Data and Metadata. 2023; 2:139
doi: 10.56294/dm2023139

ORIGINAL

Achieving Organizational Effectiveness through Machine Learning Based Approaches for Malware Analysis and Detection

Lograr la eficacia organizativa mediante enfoques basados en el aprendizaje automático para el análisis y la detección de malware

Md Alimul Haque[1], Sultan Ahmad[2,3], Deepa Sonal[4], Hikmat A. M. Abdeljaber[5], B.K.Mishra[6], A.E.M. Eljialy[7], Sultan Alanazi[2], Jabeen Nazeer[2]

[1]Department of Computer Science, Veer Kunwar Singh University. Ara, 802301, India.
[2]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University. Alkharj, 11942, Saudi Arabia.
[3]University Center for Research and Development (UCRD), Department of Computer Science and Engineering, Chandigarh University. Gharuan, Mohali 140413, Punjab, India.
[4]Department of Computer Science, Patna Women's College. Patna, India.
[5]Department of Computer Science, Faculty of Information Technology, Applied Science Private University. Amman, Jordan.
[6]Department of Physics, Veer Kunwar Singh University. Ara, 802301, India.
[7]Department of Information Systems, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, P.O.Box. 151. Alkharj, 11942, Saudi Arabia.

Cite as: Md Alimul Haque, Sultan Ahmad, Deepa Sonal, Hikmat A. M. Abdeljaber, B.K.Mishra, A.E.M. Eljialy, Sultan Alanazi, Jabeen Nazeer. Achieving Organizational Effectiveness through Machine Learning Based Approaches for Malware Analysis and Detection. Data and Metadata. 2023;2:139. https://doi.org/10.56294/dm2023139

Submitted: 11-08-2023 **Revised:** 31-10-2023 **Accepted:** 28-12-2023 **Published:** 29-12-2023

Editor: Prof. Dr. Javier González Argote

Corresponding author: Sultan Ahmad

ABSTRACT

Introduction: as technology usage grows at an exponential rate, cybersecurity has become a primary concern. Cyber threats have become increasingly advanced and specific, posing a severe risk to individuals, businesses, and even governments. The growing complexity and sophistication of cyber-attacks are posing serious challenges to traditional cybersecurity methods. As a result, machine learning (ML) techniques have emerged as a promising solution for detecting and preventing these attacks.
Aim: this research paper offers an extensive examination of diverse machine learning algorithms that have the potential to enhance the intelligence and overall functionality of applications.
Methods: the main focus of this study is to present the core principles of distinct machine learning methods and demonstrate their versatile applications in various practical fields such as cybersecurity systems, smart cities, healthcare, e-commerce, and agriculture. By exploring these applications, this paper contributes to the understanding of how machine learning techniques can be effectively employed across different domains. The article then explores the current and future prospects of ML in cybersecurity.
Results: this paper highlights the growing importance of ML in cybersecurity and the increasing demand for skilled professionals who can develop and implement ML-based solutions.
Conclusion: overall, the present article presents a thorough examination of the role of machine learning (ML) in cybersecurity, as well as its current and future prospects. It can be a valuable source of information for researchers, who seek to grasp the potential of ML in enhancing cybersecurity.

Keywords: Cybersecurity Dataset; Data Analysis; Cybercrime; ML Algorithms; Threats; Security.

RESUMEN

Introducción: a medida que el uso de la tecnología crece a un ritmo exponencial, la ciberseguridad se ha

convertido en una preocupación primordial. Las ciberamenazas son cada vez más avanzadas y específicas, lo que supone un grave riesgo para particulares, empresas e incluso gobiernos. La creciente complejidad y sofisticación de los ciberataques está planteando serios retos a los métodos tradicionales de ciberseguridad. Como resultado, las técnicas de aprendizaje automático (ML) han surgido como una solución prometedora para detectar y prevenir estos ataques.
Objetivo: este trabajo de investigación ofrece un amplio examen de diversos algoritmos de aprendizaje automático que tienen el potencial de mejorar la inteligencia y la funcionalidad general de las aplicaciones.
Métodos: el objetivo principal de este estudio es presentar los principios básicos de distintos métodos de aprendizaje automático y demostrar sus aplicaciones versátiles en diversos campos prácticos, como los sistemas de ciberseguridad, las ciudades inteligentes, la atención sanitaria, el comercio electrónico y la agricultura. Al explorar estas aplicaciones, este artículo contribuye a la comprensión de cómo pueden emplearse eficazmente las técnicas de aprendizaje automático en distintos ámbitos. A continuación, el artículo explora las perspectivas actuales y futuras del ML en ciberseguridad.
Resultados: este artículo destaca la creciente importancia del ML en la ciberseguridad y la creciente demanda de profesionales cualificados que puedan desarrollar e implementar soluciones basadas en ML.
Conclusiones: en general, el presente artículo presenta un examen exhaustivo del papel del aprendizaje automático (ML) en la ciberseguridad, así como sus perspectivas actuales y futuras. Puede constituir una valiosa fuente de información para los investigadores que deseen aprovechar el potencial del aprendizaje automático para mejorar la ciberseguridad.
Palabras clave: Conjunto de Datos de Ciberseguridad; Análisis de Datos; Ciberdelincuencia; Algoritmos ML; Amenazas; Seguridad.

