Chapter-V

MACHINE LEARNING IN MALWARE ANALYSIS AND PREVENTION

Dr. Sneha Deshmukh, Digital Forensics and Incident Response Expert, University of Mumbai, India.

Dr. Ashwin Menon, Digital Forensics and Incident Response Expert, University of Mumbai, India.

Abstract--- Strong and proactive internet security measures must be put in place since digital threats are becoming more frequent and sophisticated. The integration of AI techniques for predicting and detecting cyberattacks is examined in this study. We use unaided approaches for inconsistency discovery and administered learning calculations for break expectation on a variety of datasets, such as network logs, client behaviors, and framework exercises. To improve the precision and usefulness of danger-distincting evidence, social analysis and continuous observation frameworks are combined. Traditional detection techniques face several difficulties since malware is become more varied and complicated. This study looks at how well the Machine Learning algorithm, a potent machine learning tool, recognizes and categorizes malware samples. Emerging approaches such as behavior-based detection and semantic malware descriptions have shown promise and are deployed in commercial software. However, new techniques must be developed to keep pace with the development of malware.

Keywords--- Malware Analysis, Machine Learning, Software Defined Networks.

1. INTRODUCTION

The security landscape calls for innovative approaches to counter more sophisticated malware in a world of expanding cyberattacks. Due to their centralized and programmable nature, which offers more control and visibility over network functions, SDNs are a strategic advantage (Faruk et al., 2021). Information technology is ubiquitous and improvements to this information technology's security are needed to ensure a secure future. As critical infrastructures

increasingly rely on both public and private networks, there also exists a greater potential for widespread impact as a result of disruption or failure of such networks. To safeguard critical infrastructures means not only protecting their physical infrastructures but also the cyber aspects of the systems upon which they depend. With today's use of the Internet to facilitate the communications, monitoring, operations, and business systems supporting many critical infrastructures, there is greater opportunity for intentional, malicious compromise. Cyber-attacks are growing in number and severity. Attackers wishing to disrupt key infrastructures are motivated by various reasons and consider cyber-space as a potential tool to be able to have much more effect, e.g., to endanger humans or bring about extensive economic harm. While so far, no cyber-attack has significantly affected critical infrastructures, past attacks proved that there are wide vulnerabilities in information systems and networks that can cause serious harm. The consequences of a successful attack could involve severe economic impact through effects on key economic and industrial sectors, threats to infrastructure components like electric power, and disruptions that interfere with the response and communication abilities of first responders in emergency situations (Anderson et al., 2017; Nath & Mehtre, 2014; Saad et al., 2019).

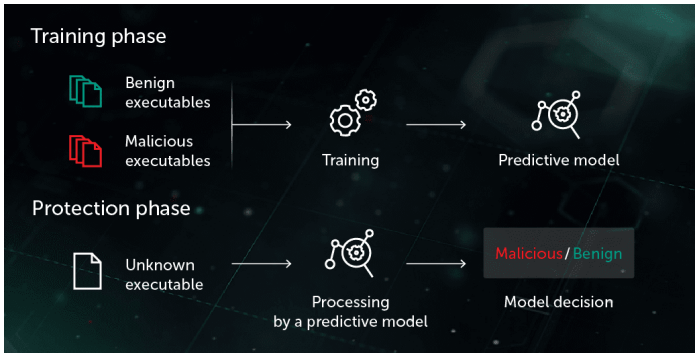


Figure 1: Machine Learning in Cybersecurity

Malware is a generic category of attack software or hardware loaded onto machines, generally unknowingly to the legitimate owner, which subverts the machine for the advantage of an attacker. Current classes of malware include viruses, worms, Trojan horses, spyware, and bot executables. Spyware is a type of malware utilized to track and/or forward information to an unauthorized third party

covertly. Bots (short for "robots") are malware applications that are secretly installed on a targeted computer, enabling an unauthorized user to take control of the infected computer remotely for a number of malicious intents. Botnets are collections of machines that have been infected by bot malware so that they are controlled by an attacker. Malware can be introduced at any stage in the system life cycle. The World Wide Web is now an important vehicle for spreading malware. Malware, for example, is remotely injected into otherwise innocent websites, where it can then go on to infect users of those so-called "trusted" sites (Dada et al., 2019; (Urooj et al., 2021).

