# SECURING PDFS: AN INNOVATIVE LSTM ALGORITHM FOR IMAGE- BASED MALWARE DETECTION

Maheshwaran. T[1], Manideep.M[2], Sai Chaitanya.K[3], Karthik.A[4]

Sri Manakula Vinayagar Engineering College, Puducherry, India smvec@smvec.ac.in

**Abstract.** The proposed system aims to enhance the current approach to combating PDF malware by addressing a key vulnerability in existing systems—specifically, the generation of evasive variants capable of bypassing machine learning based classifiers. Unlike the current system, the proposed solution leverages a hybrid algorithm and Variational Autoencoder (VAE) approach. Notably, it incorporates a pre-trained model to significantly reduce training time without compromising accuracy. This innovative combination of techniques presents an efficient and effective solution for image-based malware detection. In comparative testing, our proposed system outperforms the existing system, demonstrating superior accuracy and faster training times. By integrating hybrid algorithms and VAE, our approach provides an advanced defense against the evolving landscape of PDF based mal-
ware threats.

**Keywords:** Variational Autoencoder (VAE), PDF-based malware threats

## 1 Introduction

The term "PDF malware" refers to malicious software or code that is embedded within files using the Portable Document Format (PDF), a widely adopted format for sharing and distributing information. PDF files are commonly exploited by malicious actors who may take advantage of vulnerabilities in PDF readers or employ social engineering techniques to create deceptive PDFs containing hidden malware. This malware can be in the form of embedded scripts, links, or executable code, with the ultimate goal of exploiting vulnerabilities in the targeted system or executing unauthorized actions. When an unsuspecting user opens what appears to be a harmless PDF file, the embedded malware is triggered, potentially leading to security breaches, data theft, or compromise of the system.

In the context of combating PDF malware, it has become increasingly important to employ advanced detection techniques. These techniques include lstm algorithmic approaches and image-based analysis, which aim to identify and neutralize threats within PDF documents. By leveraging innovative algorithms and analyzing the content of PDF files at an image level, security measures can be enhanced to keep pace with the evolving sophistication of PDF-based malware.

The broader category of "malware" encompasses various types of malicious software intentionally crafted to cause harm or exploit vulnerabilities in computer systems, networks, and devices. This includes viruses, worms, trojan horses, ransomware, spyware, and adware, each designed for specific malicious purposes.

Cybercriminals commonly deploy malware through deceptive means, such as phishing emails, malicious websites, or compromised software downloads, to infiltrate and compromise the security of a target system.

## 2 .Related Work

1. Tuan Van Dao, [1] focuses on In the face of the escalating number and sophistication of malware, there is a pressing need for efficient detection and neutralization methods. Traditional approaches, relying on malware signatures or behaviors, often demand substantial computational resources for feature engineering. Recognizing the potential of machine learning in addressing these challenges, recent studies have explored the application of various techniques to identify and classify malware families. Although combining multiple state-ofthe-art methods has gained popularity, the challenge lies in determining an optimal and efficient combination. Complex neural network architectures have demonstrated improved classification performance, but they come at the cost of increased resource requirements.

2. Husam Kinawi.[2], In the realm of cybersecurity, the classification of malware holds paramount importance, providing crucial insights into the nature of threats and facilitating the development of effective countermeasures. The challenge intensifies in the context of realtime malware classification, particularly considering the high network throughputs of modern networks. Striking a balance between achieving high classification accuracy and maintaining low inference latency becomes a critical objective. This paper introduces two self-attention transformer-based classifiers, namely SeqConvAttn and ImgConvAttn, as innovative alternatives to the prevailing Convolutional Neural Network (CNN) classifiers. Going beyond this, the paper proposes a file-size-aware two-stage framework that integrates these transformer-based models, allowing for a nuanced control of the tradeoff between accuracy and latency in real-time classification scenarios.

3. Liang Ge,The escalating growth in the volume of malware, coupled with the deployment of sophisticated evasion and obfuscation techniques, has significantly impeded traditional signature-based approaches to cybersecurity. Machine learning-based methods have emerged as a promising alternative, offering faster analysis times and greater resilience against evasion tactics. In this paper, we present a semi-supervised malware detection and classification system centered around energy. The unique aspect of this system lies in its reliance on anomalyfree training data, aiming to detect and identify diverse malware types in test data. The system utilizes density estimates from the energy-based model as normalcy scores, effectively discerning normal code from potentially malicious counterparts.