INTRODUCCIÓN

In today's technological landscape, cyberattacks stand out as the foremost concern. These attacks encompass the exploitation of vulnerabilities within systems for malicious intents, ranging from theft and unauthorized alterations to complete destruction. One prevalent form of cyberattack is the deployment of malware, which exemplifies the malicious intent behind these actions.(1) Malware refers to a collection of programs or instructions intentionally created to cause harm to computer systems, users, businesses, or individual computers.(23)

This research project focuses on demonstrating the practicality of detecting malicious network traffic within computer systems, effectively bolstering the security of computer networks. Achieving this objective involves utilizing machine learning algorithms and leveraging the results of malware analysis to calculate differential correlation symmetry integrals. The role of malware detection modules involves analyzing the data they have gathered and been trained on to assess whether a particular software or network connection poses a security risk.(4) To illustrate, let's take a machine learning system that possesses the capability to articulate the fundamental principles underlying the patterns it has observed.(56) By leveraging feedback on their performance in past tasks, algorithms trained by machine learning systems can enhance their predictive capacity, incorporating this valuable information to make necessary adjustments.

On a global scale, the endeavors of cyber malefactors pose a formidable challenge to enterprises, academic institutions, governmental bodies, and individuals alike.(7) Their tactics encompass the dissemination of malicious software and the unlawful acquisition of confidential data, casting a shadow of vulnerability and apprehension.(8) On a daily basis, numerous malicious actors utilize malicious software with the intention of breaching networks, pilfering data, or carrying out illicit money transfers. Consequently, safeguarding sensitive information has emerged as an immediate priority within the scientific community.(9) The objective of this study was to present an all-encompassing framework that utilizes data mining and machine learning classification techniques to identify malicious programs and fortify the security of private information against hackers.

In recent times, the prevalence and complexity of modern malware have risen in the recent times, the prevalence and complexity of modern malware have risen significantly, presenting a substantial risk to the security of contemporary websites. Malware, a form of software specifically designed to inflict damage upon computers or networks, plays a prominent role in these malicious activities.(10) The surge in malware attacks is on a consistent rise, transcending the conventional realms of impact. Beyond the customary targets, this menace now extends its reach to encompass diverse domains, from IoT devices and medical apparatus to industrial control systems and environmental infrastructure.(119) The contemporary landscape of spyware compounds the predicament, demonstrating a formidable capacity to elude detection through its continuous code and behavioral adaptation. Amid this widespread proliferation of malware, traditional defense mechanisms reliant on signature-based methods find themselves ill-equipped to contend with the evolving threat landscape. Hence, it is crucial to embrace a more comprehensive array of defensive measures.(12)

In today's cyber landscape, cyberattacks have ushered in an era of cyberwarfare, characterized by boundless and relentless cyber espionage as shown in table1. Within this context, one of the most potent defenses against cyberwarfare revolves around the utilization of advanced modern malware. Based on the Kaspersky Security

Bulletin (KSB), it was reported that in 2022, cybercriminals launched around 400 000 new malicious files daily, highlighting the severity of the threat landscape. Notably, Kaspersky's security researchers observed a significant rise of 181 % in the daily share of encountered ransomware compared to 2021, reaching a staggering 9 500 encrypted files per day. These statistics underscore the escalating challenges posed by cyber threats in the digital landscape. As per Symantec's findings, over 50 % of newly identified malware are found to be variations or variants of existing malware.(13) This insight highlights the trend of attackers modifying and repurposing existing malware to create new threats, potentially making them harder to detect using conventional methods. In 2017, Russia conducted a cyberattack on Ukrainian electricity infrastructure as part of its ongoing efforts to disrupt its neighboring countries. This significant attack marked the first demonstration of Russia's capability for conducting large-scale cyber warfare.

Within the domain of cyber onslaughts, malefactors perpetually strive to exploit the interplay of acknowledged and concealed vulnerabilities, skillfully leveraging these openings to orchestrate potent infections. As the number of connected devices proliferates rapidly, it becomes apparent that these devices can become prime targets for large-scale malware attacks. Addressing this significant challenge is of utmost importance in maintaining cybersecurity and protecting the integrity of these interconnected objects. Creating robust defense mechanisms to counter cyberattacks, both originating from known and unknown malware, becomes a crucial and pressing necessity. The development of resilient and adaptable strategies is paramount in safeguarding against the constantly evolving threats posed by malicious software in the digital landscape.

The vast number of malware samples and families presents a significant challenge in swiftly and automatically responding to modern malware incidents in real-time. To effectively address this issue, the integration of advanced artificial intelligence (AI) techniques becomes imperative. Machine learning (ML), a proven and valuable tool in various domains, emerges as a viable solution for tackling these complexities.

Literature Survey

With the widespread use of computers, smartphones, and various Internet-connected devices, the global landscape becomes increasingly susceptible to cyberattacks. To counter the surge in malware incidents, numerous approaches for malware detection have emerged. Researchers employ a range of big data tools and machine learning techniques to discern and analyze malicious code, striving to enhance their ability to identify and combat cyber threats. Conventional machine learning techniques utilized for malware detection often require substantial processing time, yet they prove to be effective in detecting emerging malware strains. Yet, in the wake of the pervasive presence of cutting-edge machine learning algorithms such as deep learning, conventional techniques of feature engineering might confront obsolescence. These modern algorithms offer promising avenues for more efficient and accurate malware detection, potentially revolutionizing the field of cybersecurity. Within this investigation, an array of techniques for detecting and classifying malware were explored. Scholars have devised methodologies leveraging machine learning and deep learning to scrutinize samples for indications of malicious intent.(19) Through the utilization of these innovative approaches, the field of cybersecurity strives to enhance its ability to identify and mitigate the risks posed by malicious software.

