgio Giac- into. A Structural and Content-based Approach for a Precise and Robust Detection of Malicious PDF Files. In *Proceedings of the International Conference on Information Systems Security and Privacy (ICISSP)*, 2015.
9. DavideMaiorca, DavideAriu, IginoCorona, andGiorgioGiacinto. An Evasion Resilient Approach to the Detection of Malicious PDF Files. In Proceedings of the International Conference on Information Systems Security and Privacy (ICISSP), 2016.
10. Davide Maiorca,IginoCorona,andGiorgioGiacinto.Lookingat the Bag is not Enough to Find the Bomb: An Evasion of Structural Methods for Malicious PDF Files Detection. In Proceedings of the 8th ACM Symposium on Information, Computer and Commu- nications Security (ASIACCS), Hangzhou, China, March 2013.
11. Davide Maiorca,Giorgio Giacinto,and Igino Corona. APattern Recognition System for Malicious PDF Files Detection. In *Pro- ceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM)*, 2012.
12. Cristina Vatamanu, Drago¸s Gavrilu ¸ T, and R˘azvan Benchea. A Practical Approach on Clustering Malicious PDF Documents. Journal in Computer Virology, June 2012.
13. Xun Lu, Jianwei Zhuge, Ruoyu Wang, Yinzhi Cao, and Yan Chen. De-obfuscation and Detection of Malicious PDF Files with High Accuracy. In *Proceedings of the 46th Hawaii International Con- ference on System Sciences (HICSS)*, 2013.
14. Weilin Xu, Yanjun Qi, and David Evans. Automatically Evading Classifiers: A Case Study on PDF Malware Classifiers. In Proceedings of the 2016 Annual Network and Distributed System Security Symposium (NDSS), San Diego, CA, February 2016. http://evademl.org/
15. Zacharias Tzermias, Giorgos Sykiotakis, Michalis Polychronakis, and Evangelos P. Markatos. Combining Static and Dynamic Analysis for the Detection of Malicious Documents. In Proceedings of the 4th European Workshop on System Security (EUROSEC), 2011.
16. Florian Schmitt, Jan Gassen, and Elmar Gerhards-Padilla. PDF Scrutinizer: Detecting JavaScript-based Attacks in PDF Docu- ments. In *Proceedings of the 10th Annual International Confer- ence on Privacy, Security and Trust (PST)*, 2012.
17. Kevin Z. Snow, Srinivas Krishnan, Fabian Monrose, and Niels Provos. ShellOS: Enabling Fast Detection and Forensic Analysis of Code Injection Attacks. In *Proceedings of the 20th USENIX Security Symposium (Security)*, San Francisco, CA, August 2011.
18. DaipingLiu,HainingWang,andAngelosStavrou.DetectingMa- licious Javascript in PDF through Document Instrumentation. In *Proceedings of the 44th International Conference on Dependable Systems and Networks (DSN)*, Atlanta, GA, 2014.
19. Carsten Willems, Felix C. Freiling, and Thorsten Holz. Using Memory Management to Detect and Extract Illegitimate Code for Malware Analysis. In Proceedings of the Annual Computer Security Applications Conference (ACSAC), 2012.
20. Curtis Carmony, Mu Zhang, Xunchao Hu, Abhishek Vasisht Bhaskar, and Heng Yin. Extract Me If You Can: Abusing PDF Parsers in Malware Detectors. In Proceedings of the 2016 Annual Network and Distributed System Security Symposium (NDSS), San Diego, CA, February 2016
21. Meng Xu and Taesoo Kim, Georgia Institute of Technology:PlatPal: Detecting Malicious Documents with Platform Diversity . 26th USENIX Security Symposium 2017
22. VirusTotal. Free Online Virus, Malware and URL Scanner.https://www.virustotal.com/.
23. Stephan Chenette. Malicious Documents Archive for Signature Testing and Research - Contagio Malware Dump. http://contagiodump.blogspot.de/2010/08/ malicious-documents-archive-for.html.
24. Charles Smutz and Angelos Stavrou. Malicious PDF Detection using Metadata and Structural Features. In Proceedings of the Annual Computer Security Applications Conference (ACSAC), 2012.
25. Symantec 2017年安全威胁报告https://www.symantec.com/content/dam/symantec/docs/reports/istr-22-2017-en.pdf

M.Polychronakis,K.Anagnostakis,andE.Markatos.Com- prehensive shellcode detection using runtime heuristics. In Annual Computer Security Applications Conference (AC- SAC), pages 287–296, 2010.