# Information Security Management

# Malware detection like trojan horse, spyware through memory analysis by using Machine Learning

**Course Code:** BCSE354E

**Course Title:** Information Security Management

**Slot:** TB2

**Faculty:** Dr. Leki Chom Thungon

**Team:**


1. D. Nahid Shameem – 21BCE5310
2. Mandava Nidhish – 21BCE5840
3. K Ravi Teja – 21BCE5627
4. N Divya Charitha – 21BCE5299

**Abstract:**

In this new era of digital world cybersecurity is one of the major standing issue which requires a frequent updating. Sophisticated malware such as Trojan horse and Spyware pose significant challenges to the digital security. The threat level of these malware became too high and they are advanced for our traditional detection methods. In order to ensure protection from these malwares with high effectiveness we are proposing an innovative approach to detect these malwares with a hybrid algorithm combing Gradient boosting and Random Forest classifiers. This study focuses on the analysis of the target system memory to identify the polymorphic malware that often evade conventional detection mechanisms. The proposed hybrid algorithm examines the patterns and behaviours within the system's volatile memory and enhances the accuracy and efficiency of malware detection. This involves collecting and analysing memory snapshots from various digital environments with both known and unknown malware scenarios. The hybrid algorithm is trained on a diverse dataset to ensure optimal performance across different variants of trojan horse and spyware. In conclusion the proposed approach achieves a balanced accuracy and computational efficiency in the malware detection. The results of this study will be very helpful in the advancement of memory-based malware detection techniques providing a strong defence against evolving cyber threats.

**Introduction:**

The landscape cyber security is constantly evolving; malware is becoming a persistent threat to the integrity and security of digital systems throughout the world. Among the diverse type of malicious software, Torjan horses and the spywares standout for their ability to infiltrate systems undetected, wreaking havoc on sensitive data and network infrastructure. Old and traditional methods of malware detection often fall short in identifying the sophisticated threats so the need for innovative approaches that can effectively combat their proliferation.

Memory analysis has been emerging as a promising frontier in the fight against malware, offering unparalleled insights into the runtime behaviour of applications and processes. By taking the contents of system memory, security analysts can uncover subtle indicators of malicious activity that evade conventional detection mechanisms.

In this context, the integration of machine learning algorithms presents a compelling opportunity to enhance the efficacy and accuracy of malware detection, particularly for malwares like Trojan horses and spyware.

In this project, we will able to develop an adaptive malware detection system leveraging the capabilities that have of memory analysis and machine learning. Our focus centres on the utilization of a hybrid algorithm that combines the strengths of gradient boosting and random forest classifiers, two powerful techniques renowned for their ability to handle complex datasets and discern intricate patterns within them and may also choose the different algorithms for better accuracy if necessary. By taking help of the best advantages of these algorithms, we aim to enhance the resilience of our detection framework against a wide spectrum of malware threats.

The primary objectives of our research endeavour include acquiring and pre-processing real-world memory dump datasets containing instances of various malware types, with a particular emphasis on Trojan horses and spyware. We then proceed to design and implement a hybrid algorithm, drawing upon their collective strengths to bolster the accuracy and efficiency of malware detection. Through this experimentation and validation methodologies, including cross-validation and performance metric analysis, we seek to assess the efficacy of our approach in detecting and mitigating previously unseen malware samples and zero threats.

By changing the boundaries of traditional malware detection methodologies, our project endeavours to make meaningful contributions to the field of cybersecurity, equipping organizations and security professionals with the tools and insights needed to safeguard against evolving threats. Through empirical validation and iterative refinement, we aspire to pave the way for more resilient and adaptive defence mechanisms capable of combating the ever-evolving landscape of malicious software effectively.

**Problem Statement:**

The objective is to develop a sophisticated malware detection system utilizing machine learning techniques for the identification and mitigation of threats such as trojan horses and spyware through memory analysis.

**Literature survey:**

The literature encompasses a series of research endeavors aimed at tackling the rising threat of malware and ransomware attacks in the digital landscape. One key focus revolves around leveraging machine learning techniques, notably XGBoost models, to detect and combat ransomware variants such as Revil, Lockbit, and BlackCat. By constructing comprehensive datasets and delving into memory-based analysis, researchers have achieved significant milestones, boasting accuracies upwards of 97.85% with minimal false positive rates [1]. These studies highlight the paramount importance of memory traces in ransomware detection, while also underscoring challenges posed by malware obfuscation and the scarcity of adequate analysis environments.

In parallel, efforts have been directed towards addressing the intricate issue of obfuscated malware detection. Through the development of hybrid classification models like MalHyStack, which amalgamate conventional machine learning algorithms with deep learning layers, researchers have made considerable strides in enhancing accuracy rates [2]. By employing meticulous feature engineering techniques and conducting rigorous feature selection analyses, these models have demonstrated exceptional accuracies, often exceeding 99.98% for binary classification tasks and showcasing commendable performances in categorizing malware into various classes [4].