2. REVIEW OF LITERATURE

The Significant Permission Identification (SigPID) method was developed by Li et al. and employs permissions as features to detect malware. In order to find the necessary rights that actually categorize the apps as safe or harmful, they mined the permission information (Udayakumar et al., 2017). The newest methods currently employed were contrasted with the authors' results. The outcomes illustrated the fact that SigPID is more efficient at identifying malware, with the accuracy of 91.4% gained by unknown malware. Zhu et al. have suggested a method (termed DroidDet) to mine APIs, permissions, permission rate, and system events as a crucial part. To develop a model for classifying whether an application is malware-infected, they employed ensemble rotation forest. The trial results proved that the proposed method was superior to other existing methods with accuracy of 88.26%. Kim and others have published a malware detection tool. The authors refine the numerous feature types they have extracted using existence-based or similarity-based feature extraction methods (Djenna et al., 2023).

The significance of Android intents and permissions as a discriminator to detect fraudulent applications has been studied by Feizollah et al. In the course of the research, 7,406 applications were utilized, out of which 1,846 were malicious and 5,560 were benign. The findings proved that the union of both the features yields a high recognition rate of 95.5% compared to solo characteristics. Agrawal and Trivedi have used various machine learning classifiers to study various malware detection methods. The outcomes demonstrated that RF outperformed other ML classifiers. Using dynamic behaviour features, Feng et al. have developed an effective dynamic

framework called EnDroid that can identify malware with high accuracy. To extract key features and remove unnecessary and noisy ones, they used a feature selection technique. Additionally, EnDroid used a stacking ensemble technique to differentiate between malicious and benign programs. The studies' findings demonstrated that stacking enhanced performance and provided a workable technique for virus identification (Haque et al., 2023). Mahindru and Sangal introduced a system for detecting malware on mobile devices called ML-Droid. This system used dynamic analysis to find mobile malware. Furthermore, a number of machine learning approaches incorporate dynamic features to make model creation easier. In the experiment, five million Android apps were used. The results demonstrated that the recommended approach had a 98.8% accuracy rate.

Qaisar and Li have offered a multimodal study of hazardous applications. The authors employed information fusion to identify the malicious programs based on their visual, static, and dynamic properties. They use a semi-supervised method for classifying and identifying malware. The findings demonstrated that their method outperformed other conventional methods with a 95% accuracy rate. Mad4a is a hybrid malware analysis method developed by Kabakus and Dogru. This strategy makes use of both static and dynamic methods. This method's significance lies in exposing the covert actions of Android malware (Faruk et al., 2022).

Ibrahim et al. presented a method based on static analysis and the key components of Android applications, such as two recently suggested features. An API DL model created especially for this use case then uses these features as input. The proposed method is implemented and evaluated on a classified dataset of Android apps. They focused on extracting features related to broadcast receivers, opcode sequences, services, permissions, and API requests. They also included two new static features: application size and fuzzy hash.

Patil and Deng demonstrated a neural networks-based malware analysis framework with high accuracy, demonstrating how using DL networks instead of conventional ML models may increase accuracy. The results of the experiment showed how accurately the DL-based malware classification method classifies malware. Furthermore, they suggested that DL's gradient descent and backpropagation processes enhance accuracy, reduce false positives, and raise true positives.

2.1. Malware Detection Methods

The most popular approach among all antivirus software is the signature-based approach. This method compares the signature of the application to an existing database signature. A disadvantage of this approach is its incapacity to detect new, undiscovered malware, sometimes known as zero-day malware. Researchers have started using machine learning approaches that use both static and dynamic malware analysis to get around this method's drawbacks.

2.2. Machine Learning

Computers can learn without explicit programming thanks to machine learning. It has been used in many different industries and has been a popular area of study in recent years. Medical data processing, false news categorization, and speech analysis or recognition are a few examples of these domains. Researchers start using machine learning techniques with both static and dynamic malware analysis (Sharma et al., 2023).

2.3. Types of Machine Learning

Three categories of machine learning exist: reinforcement learning, supervised learning, and unsupervised learning. The different ML categories are displayed in Figure 2.

