iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** II **Month of publication:** February 2024

DOI: https://doi.org/10.22214/ijraset.2024.58295

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue II Feb 2024- Available at www.ijraset.com

Malicious Code Detection Using Machine Learning

Dr. Pradeep Kumar[1], Mrs. Kakoli Banarjee[2], Mr. Ajay Kumar[3], Rajvi Nanda[4], Reeti Agarwal[5], Ritika Garg[6]
Department of Computer Science & Engineering JSS Academy of Technical Education, Noida, India

Abstract: ***Malware, derived from "Malicious software," is a comprehensive term encompassing any software intentionally crafted to disrupt, damage, or illicitly access computer systems. It's critical to determine whether a file includes malware. The increase in malware is causing a lot of problems for businesses, including data loss and other problems. Malware can swiftly inflict significant damage to a system by slowing it down and encrypting a sizable amount of data on a personal computer. This suggests that lowering the number of false positives is important. A comprehensive description of the adaptable framework for machine learning algorithms may be found in this study. It is possible to detect malware with these methods. The justification for this is that these algorithms simplify the process of distinguishing between files that are infected with malware and those that are not. Modern antivirus and anti-malware tools offer effective protection against various malware attacks. Nevertheless, due to the constantly changing landscape of malicious activities, it is imperative to curate an up-to-date database of previous malware instances. This repository serves as a valuable resource for anticipating the characteristics of future attacks and facilitating swift responses. Different machine learning methods, including decision trees and random forests, are used in our malware detection process. The method with the highest accuracy is chosen, giving the system an excellent detection ratio. Additionally, the confusion matrix is used to calculate the false positive and false negative rates, which is how the system's performance is determined.***

I. INTRODUCTION

Despite major advancements in security procedures, malware continues to evolve relentlessly and poses a strong danger in the ever-changing field of cybersecurity. Malware analysis is an essential part of cybersecurity that uses methods from network and programme analysis to break down malicious samples. The goal of this study is to do a thorough investigation of the corpus of work that uses machine learning to analyse malware. This study explores the difficulties encountered in the never-ending game of cat and mouse between malware developers and analysts and is specifically designed for security analysts, reverse engineers, and software developers. The never-ending game of cat and mouse that is the cybersecurity scene highlights how quickly virus creators react to improvements in security protocols. Conventional detection techniques, especially those that depend on the MD5 hash of detected malware, are vulnerable to escape through techniques like obfuscation. Polymorphism and metamorphism are sophisticated approaches that modify the binary code while maintaining consistent harmful behaviour, making detection even more difficult. The study emphasises the significance of detection criteria that take into account the semantics of bad samples. This makes it more difficult for malware makers to evade detection since they would have to make complex changes.

Machine learning is an obvious choice to help with knowledge extraction in malware analysis, given the need for novel approaches in this field. Numerous researchers have embraced machine learning in the literature, using a variety of techniques with a range of goals and results. Machine learning has the potential to discover new features that can improve security protocols and raise the bar for the evasion strategies used by malware creators. This study paper discusses the ongoing threat that malware poses in the constantly changing field of cybersecurity, even with major improvements made to security procedures. The main goal is to provide a comprehensive study and methodical arrangement of the body of literature that addresses the incorporation of machine learning into malware research. This survey is designed with security analysts, reverse engineers, and software developers in mind. It is a tactical manual for professionals who want to use machine learning to automate critical steps in malware research, which will reduce their burden. Acknowledging the ongoing game of cat and mouse between security experts and virus creators, the study investigates the difficulties encountered in the present environment. It uses a variety of studies from the literature that highlight creative methods and a range of goals to support the use of machine learning as a key tactic. The well-considered content offers recommendations for setting standards for machine learning-based malware analysis and offers insights that may be put into practice. Through the focus on investigating new patterns and the introduction of the innovative idea of malware analysis economics, this study provides professionals with a thorough understanding and useful tools to strengthen their defences in the ever-changing and demanding world of cybersecurity.

