ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

«ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ
АКАДЕМИЯСЫ» РҚБ

«ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ АКАДЕМИЯСЫ» РҚБ

ХАБАРЛАРЫ

ИЗВЕСТИЯ

РОО «НАЦИОНАЛЬНОЙ
АКАДЕМИИ НАУК РЕСПУБЛИКИ
КАЗАХСТАН»

NEWS

OF THE ACADEMY OF SCIENCES
OF THE REPUBLIC OF
KAZAKHSTAN

PHYSICO-MATHEMATICAL SERIES

4 (352)

OCTOBER – DECEMBER 2024

PUBLISHED SINCE JANUARY 1963
PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

«ҚР ҰҒА Хабарлары. Физика және информатика сериясы».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Меншіктеуші: «Қазақстан Республикасының Ұлттық ғылым академиясы» РҚБ (Алматы қ.). Қазақстан Республикасының Ақпарат және қоғамдық даму министрлігінің Ақпарат комитетінде 14.02.2018 ж. берілген **No 16906-Ж** мерзімдік басылым тіркеуіне қойылу туралы куәлік.

Тақырыптық бағыты: *физика және ақпараттық коммуникациялық технологиялар сериясы.*

Қазіргі уақытта: *«ақпараттық технологиялар»бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*

Мерзімділігі: *жылына 4 рет.*

Тиражы: *300 дана.*

Редакцияның мекен-жайы: *050010, Алматы қ., Шевченко көш., 28, 219 бөл., тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

«Известия НАН РК. Серия физика и информатики».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Собственник: *Республиканское общественное объединение «Национальная академия наук Республики Казахстан» (г. Алматы).*
Свидетельство о постановке на учет периодического печатного издания в Комитете информации Министерства информации и общественного развития Республики Казахстан **No 16906-Ж** выданное 14.02.2018 г.
Тематическая направленность: *серия физика и информационные коммуникационные технологии.*
В настоящее время: *вошел в список журналов, рекомендованных ККСОН МОН РК по направлению «информационные коммуникационные технологии».*
Периодичность: *4 раз в год.*
Тираж: *300 экземпляров.*
Адрес редакции: *050010, г. Алматы, ул. Шевченко, 28, оф. 219, тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

News of the National Academy of Sciences of the Republic of Kazakhstan.
Series of physics and informatics.
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Owner: RPA «National Academy of Sciences of the Republic of Kazakhstan» (Almaty). The certificate of registration of a periodical printed publication in the Committee of information of the Ministry of Information and Social Development of the Republic of Kazakhstan **No. 16906-Ж,** issued 14.02.2018

Thematic scope: *series physics and information technology.*

Currently: *included in the list of journals recommended by the CCSES MES RK in the direction of «information and communication technologies».*

Periodicity: *4 times a year.*

Circulation: *300 copies.*

Editorial address: *28, Shevchenko str., of. 219, Almaty, 050010, tel. 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

NEWS OF THE NATIONAL ACADEMY OF SCIENCESOF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4. Namber 352 (2024). 29–41
https://doi.org/10.32014/2024.2518-1726.305

IRSTI 28.23.25
UDC 004.49

© G. Aksholak*, A. Bedelbayev, R. Magazov, 2024.
Kazakh National University named after Al-Farabi, Almaty, Kazakhstan.
E-mail: *gaksholak@gmail.com*

ANALYSIS AND COMPARISON OF MACHINE LEARNING METHODS FOR MALWARE DETECTION

Aksholak Gulnur – PhD student, Kazakh National University named after Al-Farabi, Almaty, Kazakhstan, E-mail: gaksholak@gmail.com, https://orcid.org/0000-0001-8292-6939;
Bedelbayev Agyn – candidate of sciences in physics and mathematics, associate professor of the Department "Information Systems", Kazakh National University named after Al-Farabi, Almaty, Kazakhstan, agyn08@yandex.ru, https://orcid.org/0000-0001-9839-4156;
Magazov Raiymbek – PhD student, Kazakh National University named after Al-Farabi, Almaty, Kazakhstan, E-mail: Magazovraiko@gmail.com, https://orcid.org/0009-0000-4105-2331.

Abstract. Our study aims to analyze and evaluate modern machine learning methods for detecting malware, a critical challenge given the increasing complexity and volume of cyber threats. Traditional approaches often fail to cope with new types of malware, so the use of machine learning allows you to increase the effectiveness of protection by identifying abnormal behaviors and unknown threats in real time. Machine learning methods open up new opportunities for threat detection by analyzing behavioral signs of files and network activities. In addition, the use of Machine learning methods makes it possible to adapt to new types of threats in real time, which significantly increases the level of security and reduces risks for users and organizations. We explored various algorithms, including Support Vector Machines, Random Forest, Logistic Regression, and Decision Trees, comparing their effectiveness in identifying and classifying malware. Our methodology combines static, dynamic, and memory-based analysis techniques, offering a comprehensive approach to understanding malware behavior.

Key findings reveal that Decision Trees and Random Forests demonstrate impressive accuracy in both binary and multi-class classification tasks. We also highlight novel methods such as the Self-Organizing Incremental Neural Network, which effectively handles evolving malware threats. The integration of static and dynamic analysis methods deepens insights into malware behavior.

This research underscores the importance of advancing machine learning techniques to enhance cybersecurity measures against evolving global malware threats, offering valuable insights for future research directions.

Keywords: malware, classification accuracy, feature extraction, machine learning, security analytics, cyber threats.

© Г.И. Ақшолақ*, А.А. Бедельбаев, Р.С. Мағазов, 2024.
Әл-Фараби атындағы қазақ Ұлттық университеті, Алматы, Қазақстан.
E-mail: *gaksholak@gmail.com*

ЗИЯНДЫ БАҒДАРЛАМАЛАРДЫ АНЫҚТАУҒА АРНАЛҒАН МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН ТАЛДАУ ЖӘНЕ САЛЫСТЫРУ

Ақшолақ Гүлнұр – PhD докторант, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан, E-mail: gaksholak@gmail.com, https://orcid.org/0000-0001-8292-6939;
Беделбаев Ағын – физика-математика ғылымдарының кандидаты, Әл-Фараби атындағы Қазақ ұлттық университетінің «Ақпараттық жүйелер» кафедрасының қауым. профессоры, Алматы, Қазақстан, E-mail: agyn08@yandex.ru, https://orcid.org/0000-0001-9839-4156;
Мағазов Райымбек – PhD докторант, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан, E-mail: Magazovraiko@gmail.com, https://orcid.org/0009-0000-4105-2331.

Аннотация. Біздің зерттеуіміз киберқауіптердің өсіп келе жатқан күрделілігі мен көлемін ескере отырып, маңызды мәселе болып табылатын зиянды бағдарламаларды анықтауға арналған машиналық оқытудың заманауи әдістерін талдауға және бағалауға бағытталған. Дәстүрлі тәсілдер зиянды бағдарламалардың жаңа түрлерін анықтай алмайды, ал машиналық оқытуды қолдану нақты уақыт режимінде қалыптан тыс мінез-құлық пен белгісіз қауіптерді анықтау арқылы қорғаныс тиімділігін арттыруға мүмкіндік береді. Машиналық оқыту әдістері файлдардың мінез-құлық белгілерін және желілік әрекеттерді талдау арқылы қауіптерді анықтаудың жаңа мүмкіндіктерін ашады. Сонымен қатар, Машиналық оқыту әдістерін қолдану нақты уақыт режимінде қауіптің жаңа түрлеріне бейімделуге мүмкіндік береді, бұл қауіпсіздік деңгейін едәуір арттырады және пайдаланушылар мен ұйымдар үшін тәуекелдерді азайтады. Біз әртүрлі алгоритмдерді, соның ішінде тірек векторлық әдістерді, кездейсоқ орманды, логистикалық регрессияны және шешім ағаштарын зерттейміз, олардың зиянды бағдарламаларды анықтау және жіктеудегі тиімділігін салыстырдық. Біздің әдістеме зиянды бағдарлама әрекетін түсінуге кешенді тәсілді ұсыну үшін статикалық, динамикалық және жадқа негізделген талдау әдістерін біріктіреді.

Негізгі нәтижелер шешім ағаштары мен кездейсоқ ормандардың екілік және көп класты жіктеу мәселелерінде әсерлі дәлдік көрсететінін көрсетеді. Біз сондай-ақ дамып келе жатқан зиянды бағдарлама қауіптерімен тиімді күресетін Self-Organizing Incremental Neural Network сияқты жаңа әдістерді атап өтеміз. Статикалық және динамикалық талдау әдістерін біріктіру зиянды бағдарлама әрекетін түсінуді тереңдетеді.

Бұл зерттеу болашақ зерттеу бағыттары үшін құнды түсініктерді ұсына отырып, дамып келе жатқан жаһандық зиянды бағдарламалар қауіптеріне қарсы киберқауіпсіздік шараларын жақсарту үшін машиналық оқыту әдістерін әзірлеудің маңыздылығын көрсетеді.

Түйін сөздер: зиянды бағдарламалар, классификация дәлдігі, ерекшеліктерді алу, машиналық оқыту, қауіпсіздік аналитикасы, киберқауіптер.

© Г.И. Акшолак*, А.А. Бедельбаев, Р.С. Магазов, 2024.
Казахский национальный университет имени аль-Фараби, Алматы, Казахстан.
E-mail: *gaksholak@gmail.com*

АНАЛИЗ И СРАВНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБНАРУЖЕНИЯ ВРЕДОНОСНОГО ПО

Акшолак Гулнур – PhD докторант, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан, E-mail: gaksholak@gmail.com, https://orcid.org/0000-0001-8292-6939;
Бедельбаев Аган – кандидат физико-математических наук, ассоц. профессор кафедры «Информационных систем» Казахского национального университета имени аль-Фараби, E-mail: agyn08@yandex.ru, https://orcid.org/0000-0001-9839-4156;
Магазов Райымбек – PhD докторант, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан, E-mail: Magazovraiko@gmail.com, https://orcid.org/0009-0000-4105-2331.

Аннотация. Наше исследование направлено на анализ и оценку современных методов машинного обучения для обнаружения вредоносных программ, что является критической проблемой, учитывая растущую сложность и объем киберугроз. Традиционные подходы часто не справляются с новыми типами вредоносных программ, поэтому использование машинного обучения позволяет повысить эффективность защиты за счет выявления аномального поведения и неизвестных угроз в режиме реального времени. Методы машинного обучения открывают новые возможности для обнаружения угроз путем анализа поведенческих признаков файлов и сетевой активности. Кроме того, использование методов машинного обучения позволяет адаптироваться к новым типам угроз в режиме реального времени, что значительно повышает уровень безопасности и снижает риски для пользователей и организаций. Мы исследовали различные алгоритмы, включая методы опорных векторов, случайный лес, логистическую регрессию и деревья решений, сравнивая их эффективность при выявлении и классификации вредоносных программ. Наша методология объединяет статические, динамические и основанные на памяти методы анализа, предлагая комплексный подход к пониманию поведения вредоносных программ.

Основные результаты показывают, что деревья решений и случайные леса демонстрируют впечатляющую точность как в бинарных, так и в многоклассовых задачах классификации. Мы также выделяем новые методы, такие как Self-Organizing Incremental Neural Network, которая эффективно справляется с развивающимися угрозами вредоносных программ. Интеграция статических и динамических методов анализа углубляет понимание поведения вредоносных программ.

Это исследование подчеркивает важность развития методов машинного обучения для улучшения мер кибербезопасности против развивающихся глобальных угроз вредоносных программ, предлагая ценную информацию для будущих направлений исследований.

Ключевые слова: вредоносное ПО, точность классификации, извлечение признаков, машинное обучение, аналитика безопасности, киберугрозы.

Introduction. The rapid escalation of cyber threats, particularly malware, poses significant challenges to global cybersecurity. Malware, encompassing a variety of malicious software designed to disrupt, damage, or gain unauthorized access to systems, has become increasingly sophisticated, making traditional detection methods less effective. Malware is not merely software that operates without the consent or knowledge of system administrators, as stated in the literature (Or-Meir, et al, 2019). Instead, it encompasses a broad range of software types, including viruses, worms, trojan, horses, ransomware, spyware, adware, and others, each with the primary intent of inflicting damage (Bedelbayev, et al, 2023). The sophistication and variety of malware necessitate a comprehensive understanding and robust defense mechanisms to protect against these pervasive cyber threats. In this context, machine learning (ML) has emerged as a powerful tool, offering enhanced capabilities to detect both known and novel threats by analyzing vast datasets of benign and malicious files.

Despite the advancements in ML-based malware detection, there remain significant gaps in the research. Most notably, existing studies often rely on static models that struggle to keep pace with the evolving nature of malware. This research aims to address the gap by exploring adaptive machine learning models that can learn in real-time and respond to emerging threats more effectively.

This review article proposes to evaluate and compare various machine learning algorithms, such as Support Vector Machines, Random Forests, and Decision Trees, in their ability to detect and classify malware. By conducting a series of experiments on contemporary datasets, we seek to determine which models offer the highest accuracy and robustness in a dynamically changing threat landscape. By synthesizing the findings from various studies, this review seeks to identify current trends, gaps in the research, and potential directions for future exploration. Ultimately, our goal is to offer insights that can inform the development of more effective and resilient malware detection systems.

Materials and Methods

Malware analysis is indeed a critical step in understanding and detecting malicious software. The process involves examining the characteristics, behavior, and functionality of malware to develop effective countermeasures.

Sihwail et al. in their work presented a comprehensive classification of malware analysis techniques, categorizing them into static, dynamic, hybrid and memory-based analysis (Sihwail, et al, 2018). They also reviewed various research studies that utilized machine learning methods for malware detection, offering insights into the application of these techniques in the field. Figure 1, shows malware analysis techniques and their common features.

Malware analysis	Static	API calls
		CFG
		OPcode
		N-Gram
	Dynamic	Function parametrs
		Funiction call
		Instruction traces
		Instruction flow
	Hybrid	Combines static/dynamic
	Memory	Process/Service
		DLL
		Registry keys
		Network Connections

Figure 1. Malware analysis techniques and their common features

Static Analysis

When a software or piece of code is analyzed without executing, this kind of analysis is called static analysis or code analysis. Static code analysis involves studying the binary file and looking for patterns in its structure that might be indicative of malicious behaviour without ever actually running the binary. Various static features, such as N-grams, opcodes, strings, and PE header information, are extracted for analysis. These features are then utilized in designing malware detection software like antivirus programs and IDSs. The analysis can be performed with or without applying reverse engineering on the malware samples.

Dynamic Analysis

Dynamic analysis is particularly useful for files that have not been adequately disassembled or examined through static analysis.

While traditional malware classification techniques rely on static or dynamic analysis, Zelinka et al. (2023) introduce a fractal geometry-based method, which visualizes malware behavior in a visually distinctive manner, potentially improving classification accuracy through deep learning models (Zelinka, et al, 2023).

Hybrid analysis

The study confirms that traditional analysis methods, such as static and dynamic analysis, have inherent limitations, underscoring the need for the development of more accurate and efficient techniques based on a hybrid approach.

The survey by Aboaoja et al. underscores the pressing need for hybrid detection methods that combine static, dynamic, and heuristic approaches to effectively combat the sophisticated tactics used by modern malware, such as code reordering and encryption (Aboaoja, et al, 2022).

Memory Analysis

Process/Service: Reviewing active processes and services in memory to detect any malicious activity or unexpected behavior.

DLL (Dynamic Link Libraries): Identifying and analyzing dynamically loaded libraries, which could be used by malware to perform various actions or hide its presence.

Registry Keys: Examining the Windows Registry for unauthorized changes or entries that may be added by malware to maintain persistence or configure execution.

Network Connections: Monitoring incoming and outgoing network traffic to identify communication with command-and-control servers or other malicious network activity.

Memory analysis is crucial as it can reveal the presence of malware that is actively running in the system memory, which might not be detected through static or dynamic analysis alone. It can help to identify rootkits and other forms of stealthy malware that are designed to hide their presence on a system. Memory-based analysis provides insights into how malware interacts with the system at runtime, which can be essential for developing effective countermeasures.

Malware Detection Techniques

Malware detection methods can generally be categorized into three main types: signature-based, heuristic-based (also called as behavior or anomaly-based detection), and specification-based approaches (Figure 2). These techniques identify and detect malware and take countermeasures against those malwares for the safety of computer systems from a potential loss data and resources.

Signature-Based Detection

Signature-based detection relies on known patterns or signatures of known malware. These signatures are unique identifiers derived from the characteristics of specific malware strains.

Signature-based detection is described as a widely used approach in commercial antivirus software that is fast and efficient in detecting known malware. However, it is highlighted that this approach has significant limitations, especially in detecting unknown or new generation malware. The paper (Aslan, et al, 2020) points out that malware from the same family can often evade detection by using obfuscation techniques, making signature-based methods less effective against sophisticated threats.

Heuristic-Based Detection

Aslan, et al. provide a detailed overview of various malware detection approaches, emphasizing that while signature-based methods are effective for known threats, they fall short when detecting new and sophisticated malware types, underscoring the need for hybrid or more advanced detection techniques (Aslan, et al, 2020).

Specification-Based Detection (Anomaly Detection)

Specification-based detection involves defining a set of rules or specifications for normal system behavior. Any deviation from these specifications is flagged as potentially malicious.

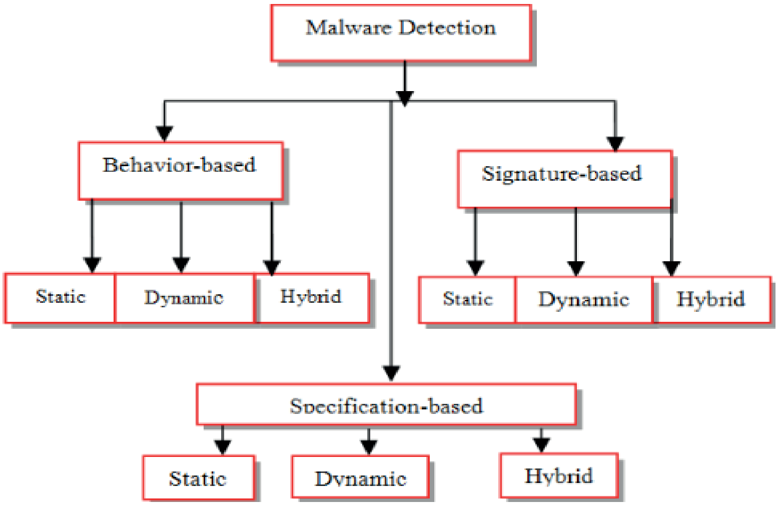

Figure 2. Categories Malware detection techniques

The system establishes a baseline of normal behavior and alerts administrators if there are significant deviations. This can include anomalies in file access patterns, network traffic, or system resource usage. This method does not rely on known signatures but rather on rules or algorithms to predict malicious intent based on certain characteristics or actions.

Each detection method has its own strengths and weaknesses, as shown in Table 1.

Table 1 – Comparison of Malware Detection Techniques

Malware detection techniques	Advantages	Disadvantages
Signature based	It can detect known instances of malware accurately, less amount of resources are required to detect the malware and it mainly focus on signature of attack	It can't detect the new, unknown instances of malware as no signature is available for such type of malware
Heuristic based	It can detect known as well as new, unknown instances of malware and it focuses on the behavior of system to detect unknown attack.	It needs to update the data describing the system behavior and the statistics in normal profile but it tends to be large. It need more resources like CPU time, memory and disk space and level of false positive is high.
Specification based	It can detect known and unknown instances of malware and level of false positive is low but level of false negative is high.	It is not as effective as behavior based detection in detecting new attacks; especially in network probing and denial of service attacks.

Machine learning techniques have transformed malware detection, offering enhanced capabilities to identify both known and novel threats by analyzing large datasets of benign and malicious files.

While traditional signature-based methods are limited in detecting new malware, El Merabet and Hajraoui highlight the advantages of machine learning classifiers, such as support vector machines and neural networks, which excel at generalizing from training data to accurately detect previously unseen malware (El Merabet, et al, 2019).

Bharadiya provides an insightful overview of the applications of machine learning in cybersecurity, emphasizing its critical role in addressing complex challenges such as phishing detection, malware identification, and intrusion detection systems (Bharadiya, 2023).

In Figure 3 shows a malware detection system using machine learning.

The figure provided appears to be a schematic representation of a machine learning-based malware detection system. It illustrates the process flow from training to testing phases.

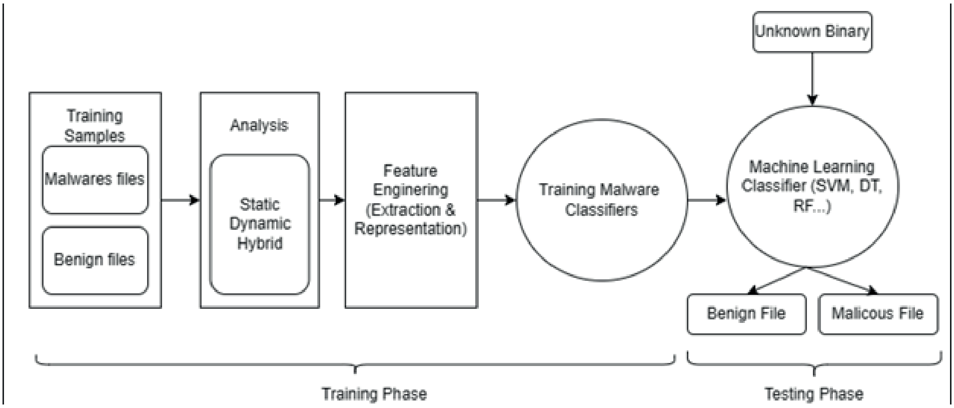

Figure 3. Schematic Framework of Malware Detection System using ML

Training Samples: This step involves collecting a dataset composed of both malicious files (malwares) and benign files. The quality and diversity of these files are crucial for building an effective classifier.

Analysis: In this step, the dataset undergoes analysis, which can be static, dynamic, or a hybrid combination of both. Static analysis involves examining the malware without executing it, while dynamic analysis involves running the malware in a controlled environment to observe its behavior. Hybrid analysis combines elements of both static and dynamic analysis to provide a comprehensive overview.

Feature Engineering (Extraction & Representation): This is a critical step where information features of the malware are extracted. Features could include API calls, binary data, control flow graphs, and other relevant data points that help in distinguishing between benign and malicious files.

Training Malware Classifiers: The extracted features are then used to train machine learning classifiers such as Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and others. The classifier learns to identify patterns and characteristics that are indicative of malware.

Testing Phase: In this phase, an unknown binary file is given to the trained classifier, which then predicts whether the file is benign or malicious based on the learned patterns during the training phase.

Results and Discussion

The main task of machine learning for detection or classification of malware is the output returned by the system implemented. On the one hand, a malware detection system outputs a single value y = f(x), in the range from 0 to 1, which indicates the maliciousness of the executable. On the other hand, a classification system outputs the probability of a given executable belonging to each output class or family, $y \in \mathbb{R}N$, where N indicates the number of different families (Gibert, et al, 2020).

Kamboj et al. conducted a comprehensive study comparing various machine learning models for malware detection, concluding that the Random Forest classifier achieved the highest accuracy at 99.99%, making it highly effective in identifying malicious files (Kamboj, et al, 2023).

The authors conducted their research by employing a comprehensive methodology that included collecting and analyzing a significant dataset of both malicious and benign files. They utilized advanced machine learning models, notably Random Forest and XGBoost, to classify and identify malware types accurately. The study focused on various malware categories, such as Adware, Trojan, Backdoors, and others, using features like MD5 hash size and Optional Header size for detection. The effectiveness of each model was evaluated based on their accuracy in distinguishing between malicious and benign files, leading to the identification of Random Forest as the most accurate model (Kamboj, et al, 2023).

Falana et al. introduce an innovative visualization-based approach to malware detection, where malware binaries are converted into RGB images and analyzed using a deep convolutional neural network (DCNN), demonstrating superior accuracy compared to traditional methods (Falana, et al, 2022).

Ihab Shhadat et al. demonstrated high accuracy in malware classification using a benchmark dataset. They determined the accuracy metrics for malware detection and classification, achieving a high accuracy of 98.2% for binary classification with Decision Trees and 95.8% for multi-class classification with Random Forest. Performance evaluations were conducted on various types of malware, including Dridex, Locky, TeslaCrypt, Vawtrak, Zeus, DarkComet, CyberGate, CTB-Locker, and Xtreme. The datasets used for these experiments consisted of 1156 files, with 984 malicious files and 172 benign files in formats such as .exe, .pdf, and .docx (Shhadat, et al, 2020).

Mohammed Chemmakha et al. improved model performance and computational efficiency through feature selection, using embedded methods to identify the 10 most relevant features. This method yielded a 99.47% accuracy for Random Forest and 99.02% for XGBoost. The dataset used contains 13,8048 lines, including 41323 malicious and 96742 harmless files, and covers 57 features. This data set is presented in the Portable Executable Header (PE HEADER) format (Chemmakha, et al, 2022).

Mushtaq E. et al. used Kaggle's Malware Detection dataset, balanced out of 50,000 malware and 50,000 benign samples across 35 functions. The Random Forest and XGBoost models are superior to others, with Random Forest achieving the highest accuracy of 99.96%. And KNN achieved the highest accuracy of 85.39% using a binary dataset with 16 objects. The file presents a comparative study of machine learning models for malware detection with an emphasis on an accuracy metric for evaluating performance (Mushtaq, et al, 2022).

The study presents a comparative analysis of malware detection techniques using machine learning algorithms, focusing on Decision Tree (DT), K-Nearest Neighbors (K-NN), and Support Vector Machine (SVM). The study employs a dataset comprising 305 types of malware and 236 types of benign software, all in Windows PE format. Decision Tree emerged as the most accurate model with a detection accuracy of 99% and a False Positive Rate (FPR) of 0.021%, indicating its superior performance in classifying malware from benign files in this context (Selamat, et al, 2019).

The study introduces the ANTE system for early bot detection in IoT networks, utilizing Autonomous Machine Learning (AutoML) to select the optimal ML pipeline for identifying various botnet types. It achieves an average detection accuracy of 99.06% and a bot detection precision of 100% across four datasets: ISOT HTTP Botnet, CTU-13, CICDDoS2019, and BoT-IoT. These results were obtained by comparing ANTE's performance to existing literature, showcasing its ability to adapt and select the most suitable ML pipeline for different scenarios and botnet types (Araujo, et al, 2022).

In the study (Azeem, et al, 2024), the authors used the UNSWNB15 dataset, focusing on network security. Machine Learning (ML) methods like K-Nearest Neighbors (KNN), Extra Tree (ET), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), and Multilayer Perceptron (nnMLP) were applied. Random Forest achieved the highest accuracy of 97.68%. The dataset contains real-time normal and abnormal network events, divided into four CSV files, totaling over 2.5 million records. The study aimed to enhance malware detection through effective feature selection and ML classification techniques.

In the paper (Baptista, et al, 2019), authors introduce a cutting-edge malware detection approach that leverages binary visualization and self-organizing incremental neural networks (SOINN) to efficiently identify malicious payloads in various file formats. Their method stands out by converting a file's binary data into a visual image and applying SOINN for analysis, which shows improved detection capabilities, especially for obfuscated codes. The technique emphasizes the transformation of binary data into color-coded images using Hilbert space-filling curves for optimal data clustering. This visualization aids in highlighting unusual patterns that may indicate malware, significantly when obfuscation techniques are used to disguise malicious code.

In the process of reviewing the work on machine learning algorithms, the following comparative table was created (Table 2).

Table 2 – Comparison of machine learning algorithms for malware detection

Machine learning algorithms for malware detection	Strengths	Weaknesses	Use Case
Support Vector Machine (SVM)	Effective in high-dimensional spaces; robust to overfitting when the number of dimensions is greater than the number of samples.	Not suitable for large datasets due to high computational cost; choice of kernel can significantly impact performance.	Ideal for scenarios where the feature space is large and well-defined, and computational resources are sufficient.
Random Forest (RF)	Handles large datasets efficiently; reduces overfitting by averaging multiple decision trees; robust to noise and outliers.	Can be less interpretable than single decision trees; might require significant computational resources for large forests.	Suitable for environments where interpretability is less critical than accuracy and robustness, such as large-scale malware classification tasks.
Logistic Regression (LR)	Simple and interpretable; effective for binary classification problems; computationally efficient.	Assumes a linear relationship between features and the target; less effective with complex, non-linear data.	Useful for quick and interpretable binary classification, especially in preliminary malware detection stages.
Decision Tree (DT)	Simple to understand and interpret; can handle both numerical and categorical data; requires little data preprocessing.	Prone to overfitting, especially with noisy data; can create biased trees if some classes dominate.	Effective for initial exploratory data analysis and in situations where interpretability is crucial.
K-Means Clustering	Simple and fast; scalable to large datasets; useful for unsupervised learning tasks.	Requires the number of clusters to be defined beforehand; sensitive to initial cluster centroids.	Appropriate for discovering hidden patterns and groupings in unlabeled datasets, useful for detecting new and unknown malware families.
Naive Bayes (NB) classifier	Easy to implement and computationally efficient, making it suitable for real-time applications. Performs well even with noisy data.	Assumes feature independence, which is rarely true in real-world applications, leading to less accurate predictions; might be biased towards majority classes in imbalanced datasets.	Ideal for applications where interpretability and quick results are more important than absolute accuracy, such as email spam detection and preliminary malware filtering.

One Hot Encoding	Converts categorical variables into numerical format, making them usable in most machine learning models; simplicity and versatility.	Can significantly increase the number of features, leading to more complex and computationally expensive models; often results in sparse matrices, which can be inefficient to process.	Best used when categorical variables are essential to model predictions, such as in malware classification tasks where specific types of malware categories need to be encoded for detection algorithms.
Self-Organizing Incremental Neural Network (SOINN)	Can learn from new data without needing to retrain the entire model; highly adaptive to evolving malware threats; suitable for real-time applications.	More complex to implement and understand compared to traditional neural networks; performance heavily depends on correct tuning of hyperparameters.	Particularly useful in scenarios where the threat landscape is rapidly evolving, such as in the continuous monitoring of network traffic for new malware strains.

Conclusion

This review has presented a detailed analysis of modern machine learning methods applied to malware detection, highlighting the strengths and weaknesses of various algorithms such as Support Vector Machines (SVM), Random Forests (RF), Logistic Regression (LR), and Decision Trees (DT). While these algorithms have demonstrated impressive accuracy in detecting known malware, this analysis reveals several significant challenges that remain unaddressed by current research.

Firstly, the static nature of many machine learning models limits their effectiveness against rapidly evolving malware threats. Future studies should prioritize the creation of adaptive algorithms capable of continuous learning, which would significantly enhance the resilience of malware detection systems.

Moreover, the existing literature often overlooks the importance of scalable solutions. As the volume of malware continues to grow, models that can efficiently handle large-scale datasets without compromising accuracy are crucial. Addressing this scalability issue is another critical area for future research.

The review also identifies the lack of interpretability in many advanced machine learning models as a major limitation. While techniques such as Random Forests and neural networks offer high accuracy, their complexity often makes them difficult to interpret and trust in critical cybersecurity contexts. Future research should explore ways to improve the transparency of these models, ensuring that they are not only accurate but also understandable to cybersecurity professionals.

In summary, while significant progress has been made in applying machine learning to malware detection, this review highlights the pressing need for further research into adaptive, scalable, and interpretable models. By addressing these gaps, future studies can contribute to the development of more robust and effective malware detection systems, capable of meeting the challenges posed by an ever-evolving cyber threat landscape.

References

Or-Meir, O., Nissim, N., Elovici, Y., & Rokach, L. (2019). Dynamic malware analysis in the modern era—A state of the art survey. *ACM Computing Surveys (CSUR)*, *52*(5), 1-48. *https://doi.org/10.1145/3329786.*

Bedelbayev, A., Ussatova, O., Zhumabekova, A., & Höfig, E. (2023). Application of machine learning algorithm in the analysis of malicious software. *News of NAS RK. Series of Physics and mathematics*, (2), 21-31. *https://doi.org/10.32014/2023.2518-1726.182*.

Sihwail, R., Omar, K., & Ariffin, K. Z. (2018). A survey on malware analysis techniques: Static, dynamic, hybrid and memory analysis. *Int. J. Adv. Sci. Eng. Inf. Technol*, *8*(4-2), 1662-1671.

Zelinka, I., Szczypka, M., Plucar, J., & Kuznetsov, N. (2023). From malware samples to fractal images: A new paradigm for classification. *Mathematics and Computers in Simulation*. https://doi.org/10.1016/j.matcom.2023.11.032

Aboaoja, F. A., Zainal, A., Ghaleb, F. A., Al-rimy, B. A. S., Eisa, T. A. E., & Elnour, A. A. H. (2022). Malware detection issues, challenges, and future directions: A survey. *Applied Sciences*, *12*(17), 8482. *https://doi.org/10.3390/app12178482.*

Aslan, Ö. A., & Samet, R. (2020). A comprehensive review on malware detection approaches. *IEEE access*, *8*, 6249-6271.

El Merabet, H., & Hajraoui, A. (2019). A survey of malware detection techniques based on machine learning. *International Journal of Advanced Computer Science and Applications*, *10*(1).

Bharadiya, J. (2023). Machine Learning in Cybersecurity: Techniques and Challenges. *European Journal of Technology*, *7*(2), 1 - 14. *https://doi.org/10.47672/ejt.1486.*

Gibert, D., Mateu, C., & Planes, J. (2020). The rise of machine learning for detection and classification of malware: Research developments, trends and challenges. *Journal of Network and Computer Applications*, *153*, 102526. https://doi.org/10.1016/j.jnca.2019.102526.

Kamboj, A., Kumar, P., Bairwa, A. K., & Joshi, S. (2023). Detection of malware in downloaded files using various machine learning models. *Egyptian Informatics Journal*, *24*(1), 81-94. https://doi.org/10.1016/j.eij.2022.12.002

Falana, O. J., Sodiya, A. S., Onashoga, S. A., & Badmus, B. S. (2022). Mal-Detect: An intelligent visualization approach for malware detection. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), 1968-1983. https://doi.org/10.1016/j.jksuci.2022.02.026.

Shhadat, I., Hayajneh, A., & Al-Sharif, Z. A. (2020). The use of machine learning techniques to advance the detection and classification of unknown malware. *Procedia Computer Science*, *170*, 917-922. https://doi.org/10.1016/j.procs.2020.03.110.

Chemmakha, M., Habibi, O., & Lazaar, M. (2022). Improving machine learning models for malware detection using embedded feature selection method. *IFAC-PapersOnLine*, *55*(12), 771-776. https://doi.org/10.1016/j.ifacol.2022.07.406.

Mushtaq, E., Shahid, F., & Zameer, A. (2022). A comparative study of machine learning models for malware detection. In *2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST)* (pp. 677-681). IEEE.

Selamat, N., & Ali, F. (2019). Comparison of malware detection techniques using machine learning algorithm. *Indones. J. Electr. Eng. Comput. Sci*, *16*, 435.

Araujo, A. M., de Neira, A. B., & Nogueira, M. (2022). Autonomous machine learning for early bot detection in the internet of things. *Digital Communications and Networks*. https://doi.org/10.1016/j.dcan.2022.05.011.

Azeem, M., Khan, D., Iftikhar, S., Bawazeer, S., & Alzahrani, M. (2024). Analyzing and comparing the effectiveness of malware detection: A study of machine learning approaches. *Heliyon*, *10*(1). https://doi.org/10.1016/j.heliyon.2023.e23574.

Baptista, I., Shiaeles, S., & Kolokotronis, N. (2019). A novel malware detection system based on machine learning and binary visualization. In *2019 IEEE International Conference on Communications Workshops (ICC Workshops)* (pp. 1-6). IEEE.

CONTENTS

INFORMATION AND COMMUNICATION TECHNOLOGIES

МАЗМҰНЫ

АҚПАРАТТЫҚ-КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР

СОДЕРЖАНИЕ

ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ

Publication Ethics and Publication Malpracticein
the journals of the National Academy of Sciences of the Republic of Kazakhstan

For information on Ethics in publishing and Ethicalguidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the National Academy of Sciences of the Republic of Kazakhstan implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any otherlanguage, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The National Academy of Sciences of the Republic of Kazakhstan follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the National Academy of Sciences of the Republic of Kazakhstan.

The Editorial Board of the National Academy of Sciences of the Republic of Kazakhstan will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive
ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Директор отдела издания научных журналов НАН РК *А. Ботанқызы*
Редакторы: *Д.С. Аленов, Ж.Ш.Әден*
Верстка на компьютере *Г.Д.Жадыранова*

Подписано в печать 2.12.2024.
Формат 60х881/8. Бумага офсетная. Печать –ризограф.
16,0 п.л. Тираж 300. Заказ 4.

Национальная академия наук РК
050010, Алматы, ул. Шевченко, 28, т. 272-13-19