Furthermore, a significant emphasis has been placed on behaviour-based malware detection methodologies, pivotal in identifying and mitigating dynamic threats. Leveraging dynamic analysis environments and runtime features extracted from sandboxing platforms like Cuckoo, researchers have achieved remarkable accuracies using ensemble machine learning algorithms. The integration of diverse

feature sets, including printable strings, Shannon entropy, and behavioural attributes, has led to accuracies surpassing 99.54% [3]. Additionally, investigations into machine learning and deep learning techniques have yielded promising results, with accuracies reaching 96% in classifying malware into distinct families [5]. These endeavours collectively underscore the importance of innovative approaches and advanced algorithms in confronting the evolving landscape of malware threats, fostering resilience and robustness in cybersecurity protocols.

The literature surveyed encompasses a variety of approaches to address the pervasive threat of malware across different domains, including computer systems, IoT devices, and mobile platforms. Researchers have recognized the evolving nature of malware, which encompasses viruses, spyware, bots, and ransomware, each posing unique challenges to system security [6]. Traditional signature-based methods, while effective to a certain extent, fall short in accurately detecting zero-day attacks and polymorphic viruses. In response, machine learning-based detection methods have gained prominence, offering a modern and adaptive approach to malware detection. These methods, categorized into static, dynamic, or hybrid analysis techniques, have been explored extensively to enhance accuracy and effectiveness in detecting malware across various platforms [7].

Specific research efforts have focused on improving malware detection in specific environments, such as IoT devices running the Android operating system. With the proliferation of Android malware variants employing sophisticated detection avoidance techniques, including obfuscation, there's a growing need to enhance detection mechanisms. Novel approaches, such as leveraging convolutional neural networks (CNNs) trained on Markov images of app executables, demonstrate promising results in accurately detecting obfuscated malware and identifying obfuscation types [8]. The use of CNN models trained from Markov images generated using application byte code showcases a sustainable and cost-effective method for obfuscated malware detection, which surpasses traditional feature-engineering-based approaches.

Moreover, research has explored machine learning techniques to detect malware in specific contexts, such as spear-phishing attacks utilizing macro malware written in VBA. Addressing the challenge of imbalanced datasets and practical performance evaluation, methods utilizing language models like Doc2vec and Latent Semantic Indexing (LSI) alongside popular classifiers have been proposed [9]. These methods mitigate class imbalance issues and exhibit robustness in detecting various types of macro malware, irrespective of the family type. Additionally, systematic literature reviews have been conducted to provide a comprehensive taxonomy of machine learning methods for malware detection. These reviews analyse a wide range of research works, categorize machine learning algorithms, and evaluate their performance to address existing challenges and guide future research directions in malware detection [10].

The literature review presented in the provided abstracts offers a comprehensive exploration of malware detection techniques using machine learning (ML) algorithms, particularly focusing on the Android platform and hardware-assisted approaches. The first set of abstracts discusses the effectiveness of various ML algorithms in detecting malware, emphasizing the importance of accurately identifying polymorphic malware that evolves to evade traditional signature-based detection [11]. Through experimental evaluations, the authors compare the performance of algorithms like Decision Trees (DT), Convolutional Neural Networks (CNN), and Support Vector Machines (SVM), highlighting their high detection accuracy rates and low false positive rates, which are crucial for effective malware detection.

Moving on to the context of Android mobile malware detection, the second set of abstracts delves into systematic reviews of ML-based techniques specifically tailored for Android devices [12]. Given the significant market share of Android smartphones, the authors underscore the vulnerabilities inherent in the open-source nature of the platform. They emphasize the evolution of malware threats targeting Android systems and the necessity of employing ML algorithms to effectively detect and mitigate these threats. Additionally, the abstracts highlight the importance of considering contextual features alongside API calls and permissions for improved detection accuracy [13].

Lastly, the literature includes abstracts focusing on hardware-assisted malware detection, acknowledging the limitations of software-based solutions, such as antivirus programs, in combating sophisticated malware attacks. The authors explore the integration of machine learning with hardware performance counters, trace buffers, and on-chip network traffic analysis to enhance the speed and accuracy of malware detection [14]. This approach aims to address the evolving nature of malware threats and the need for real-time detection in safety-critical systems. Additionally, the authors indicate future directions for research, including the exploration of explainable machine learning for precise classification of benign and malicious programs and interpretation of detection results for identifying malicious behaviours [15].

The literature review provided offers a comprehensive overview of recent advancements in malware detection utilizing machine learning (ML) algorithms, particularly focusing on techniques applied to the Android platform and leveraging Hardware Performance Counters (HPCs) [16]. The first paper highlights the increasing utilization of HPC events by ML algorithms for malware detection, demonstrating the efficacy of Neural Network (NN) algorithms like Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), and Full Order Radial Basis Function (RBF) in achieving high accuracy rates. Furthermore, it identifies key features of HPCs exploited for recognizing abnormal activities and outlines future research directions, including the evaluation of proposed algorithms with real HPCs datasets [17].

Moving forward, the second paper introduces a novel approach for detecting different categories of Android malware by utilizing a Huffman encoding-based feature vector generation technique. This method, leveraging system call frequencies as features, significantly improves the efficiency of the detection model. Results demonstrate the superiority of the proposed model, particularly with the Random Forest classifier, in achieving a detection accuracy of 98.70% [18]. The paper also underscores the importance of incorporating dynamic features like

system calls and suggests integrating static features such as permissions and API calls for future research.

