IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

LINUX MALWARE DETECTION AND CLASSIFICATION

[1] Dr.P.Kavitha, [2] M.Marivigneshwaran, [3] G.Muthukumar, [4] T.Pavithiran, [5] R.Priyanka

[1] Associate Professor, [2] B.E student, [3] B.E student, [4] B.E student, [5] Assistant Professor
Department of computer Science and Engineering
P.S.R Engineering College, Sivakasi.

ABSTRACT:

The primary objective of this project is to create a reliable malware detection and classification system that is specifically designed for Linux environments. This system uses machine learning to provide real-time detection capabilities that quickly identify known and unknown threats. One of the system's primary features is an extensive classification mechanism that classifies malware and offers useful details about threat characteristics. Ethical concerns ensure that data handling and user privacy are given the most importance during the detection process. This initiative attempts to minimize false positives, maximize resource efficiency, and improve flexibility in response to changing threats. The project increases the understanding of responsive malware detection and strengthens the security of Linux systems by providing a contribution to the field of cybersecurity.

1. INTRODUCTION

Linux systems have become essential parts of many infrastructures in today's digitally connected world, providing to a wide range of applications from servers and embedded devices to regular computer settings. Linux-based systems have grown in popularity due to their reputation for stability and security, which makes them essential for both personal and business computing. But hackers have unavoidably become more aware of this increasing relevancy. The Linux environment, which was before thought to be resistant, is now susceptible to an increasing amount of malware threats. The requirement for efficient malware detection and classification systems is growing as the number of Linux-based intrusions keep rising. This project addresses the growing problem of Linux malware by introducing a cutting-edge technique that uses machine learning to improve the security of Linux systems. The objective is clear to develop a strong and effective malware detection and classification system that is well matched to the characteristics of the Linux environment. The system is designed to provide real-time detection capabilities, quickly identifying not just known dangers but also developing threats, such as polymorphic malware. The design of the system deals with false positives, another important issue with conventional detection techniques. Linux malware detection and classification involve a multifaceted approach aimed at identifying, analyzing, and categorizing malicious software targeting Linux-based systems. Linux malware defense involves machine learning algorithms recognizing malware patterns and behaviors, supported by anomaly detection for monitoring system behavior deviations. Custom rules and specialized tools target specific threats, particularly rootkits, renowned for stealth. Network traffic analysis, behavioral scrutiny, and regular updates to malware databases form a robust, adaptive defense against the evolving Linux threat landscape.

2. LITERATURE SURVEY

Paper Name: Malware Detection and Classification Based on Graph Convolutional Network

Author Name : Hsiang-Yu Chuang, Jiann-Liang Chen, Yi-Wei Maa **Year :** 2022

The proposed study focuses on leveraging graphical convolutional networks and function call graphs to develop a model for detecting and classifying malware. By analyzing how malware behaves in sandboxes, the study aims to establish connections between function calls and functions. This enables the creation of a graph that represents the behavioral patterns of malware, potentially enhancing the ability to identify and classify different types of malicious software.

Paper Name : Malware Classification Framework Based on Deep Learning Algorithms

Author Name : Omer Aslan and Abdullah Asim Yilmaz

Year : 2021

The research introduces a novel hybrid architecture that combines two comprehensive pre-trained network models in an optimized way. This approach aims to enhance the classification of malware. The method's efficacy was evaluated using datasets like Malimg, Microsoft BIG 2017, and Malevis. The results from experiments indicate that the suggested method significantly improves the accuracy of malware classification, demonstrating high effectiveness in identifying and categorizing various types of malicious software with a notable level of precision.

Paper Name : Intelligent Vision-Based Malware Detection and Classification **Author Name:** S.Abijah, Roseline, S.Geetha, Seifedine Kadry and Yunyoung Nam

Year : 2020

This study work takes a different approach by utilizing a visualization technique that represents malware as 2D images. Through this visualization, the study proposes a robust anti-malware solution based on machine learning. By transforming malware into visual representations, the model aims to leverage machine learning techniques effectively to identify and counter various forms of malicious software, potentially overcoming the limitations of traditional detection methods.