4. Baig, In the face of an escalating threat landscape posed by malware, safeguarding data security and confidentiality has become a paramount concern. Recognizing the growing significance of malware detection, this research introduces an efficient behavioral malware detection system designed specifically for Portable Executable (PE) files. The approach relies on machine learning classifiers to discern and categorize potential threats. The study utilizes the Blue Hexagon Open Dataset for Malware Analysis (BODMAS), a recently published dataset spanning from August 2019 to September 2020, to both train and evaluate the proposed de-
sign.

5. Abu-L-Haija ,In the landscape of digital document sharing, Portable Document Format (PDF) files stand out as one of the most widely used formats, making them an attractive target for hackers seeking to exploit them as infection vectors. Security threats often emerge when malicious actors hide harmful code within apparently innocuous PDF documents, leading to the creation of PDF malware. Addressing this issue requires advanced techniques to differentiate between benign and malicious files. Research studies have consistently shown that machine learning methods offer efficient detection mechanisms against such PDF-based threats. This paper introduces a novel detection system designed to analyze PDF documents, distinguishing between benign and malware-infected files.

6. Sultan S, In the realm of system security, modern antivirus software often falls short in providing adequate protection against malicious Portable Document Format (PDF) files, posing a potential threat to computer systems. To address this vulnerability, this paper introduces a novel PDF malware classification system based on machine learning (ML). The unique aspect of this system lies in its dual approach to inspecting PDF files, employing both statistical and dynamic analysis. This dual-method approach enhances the accuracy of identifying the true nature of a document. Crucially, the method is non-signature-based, making it potentially effective in distinguishing obscure and zero-day malware threats. The experiment involves the deployment of five different classifier algorithms, and the best-fit approach is determined by evaluating key metrics such as true positive rate (TPR), precision, false positive rate (FPR), false negative rate (FNR), and F1-score for each algorithm

7. Land yang Fan, In the era of information digitization, PDF files have become a significant carrier of malicious documents, presenting a challenge for executable file detection technologies. While machine learning-based classifiers have shown improved effectiveness in detecting PDF malware, adversaries continually develop countermeasures, such as generating adversarial examples, to evade detection. In contrast to previous works that aimed to highlight vulnerabilities in learning-based detection models, this study focuses on addressing the computational demands associated with stochastic manipulations applied in existing research.

8. R. Kavitha,The proliferation of Android apps has brought about a corresponding increase in Android malware, posing a significant threat to mobile ecosystems. With Android phones constituting a substantial 72.2 percent of all smartphone sales, the potential impact of malware operations on these devices is substantial. Hackers employ various tactics such as credential theft, eavesdropping, and malicious advertising to compromise the security of Android phones. Recognizing the growing menace, researchers have delved into Android malware detection from diverse perspectives, proposing hypotheses and methodologies. This paper specifically focuses on machine learning (ML)-based techniques as effective tools for identifying Android malware. ML approaches have proven their efficacy by enabling the creation of classifiers from sets of training cases, eliminating the need for explicit signature definitions in malware detection

9. Nasir Ghani, Ransomware, utilizing encryption methods to render data inaccessible to legitimate users, has emerged as a formidable cyber threat, causing significant damage to governments, corporations, and private users. The proliferation of various ransomware families has prompted researchers to devise and propose a multitude of detection and classification schemes. In response to the escalating nature of these

cyber threats, many of these schemes leverage advanced machine learning techniques for processing and analyzing realworld ransomware binaries and action sequences. This paper conducts a comprehensive survey of the ransomware detection and classification landscape, categorizing existing solutions into several key categories. These categories include network-based approaches, host-based approaches, forensic characterization methods, and authorship attribution techniques.

10. Sarvar shah khan, As the transfer of files becomes more prevalent, the proliferation of maliciously coded documents has given rise to sophisticated cyber-attacks. Portable Document Format (PDF) files have become a significant vector for malware due to their adaptability and widespread usage. The challenge in detecting malware within PDF files lies in their capacity to incorporate various harmful elements such as embedded scripts, exploits, and malicious URLs. This paper addresses this challenge through a comparative analysis of machine learning (ML) techniques, specifically Naive Bayes (NB), K-Nearest Neighbor (KNN), Average One Dependency Estimator (A1DE), Random Forest (RF), and Support Vector Machine (SVM) for PDF malware detection.