Notwithstanding the advancements, there are nonetheless issues with cybersecurity. Notwithstanding the growing danger of mobile viruses, Windows continues to be the major emphasis across all platforms. Traditional rule-based systems, which are predicated on pre-established rules and patterns, struggle to keep up with new and developing fraud tactics, which can result in a great deal of false negatives and possible financial losses.

Overcoming modern anti-analysis methods—particularly encryption—is the first task discussed. The second concern is how accurately malware behaviour is modelled, which is affected by the processes that are chosen to be analysed. The third problem pertains to obsolete or non-available datasets, which may affect the applicability and repeatability of results in malware analysis using machine learning. To overcome these problems, the study provides guidelines for creating suitable benchmarks in malware analysis using machine learning. These benchmarks aim to increase the rigour of assessments and ensure that security policies may be modified to tackle contemporary threats. Additionally, it is emphasised that research into emerging trends, such as malware attribution and prioritisation, is necessary to develop a holistic approach for combating malware.

This research introduces a novel concept called "malware analysis economics," which recognises the inherent trade-offs between analysis accuracy, time, and cost. Carefully balancing these components is necessary to set up a malware analysis environment. Navigating the complex terrain of viral analysis requires an understanding of the importance of efficiency.

In summary, this paper provides a comprehensive analysis and overview of the literature on the application of machine learning to malware analysis. It outlines the challenges faced, offers guidelines for improvement, and introduces the concept of malware analysis economics. By applying machine learning and resolving important problems, organisations may strengthen their cybersecurity defences and keep a step ahead of ever-changing malware varieties.

II. RELATED WORK

The table described below includes the various works of field research on malware detection and their main purposes, findings, and limitations.

TABLE 1

S.No	Author/Yr	Purpose	Findings	Limitations
1.	Rushiil Deshmukh et al. (2021)	Utilizing Machine Learning and Deep Learning for the Classification of Malware	Examines two methodologies for classifying malware: Utilizes a machine-learning approach to predict the specific class of malware to which each data point belongs among the nine available classes.	The availability of increased computing number of epochs. The model's accuracy can potentially be enhanced by incorporating assembly (ASM) files.
2	Oladimeji Kazeem et al. (2023)	Enhancing Fraud Detection through Machine Learning	Explore the application of machine learning algorithms in the realm of fraud detection and prevention. Create and implement a real-time monitoring system designed for the purpose of detecting fraud.	Utilizing combined and aggregated models, such as Gradient Boosting or Random Forest Machines, has the capacity to elevate the accuracy of fraud detection. Enhance the quality of the training dataset by implementing more sophisticated data preprocessing procedures to effectively handle missing or noisy data.
3.	IJRASET,et al.(2022)	Comprehensive Survey on Malware Detection Using Machine Learning.	The goal is to uncover a machine-learning-driven solution that tackles the issues presented by malware.	Signature-based techniques encounter two significant challenges: they are ineffective in detecting new or unknown malware, and they can be easily evaded by malware variants. Implementing dynamic techniques offers flexibility but can be a time-consuming process.