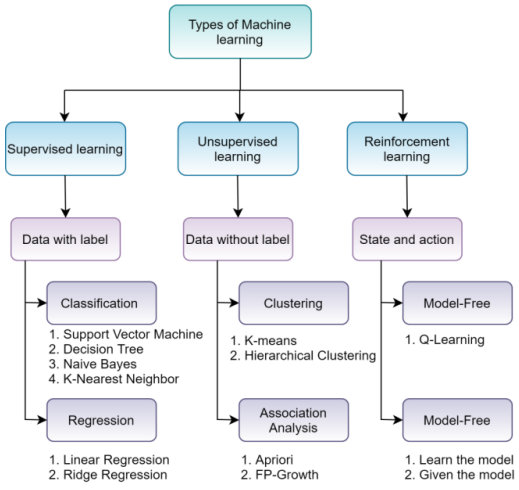
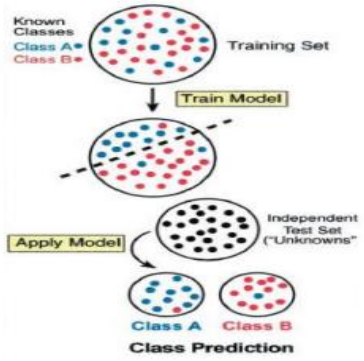


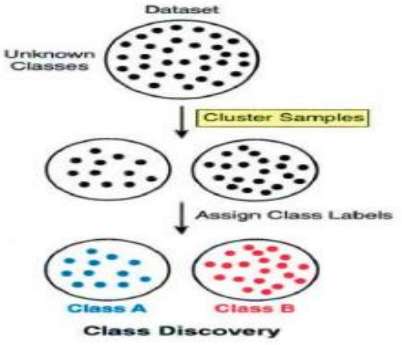
Figure 2: Various categories of ML

1) Supervised learning is divided into two categories: regression and categorization.
2) Classification- It involves classifying the cases into distinct groups. There are categorical values in the output attribute.

- Binary classification: This divides a set of instances into two categories, such as "malware" and "benign."”
- Multiclass classification: This divides a set of occurrences into two or more classes, such as "malware family classification."”
- Regression: Real values, such as "weights" and "heights," make up the output attribute.”



(a)



(b)

Figure 3: Unsupervised and Supervised Classification

- Another name for unsupervised learning is "learn without examples." This learning does not require any prior knowledge of the predicted attribute. The two kinds of this learning are association and clustering.
- Association: This problem is based on market analysis. It presents the collection of objects' association rule. For instance, what is the likelihood that item B will be purchased along with item. The Apriori algorithm is one instance of association rule mining.
- The practice of organizing data points into groups whose members are similar in some way is known as clustering. K-means, partitioned clustering, hierarchical clustering, and others are a few instances of clustering.
- Reinforcement learning: This dynamic programming technique uses a system of rewards and penalties to train the model. It is employed to specify the optimal course of action that enables the agent to maximize a reward while resolving an issue. Q-learning and R-learning are two instances of reinforcement learning.

2.4. Motivation

Smartphones are getting more and more common in our daily lives in the modern era. Many people use smartphones for a number of things, such as banking, shopping, entertainment, and gaming. Numerous operating systems are available on the market, such as Windows Phone, iOS, Android, and BlackBerry. Of these, Android is the most widely used (Alraizza & Algarni, 2023). With over 3.04 million apps, it holds an 85% market share. There were one billion mobile subscribers as of December 2019, but during the COVID-19 epidemic, that number rose to seven billion. Users now use smart devices to access a variety of services due to the exponential rise of mobile technologies. The growth of the upcoming economy and mobile Internet is significantly influenced by the proliferation of Android apps. Attackers have been interested in Android apps due to their growing usage. Recently, the rapid development of mobile technology has given rise to a number of dangers, including financial loss, information leakage, and system damage. According to the MacAfee research, there will be about 121 million more Android malware cases in 2020. The rise in Android malware has made manually managing the malware samples more difficult. To avoid this issue, a more efficient and automated method of malware detection must be developed (Akhtar & Feng, 2022).

The signature-based method is the foundation of the conventional malware detection techniques, which are ineffective at identifying novel viruses. In the past, the malware was created with straightforward goals in mind. It was therefore easier to identify. Conventional malware is the term for this kind of malware. Attackers are developing sophisticated malware that is more difficult to detect and can run in kernel mode. New generation malware is the term used to describe this kind of malware. The latest malware that is coming out is really intricate and smart. As a result, complex and sophisticated malware cannot be swiftly and precisely detected using traditional approaches. Therefore, techniques for enhancing malware identification and categorization need to be created.

