Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 4, pp. 925–934
DOI : 10.47974/JIOS-1616

A relational rule-based system for PDF malware detection

Geet C. Salame *
Nirlepa T. Shinde †
Prajakta P. Baad §
Deepak D. Kshirsagar ‡
Department of Computer Science and Engineering
School of Computational Sciences
COEP Technological University
Pune 411005
Maharashtra
India

Surendra Singh Tanwar @
Department of Electrical Engineering
Engineering College Bikaner
Bikaner 334004
Rajasthan
India

Abstract

Malware attacks are one of the significant issues in the cyber security domain. Many sectors, including financial, healthcare, IoT, etc., are affected by malicious activities through malware attacks. Every day, new malware variants are being released, and traditional solutions cannot detect new malware variants.

This paper proposes a methodology using relational rules to detect Portable Document Format (PDF) malware. The methodology includes data pre-processing, relational rules, and PDF malware detection. The proposed methodology obtains 89 relational rules from PART, OneR, and JRip rule-based models. These 89 rules are applied, and the performance is tested on Evasive-PDFMal2022. The methodology achieves a higher accuracy of 99.1223%

* *E-mail:* geetcs19.comp@coeptech.ac.in (Corresponding Author)
† *E-mail:* shindent19.comp@coeptech.ac.in
§ *E-mail:* baadpp20.comp@coeptech.ac.in
‡ *E-mail:* ddk.comp@coeptech.ac.in
@ *E-mail:* sstanwar@ecb.ac.in

with 89 relational rules on Evasive-PDFMal2022. Further, the proposed methodology is also compared with traditional malware detection systems.

Subject Classification: *Primary 68T99.*

Keywords: *PDF malware, Malware detection, Relational rules, Machine learning.*

1. Introduction

Malware is a significant problem at a global level since the nature of malware always keeps changing. Malware attacks [1] in the financial, healthcare [2], and IoT sectors are increasing. According to a survey [3] in Q3 2023, 169,194,807 links are identified as malicious. Social engineering is the biggest threat to organizations, which includes various emails, websites, phishing[4], social networks, message apps, and phone calls. 61%,74%, and 75% of organizations experienced malware activities [5] in 2020, 2021, and 2022, respectively. Two techniques [6] are used to detect malware i.e. static and dynamic analysis. Hackers have improved their attacking methods, which these traditional detection techniques do not detect easily.

Most of the people share their data with the help of PDF files. Cybercriminals use the documents in .pdf format extension to bypass the system's security. In the first half of 2022, increase of 30% malware [5] involved PDFs and Microsoft Office files. The use of malicious PDF files has increased fivefold since 2023 began. Hidden phishing is provided by adding vulnerability in a PDF file. Hackers use MalDoc techniques [7] to insert malware in PDFs. Hence, this research paper aims to study PDF malware detection.

PDFs have a flexible file structure that makes it easier to hackers [8] to insert malicious content. Attackers may exploit its sophisticated features and complicated structure. The popularity of PDF and the growing number of vulnerabilities in major PDF viewer software make it a favored choice for malware attack vectors.

The structure of a PDF consists of a header, trailer, body, and cross-reference table. The body section includes the document's contents, including embedded data like text and images. When a PDF file is viewed, the reader analyzes each object mentioned [9] in the cross-reference table and gradually renders the file's components.

The contribution of this paper is summarized as follows:

1. It proposes a relational rule-based methodology to detect PDF malware.
2. The method obtains relational rules from rule-based ML models and performs better for PDF malware detection.
3. Finally, the proposed method is compared with traditional malware detection systems.

The remainder of the paper is organized as follows. Section 2 describes the review of traditional malware detection systems. Section 3 presents the proposed methodology to detect PDF malware, and its implementation with results are illustrated in Section 4. Section 5 compares traditional malware detection systems with the proposed methodology, and the conclusion with the future scope is mentioned in Section 6.