4.	Muhammad Shoaib Akhtar (2022)	Advancements in Malware Analysis and Detection through Machine Learning Algorithms	This research paper demonstrates that DT, CNN, and SVM exhibit strong performance in terms of detection accuracy.	The accuracy can be improved further.
5.	S. Soja Rani et al. (2020)	A comprehensive survey of various methodologies for malware detection employing machine learning techniques.	This research paper presents a comprehensive examination of the evolution of concealment techniques.	Classification-specific methods depend on the combination of various feature selection techniques.
6.	Mohammed Altay (2023)	Utilizing Deep Learning Algorithms for Malware Detection	This research paper aims to identify malware by utilizing the dataset generated by CTU University in 2011, which combines certified botnet traffic with normal network activity.	Difficulty in training Deep Learning Models.
7	Malak Aljabri et al. (2022)	Utilizing Machine Learning Techniques for the Detection of Malicious URLs	This paper concentrates on reviewing research studies concerning the utilization of machine learning algorithms for the detection of malicious URLs. The article introduces various taxonomies and provides comparative results, contributing valuable insights to the field of malicious URL detection.	The SVM algorithm employed faces limitations in handling large or noisy datasets.
8	Ferhat Ozgur et al. (2020)	Utilizing Machine Learning for the Detection of Malicious URLs	This survey suggests enhancing classifier performance in the detection of malicious websites by incorporating host-based and lexical features from the associated URLs. Random Forest models and Gradient Boosting classifiers are employed to develop a URL classifier, utilizing URL string attributes as features.	By employing J48, SVM, KNN, and NB machine learning algorithms, a comprehensive comparison can be conducted to evaluate their processing time and accuracy performance in the detection of malicious URLs. This comparative analysis aims to assess the strengths and weaknesses of each algorithm, providing insights into their efficiency and effectiveness for the specific task of identifying malicious URLs.
9	Gwanghyun Ahn et al. (2022)	Machine Learning-Based Method for Detecting Malicious Files	In this study, the enhancement of detection accuracy was achieved by leveraging the benefits of dynamic analysis for the detection of malicious files.	The drawbacks of both static and dynamic analysis methods are addressed, enhancing the ability to accurately detect malicious code while improving the speed at which this detection is conducted.
10	Mohamed Baset(2016)	Applying Machine Learning for the Detection of Malware	The survey's findings indicated that the accuracy of the results claimed by the authors may be influenced by certain factors, such as sample size, the number of features, or the technique employed for feature extraction.	The evaluation results lack coverage of string and byte n-grams features.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue II Feb 2024- Available at www.ijraset.com

III. COMPARISON OF MALICIOUS CODE DETECTION MODELS

The table described below shows the advantages and disadvantages of different machine learning models.

TABLE 2

Model Name	Advantages	Disadvantages
Naïve Bayes	It enables rapid and highly scalable model building and scoring, with a linear scalability observed concerning the number of predictors and rows.	In cases where the test data contains a categorical variable representing a category not included in the training dataset, the Naïve Bayes model assigns a zero probability to that category. Consequently, the model is
		unable to generate predictions for this particular category.
Logistic Regression	It anticipates the correlation between input features and the likelihood of a transaction being fraudulent. Its preference for fraud detection arises from its straightforward legibility and simplicity.	Logistic regression assumes a linear relationship between the independent variables and the log-odds of the dependent variable. This assumption may limit its applicability in scenarios with complex, non-linear relationships.
Decision Tree	Capable of handling non-linear correlations between features and the target variable, decision trees offer the advantage of being well-suited for discerning intricate patterns in fraud detection.	Decision trees are prone to overfitting, especially when the tree is deep and captures noise or outliers in the training data. Pruning or limiting the tree depth can help mitigate this issue.
Support Vector Machines	Capable of managing high-dimensional data and non-linear relationships.	The model might exhibit bias toward the majority class (legal transactions) when dealing with unbalanced datasets, resulting in decreased performance in accurately identifying the minority class (fraudulent transactions).
Random Forest Classifier	Integrates multiple decision trees to enhance accuracy and address intricate fraud patterns.	Might not yield favorable results for small datasets (those with few features) as the impact of randomness is significantly diminished.