Paper Name : Machine Learning Approach for Linux Malware Detection **Author Name:** Asmitha K and Vinod P

Year : 2014

The machine learning approach for Linux malware detection centers on employing advanced algorithms to scrutinize both the behavioral characteristics and code patterns specific to Linux systems. By doing so, this method aims to fortify cybersecurity measures by proactively detecting and mitigating threats posed by Linux-specific malware. This proactive identification and counteraction serve to protect Linux-based systems, preventing potential security breaches that might exploit vulnerabilities within the Linux environment, thus bolstering the overall security posture of these systems.

Paper Name : Malware Detection Using Machine Learning Algorithms

Author Name: Shyam sundar, K.S. Easwara Kumar

Year : 2012

Malware detection heavily depends on the utilization of machine learning algorithms that specialize in dissecting data patterns, unusual behaviors, and anomalies. As a crucial element in contemporary defense strategies within the constantly changing digital realm, this approach empowers the early identification and mitigation of diverse malware, ensuring a proactive stance against the evolving landscape of cyber threats.

3. PROPOSED SYSTEM

The "Linux Malware Detection and Classification" project is developed using the methods of machine learning to identify and safeguard Linux computers from a wide range of malware threats. This project involves a systematic approach with multiple phases, such as gathering data, extracting features, identifying malware, and categorising the results. An comprehensive procedure of data collection is an essential process of this work. To efficiently train and assess the machine learning models, a large diverse dataset of Linux system files is collected. This dataset serves as the foundation for the system's ability to understand and recognize the subtle intricacies that distinguish benign files from malicious files. The collection of samples, which includes samples that contain trojans, elements of distributed denial-of-service (DDoS), and covert backdoors, enables the system to counter a variety of possible threats. Feature extraction. It entails the laborious process of analysing and extracting useful features from the gathered Linux system files. These characteristics capture the key features of the behaviour and structure of the files, giving the machine learning model the knowledge it needs to determine the security of the files with accuracy. The most important part of the project is the machine learning techniques employed for malware detection and classification. The decision tree algorithm is utilized by the project by using the features that have been taken from the Linux

system files. This implementation reduces the possible threat created by malware by ensuring that the system can quickly and accurately detect and classify files. This system has the potential to protect vital Linux infrastructures from the various and constantly changing threats of the digital world because of its systematic approach to data collection, feature extraction, and sophisticated machine learning techniques. The proposed Linux Malware Detection and Classification system is a versatile and robust solution for identifying and categorizing malware threats. It begins with comprehensive data collection, encompassing both benign and malicious Linux system files. Feature extraction extracts key attributes, such as system calls, which are crucial for analysis. Machine learning takes center stage, enabling the system to differentiate between benign and malware by learning from the dataset. The system's classification capabilities extend beyond detection, encompassing various malware types like trojans, DDoS attacks, and backdoors. It prioritizes high accuracy, utilizing machine learning and selective feature extraction. In essence, this system integrates data collection, feature extraction, machine learning, and rigorous testing to provide a robust solution for Linux malware detection and classification, equipping it to effectively combat evolving cybersecurity threats.

4.SYSTEM IMPLEMENTATION AND MODULES DESCRIPTION:

ELF Mining: The ELF (Executable and Linkable Format) Mining module serves as the initial step in the data processing. As the dataset primarily comprises files in the ELF format, this module is responsible for extracting and transforming these ELF files into a format that can be ingested and processed by the subsequent stages of the project. ELF files are commonly used for executables, object code, shared libraries, and even core dumps. The ELF Mining module ensures that the entire dataset, consisting of Linux system files, is made compatible with the subsequent stages of the project's workflow.

Decision Tree Model: The most important stage of the project's development is represented by the Train Model module. Here, the model is meticulously trained using the vast and diverse dataset of malware samples. The primary objective is to equip the model with the ability to comprehend the underlying patterns, behaviors, and characteristics that differentiate benign files from malicious ones. The ultimate goal is to create a model that can consistently and accurately predict the security status of files, making it a valuable asset in proactively identifying potential threats in a Linux system.