2. Related work

The approach presented in [8] uses a stacking ensemble model for PDF malware detection. This approach is tested on Contagio and Evasive-PDFMal2022 datasets. The approach produces higher accuracy of 99.89% on Contagio and 98.69 % on Evasive-PDFMal2022 datasets with an ensemble model.

The approach in [9] proposes an enhanced feature set based on anomalies to detect evasive PDF malware. The approach adds six new features to the structural feature set. The approach was tested on Evasive-PDFMal2022 and achieved more than 98% accuracy with ensemble models.

The study proposes [10] model architecture to detect PDF malware with the Optimizable Decision tree (O-DT). The architecture is tested on Evasive-PDFMal2022, and the O-DT model produces higher recall, accuracy, and precision of 98.90%, 98.84%, and 98.80%, respectively.

The model presented in [11] utilizes Invasive Weed Optimization (IWO) to tune the parameters of Stacked Long Short-Term Memory (S-LSTM) to identify PDF malware. The model is tested on Contagio and Evasive-PDFMal2022 datasets. The IWO-S-LTM technique obtains 98.20% accuracy on the Evasive-PDFMal2022 dataset and 98.47% on the Contagio dataset.

Generative Adversarial Networks (GANs) are used [12] to identify malware in PDF files. The GAN model provides higher recall of 94.74% on PDF malware dataset.

The architecture presented in [13] extracts feature from Contagio, VirusTotal, and Evasive- PDFMal2022 and derived features for PDF malware detection. The ML and DL models are implemented on a combined dataset of Contagio, VirusTotal, and Evasive-PDFMal-2022. The Random Forest achieves a higher accuracy of 99.24% using 28 standard and seven derived features.

The methodology presented in [14] uses a classifier attribute evaluator technique with ZeroR for feature selection. The method selects the top-ranked 21 features and tests onEvasive-PDFMal2022 with ML models. The K-Nearest Neighbor (KNN) model produces 99.8599% accuracy.

The work mentioned in [15] uses image processing and visualization for PDF malware detection. The PDF malware is converted into grayscale images with Byte and Markov plot. The random forest classifier provides an F1 score of 99.48% compared to the decision tree and KNN on Contagio dataset.

The study [16] uses byte streams and a Convolutional Neural Network (CNN) to detect PDF malware. The study is tested on the dataset collected from different antivirus companies and achieves higher recall and F1 scores of 97.37% and 93.20%, respectively, using CNN.

The work [17] proposed a Multilayer perceptron with backpropagation for detection of PDF malware. The work is tested on extracted records from real-world datasets and obtains a higher true positive rate of 95.12%.

The literature review indicates that the latest types of malware are detected, but the detection performance could be better. Some of the methodologies are concentrated on real-time datasets, but the false alarm rate is high, implying the future scope of the proposed methods. It has motivated us to work in this domain.

3. Proposed Method

The proposed method used to detect PDF Malware is shown in Figure 1. The proposed method includes the identification of the dataset, data pre-processing, relational rules, and malware detection.

The captured dataset comprises normal and malware records associated with structural and general features. The dataset may consist of noise that degrades the performance of malware detection. So, data pre-

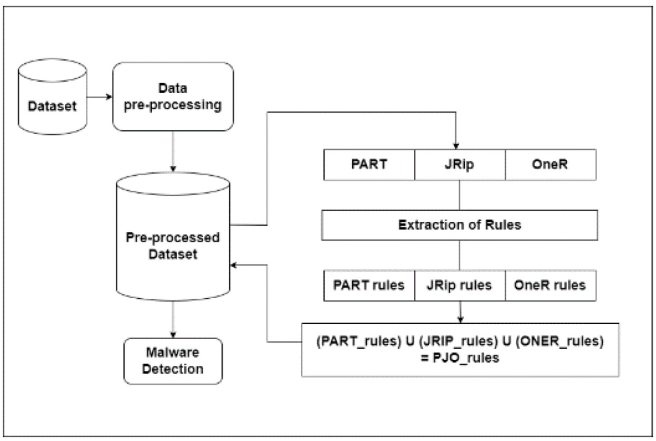

Figure 1

Proposed Method for PDF Malware Detection

processing is performed on the noisy dataset, and a pre-processed dataset is obtained.

The available rule-based models from the collection of machine learning are applied to the pre-processed dataset. The rule-based models create relational rules and perform malware detection. The created rules by three rule-based models such as PART, JRip, and OneR are were further used to detect PDF Malware. The relational rules created by PART, JRip, and OneR are combined, and resultant relational rules are applied to the pre-processed dataset. The resultant relational rules are used to detect PDF Malware, and performance analysis is performed. The proposed method used to detect PDF Malware is shown in Figure 1. The proposed method includes the identification of the dataset, data pre-processing, relational rules, and malware detection.

4. Implementation and Results

The method is implemented on a machine with 8.00 GB RAM and a 64-bit OS. The methodology is implemented [19] using the Python script and WEKA tool.

The proposed methodology uses Evasive-PDFMal2022 [8] to implement and detect PDF Malware. The dataset includes redundant features and missing values, some of which are associated with string type. The redundant features from the dataset are removed, and missing values are replaced with zero. Some of the features associated with string

type are converted into a numeric format using the to_numeric() function available in the Pandas library, and a pre-processed dataset [18] is obtained. The raw dataset includes 32 features, excluding labels. After the data pre-processing was performed on the raw dataset, the pre-processed dataset included 31 features with labels. The dataset includes benign samples of 4468 and malicious PDF samples of 5557.

The rule-based models available in machine learning are applied to the pre-processed Evasive-PDFMal2022[8] dataset. The PART, JRip, and OneR models [20] are applied, and the relational rules created by the models are extracted. The OneR model creates a single rule by choosing the best feature. The PART and JRip model work with 37 and 15 rules during training the model. The relational rules created by the OneR model are broken down into 39 rules and referred to as OneR_rules. The single default relational rule present in PART and JRip is removed. The remaining relational rules created by the PART and JRip model are referred to as PART_rules and JRip_rules.

Further, the PART_rules and JRip_rules include 36 and 14 rules, respectively. The rules created by PART, JRip, and OneR are applied to detect malware on the pre-processed dataset. The union operation is performed on OneR_rules, JRip_rules, and PART_rules. The union operation on three relational rule sets obtains 89 rules for PDF malware detection.

The relational 89 rules obtained from the OneR_rules, JRip_rules, and PART_rules are referred to as PJO_rules and applied to the pre-processed datasets. The methodology performs PDF Malware detection and performance analysis, as shown in Table 1.

Table 1 shows the performance analysis of relational rules created by OneR, PART, and JRip models. It also shows the performance analysis of

Table 1

Performance Analysis using Relational Rules

Individual Models	No of Rules	Accuracy (%)	Precision (%)	F1-score (%)	Recall (%)	FAR	Detection Time(sec)
OneR	39	87.6521	89.9306	86.6537	83.6072	0.086223	156.2
PART	36	97.9254	98.5679	97.6935	96.8345	0.011685	421.0
JRip	14	99.3417	99.4853	99.2632	99.0421	0.004154	213.6
Proposed Method	89	99.1223	99.5077	99.0203	98.5376	0.003991	567.2

89 rules obtained by union operation on the OneR, PART, and JRip rule sets. The single rule created by OneR is rewritten in 39 rules and achieves a lower accuracy of 87.652% with a lesser detection time of 156.2 seconds. The 36 rules created by the PART achieved 97.9254% accuracy with 421.0 seconds of detection time. The 14 relational rules created by JRip produce the highest accuracy of 99.3417% with a detection time of 213.6 seconds. The obtained 89 relational rules of PJO_rules achieve higher accuracy, precision, F1-score, and recall of 99.1223%, 99.5077%, 99.0203%, and 98.5376%, respectively, compared to rules of PART and OneR. The PJO_rules produces a higher precision of 99.5077% and a lower FAR of 0.003991 compared to JRip_rules with 567.2 seconds detection time.

5. Method comparison

A comparative analysis is performed with rule-based ML models and traditional malware detection systems. The rule-based ML models such as PART, OneR, and JRip are applied to the pre-processed Evasive-PDFMal2022 dataset and performance analysis is performed.

The performance analysis of the proposed method and rule-based ML models is shown in Table 2.

Table 2 shows that OneR produces lower accuracy and recall of 86.1460% and 82.2446% with a lower detection time of 1.1 seconds compared to the proposed method and the remaining two ML models. The 89 relational rules obtained from union operation on OneR, JRip, and PART rules obtain higher accuracy, precision, and F1-score of 99.1223%, 99.5077%, and 99.0203%, respectively, compared to PART, JRip, and OneR. The methodology obtains a higher recall of 98.5376% compared to OneR and a lower FAR of 0.003991 compared to PART and OneR, with a detection time of 567.2 seconds.

Table 2

Comparison of proposed method with ML models

Models/ Methods	Accuracy (%)	Precision (%)	F1-score (%)	Recall (%)	FAR	Detection Time (sec)
PART	99.0325	98.9483	98.9151	98.8819	0.008462	14.9
JRip	98.8630	98.7469	98.7248	98.7028	0.010081	31.5
OneR	86.1460	87.8944	84.9757	82.2446	0.103048	1.10
Proposed Method	99.1223	99.5077	99.0203	98.5376	0.003991	567.2

Table 3
Comparison with existing malware detection systems

Research	Model/ Method	Accuracy (%)	Precision (%)	F1-score (%)	Recall (%)
[8]	Stacking learning	98.69	98.88	98.77	98.87
[10]	O-DT	98.84	98.80	98.85	98.90
[11]	IWO-S-LSTM	98.20	98.65	96.4	94.43
Proposed Method	Relational rules	99.12	99.51	99.02	98.54

Finally, the proposed methodology used for PDF Malware detection is compared to traditional malware detection systems, as shown in Table 3.

Table 3 shows that the proposed method of relational rules produces higher accuracy, precision, and F1-score of 99.12%, 99.51%, and 99.02%, respectively, compared to the systems [8], [10], and [11]. The relational rules used in the proposed methodology achieve a higher recall of 98.54% compared to [11].

6. Conclusion

The presented method for PDF malware detection uses the relational rules generated by rule-based ML models. The rules generated by individual ML models and obtained 89 relational rules by union operation are tested on Evasive-PDFMal2022, and performance analysis is performed. The proposed methodology achieves higher accuracy and precision of 99.12% and 99.51%, using 89 relational rules. The proposed methodology produces higher accuracy and FAR than PART, JRip, and OneR models but requires more detection time.

The presented methodology will be extended with a feature selection method in the future to detect PDF malware, and detection time will be minimized.

References

[1] Abhay Pratap Singh Mahendra Singh, "Real time malware detection in encrypted network traffic using machine learning with time based features", *Journal of Discrete Mathematical Sciences & Cryptography*, Vol. 26 (2023).

[2] Morolong, Mamoqenelo P., Shava, Fungai Bhunu & Gamundani, AttleeM. Cloud computing security in health cyber physical systems, J*ournal of Discrete Mathematical Sciences and Cryptography*, 26:5, 1553–1568 (2023), DOI: 10.47974/JDMSC-1821.

[3] IT threat evolution in Q3 2023. Non-mobile statistics, SECURELIST by Kaspersky, Available at- https://securelist.com/it-threat-evolution-q3-2023-non-mobile-statistics/111228/ (2023).

[4] Rutuja R. Patil, Gagandeep Kaur, Himank Jain, Ayush Tiwari, Soham Joshi, Keshav Rao, "Machine learning approach for phishing website detection: A literature survey" *Journal of Discrete Mathematical Sciences & Cryptography,* Vol. 25 (2022).

[5] Sharma, S., Raja, L., Bhatnagar, V., Sharma, D., Bhagirath, S. N., & Poonia, R. C. Hybrid HOG-SVM encrypted face detection and recognition model. *Journal of Discrete Mathematical Sciences and Cryptography,* 25(1), 205–218 (2022).

[6] Roy, A., Razia, S., Parveen, N., Rao, A. S., Nayak, S. R., & Poonia, R. C. Fuzzy rule based intelligent system for user authentication based on user behaviour. *Journal of Discrete Mathematical Sciences and Cryptography,* 23(2), 409–417 (2020).

[7] Cybersecurity threatscape: Q3 2023, by Positive technologies, available at - https://www.ptsecurity.com/ww-en/analytics/cybersecurity-threatscape-2023-q3/ (2023).

[8] Issakhani, Maryam, Princy Victor, Ali Tekeoglu, and Arash Habibi Lashkari. "PDF Malware Detection based on Stacking Learning." In ICISSP, pp. 562-570 (2022).

[9] Yerima, Suleiman Y., and Abul Bashar. "Explainable Ensemble Learning Based Detection of EvasiveMalicious PDF Documents." *Electronics* 12, no. 14: 3148 (2023).

[10] Abu Al-Haija, Qasem, Ammar Odeh, and Hazem Qattous. "PDF Malware Detection Based on Optimizable Decision Trees." *Electronics* 11, no. 19 : 3142 (2022).

[11] Chandran, P., Rajini Hema, and M. Jeyakarthic. "Invasive weed optimization with stacked long short-term memory for pdf malware detection and classification." *International Journal of Health Sciences :* 4187-4204 (2022).

[12] PM, Pruthvi Priya, and P. Hemavathi. "PDF Malware Detection System based on Machine Learning Algorithm." In 2022 International

Conference on Automation, Computing and Renewable Systems (ICACRS), pp. 538-542. IEEE (2022).

[13] Hossain, GM Sakhawat, Kaushik Deb, Helge Janicke, and Iqbal H. Sarker. "PDF Malware Detection: Towards Machine Learning Modeling with Explainability Analysis." IEEE Access (2024).

[14] Khan, Bilal, Muhammad Arshad, and Sarwar Shah Khan. "Comparative analysis of machine learning models for PDF malware detection: Evaluating different training and testing criteria." *Journal of Cybersecurity* 5 : 1-11 (2023).

[15] Corum, Andrew, Donovan Jenkins, and Jun Zheng. "Robust PDF malware detection with image visualization and processing techniques." In 2019 2nd International Conference on Data Intelligence and Security (ICDIS), pp. 108-114. IEEE (2019).

[16] Jeong, Young-Seob, Jiyoung Woo, and Ah Reum Kang. "Malware detection on byte streams of pdf files using convolutional neural networks." *Security and Communication Networks* 2019 (2019).

[17] Zhang, J. "MLPdf: An Effective Machine Learning Based Approach for PDF Malware Detection. arXiv 2018." arXiv preprint arXiv:1808.06991 (2018).

[18] Kshirsagar, Deepak, and Pooja Agrawal. "A study of feature selection methods for android malware detection." *Journal of Information and Optimization Sciences* 43, no. 8: 2111-2120 (2022).

[19] Agrawal, Pooja V., Deepak D. Kshirsagar, and Anish R. Khobragade. "Symmetric uncertainty based feature selection method in android malware detection." In Recent Advances in Material, Manufacturing, and Machine Learning, pp. 934-941. CRC Press (2023).

[20] Vaidya, Atharva, and Deepak Kshirsagar. "Analysis of feature selection techniques to detect dos attacks using rule-based classifiers." In Applied Information Processing Systems: Proceedings of ICCET 2021, pp. 311-319. Springer Singapore (2022).