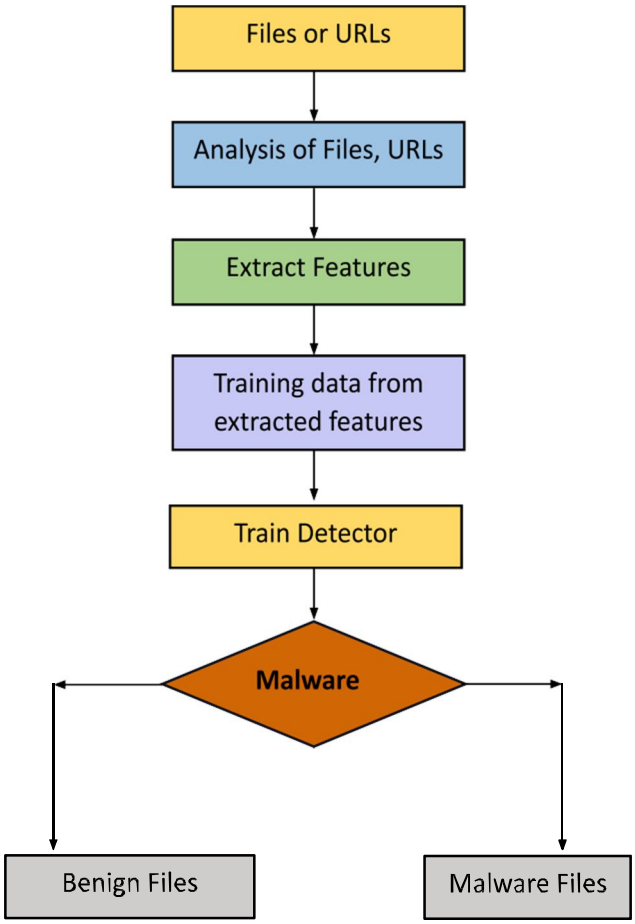
IV. METHODOLOGIES

The proposed methodology employs various machine learning algorithms to discern the authenticity of files and URLs, particularly focusing on malware classification. Data points are categorized into nine malware classes using three distinct ML models: logistic regression, Random Forest, and Multilayered Perceptron classifier. The models operate on key parameters such as file size in bytes, hex-code uni-gram, hex-code bi-gram, and the final feature matrix. Each model has its own unique classification factor, with regularization for logistic regression, number of nodes for Random Forest, and the number of hidden layers for the multi-layered Perceptron classifier.

In the alternative Deep Learning approach, the methodology leverages convolutional operations. This involves the application of a filter of a predetermined size across groups of pixels within an image. The resulting output is generated by the interactions among these pixels, producing an image that faithfully represents the spatial and temporal dependencies embedded in the data.

In the end, after applying different ML models to the given datasets, we can determine whether the given file or URL is benign or malicious.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue II Feb 2024- Available at www.ijraset.com



A. Datasets

The first task is to collect data points in order to train, test, and validate an ML model. These data points consist of input features (attributes or variables) and corresponding output labels. The primary goal is to use the datasets to train a model that can learn patterns, relationships, or trends. Datasets in ML are typically divided into three subsets: Training set, Validation set, and Testing set. The performance and capacity for generalisation of machine learning algorithms are strongly influenced by the selection and calibre of datasets. When it comes to datasets in machine learning, keep the following points in mind:

1) *Instructional Dataset:* The ML model is trained using the training dataset. It is made up of labelled examples in which the target variable or related outcome is coupled with the input data. To guarantee that the model learns strong patterns and can generalise well to new data, it is imperative to have a varied and representative training dataset.
2) *Validation Set:* A different validation dataset is frequently used in the training phase in order to adjust hyperparameters and avoid overfitting. It guides changes to increase generalisation by evaluating the model's performance on data that it hasn't seen previously.
3) *Exam Dataset:* For assessing the model's performance following training, the testing dataset is essential. It must be different from the training and validation sets in order to provide an objective evaluation of the model's capacity to generalise to fresh, untested data.
4) *Annotation and Labeling:* Accurate data labelling is crucial, particularly for supervised learning. Labels give the model the ground truth it needs to identify patterns. Sometimes manual labelling or annotation is required, especially for complicated tasks like sentiment analysis or object detection
5) *Preparing Data:* To improve model performance, preprocessing techniques, including cleaning, normalisation, and feature engineering are frequently applied to datasets. The model's capacity to discover significant patterns is directly impacted by the calibre of preprocessing.