Test Model : The Test Model module is a important component of the project, primarily serving as the validation and evaluation stage. It evaluates the model's ability to generalize from the knowledge it acquired during training and apply it to new data. By feeding the model with a variety of files, it assesses the accuracy and effectiveness of the model's predictions. The model is expected to demonstrate its capacity to accurately classify files into categories such as benign or various types of malware. The project's workflow is meticulously orchestrated through these three modules. The ELF Mining module acts as the data preprocessing stage, converting ELF files into a usable format. The Train Model module empowers the model with the ability to classify files based on its training with a vast dataset of malware samples. The Test Model module assesses the model's proficiency in making accurate predictions when confronted with previously unencountered data. Together, these modules constitute a robust system for Linux malware detection and classification, offering a proactive approach to safeguarding Linux systems from the ever-evolving landscape of security threats.

Feature Extraction: Feature extraction is the process of converting raw data, such as Linux system files in the ELF format, into a structured format suitable for machine learning-based classification. It involves selecting, processing, and representing relevant data elements as features in a feature vector. Feature extraction ensures that the data is prepared for analysis, allowing machine learning models to work effectively. The resulting feature vectors serve as input to machine learning algorithms, aiding in the detection and classification of malware in Linux systems.

FUNCTIONAL DIAGRAM:

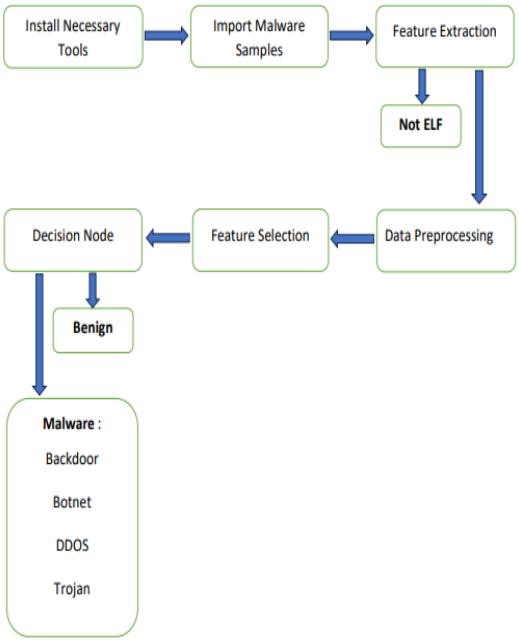

Figure: Flow Diagram For LMDC

Feature Selection: Feature selection in Linux malware classification involves identifying and choosing the most relevant attributes or characteristics from data samples to enhance classification accuracy, reducing computational complexity, and improving model performance by focusing on the most informative features. This process helps in distinguishing between benign and malicious software more effectively.

Data Preprocessing: Data preprocessing involves cleaning and organizing raw data to make it suitable for analysis. This process includes handling missing values, correcting errors, removing outliers, and transforming data into a consistent format. It may also involve reducing data dimensionality, integrating data from various sources, and formatting data appropriately. By preparing data effectively, ensure that machine learning models can work with high-quality, accurate data, leading to better results.

Benign: Benign is referred to as "good" or "legitimate" file, that is designed and used for legitimate and lawful purposes. Examples of benign file include operating systems, word processors, web browsers. These files are not intended to harm a computer or its data.

Malware: Malware is a category of file designed with malicious intent. It includes a wide range of file that aims to disrupt, damage, steal, or gain unauthorized access to computer systems and data. Common types of malware include viruses, worms, Trojans, spyware, adware, ransomware, and more. Malware can cause harm to computer systems, compromise user privacy, and lead to financial or data loss. Some malwares are,

DDoS (Distributed Denial of Service): A DDoS attack floods a target server or network with a massive volume of traffic, overwhelming it and causing a disruption in its normal functioning.

Backdoor: A backdoor is a hidden and unauthorized access point in a computer system or software that allows an attacker to gain control or access without typical authentication methods.

Trojan (or Trojan Horse): A Trojan is malicious software disguised as legitimate software. Once installed, it often provides unauthorized access to a computer or steals data without the user's knowledge.

Virus: A computer virus is a type of malware that attaches itself to legitimate programs or files. It can replicate and spread to other files, potentially causing harm to a computer system or its data.

5.SAMPLE OUTPUT:

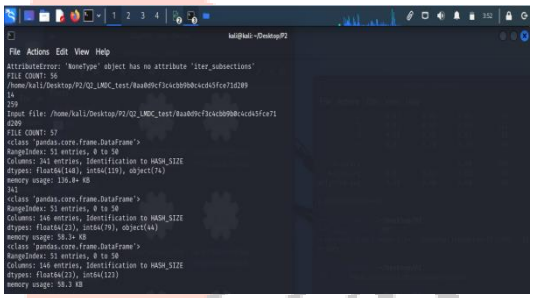

Figure: Training

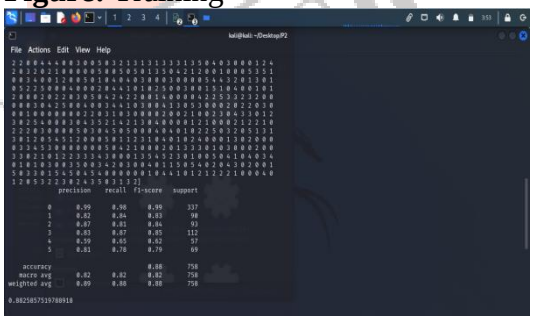

Figure: Testing

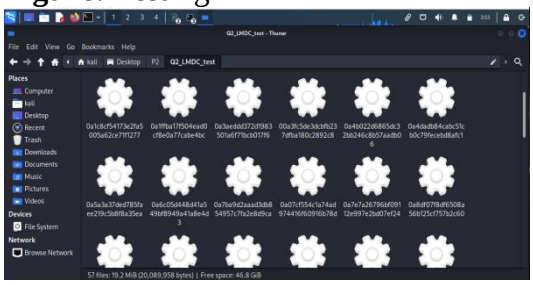

Figure: Malware Files

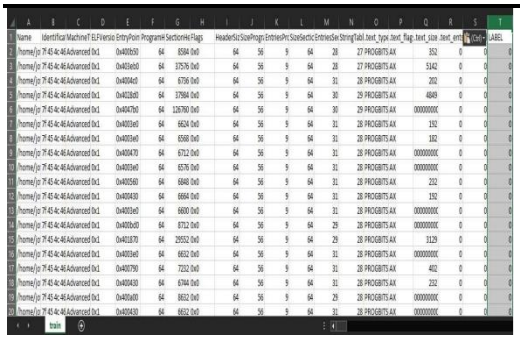

Figure: Training Data

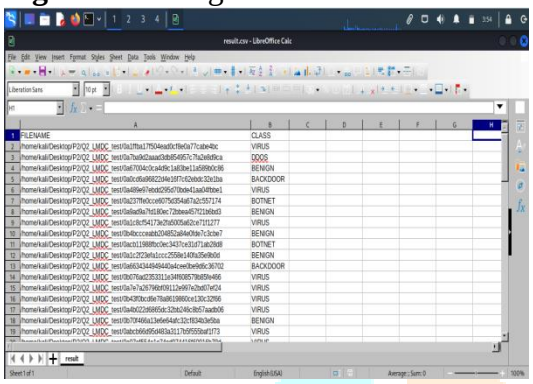

Figure: Malware Classification

6. CONCLUSION:

In this project, a Linux malware detection and classification system based on machine learning techniques was developed. For that, a large dataset of Linux malware samples was collected, then analyzed and various behavioural features extracted. The "Linux Malware Detection and Classification System" is a significant project with the objective of improving the security of Linux systems, which are essential parts of our digital infrastructure. A customizable, efficient malware detection system developed for Linux is more important than ever in the constantly evolving digital world, where cyber threats continue to increase. The primary objective of this project was to develop an improved malware detection and classification system by utilizing machine learning. The number of malware types targeting Linux is constantly growing, and conventional detection techniques are no longer sufficient. The system developed for this research has the ability to identify risks in real time, including malware threats, and it can respond quickly to polymorphic malware, protecting Linux security. The extensive classification system the project uses is one of its most important feature. This technique offers a complete understanding of the properties and behaviour of malware, going beyond simply detection. This not only makes it easier to identify threats, but it also gives organizations the ability to understand the risks they admit and make smarter decisions. The decrease in false positives is one of this project's significant results. Security workers and assets may be severely compromised by false alarms. The system becomes more reliable as well as effective, which improves the use of resources by reducing these inaccuracies. A significant step in strengthening Linux system security is the "Linux Malware Detection and Classification" project. By providing a comprehensive and ethical response to the constantly present problem of malware, it improves the technique of adaptable malware detection and categorization. As the backbone of modern technology in a period where digital infrastructure is most important, this project is essential to protecting all essential systems, data, and information. The system can achieve high accuracy and efficiency in detecting and classifying Linux malware and can also provide useful insights into the characteristics and behaviours of different malware types. It can be a valuable tool for security researchers and practitioners who deal with Linux malware analysis and prevention.