3. PROPOSED METHODOLOGY

The focus is on addressing the shortcomings of traditional approaches and improving the overall hazard recognition and response capabilities while putting forth a high-level network safety framework that makes use of AI. Create a crossover strategy by utilizing a variety of AI models, such as controlled and solo learning computations. This might enhance general accuracy and lessen the drawbacks of particular models. Put mechanisms in place to make AI models more resilient to hostile attacks. To improve control efforts, adversarial preparatory techniques should be employed and models should be updated often. reaction systems that use posteriorizing to quickly address dangers that have been discovered. Computerized solutions that reduce reaction time and lessen the impact of safety events include disengaging compromised frameworks and altering security arrangements.

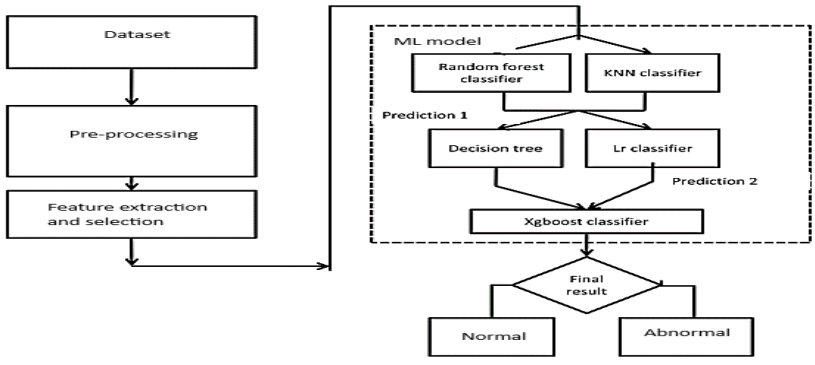


Figure 4: Basic Block diagram

Data Collection

It is typical practice to handle information assortment by obtaining testing of known malware variants from threat knowledge feeds and malware archives. Next, those examples are examined to remove structures that might be used to type and describe the behavior of malware, including byte sequences, procedure calls, page credits, and record hashes.

Scientists are also able to observe how malicious software tests perform in controlled experiments using dynamic probing methods like emulation and sandboxing. This allows experts to discover runtime exercises and obtain safe programming exams, which leads to a proper dataset for scheduling and evaluation. Even after static record inspection, malware recognition at times incorporates information collection, such as record changes, library modifications, network exchanges, and cycle injections (Lee et al., 2019).

Data Processing

Malware detection data processing is the methodical examination and modification of raw data to derive useful information for detection and prevention of malicious software threats.

Furthermore, labelling data helps in the effective training and assessment of supervised systems by labelling classes to specify whether a sample is benign or malignant. Operations such as data augmentation to extend the variety of datasets and separating data into train, validation, and test sets to accurately analyze model performance also fall under the category of data processing.

Feature Extraction

The process of identifying and removing relevant features or attributes from malware testing to assist in identification and description of malicious programming is referred to as highlight extraction for malware recognition.

These highlights allow AI systems to make the distinction between safe and malicious material by giving them important information about the features, patterns, and structure of malware.

Highlight extraction techniques for malware detection may vary based on the type of data being scanned, such as dynamic rules of behaviour, organization traffic,

and static record credits. Common highlights extracted from malware tests are such as text, byte sequences, code development, metadata, direct programming interface calls, document size, record type, and record hashes.

Detection

The process of developing detection models which can differentiate between unwanted and harmless software using labelled datasets for training machine learning algorithms is referred to as model training in malware detection.

From the input features captured in the course of data processing, the model learns basic malware sample patterns and characteristics as the model trains. This involves exposing the algorithm to the training dataset and successively modifying its inbuilt configurations for the purposes of reducing the loss function or predictive error. Neural networks, support vector machines, decision trees, and random forests are some of the machine learning algorithms that are often used for malware detection (Akhtar & Feng, 2023).

Prediction

The GBA model can classify new unseen samples once it is trained on labelled data and tuned to minimize prediction errors. The trained GBA model is given the features which were excluded from the input sample in the prediction phase. The model then calculates the probability or confidence level with which the sample belongs to the malicious class. With the help of a judgment threshold, the sample is classified as benign or malicious depending on this score (Buriro et al., 2023).