6) *Unbalanced Collections:* Challenges may arise from imbalanced datasets, in which some classes are underrepresented. To avoid biased model results, managing this issue carefully is necessary. Methods including under- or oversampling, as well as the application of certain algorithms, are used.
7) *Public Datasets:* A wide range of public datasets are available for machine learning tasks. Sites such as Kaggle, UCI Machine Learning Repository, and others offer access to a variety of datasets, promoting cooperation and benchmarking between researchers and practitioners.
8) *Ethical Considerations:* It is important to make sure that datasets are representative and unbiased in order to prevent the perpetuation of societal biases. Privacy and fairness are two ethical considerations that should be taken into account when selecting and managing datasets.
9) *Continuous Learning:* Over time, ML models may come across new scenarios or data patterns. By regularly updating and expanding datasets, models can adapt and continue to be effective in changing environments.
10) *Domain-Specific Information:* Domain-specific datasets are needed for some ML applications. For example, datasets including annotated medical pictures may be necessary for medical image analysis, while text corpora unique to particular sectors may be needed for natural language processing.

B. Pre-Processing

The second step involves the meticulous process of cleaning, transforming, and organizing raw data into a format suitable for training a Machine Learning (ML) model. Effective pre-processing is crucial, as the quality of the input data directly influences the accuracy and reliability of the trained model. Pre-processing raw data is a critical step in the complex process of creating a strong Machine Learning (ML) model. Data must be carefully cleaned, transformed, and arranged in order for the input to be shaped in a manner that will support efficient model training. The quality of this pre-processing step is closely related to the final ML model's accuracy and dependability.

1) *Data Cleaning:* Find and fix missing numbers, inconsistent patterns, and errors in the dataset. This helps to improve learning outcomes by ensuring that the model is not impacted by noise or unrelated information.
2) *Data Conversion:* To improve the dataset's interoperability with the selected ML methods, alter or convert it. In order to maximise the data for effective model training and minimise the effects of different scales, this may entail normalising, scaling, or encoding categorical variables.
3) *Organization of Data:* Organise the data according to a consistent, logical structure that complies with the demands of the machine learning model. Feature engineering is the process of selecting or creating pertinent characteristics to capture important patterns and relationships.

C. Feature Extraction

Feature extraction is an essential step in the complex process of creating Machine Learning (ML) models. Right now, the main goal is to turn raw data into a condensed representation, which is a collection of features that contains the most important information for the learning process. This complex procedure includes choosing, merging, and modifying input variables in order to improve the machine learning model's effectiveness and performance.

1) *Simplified Illustration:* The main objective is to extract the most pertinent information from raw data and condense it into a small set of features. Understanding and generalising the model is made easier by this concise representation.
2) *Pertinent Data:* By choosing relevant characteristics, one may make sure the model concentrates on the factors that have a major impact on the learning task. To simplify the process, unnecessary or redundant elements are removed.
3) *Boosting Performance:* Improving the selected features' overall performance is the goal of feature transformation and optimisation. To meet the unique needs of the learning job, this can entail building new features, scaling them, or normalising existing ones.

D. Feature Selection

1) *Introduction:* Feature selection becomes more important in the complex dance of creating a strong Machine Learning (ML) model. Hand-selecting a subset of characteristics from the original set is the focus of this crucial stage, which is a purposeful curation meant to maximise relevance and significance and eventually improve model performance. The primary goals are to reduce overfitting and improve interpretability by bringing the model into line with the many subtleties of the underlying data.