FUTURE ENHANCEMENT :

1. **Behavioral Analysis :** Implement more advanced analysis techniques to observe and identify malicious behavior in real-time, such as monitoring File transaction, File downloading, Process activities and enhancing the ability to detect the unknown or zero-day threats.

2. Real-time Updates and Response : Implement a system that can continuously update and response in real-time to detected threats, potentially including automated responses or quarantining of suspicious files.

3.Antivirus Application : To develop it as a Anti-virus application for the Linux operating system to prevent from malicious files, providing realtime security, a user friendly interface, continuous updates, threat intelligence and performance optimization.

REFERENCES

1. F. Shahzad, S. Bhatti, M. Shahzad, M. Farooq, "In-Execution Malware Detection using Task Structures of Linux Process," in: IEEE International Conference on Communication, 2011.

2. F. Shahzad, M. Farooq, "Elf-miner: Using structural knowledge and data mining methods to detect new (linux) malicious executables,"

Knowledge and Information Systems, 2011, pp. 1-24.
3. F. Shahzad, M. Shahzad, M. Farooq, "In-Execution Dynamic Malware Analysis and Detection by Mining Information in Process Control Blocks of Linux OS," Information Sciences [online],2013, pp. 45-63.
4. J.-S. Luo and D. C.-T. Lo, "Binary malware image classification using machine learning with local binary pattern," in Proc. IEEE International Conference, 2017.
5. J. Zhang, Z. Qin, H. Yin, L. Ou, S. Xiao, and Y. Hu, "Malware variant detection using opcode image recognition with small training sets," in Proc. 25th Int. Conf. Comput. Commun. Netw. (ICCCN), Aug. 2016, pp. 1–9.
6. R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, and S. Venkatraman, "Robust intelligent malware detection using deep learning," IEEE Access, vol. 7, 2019, pp. 46717–46738.
7. Ö. Aslan and R. Samet, "A comprehensive review on malware detection approaches," IEEE Access, vol. 8, 2020, pp. 6249–6271.
8. H.-J. Li, Z. Wang, J. Pei, J. Cao, and Y. Shi, "Optimal estimation of low-rank factors via feature level data fusion of multiplex signal systems," IEEE Trans. Knowl. Data Eng., early access, Aug. 13, 2020.
9. H.-J. Li, L. Wang, Y. Zhang, and M. Perc, "Optimization of identifiability for efficient community detection," New J. Phys., vol. 22, no. 6, Jun. 2020, Art. no. 063035, doi: 10.1088/1367-2630/ab8e5e.
10. R. Komatwar and M. Kokare, "A survey on malware detection and classification," J. Appl. Secur. Res., 2020, pp. 1–31, Aug.
11. Microsoft Malware Classification Challenge (Big 2015). Accessed: Apr. 20, 2021. [Online]. Available: https://www.kaggle.com/c/malware-classification.
12. W. Han, J. Xue, Y. Wang, F. Zhang, and X. Gao, "APTMalInsight: Identify and cognize APT malware based on system call information and ontology knowledge framework," Inf. Sci., vol. 546, 2021, pp. 633–664.
13. N. Usman, S. Usman, F. Khan, M. A. Jan, A. Sajid, M. Alazab, and P. Watters, "Intelligent dynamic malware detection using machine learning in IP reputation for forensics data analytics," Future Gener. Comput. Syst., vol. 118, 2021, pp. 124–141.
14. Xiaohui Chen; Ying Tong; Chunlai Du; Yongji Liu "MalPro: Learning on Process-Aware Behaviors for Malware Detection," IEEE Symposium on Computers and Communications (ISCC), 2022.