Baset(14) addresses the need for more effective malware detection methods in the face of evolving cyber threats. The study focuses on the utilization of machine learning algorithms to detect and classify malware samples accurately. Baset(14) examines various machine learning techniques, including decision trees, support vector machines, neural networks, and ensemble methods, to identify the most effective approach for malware detection. One limitation of the Baset(14) research is its focus on a specific time frame, as the field of malware detection continues to evolve rapidly. Including more recent studies and considering emerging techniques could enhance the dissertation's relevance and applicability to current cybersecurity challenges. Overall, Baset's "Machine Learning for Malware Detection" makes a significant contribution to the field of cybersecurity by exploring the potential of machine learning algorithms in detecting and classifying malware. The research findings and insights offer valuable knowledge for researchers and practitioners seeking to improve the efficacy of malware detection systems.

Chowdhury et al.(15), focus on addressing the increasing threat of malware by leveraging data mining and machine learning algorithms. They present a comprehensive framework for analyzing malware samples and detecting their presence using classification techniques. The study explores the effectiveness of different machine learning algorithms and features in malware detection. The authors demonstrate a clear understanding of the subject matter and the significance of integrating these technologies for effective malware detection. However, one limitation of the paper is the lack of discussion on the scalability and applicability of the proposed methodology to real-world scenarios. The authors could have addressed the challenges and limitations of implementing the proposed framework in large-scale environments with diverse types of malware. Author approach to malware detection and the evaluation of various techniques contribute to the advancement of malware analysis methodologies.

A study conducted a comprehensive evaluation and assessment of diverse models to determine their accuracy. The functionality of any digital platform application heavily relies on the availability of data, as highlighted by the author. Given the multitude of cyber risks, it becomes imperative to adopt precautionary measures to ensure the protection of data.

While developing any model, feature selection can be a challenging task. However, machine learning stands out as an advanced technique that enables accurate prediction. To address the complexity of non-standard data, an adaptable workaround is required. The analysis of malware and the establishment of new rules and patterns, play a crucial role in effectively managing and preventing future attacks.[16]

IT security professionals employ malware analysis tools to identify patterns and analyze malicious software. The presence of advanced technologies for examining malware samples and assessing their level of malignancy is highly advantageous for the cybersecurity industry. These tools play a crucial role in monitoring security alerts and proactively preventing malware attacks. When confronted with dangerous malware, it is imperative to eradicate it promptly to halt the spread of infection.

The popularity of malware analysis is on the rise due to its effectiveness in mitigating the impact of the escalating malware threats and the evolving tactics employed by attackers. Businesses are recognizing the importance of this practice in combating the growing complexity of malware-driven attacks.[17]

Table 1. Malware Dataset file types

	File Type	No. of Files
Malware	Backdoor	3555
	Rootkit	2735
	Trojan	2464
	Work	822
	Exploit	553
	Other	3239
	Cleanware	2612
	Total	15980

In Chowdhury et al.[15] research, a practical approach for detecting malware using machine learning classification was proposed. The study aimed to investigate the impact of parameter adjustments on the accuracy of malware classification. Our approach incorporated N-gram and API call capabilities, which were found to be effective and reliable in our experimental evaluation. To further improve detection precision and reduce false positives, future work will involve integrating a larger number of features. Performance results of competing approaches, as presented in table 1, clearly demonstrate the superiority of our Chowdhury et al.[15] approach.

Table2. Comparisons between various classifiers output

Methods	Accuracy(%)	TPR(%)	FPR(%)
KNN	96,03	97,18	4,52
CNN	98,66	99,33	4,98
Naïve Byes	90,72	90,11	14
Random Forest	93,02	96,8	7,5
Support Vector Machine	97,42	98,08	5,64
Decision Tree	99,01	99,06	2,03

In contemporary times, researchers within the realm of malware detection have redirected their focus towards delving into innovative strategies involving machine learning algorithms. Within this research paper, a novel defensive mechanism is unveiled, tailored to evaluate three distinct machine learning algorithm methodologies for the purpose of malware detection. By employing rigorous statistical analysis, the outcome of this investigation showcases that the decision tree approach emerges as the most adept, boasting an exceptional detection accuracy rate of 99,01 % and the notable distinction of possessing the lowest false positive rate (FPR) at a mere 0,023 %, all validated on a constrained dataset.

The authors start by discussing the increasing prevalence and complexity of malware threats in the digital landscape. They highlight the need for robust malware analysis techniques to mitigate the risks associated with these malicious programs. The paper then delves into the details of machine learning and deep learning algorithms, explaining their potential in detecting and analyzing malware samples. The paper highlights the significance of leveraging machine learning and deep learning techniques for effective malware analysis. It sheds light on the advancements in this field and provides valuable insights into the performance and capabilities of these computational approaches. The findings presented in this research contribute to the ongoing efforts to enhance cybersecurity measures and combat malware threats.