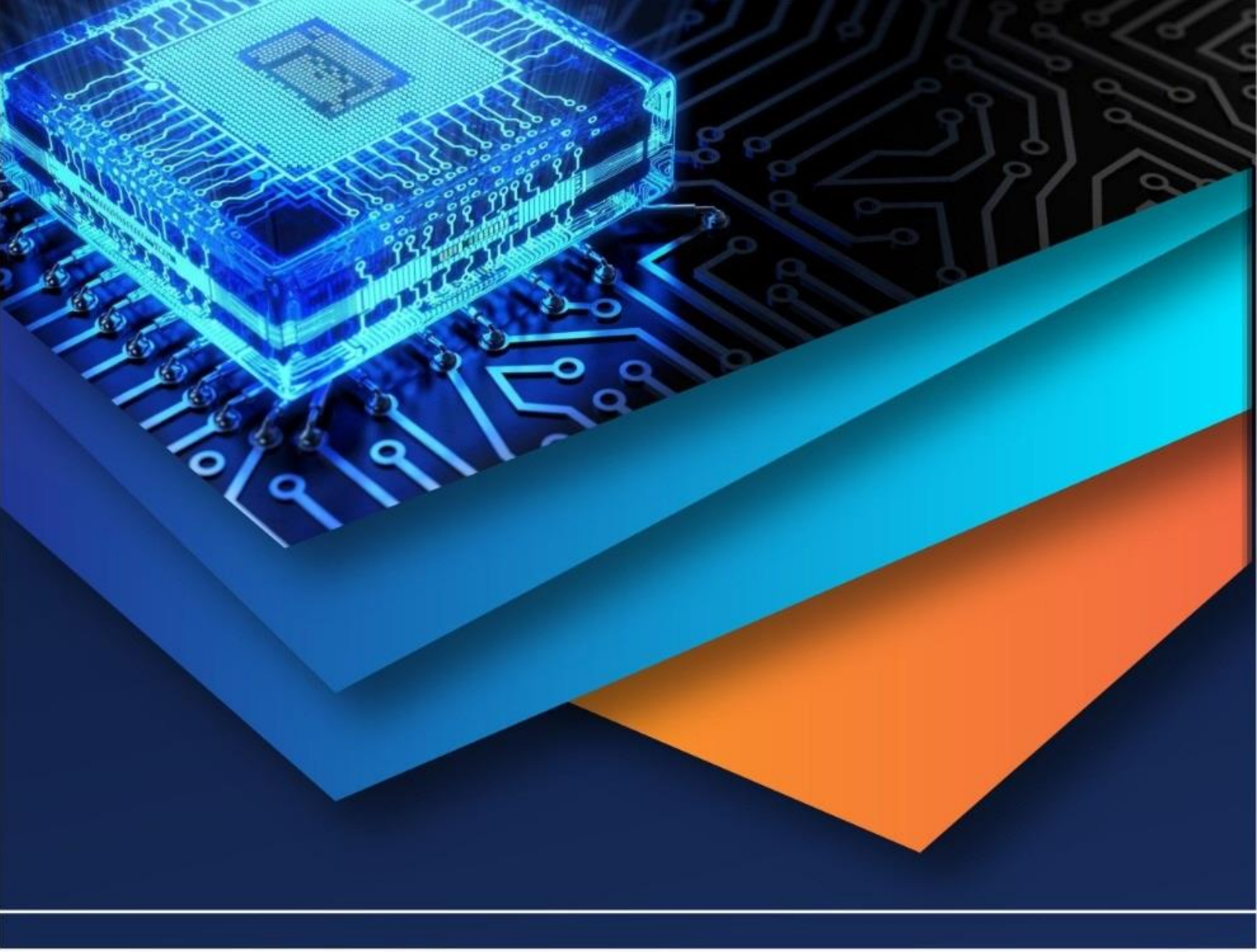
2) *Pertinence and Importance:* Finding and keeping features that are important for the learning job at hand is the key to feature selection. By doing this, the model is guaranteed to concentrate on factors that actually improve prediction accuracy.
3) *Boosting Performance:* Enhancing the ML model's overall performance is the aim of focusing on a more refined set of features. By doing this, overfitting is avoided, in which case the model may identify noise in the data instead of actual patterns.
4) *Interpretability:* A more interpretable model benefits from having a more condensed set of features. Knowing how certain attributes affect things helps to develop insights into how decisions are made, which is important for the validation and credibility of the model.