4. EXPERIMENTAL ANALYSIS

As was mentioned in the preceding section, some research publications suffered from unbalanced datasets, while others employed very large datasets. This problem stems from the researchers' conviction that bigger datasets improve accuracy and lessen bias. Even though this is a widely held belief, it's crucial to keep in mind that extremely huge datasets could present some difficulties.

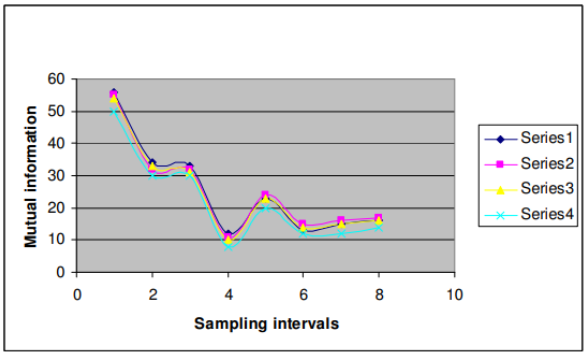


Figure 5: Normal Data

The time and processing power needed to train the model can be decreased by using a dataset of a manageable size. The use of current datasets is another issue that must be resolved in order to preserve the study's relevance and applicability in real-time circumstances.

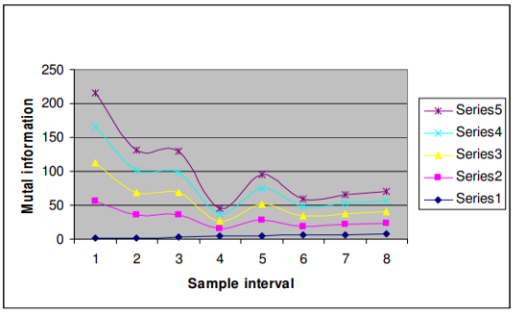


Figure 6: Attack Data

One popular preprocessing technique for dealing with unpredictable image sizes is resizing photos to a fixed size. Many machine learning models depend on the input photos having a consistent size and format, which is ensured by this step. But one must be aware of how information loss is affected by resizing. Therefore, more research is required to create resizing methods that reduce information loss while preserving an appropriate and consistent image size for processing (Awais et al., 2023).

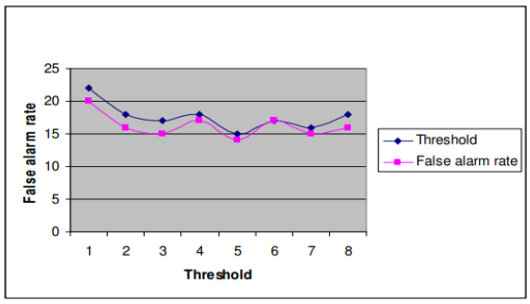


Figure 7: False Alarm Rate

Researchers may think about using distributed computing and parallelization approaches to expedite training and shorten the time needed to handle massive datasets.

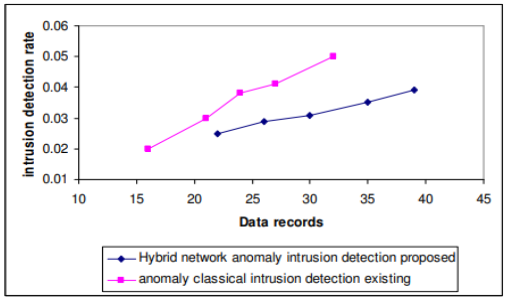


Figure 8: Performance of Anomaly Intrusion Detection in Combined Network

In order to train the model in parallel, this entails dividing the task among various computer resources, such as numerous PCs or GPUs. The time and memory resources needed for training can be greatly decreased with parallelization (Masum et al., 2022).

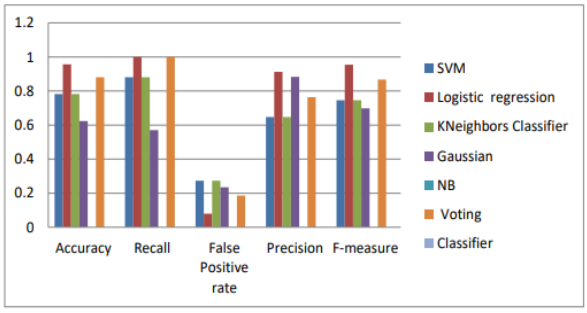


Figure 9: Performance Matrices of ML Algorithms on Mendeley Data

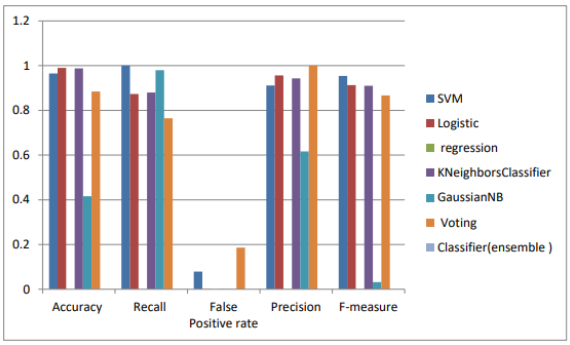


Figure 10: Performance Matrices of ML Algorithms on Figshare Dataset

Numerous trends in malware study and detection have surfaced over time, each having unique disadvantages. As a result, researchers have focused on other techniques that aim to increase detection and classification precision while identifying malware in real-time with few false positives (Bearden & Lo, 2017; (Cuan et al., 2018).

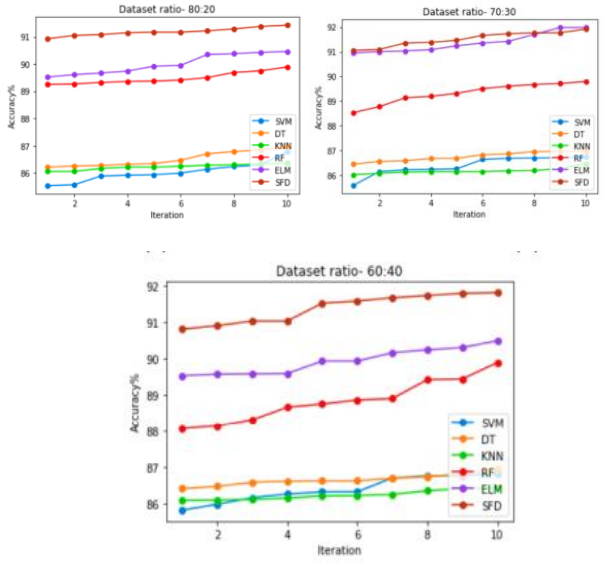


Figure 11: Evaluation of the Accuracy of Various Benign-to-malicious URL Ratios (a) 80:20, (b) 70:30, and (c) 60:40.

In any experiment involving classification or prediction, the dataset is crucial. The dataset includes information that gives the model a high degree of

comprehension. To improve model training and produce the best possible output, only pertinent information from the input dataset can be utilized. Therefore, a dataset is essential for developing and evaluating the suggested techniques.

5. CONCLUSION

With the high complexity and occurrence of malicious software, malware analysis forms a critical component of cybersecurity. Creating good protection involves knowledge about malware. Malware analysis enhances the detection and prevention of future attacks by classifying and recognizing different types of malwares. Since machine learning (ML) can analyse enormous amounts of data and detect complex patterns, it is critical to malware analysis. We provide an overview of the latest trends in machine learning-based malware analysis in this paper, explaining each of them. The works analysed demonstrate the merits and efficacy of using XML, DL, and transfer learning methods in malware studies. These methods contribute to the accuracy, explainability, and transparency of malware analysis and detection. Each trend's limitation and challenge are also addressed. We further provide some directions of research that could influence malware analysis in the future based on the survey outcome. These areas are capable of showing interesting opportunities for the field to continue to push itself in the way that will enable it to get past the challenges the researchers are facing. Although this study provides useful insight, it is important to keep in mind that it has some limitations. The findings might not be as generalizable as they potentially could because of the infinitesimal sample size. Repeating these findings with larger samples should be aimed at by later research.

REFERENCES

[1] Faruk, M. J. H., Shahriar, H., Valero, M., Barsha, F. L., Sobhan, S., Khan, M. A., ... & Wu, F. (2021, December). Malware detection and prevention using artificial intelligence techniques. In *2021 IEEE international conference on big data (big data)* (pp. 5369-5377). IEEE.

[2] Anderson, H. S., Kharkar, A., Filar, B., & Roth, P. (2017). Evading machine learning malware detection. *black Hat, 2017*, 1-6.