Lastly, the third paper focuses on the evaluation of machine learning algorithms for Android malware detection using features extracted from Android manifest file permissions. It evaluates the effectiveness of various algorithms, with Random Forest emerging as the top performer in terms of precision, accuracy, and F1-score [19]. The paper emphasizes the inadequacy of commercial anti-virus tools and highlights the potential of machine learning algorithms in improving malware detection on Android platforms. Future research directions include the development of novel algorithms tailored to malware detection and the exploration of ensemble approaches and additional ML algorithms for static detection using manifest permission features [20].

**References:**


1. Aljabri, M.S., Alhaidari, F.A., Albuainain, A., Alrashidi, S., Alansari, J., Alqahtani, W., & Alshaya, J. (2024). Ransomware detection based on machine learning using memory features. Egyptian Informatics Journal.

2. Roy, K.S., Ahmed, T., Udas, P.B., Karim, M.E., & Majumdar, S. (2023). MalHyStack: A hybrid stacked ensemble learning framework with feature engineering schemes for obfuscated malware analysis. Intell. Syst. Appl., 20, 200283.

3. Jia, L., Yang, Y., Tang, B., & Jiang, Z. (2023). ERMDS: A obfuscation dataset for evaluating robustness of learning-based malware detection system. BenchCouncil Transactions on Benchmarks, Standards and Evaluations.

4. Singh, J., & Singh, J. (2020). Detection of malicious software by analyzing the behavioral artifacts using machine learning algorithms. Inf. Softw. Technol., 121, 106273.

5. Bokolo, B.G., Jinad, R., & Liu, Q. (2023). A Comparison Study to Detect Malware using Deep Learning and Machine learning Techniques. 2023 IEEE 6th International Conference on Big Data and Artificial Intelligence (BDAI), 1-6.

6. Al-Janabi, M., & Altamimi, A.M. (2020). A Comparative Analysis of Machine Learning Techniques for Classification and Detection of Malware. 2020 21st International Arab Conference on Information Technology (ACIT), 1-9.

7. A, D.K., P, V., Yerima, S.Y., Bashar, A., David, A., T, A.N., Antony, A., Shavanas, A.K., & T., G.K. (2023). Obfuscated Malware Detection in IoT Android Applications Using Markov Images and CNN. IEEE Systems Journal, 17, 2756-2766.

8. Mimura, M. (2020). An Improved Method of Detecting Macro Malware on an Imbalanced Dataset. IEEE Access, 8, 204709-204717.

9. Gorment, N.Z., Selamat, A., Cheng, L.K., & Krejcar, O. (2023). Machine Learning Algorithm for Malware Detection: Taxonomy, Current Challenges, and Future Directions. IEEE Access, 11, 141045-141089.

10. Chowdhury, M.N., Haque, A., Soliman, H.S., Hossen, M.S., Fatima, T., & Ahmed, I. (2023). Android Malware Detection using Machine learning: A Review. ArXiv, abs/2307.02412.

11. Akhtar, M.S., & Feng, T. (2022). Malware Analysis and Detection Using Machine Learning Algorithms. Symmetry, 14, 2304.

12. Senanayake, J.M., Kalutarage, H.K., & Al-Kadri, M.O. (2021). Android Mobile Malware Detection Using Machine Learning: A Systematic Review. Electronics.

13. AlJarrah, M.N., Yaseen, Q.M., & Mustafa, A.M. (2022). A Context-Aware Android Malware Detection Approach Using Machine Learning. Inf., 13, 563.

14. Lee, J., Jang, H., Ha, S., & Yoon, Y. (2021). Android Malware Detection Using Machine Learning with Feature Selection Based on the Genetic Algorithm. Mathematics.

15. Pan, Z., Sheldon, J., Sudusinghe, C., Charles, S., & Mishra, P. (2021). Hardware-Assisted Malware Detection using Machine Learning. 2021 Design, Automation & Test in Europe Conference & Exhibition (DATE), 1775-1780.

16. Bawazeer, O.M., Helmy, T., & Al-Hadhrami, S. (2021). Malware Detection Using Machine Learning Algorithms Based on Hardware Performance Counters: Analysis and Simulation. Journal of Physics: Conference Series, 1962.

17. Manzil, H.H., & Manohar Naik, S. (2023). Android malware category detection using a novel feature vector-based machine learning model. Cybersecurity, 6, 1-11.

18. McDonald, J.T., Herron, N., Glisson, W.B., & Benton, R. (2021). Machine Learning-Based Android Malware Detection Using Manifest Permissions. Hawaii International Conference on System Sciences.

19. Agrawal, R., Shah, V., Chavan, S.S., Gourshete, G., & Shaikh, N. (2020). Android Malware Detection Using Machine Learning. 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 1-4.

20. Harshalatha, P., & Mohanasundaram, R. (2020). Classification Of Malware Detection Using Machine Learning Algorithms: A Survey. International Journal of Scientific & Technology Research, 9, 1796-1802.