**Table 1.** Analysis Table

| S.No | Paper Title - Author Name | Dataset | Merits | Demerits |
|---|---|---|---|---|
| 1. | Learn2Evade: Learning- Based Genera-tive Model for Evading PDF Malware Classifiers | PDFrate-v2 | he effectiveness of generated evasive PDF malware variants | it primarily targets specific classifiers |
| 2. | An Attention Mechanism for Combination of CNN and VAE for Image-Based Malware Classification | unbalanced and balanced Malimg da-tasets | complex malware variants can be detected | Balancing complexity and performance is a challenge |
| 3. | Self-Attentive Models for Real-Time Malware Classification | BODMAS PE malware dataset, BODMAS-11 and BODMAS-49. W | flexibility and improved performance | efficiency and low latency are crucial |
| 4. | PDF Malware Detection Based on Optimizable | EvasivePDF-Mal20 22 | the effectiveness of generated evasive PDF malware vari-ants | the need more adaptive and resilient detection strategies |

| | | | | |
|---|---|---|---|---|
| | Decision Trees | | | |
| 5. | LinRegDroid: Detection of Android | Drebin dataset, | enhanced accuracy and efficiency | the complex and non-linear characteristics |

## 3 Conclusion

In conclusion, the project aims to address the escalating threat of malware within PDF files by proposing an advanced malware classification system. By integrating a lightweight pre-trained model, Variational Autoencoder (VAE), and attention mechanisms, the project seeks to enhance the accuracy and efficiency of malware detection. Leveraging the "pdfmal-2022" dataset from the University of New Brunswick's Canadian Institute for Cybersecurity ensures the system is trained on diverse and real-world samples. Feature extraction plays a pivotal role in distilling relevant information from PDF files, while the use of a pre-trained model accelerates training and improves adaptability. The attention mechanisms and VAE further refine the system's focus on crucial features. Prediction, the culmination of the project, involves the model utilizing its learned knowledge to make informed decisions about the presence of malware in unseen PDF files. Through this comprehensive approachh, the project aims to contribute to the ongoing efforts to fortify cybersecurity measures against the dynamic and sophisticated landscape of PDF-based malware threats. The success of this endeavor would signify a valuable advancement in the field of malware detection, offering a resilient solution to counteract evolving cyber
threats.

## References

[1] J. Saxe and K. Berlin, "Deep neural network based malware detection using two dimensional binary program features," in Proc. 10th Int Conf. Malicious Unwanted. Softw., 2015, pp. 11–20.

[2] Z. Yuan, Y. Lu, Z. Wang, and Y. Xue, "Droid-Sec: Deep learning in android malware detection," ACM SIGCOMM Comput. Commun. Rev., vol. 44, no. 4, pp. 371–372, 2014.

[3] U. Bayer, P. M. Comparetti, C. Hlauschek, C. Kruegel, and E. Kirda, "Scalable, behavior-based malware clustering," in Proc. Netw. Distrib. Syst. Secur. Symp., 2009, vol. 9, pp. 8–11.

[4] J. Jang, D. Brumley, and S. Venkataraman, "Bitshred: Feature hashing malware for scalable triage and semantic analysis," in Proc. 18th ACM Conf. Comput. Commun. Secur., 2011, pp. 309–320.

[5] M. Cova, C. Kruegel, and G. Vigna, "Detection and analysis of drive-bydownload attacks and malicious javaScript code," in Proc. 19th Int Conf. World Wide Web, 2010, pp. 281–290.

[6] M. A. Rajab, L. Ballard, N. Lutz, P. Mavrommatis, and N. Provos, "CAMP: Contentagnostic malware protection," in Proc. Netw. Distrib. Syst. Secur. Symp., 2013.

[7] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "COMPA: Detecting compromised accounts on social networks," in Proc. Netw. Distrib. Syst. Secur. Symp., 2013.

[8] G. Stringhini, C. Kruegel, and G. Vigna, "Shady paths: Leveraging surfing crowds to detect malicious web pages," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., 2013, pp. 133–144.

[9] J. Schlumberger, C. Kruegel, and G. Vigna, "Jarhead analysis and detection of malicious Java applets," in Proc. 28th Annu. Comput. Secur. Appl. Conf., 2012, pp. 249–257.

[10] P. Laskov and N. Šrndi´c, "Static detection of malicious JavaScript-bearing PDF documents," in Proc. 27th Annu. Comput. Secur. Appl. Conf., 2011, pp. 373–382.