[3] Nath, H. V., & Mehtre, B. M. (2014, March). Static malware analysis using machine learning methods. In *International Conference on Security in Computer Networks and Distributed Systems* (pp. 440-450). Berlin, Heidelberg: Springer Berlin Heidelberg.

[4] Saad, S., Briguglio, W., & Elmiligi, H. (2019). The curious case of machine learning in malware detection.

[5] Akhtar, M. S., & Feng, T. (2022). Malware analysis and detection using machine learning algorithms. *Symmetry*, *14*(11), 2304. https://doi.org/10.3390/sym14112304

[6] Dada, E. G., Bassi, J. S., Hurcha, Y. J., & Alkali, A. H. (2019). Performance evaluation of machine learning algorithms for detection and prevention of malware attacks. *IOSR Journal of Computer Engineering*, *21*(3), 18-27.

[7] Urooj, U., Al-Rimy, B. a. S., Zainal, A., Ghaleb, F. A., & Rassam, M. A. (2021). Ransomware Detection Using the Dynamic Analysis and Machine Learning: A survey and Research Directions. *Applied Sciences*, *12*(1), 172. https://doi.org/10.3390/app12010172

[8] Udayakumar, N., Anandaselvi, S., & Subbulakshmi, T. (2017, December). Dynamic malware analysis using machine learning algorithm. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 795-800). IEEE.

[9] Djenna, A., Bouridane, A., Rubab, S., & Marou, I. M. (2023). Artificial intelligence-based malware detection, analysis, and mitigation. *Symmetry*, *15*(3), 677. https://doi.org/10.3390/sym15030677

[10] Haque, M. A., Ahmad, S., Sonal, D., Abdeljaber, H. A., Mishra, B. K., Eljialy, A. E. M., ... & Nazeer, J. (2023). Achieving organizational effectiveness through machine learning based approaches for malware analysis and detection. *Data and Metadata, 2*, 139-139.

[11] Faruk, M. J. H., Masum, M., Shahriar, H., Qian, K., & Lo, D. (2022, June). Authentic learning of machine learning to ransomware detection and prevention. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)* (pp. 442-443). IEEE.

[12] Sharma, P., Kapoor, S., & Sharma, R. (2023). Ransomware detection, prevention and protection in IoT devices using ML techniques based on dynamic analysis approach. *International Journal of System Assurance Engineering and Management*, *14*(1), 287-296.

[13] Alraizza, A., & Algarni, A. (2023). Ransomware detection using machine learning: A survey. *Big Data and Cognitive Computing*, *7*(3), 143. https://doi.org/10.3390/bdcc7030143

[14] Akhtar, M. S., & Feng, T. (2023). Evaluation of machine learning algorithms for malware detection. *Sensors*, *23*(2), 946. https://doi.org/10.3390/s23020946

[15] Lee, K., Lee, S. Y., & Yim, K. (2019). Machine learning based file entropy analysis for ransomware detection in backup systems. *IEEE access*, *7*, 110205-110215.

[16] Buriro, A., Buriro, A. B., Ahmad, T., Buriro, S., & Ullah, S. (2023). MalwD&C: a quick and accurate machine learning-based approach for malware detection and categorization. *Applied Sciences*, *13*(4), 2508. https://doi.org/10.3390/app13042508

[17] Awais, M., Tariq, M. A., Iqbal, J., & Masood, Y. (2023, February). Anti-ant framework for android malware detection and prevention using supervised learning. In *2023 4th International Conference on Advancements in Computational Sciences (ICACS)* (pp. 1-5). IEEE.

[18] Masum, M., Faruk, M. J. H., Shahriar, H., Qian, K., Lo, D., & Adnan, M. I. (2022, January). Ransomware classification and detection with machine learning algorithms. In *2022 IEEE 12th annual computing and communication workshop and conference (CCWC)* (pp. 0316-0322). IEEE.

[19] Bearden, R., & Lo, D. C. T. (2017, December). Automated microsoft office macro malware detection using machine learning. In *2017 IEEE international conference on big data (Big Data)* (pp. 4448-4452). IEEE.

[20] Cuan, B., Damien, A., Delaplace, C., & Valois, M. (2018, July). Malware detection in pdf files using machine learning. In *SECRYPT 2018-15th International Conference on Security and Cryptography* (p. 8p).