Among the various classifiers we evaluated, the DT machine learning method stood out with an impressive accuracy of 99,01 %, as shown in table 2. This outcome underscores its remarkable performance and positions it as the most successful classifier in our study. The conducted experiment illuminated the untapped potential within static analysis, harnessing the wealth of PE information and meticulously chosen key data features. This strategic approach not only resulted in an elevated level of detection accuracy but also unveiled a remarkably precise portrayal of malware intricacies.

With the evolution of the Internet, the prevalence and complexity of malicious programs, commonly known as malware, have surged. The widespread connectivity offered by the Internet has granted malware authors access to a plethora of tools for generating diverse forms of malware. As each day passes, malware continues to expand its reach and enhance its sophistication. The objective of this study was to analyze and evaluate the performance of classifiers, aiming to gain deeper insights into the functioning of machine learning in combating malware.

Through experimental analysis, it was confirmed that the random forest approach outperformed other methods in accurately categorizing data, achieving an impressive accuracy rate of 99,4 percent. These findings emphasized the compatibility of the PE library with static analysis techniques and highlighted the effectiveness of focusing on selected properties to enhance malware detection and characterization. One significant advantage of this approach is the reduced likelihood of unintentionally installing malicious software, as users are empowered to verify the validity of a file before opening it.

METHODS

Research Problem

The potential harm caused by malware is a significant research problem in the field of cybersecurity. Malware, which refers to malicious software designed to disrupt, damage, or gain unauthorized access to computer systems, poses a multitude of threats to individuals, organizations, and society as a whole. Malware continues to evolve and adapt to new security measures, making it challenging for traditional detection and prevention methods to keep up. The potential impact of malware on different sectors, including individuals, businesses, governments, and critical infrastructure. The consequences of malware attacks can range from financial losses and data breaches to operational disruptions and even threats to public safety.

Author aims to explore and develop effective strategies for malware detection, classification, and mitigation. This involves studying the characteristics, behavior, and propagation mechanisms of different types of malware, such as viruses, worms, Trojans, ransomware, and spyware. Understanding the underlying principles and patterns of malware can aid in the development of robust defense mechanisms. Furthermore, as the digital landscape expands and new technologies emerge, researchers need to anticipate future challenges associated with malware. This includes examining the potential risks posed by emerging technologies such as Internet of Things (IoT), cloud computing, and artificial intelligence, which may create new attack vectors and vulnerabilities. The research problem of malware's potential harm encompasses the need to understand, detect, and mitigate the ever-evolving threats posed by malicious software. By addressing this problem, researchers can contribute to the development of robust cybersecurity solutions and safeguard digital ecosystems from the detrimental effects of malware.

It is crucial to develop solutions that can effectively detect previously unknown malware while significantly reducing the number of required characteristics for detection. By achieving this, we can enhance the efficiency and accuracy of malware detection processes, enabling proactive identification and mitigation of emerging threats. The ability to identify new and unseen malware variants is crucial in combating the constantly evolving landscape of malicious software. Moreover, streamlining the necessary characteristics for detection can optimize resource utilization and improve the overall performance of malware detection systems. By striking a balance between novelty detection and feature reduction, we can strengthen our defenses against emerging malware threats and enhance the resilience of our cybersecurity measures.[18]

Hypothesis1

The accuracy levels of three machine learning (ML) methods, namely decision tree (DT), convolutional neural network (CNN), and support vector machine (SVM), for detecting malware.

Research Methodology

In this article, a comprehensive overview is presented of the different stages and elements involved in a typical workflow for detecting and classifying malware using machine learning. The paper delves into the inherent challenges and constraints associated with such workflows and evaluates the latest advancements and emerging trends in the field, focusing specifically on deep learning techniques. For a comprehensive grasp of the machine learning technique proposed for malware detection, a visual representation of the entire workflow process is presented in figures 3 and 4. These figures elucidate the step-by-step progression, providing a more holistic understanding of the methodology from its initiation to its completion.

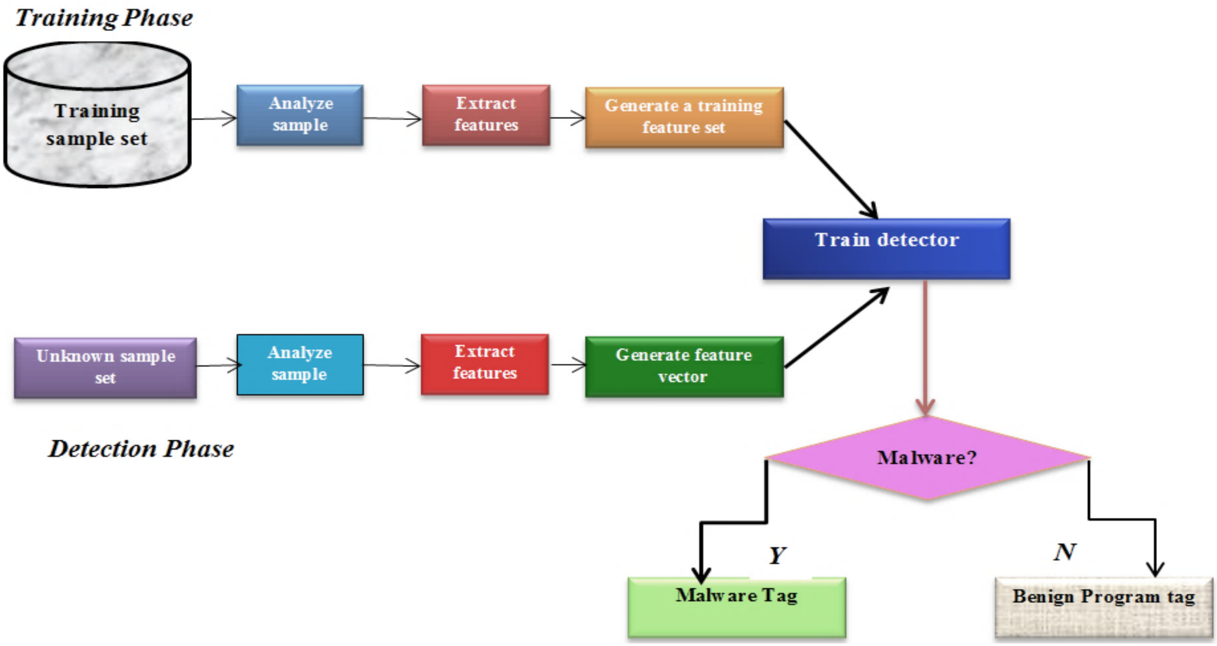

Figure 1. Proposed ML malware detection system

Proposed Work and Malware Dataset

Annually, numerous companies and organizations become targets of malware attacks, facing dire consequences such as financial losses. The traditional antivirus scanners prove inadequate to address the evolving cyber threats, leaving millions of hosts vulnerable to attacks. Extensive research has been conducted on malware detection systems, which hold a vital and indispensable role in the realm of cybersecurity. Our study aims to present a methodical approach for classifying contemporary malware using Machine Learning algorithms on the newly curated CICAndMal2017 dataset, generously provided by the Canadian Institute for Cybersecurity (CIC).(19)The effectiveness of our detection methods was thoroughly assessed, considering various evaluation measures. Our research utilized the CICAndMal2017 dataset, an advanced malware dataset, to train robust malware detection classifiers.(12) The log features retrieved from the samples offer a diverse set of training possibilities for various models. A total of approximately 62 distinct malware families were identified within the dataset, comprising more than 15 980 data points from various locations. The dataset itself contained 281 columns and 15 980 rows, providing a rich and extensive resource for analysis. Within our dataset, each category of malware is represented by a collection of distinct malware families, each consisting of various data instances. For a more comprehensive view of the dataset, as refer to table 2, which provides specific details about its contents.

Data Analysis and Exploration

The process of data analysis and exploration holds utmost significance in our study. It involves reorganizing the dataset and comprehending the diverse variables to establish an effective modeling strategy. These crucial steps lay the foundation for the success of our research.

Data Preprocessing

AI algorithms derive their knowledge from the data they are fed. Hence, if the dataset possesses inadequate quality, inaccuracies, or lacks completeness, the ensuing algorithm's classification performance is bound to falter. It is vital to ensure that the data used for training AI models is accurate and comprehensive, as the algorithm's performance is reliant on the information it learns from the provided data. To ensure effective performance, it is essential to thoroughly prepare the data before feeding it into the machine. This process involves data preprocessing, which includes cleaning, filtering, and normalization. Our approach predominantly relies on machine learning techniques, where data preprocessing plays a crucial role in feature transformation, normalization, extraction, and encoding to enhance the overall model performance. The data utilized in this study was stored in the file system as binary code, represented by unprocessed executables. Prior to conducting the research, necessary preparations were made to handle the data appropriately. Unpacking the executables necessitated a secure environment, typically facilitated through a virtual machine (VM). To automate the process of unpacking compressed executables, PEiD software was employed.(20)

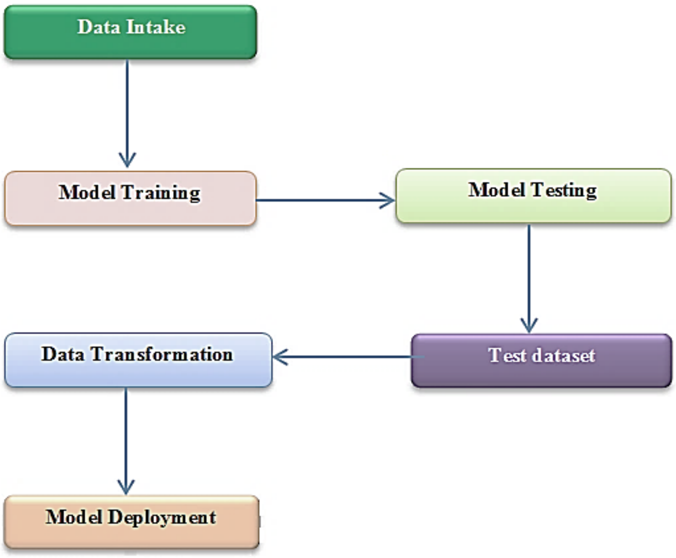

Figure 2. Research framework process

Features Extraction

In the realm of addressing binary and multi-class classification tasks, our approach entailed the utilization of embedded methods to skillfully train the raw data through the implementation of a machine learning algorithm. This choice was made due to its exceptional performance and accuracy in determining the relevance of attributes within the dataset. Datasets in the modern era often consist of a vast number of features, reaching tens of thousands. However, as the number of features has increased over time, it has become evident that the resulting machine learning models tend to suffer from overfitting issues.(21) To tackle this challenge, we adopted a strategy of constructing a reduced set of features from the original larger set. This approach is commonly employed to retain a comparable level of accuracy while utilizing a smaller number of features.(21) The primary objective of this research was to enhance the existing dataset of dynamic and static features by selecting and retaining the most informative ones while discarding those that did not contribute significantly to data analysis.(22) The aim was to streamline the dataset and focus on the key features that would provide meaningful insights and improve the overall effectiveness of the analysis process.

Features Selection

Once the feature extraction phase was completed, which entailed identifying additional features, the subsequent step involved feature selection. This iterative procedure assumed a pivotal role in amplifying precision, refining the model's structure, and curtailing the risks of overfitting. Through meticulous curation, the most pertinent attributes were gleaned from a wealth of recently identified characteristics, thereby contributing to the overall optimization process. The objective was to optimize the model's performance by focusing on the most informative and discriminative features. In previous research, numerous feature classification strategies have been employed by researchers to detect malicious code within software. Amidst these various approaches, the efficacy of the feature rank technique shines brightly as a method that excels in pinpointing the most pertinent attributes for constructing malware detection models of remarkable accuracy. Consequently, in this study, the feature rank technique was extensively utilized, drawing on its effectiveness and reliability in identifying the crucial features associated with malware. The objective was to optimize the model's performance by leveraging the power of feature selection.(23)

RESULTS AND DISCUSSION

The classification process consisted of two primary phases: training and testing. In the training phase, the system was provided with a diverse set of files, including both harmful and safe ones, to develop its understanding of different patterns and characteristics. Automated classifiers were then trained using sophisticated learning algorithms, enabling them to acquire the necessary knowledge and discern patterns in the data. This training process was instrumental in equipping the classifiers with the ability to differentiate between harmful and safe files accurately.[24] Through the process of annotating each set of data, every classifier (including KNN, CNN, NB, RF, SVM, and DT) progressively improved its ability to make accurate predictions. In the subsequent testing phase, a collection of new files, comprising both potentially harmful and benign ones, was provided to the classifier.[25,26,27] Leveraging its learning and knowledge gained during the training phase, the classifier then evaluated and determined the malicious or clean nature of each file, thereby demonstrating its capability to effectively identify and classify malware.

Confusion Matrix

The outcome of our research is depicted in the confusion matrix, figure 5 visually represents the performance metrics of the different machine learning classifiers, including DT, KNN, CNN, NB, RF, and SVM. Notably, DT exhibited the highest accuracy rate at 99,01 % and achieved a commendable True Positive Rate (TPR) of 99,06 %. Additionally, DT displayed the lowest False Positive Rate (FPR) with an accuracy of 2,03 %. A comprehensive analysis of the confusion matrix further confirms that DT outperformed all other classifiers,[28] including KNN, CNN, NB, RF, and SVM, in terms of accuracy and overall effectiveness.[29]

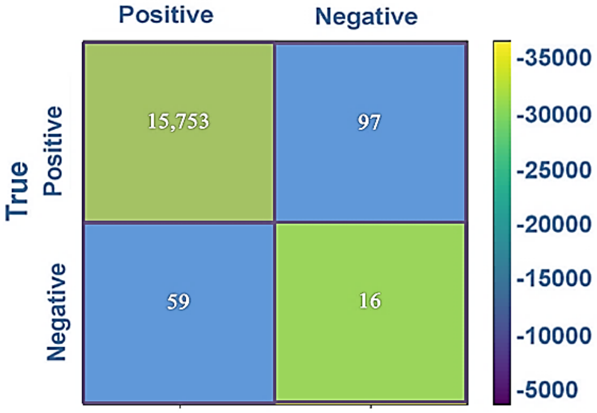

Figure 3. Confusion Matrix with binary classification

The proposed approach for classifying and detecting malware underwent thorough experimental evaluation using a comprehensive dataset consisting of both malware and clean ware samples. To analyze and characterize the malware, we employed a range of supervised machine learning algorithms or classifiers, including KNN, CNN, NB, RF, SVM, and DT. These classifiers were instrumental in accurately categorizing and differentiating malicious software from benign files, contributing to the overall effectiveness of our suggested method.[30]

By conducting a statistical analysis of the outcomes presented in table 3, we arrived at the conclusion that the accuracy rates of the classifiers, namely KNN (96,03 %), CNN (98,66 %), Naïve Bayes (90,72 %), Random Forest (93,02 %), SVM (97,42 %), and DT (99,01 %), clearly indicate that DT is the most suitable model for our malware detection strategy. These results highlight the superior performance of the DT classifier in accurately identifying and classifying malware instances. Upon examining the true positive rates (TPRs) of the classifiers, expressed as percentages, namely KNN (97,18 %), CNN (99,33 %), Naïve Bayes (90,11 %), Random Forest (96,8 %), SVM (98,08 %), and DT (99,06 %), it is evident that CNN emerges as the second-best model for the effective detection and identification of malware. Additionally, SVM exhibits promising performance as the third optimal model for detecting malware instances. Table 3 provides an overview of the false positive rates (FPRs) expressed as percentages for each classifier: KNN (4,52 %), CNN (4,98 %), Naïve Bayes (14 %), Random Forest (7,5 %), SVM (5,64 %), and DT (2,03 %). Based on these results, it can be inferred that CNN, SVM, DT, and KNN classifiers exhibit similar high levels of accuracy and performance across various scenarios and objectives. The comparison of the three most optimal algorithms (DT = 99,01 %, SVM = 97,42 %, and CNN = 98,66 %) clearly indicates their superior performance in terms of both accuracy and true positive rate (TPR). Among these algorithms, DT demonstrates the highest accuracy, making it the preferred choice for effective malware detection and identification.

CONCLUSIONS

In recent times, there has been a noticeable increase in the academic community's attention towards utilizing ML algorithms for malware identification. This research article adds to the expanding sphere of interest by presenting a defensive mechanism that systematically assesses three distinct machine learning algorithm strategies for the purpose of malware detection. The selection process is rooted in their respective performance and effectiveness, culminating in the identification of the most apt approach. The findings indicate that DT (99,01 %), CNN (98,66 %), and SVM (97,42 %) achieved commendable performance in terms of detecting accuracy when compared to alternative classifiers. The effectiveness of DT, CNN, and SVM algorithms in identifying malware was also assessed based on their performance in maintaining a low FPR (DT = 2,03 %, CNN = 4,98 %, and SVM = 5,64 %) within a specific dataset. Within the context of this research endeavor, an empirical investigation was undertaken to evaluate and gauge the detection accuracy of a machine learning (ML) classifier. The approach entailed the utilization of static analysis, employed to extract features from PE data and subsequently ascertain the classifier's performance. This evaluation involved comparing the performance of this ML classifier with two other ML classifiers. Through our diligent efforts, machine learning algorithms have reached a significant milestone in distinguishing between harmful and benign data. Among all the classifiers examined, the DT machine learning method demonstrated the highest accuracy, achieving an impressive 99 % success rate. These paragraphs underline the advantages associated with employing static analysis in the realm of cybersecurity. Static analysis involves the examination of code without its execution, allowing for the extraction of meaningful features that can aid in detecting and characterizing malware. The utilization of PE (Portable Executable) information, which pertains to the file format used for executable programs in Windows operating systems, adds an extra layer of specificity. Additionally, the careful selection of data ensures that the analysis focuses on the most relevant attributes, contributing to the accuracy of the results obtained. The experimental findings indicate that this approach holds promise in terms of both detecting malware more accurately and gaining a deeper understanding of its intricacies. The dataset, sourced from the esteemed Canadian Institute for Cybersecurity, formed the bedrock for conducting various essential phases of this study. It served as the cornerstone for training, rigorous testing, and the subsequent comparative analysis of the three distinct machine learning models, namely Decision Trees (DT), Convolutional Neural Networks (CNN), and Support Vector Machines (SVM). The three specific machine learning models mentioned - Decision Trees (DT), Convolutional Neural Networks (CNN), and Support Vector Machines (SVM) - represent distinct algorithmic approaches used in machine learning. These models were subjected to the dataset for the purposes of evaluation and comparison, ultimately forming a significant part of the study's methodology. The paper thus highlights the dataset's source and its integral role in shaping the study's experimental framework, while also providing context about the machine learning models selected for analysis.

REFERENCES

1. Haque MA, Haque S, Kumar K, Singh NK. A Comprehensive Study of Cyber Security Attacks, Classification, and Countermeasures in the Internet of Things. Digital Transformation and Challenges to Data Security and Privacy, IGI Global; 2021, p. 63-90.

2. Nikam U V, Deshmuh VM. Performance evaluation of machine learning classifiers in malware detection. 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), IEEE; 2022, p. 1-5.

3. Haque MA, Ahmad S, John A, Mishra K, Mishra BK, Kumar K, et al. Cybersecurity in Universities: An Evaluation Model. SN Computer Science 2023;4:569. https://doi.org/10.1007/s42979-023-01984-x.

4. Sethi K, Kumar R, Sethi L, Bera P, Patra PK. A novel machine learning based malware detection and classification framework. 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), IEEE; 2019, p. 1-4.

5. Whig V, Othman B, Haque MA, Gehlot A, Qamar S, Singh J. An empirical analysis of artificial intelligence (AI) as a growth engine for the healthcare sector. 2022 2nd international conference on advance computing and innovative technologies in engineering (ICACITE), IEEE; 2022, p. 2454-7.

6. Tao F, Akhtar MS, Jiayuan Z. The future of artificial intelligence in cybersecurity: A comprehensive survey. EAI Endorsed Transactions on Creative Technologies 2021;8:e3-e3.

7. Haque MA, Haque S, Zeba S, Kumar K, Ahmad S, Rahman M, et al. Sustainable and efficient E-learning internet of things system through blockchain technology. E-Learning and Digital Media 2023;0(0):1-20. https://doi.org/10.1177/20427530231156711.

8. Haque MA, Bokhari MU, Sinha AK, Singh NK. Comparative study on Wireless threats and their Classification.

INDIACom-2017; IEEE Conference ID: 40353 2017 4th International Conference on "Computing for Sustainable Global Development", 01st - 03rd March, 2017 BVICAM, 2017, p. 5057-9.

9. Kumar D, Haque A, Mishra K, Islam F, Mishra BK, Ahmad S. Exploring the Transformative Role of Artificial Intelligence and Metaverse in Education: A Comprehensive Review. Metaverse Basic and Applied Research 2023; 2: 55 s. f.

10. Ahmad S, Jha S, Alam A, Alharbi M, Nazeer J. Analysis of Intrusion Detection Approaches for Network Traffic Anomalies with Comparative Analysis on Botnets (2008–2020). Security and Communication Networks 2022;2022.

11. Ahmad S, Afzal MM. A Study and Survey of Security and Privacy issues in Cloud Computing. International Journal of Engineering Research & Technology (IJERT), ISSN s. f.:181-2278.

12. Akhtar MS, Feng T. Malware Analysis and Detection Using Machine Learning Algorithms. Symmetry 2022;14:2304.

13. Citron DK. The fight for privacy: Protecting dignity, identity and love in the digital age. Random House; 2022.

14. Baset M. Machine learning for malware detection 2016.

15. Chowdhury M, Rahman A, Islam R. Malware analysis and detection using data mining and machine learning classification. International conference on applications and techniques in cyber security and intelligence: applications and techniques in cyber security and intelligence, Springer; 2018, p. 266-74.

16. Akhtar MS, Feng T. Deep learning-based framework for the detection of cyberattack using feature engineering. Security and Communication Networks 2021;2021:1-12.

17. Altaher A. Classification of android malware applications using feature selection and classification algorithms. VAWKUM Transactions on Computer Sciences 2016;10:1-5.

18. Sethi K, Chaudhary SK, Tripathy BK, Bera P. A novel malware analysis for malware detection and classification using machine learning algorithms. Proceedings of the 10th International Conference on Security of Information and Networks, 2017, p. 107-13.

19. Canadian Institute for Cybersecurity. s. f.

20. Saad S, Briguglio W, Elmiligi H. The curious case of machine learning in malware detection. arXiv preprint arXiv:190507573 2019.

21. Selamat N, Ali F. Comparison of malware detection techniques using machine learning algorithm. Indones J Electr Eng Comput Sci 2019;16:435.

22. Firdausi I, Erwin A, Nugroho AS. Analysis of machine learning techniques used in behavior-based malware detection. 2010 second international conference on advances in computing, control, and telecommunication technologies, IEEE; 2010, p. 201-3.

23. Hamid F, Joshi S. Enhancing malware detection with static analysis using machine learning. Int J Res Appl Sci Eng Technol 2019;7:38-42.

24. Kumar P, Gupta GP, Tripathi R. A distributed ensemble design based intrusion detection system using fog computing to protect the internet of things networks. Journal of ambient intelligence and humanized Computing 2021;12:9555-72.

25. Inastrilla CRA. Big Data in Health Information Systems. Seminars in Medical Writing and Education 2022;1:6-6. https://doi.org/10.56294/mw20226.

26. Jinchuña-Huallpa J, Fernández-Sosa LE. Normativa de la estructura de control interno que afecta la

calidad de gestión en la etapa de liquidación de obras del Gobierno Regional de Tacna. Sincretismo 2020;1.

27. Contreras JG, Rodríguez AU, Gaviño AS. Comportamiento Organizacional para el Balance Integral Humano desde la NOM-035 en escenario post-pandemia COVID-19. Revista Científica Empresarial Debe-Haber 2023;1:41-57.

28. Canova-Barrios C, Machuca-Contreras F. Interoperability standards in Health Information Systems: systematic review. Seminars in Medical Writing and Education 2022;1:7-7. https://doi.org/10.56294/mw20227.

29. Kumar P, Tripathi R, P. Gupta G. P2IDF: A privacy-preserving based intrusion detection framework for software defined Internet of Things-fog (SDIoT-Fog). Adjunct Proceedings of the 2021 International Conference on Distributed Computing and Networking, 2021, p. 37-42.

30. Kumar P, Gupta GP, Tripathi R. PEFL: Deep privacy-encoding-based federated learning framework for smart agriculture. IEEE Micro 2021;42:33-40.

FUNDING

This study is supported via funding from Prince Sattam Bin Abdulaziz University project number (PSAU/2023/R/1444).

ACKNOWLEDGMENT

We thank the Deanship of Scientific Research, Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia for help and support.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHOR CONTRIBUTIONS

Conceptualization: Md. Alimul Haque; Sultan Ahmad; Deepa Sonal.

Investigation: Md Alimul Haque; Sultan Ahmad.

Methodology: Md. Alimul Haque; Sultan Ahmad; Deepa Sonal; Hikmat A. M. Abdeljaber; B.K.Mishra; A.E.M. Eljialy; Sultan Alanazi; Jabeen Nazeer.

Writing - original draft: Md. Alimul Haque; Sultan Ahmad; Deepa Sonal; Hikmat A. M. Abdeljaber; B.K.Mishra; A.E.M. Eljialy; Sultan Alanazi; Jabeen Nazeer.

Writing - review and editing: Md. Alimul Haque; Sultan Ahmad; Deepa Sonal; Hikmat A. M. Abdeljaber, B.K.Mishra; A.E.M. Eljialy; Sultan Alanazi; Jabeen Nazeer.

All authors contributed to design and development of the system as well as the manuscript. All authors have read and approved the final manuscript.