V. CONCLUSION

This research shows that machine learning (ML) algorithm solutions for malware identification have garnered increasing attention from academia in recent times. Overcoming modern anti-analysis methods, like encryption, is the first priority. The second concern is how accurately malware behaviour is modelled, and this depends on the processes that are chosen to be analysed. The third difficulty is the ageing and unavailability of data sets used in assessments, which affects the relevance and repeatability of findings. These factors become essential while creating an environment for malware analysis. We suggest a number of principles for creating suitable benchmarks for malware analysis using machine learning in order to overcome these problems. We also highlight important new phenomena that merit further research, like malware attribution and triage. Furthermore, we acknowledge the current trade-offs between analysis accuracy, time, and cost and present the novel idea of malware analysis economics. We propose several guidelines for preparing appropriate benchmarks for malware analysis through machine learning.

REFERENCES

[1] Rushiil Deshmukh, Angelo Vergara, Debtanu Bandyopadhyay, Kevin Huang, et al. “Malware Classification using Machine Learning and Deep Learning.”
[2] Oladimeji Kazeem, “Fraud Detection using Machine Learning.” DOI:10.13140/RG.2.2.12616.29441
[3] Pritam Ahire, Mohanki Shreya, Shreya Shinde, Preeti Pisal, and Manasi Manikumar “A Survey on Malware Detection Using ML”. ISSN: 2321-9653 https://doi.org/10.22214/ijraset.2022.39813
[4] Akhtar, M.S.; Feng, T. “Malware Analysis and Detection Using Machine Learning Algorithms.” Symmetry **2022**, 14, 2304. https://doi.org/10.3390/sym14112304
[5] Rani, S. & S R, Reeja. (2020). A Survey on Different Approaches for Malware Detection Using Machine Learning Techniques. 10.1007/978-3-030-34515-0_42.
[6] Mohammed Altay, “Malware Detection Using Deep Learning Algorithms." AURUM Journal of Engineering Systems and Architecture, 2023, № 1, p. 11-26; https://doi.org/10.53600/ajesa.1321170
[7] M. Aljabri et al., "Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions," in IEEE Access, vol. 10, pp. 121395-121417, 2022, doi: 10.1109/ACCESS.2022.3222307.
[8] Catak, Ferhat Ozgur & Şahinbaş, Kevser & Dortkardes, Volkan. (2020). “Malicious URL Detection Using Machine Learning.” 10.4018/978-1-7998-5101-1.ch008.
[9] Ahn, G.; Kim, K.; Park, W.; Shin, D. “ Malicious File Detection Method Using Machine Learning and Interworking with MITRE ATT&CK Framework.” Appl. Sci. **2022**, 12, 10761. https://doi.org/10.3390/app122110761
[10] Baset, Mohamad. (2016).” MACHINE LEARNING FOR MALWARE DETECTION.” 10.13140/RG.2.2.18107.00801.
[11] Nikam, U.V.; Deshmuh, V.M. “Performance evaluation of machine learning classifiers in malware detection.” In Proceedings of the 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, 23–24 April 2022; pp. 1–5.
[12] Akhtar, M.S.; Feng, T.” IOTA-based anomaly detection machine learning in mobile sensing.” EAI Endorsed Trans. Create. Tech. 2022, 9, 172814.
[13] P. Singh, S. Kaur, S. Sharma, G. Sharma, S. Vashisht, and V. Kumar, "Malware Detection Using Machine Learning," 2021 International Conference on Technological Advancements and Innovations (ICTAI), Tashkent, Uzbekistan, 2021, pp. 11–14, doi: 10.1109/ICTAI53825.2021.9673465.
[14] Gavriluț, Dragoș, Cimpoesu, Mihai, Anton, D., and Ciortuz, Liviu. (2009). “Malware detection using machine learning.” 4. 735–741. 10.1109/IMCSIT.2009.5352759
[15] Akshit Kamboj, Priyanshu Kumar, Amit Kumar Bairwa, Sandeep Joshi, ”Detection of malware in downloaded files using various machine learning models.” Egyptian Informatics Journal, Volume 24, Issue 1, 2023, Pages 81-94,ISSN 1110-8665, https://doi.org/10.1016/j.eij.2022.12.002.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY